SIAM J. CONTROL
Vol. 11, No. 1, February 1973

INVERSE PROBLEM OF LINEAR OPTIMAL CONTROL*

ANTONY JAMESON AND ELIEZER KREINDLER†

Abstract. Necessary and sufficient conditions are derived such that a multi-input, time-varying, linear state-feedback system minimizes a quadratic performance index (the inverse linear optimal control problem). A procedure for determining all such equivalent performance indices that yield the same feedback matrix is indicated.

1. Introduction. Consider a linear system given by

$$\dot{x} = Ax + Bu, \qquad x(t_0) = x_0, \tag{1.1}$$

$$u = Dx, \tag{1.2}$$

and a performance index

$$I = \frac{1}{2}x^T(t_1)Fx(t_1) + \frac{1}{2}\int_{t_0}^{t_1} (x^TQx + u^TRu)\,dt, \tag{1.3}$$

where x is the n-dimensional state, u the m-dimensional control, t_1 is a fixed terminal time, and the superscript T denotes matrix transpose. The matrices A, B, D, Q and R may be time-varying, and are assumed to be uniformly bounded and continuous on $[t_0, t_1]$.

The inverse problem of linear optimal control is to find necessary and sufficient conditions on the system matrices A, B and D so that some performance index of the type (1.3) is minimized, and to determine all such R, Q and F.

The direct problem and its solution are of course well known. The feedback matrix D such that the system (1.1), (1.2) minimizes (1.3) is given by

$$D = -R^{-1}B^TP, \tag{1.4}$$

where P is the solution of the matrix Riccati equation

$$-\dot{P} = PA + A^TP - PBR^{-1}B^TP + Q, \qquad P(t_1) = F. \tag{1.5}$$

For the existence of a unique minimizing u it is assumed that R is positive definite (denoted by $R > 0$), and as a sufficient condition for the existence of a solution $P(t)$ of (1.5) it is usually assumed that Q and F are nonnegative definite ($Q \geqq 0, F \geqq 0$). The minimal value I_* of I is then nonnegative for all x_0 and t_0, and since

$$I_* = \tfrac{1}{2}x_0^TP(t_0)x_0, \tag{1.6}$$

P is nonnegative definite. The case where $t_1 \to \infty$ is of particular interest when A and B are constant, because the performance index

$$I = \frac{1}{2}\int_0^{\infty} (x^TQx + u^TRu)\,dt, \tag{1.7}$$

* Received by the editors September 29, 1970, and in revised form April 18, 1972.

† Research Department, Grumman Aerospace Corporation, Bethpage, New York 11714.

with Q and R constant, results in a constant feedback matrix D. If (1.1) is completely controllable and $Q = C^TC$ is such that $[A, C]$ is completely observable, then the matrix P in (1.4) is the asymptotically stable equilibrium point of the Riccati equation

$$\dot{P} = PA + A^TP - PBR^{-1}B^TP + Q, \qquad P(0) = P_0 = P_0^T \geqq 0, \tag{1.8}$$

i.e., it is the positive definite solution of the matrix equation

$$0 = PA + A^TP - PBR^{-1}B^TP + Q. \tag{1.9}$$

The present paper solves the general time-varying case with the performance index given by (1.3), as well as the time-invariant case with the performance index (1.7). The plan of attack is as follows.

In § 2 we determine necessary conditions for the existence of real symmetric matrices $R > 0$ and P (also for $P \geqq 0$ and $P > 0$), such that (1.4) is satisfied; the most restrictive of these is that DB have real eigenvalues. The sufficiency of these conditions is demonstrated in § 3 by producing general formulas for such R and P. In § 4 we give complete sets of necessary and sufficient conditions for solutions $P \geqq 0$ and $P > 0$; in view of (1.6), these conditions are necessary and sufficient for I_* to be nonnegative and positive for all x_0 and all $t_0 < t_1$, and they are necessary for construction of $Q \geqq 0$—hence their importance.

The solutions of (1.4) for R and P are pointwise in time, but P can be constructed to be differentiable, so that we have $\dot{P}$. Then $F = P(t_1)$, and Q is given by

$$Q = -\dot{P} - PA - A^TP + D^TRD. \tag{1.10}$$

We remark that Q so determined may not be nonnegative definite even if P is positive definite; this is true also in the time-invariant case. We now have the entire class of matrices $\{R, Q, F, P\}$ that satisfy (1.4) and (1.5), and we show in § 5 that each member of this class of performance indices is actually minimized by the given control law (1.2), thus solving the inverse problem (Theorem 5.1).

The inverse problem was first posed and partly solved by Kalman [1] who considered the time-invariant single-input case where R reduces to a scalar. He showed that the satisfaction of a particular inequality, the sensitivity inequality, implies that there exists a performance index (1.7) with a nonnegative definite Q and $R = 1$ which is minimized. This result was generalized by Anderson [2] to the multi-input, time-invariant case. Our approach and results, which do not necessarily produce a nonnegative definite Q, are different. Results for the case where Q is nonnegative definite, and relations with the Kalman–Anderson results, will be presented in a sequel to this paper.

The generality of the characterization of R and P (Theorems 3.1–3.4) gives them independent value; they provide new insight into the already extensively researched linear optimal control problem and are bound to find many applications, particularly in the area of equivalent loss functions [3]. We note that the results of this paper are important for the local treatment of the general nonlinear inverse problem. The fact that our results do not require $Q \geqq 0$ is then significant; this requirement, usually natural for the direct linear optimal problem, is unduly restrictive for the nonlinear case (where Q is replaced by H_{xx}, the second partial of the Hamiltonian).

We remark in conclusion that the conditions for optimality of (1.1), (1.2) derived here are in general no longer necessary when a cross-product term u^TSx is added in the integrand of (1.3) or (1.7). In fact, it is shown in [4] that *every* system (1.1), (1.2) minimizes a performance index (1.3) with a cross-product term, and that there are many ways such a performance index can be constructed. If the performance index is further generalized to include derivatives of the control, dynamic feedback and feedforward controllers can be included, and thus *every linear, finite-dimensional dynamic feedback system minimizes some sufficiently general quadratic performance index* [4].

2. Compatibility conditions. First we note that only the symmetric parts of R, Q and F appear in (1.4) and in the Riccati equation (1.5). Thus, the existence of symmetric P, R, Q and F satisfying (1.4) and (1.5) is a necessary condition for a closed-loop system (1.1), (1.2) to be optimal with respect to (1.3).

From (1.4),

$$B^TP = -RD, \tag{2.1}$$

and our objective is to solve this equation for all real, symmetric and positive definite R, $R = R^T > 0$, and all real and symmetric P. In this section we derive several necessary conditions for the existence of such solutions; in the next section we show constructively that these are also sufficient. We recall that B is $n \times m$, D is $m \times n$, and they do not necessarily have full rank. We have the following lemma.

LEMMA 2.1. *Necessary conditions for* (2.1) *to have real symmetric solutions P, and real, symmetric and positive definite solutions R, are*:

(i) *for any P, R must be such that the compatibility condition holds*:

$$B^TB^{\ddagger T}RD = RD, \tag{2.2}$$

where $B^\ddagger$ is any matrix (e.g., the Penrose generalized inverse $B^\dagger$) *such that* $BB^\ddagger B = B$;

(ii) *for P to be symmetric, the symmetry condition must hold*:

$$RDB = B^TD^TR; \tag{2.3}$$

(iii) *for R to be positive definite, a rank condition on BD must hold*:

$$\operatorname{rank} BD = \operatorname{rank} D. \tag{2.4}$$

Proof. (i) Premultiplying (2.1) by $B^TB^{\ddagger T}$ and using the identity $B^TB^{\ddagger T}B^T = B^T$, we have

$$B^TB^{\ddagger T}B^TP = B^TP = -RD = -B^TB^{\ddagger T}RD.$$

(ii) Postmultiplying (2.1) by B we have

$$B^TPB = -RDB, \tag{2.5}$$

whence the symmetry of RDB is necessary for the symmetry of P.

(iii) From (2.1),

$$PBD = -D^TRD.$$

We first observe that since R is positive definite,

$$\text{rank } D^TRD = \text{rank } D \tag{2.6}$$

because for every ρ such that $D^TRD\rho = 0$ we have

$$\rho^TD^TRD\rho = (D\rho)^TR(D\rho) = 0,$$

whence $D\rho$ must be zero. Thus, by (2.6),

$$\text{rank } BD \geqq \text{rank } PBD = \text{rank } D^TRD = \text{rank } D.$$

But since rank $BD >$ rank D is impossible, (2.4) follows. Q.E.D.

That RDB must be symmetric with a real, symmetric and positive definite R, leads to the next lemma.

LEMMA 2.2. *A real $R = R^T > 0$ such that RDB is symmetric exists only if the $m \times m$ matrix DB satisfies the eigenvector condition*:

(2.7) *DB has m linearly independent real eigenvectors.*

Proof. Let $L^TL = R > 0$ be such that RDB is symmetric. Then $(L^{-1})^TRDBL^{-1} = LDBL^{-1}$ is symmetric, and it therefore has m linearly independent real eigenvectors. But since DB is similar to $LDBL^{-1}$, (2.7) follows. Q.E.D.

The conditions (2.4) and (2.7) are necessary for the desired solution of (2.1), and it is shown constructively by Theorems 3.1, 3.2 and 3.3 in the next section that they are also sufficient. We can therefore state the following theorem.

THEOREM 2.1. *Equation* (2.1) *has solutions $R = R^T > 0$ and $P = P^T$ if and only if the rank condition* (2.4) *on BD and the eigenvector condition* (2.7) *on DB hold.*

We now proceed to derive conditions for symmetric P to be nonnegative definite and positive definite. We have the following lemma.

LEMMA 2.3. *For* (2.1) *with $R = R^T > 0$ to have solutions $P = P^T \geqq 0$ it is necessary that*

$$\text{rank } DB = \text{rank } D, \tag{2.8}$$

and that all eigenvalues λ of DB be nonpositive:

$$\lambda_i \leqq 0, \quad i = 1, 2, \cdots, m; \tag{2.9}$$

for solutions $P = P^T > 0$, it is necessary that

$$\text{rank } DB = \text{rank } D = \text{rank } B. \tag{2.10}$$

Proof. From (2.1), since R is nonsingular,

$$\text{rank } D = \text{rank } RD = \text{rank } B^TP = \text{rank } PB,$$

and from (2.5),

$$\text{rank } DB = \text{rank } RDB = \text{rank } B^TPB.$$

Thus (2.8) is equivalent to

$$\text{rank } B^TPB = \text{rank } PB. \tag{2.11}$$

Since $P \geqq 0$, we may write $P = Z^TZ$, Z real. We observe, as in the proof of (2.6), that

$$\operatorname{rank}(ZB)^TZB = \operatorname{rank} ZB.$$

Thus

$$\operatorname{rank} B^TPB = \operatorname{rank}(ZB)^TZB = \operatorname{rank} ZB \geqq \operatorname{rank} Z^TZB = \operatorname{rank} PB.$$

But since rank $B^TPB >$ rank PB is impossible, (2.8) follows. If $P \geqq 0$ or $P > 0$, then from (2.5),

$$RDB \leqq 0. \tag{2.12}$$

Writing $R = L^TL$, $(L^{-1})^TRDBL^{-1} = LDBL^{-1}$ is also nonpositive definite, real, and symmetric, and therefore has nonpositive real eigenvalues. Since DB is similar to $LDBL^{-1}$, (2.9) is established. If $P > 0$, then since R is nonsingular, (2.1) shows that

$$\operatorname{rank} D = \operatorname{rank} B, \tag{2.13}$$

whence (2.10) must hold. Q.E.D.

LEMMA 2.4. *The rank condition* (2.8) *on DB together with the symmetry condition* (2.3) *on RDB imply that the compatibility condition* (2.2) *and the rank condition* (2.4) *on BD are satisfied.*

Proof. From (2.2), using the symmetry of RDB and the identity $BB^{\ddagger}B = B$, we have

$$B^TB^{\ddagger T}RDB = B^TB^{\ddagger T}B^TD^TR = B^TD^TR = RDB,$$

or

$$(B^TB^{\ddagger T}R - R)DB = 0. \tag{2.14}$$

But since rank DB = rank D, if M is a nonzero matrix such that $MDB = 0$, then also $MD = 0$. Thus, (2.14) implies (2.2). From (2.3),

$$RDBD = B^TD^TRD. \tag{2.15}$$

Let ρ be any vector such that $D\rho \neq 0$. Then in view of (2.6), $D^TRD\rho \neq 0$, and by (2.8), $B^TD^TRD\rho \neq 0$, whence by (2.15), $RDBD\rho \neq 0$. Thus $BD\rho \neq 0$ if $D\rho \neq 0$, implying (2.4). Q.E.D.

From Lemmas 2.2, 2.3 and 2.4 we see that conditions (2.7), (2.8) and (2.9) emerge as the major ones necessary for $P \geqq 0$, and in § 4 we show that they are indeed sufficient. We can therefore state the following theorem.

THEOREM 2.2. *Equation* (2.1) *has solutions* $R = R^T > 0$ *and* $P = P^T \geqq 0$ *if and only if DB has m linearly independent real eigenvectors, its rank equals that of D, and its eigenvalues are all nonpositive* (*conditions* (2.7), (2.8) *and* (2.9), *respectively*); *for* $P = P^T > 0$, *the rank of DB must also be equal to that of B* (*condition* (2.10)).

We close this section with the remark that the rank condition rank BD = rank D can be seen as a direct consequence of optimality. There is no loss of generality (see Appendix) in assuming that B is in canonical form and u is partitioned accordingly:

$$B = \begin{bmatrix} 0 & 0 \\ 0 & I \end{bmatrix}, \quad u = \begin{bmatrix} u_1 \\ u_2 \end{bmatrix}. \tag{2.16}$$

Since u_1 does not affect $x(t)$, it must minimize at all times the quadratic form u^TRu:

$$[u_1^T \quad u_2^T]\begin{bmatrix} R_{11} & R_{12} \\ R_{12}^T & R_{22} \end{bmatrix}\begin{bmatrix} u_1 \\ u_2 \end{bmatrix} = u_1^TR_{11}u_1 + 2u_1^TR_{12}u_2 + u_2^TR_{22}u_2,$$

whence the minimizing u_1 is given by

$$u_1 = -R_{11}^{-1}R_{12}u_2,$$

i.e., u_1 is proportional to u_2. Thus, if we partition D as $D^T = [D_1^T \quad D_2^T]$, then D_1 must satisfy

$$D_1 = KD_2, \tag{2.17}$$

or, rank D = rank D_2. But

$$BD = \begin{bmatrix} 0 & 0 \\ 0 & I \end{bmatrix}\begin{bmatrix} D_1 \\ D_2 \end{bmatrix} = \begin{bmatrix} 0 \\ D_2 \end{bmatrix},$$

whence rank BD = rank D_2 = rank D.

3. General representations of R and P. We first construct $R = R^T > 0$ so that RDB is symmetric. By Lemma 2.2 we must assume that the eigenvectors of DB, and hence of B^TD^T, are real and linearly independent. Thus the matrix V whose columns are the eigenvectors v of B^TD^T is real and nonsingular. We may write the set of equations $B^TD^Tv_i = \lambda_iv_i$, $i = 1, 2, \cdots, m$, as

$$B^TD^TV = V\Lambda, \tag{3.1}$$

where Λ is the diagonal matrix of eigenvalues λ of B^TD^T. Then for any real, nonsingular matrix Π that commutes with Λ we have

$$B^TD^TV\Pi = V\Lambda\Pi = V\Pi\Lambda, \tag{3.2}$$

whence the columns of $V\Pi$ are seen as a new, linearly independent set of real eigenvectors; in fact, all such sets of eigenvectors are generated by all such Π. (If all eigenvalues λ are distinct, then Π must be diagonal and it simply scales the eigenvectors v; if $\lambda_i = \lambda_j$, then any linear combination of v_i and v_j is also an eigenvector and all such combinations are generated by the now permissible off-diagonal elements π_{ij} and π_{ji} of Π.) We have the following theorem.

THEOREM 3.1. *Let the necessary eigenvector condition* (2.7) *hold. Then every given real $R = R^T > 0$ such that RDB is symmetric, is necessarily given by*

$$R = VV^T, \tag{3.3}$$

where the columns v of V are suitably chosen eigenvectors of B^TD^T. Let V be any such given matrix, and let Γ be a real matrix such that

$$\Gamma = \Gamma^T > 0 \quad \textit{and} \quad \Gamma\Lambda = \Lambda\Gamma, \tag{3.4}$$

where Λ is defined by (3.1). *Then all real $R = R^T > 0$ such that RDB is symmetric are generated by all real Γ in*

$$R = V\Gamma V^T, \tag{3.5}$$

where Γ satisfies (3.4).

Proof. Let $R = MM^T$. Then, since B^TD^TR is symmetric,

$$M^{-1}(B^TD^TR)(M^{-1})^T = M^{-1}B^TD^TM$$

is a symmetric matrix whose eigenvalues are those of B^TD^T. There exists therefore an orthogonal matrix H such that

$$H^TM^{-1}B^TD^TMH = \Lambda, \qquad HH^T = I,$$

whence the columns of MH are seen to be eigenvectors of B^TD^T. If we define $V = MH$, then

$$VV^T = MHH^TM^T = MM^T = R,$$

as claimed in (3.3). Conversely, for any V in (3.3), RDB is symmetric:

$$VV^TDB = V\Lambda V^T = B^TD^TVV^T.$$

Recalling the comment that follows (3.2), we let $\Pi\Pi^T = \Gamma$, whence (3.5) follows from (3.3). Q.E.D.

Not every $R = R^T > 0$ that satisfies the symmetry condition (2.3) automatically satisfies also the compatibility condition (2.2). For example, let

$$B = \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix}, \quad D = \begin{bmatrix} 1 & 0 \\ 1 & 0 \end{bmatrix}, \quad R = \begin{bmatrix} r_{11} & r_{12} \\ r_{12} & r_{22} \end{bmatrix}.$$

Then

$$BD = \begin{bmatrix} 0 & 0 \\ 1 & 0 \end{bmatrix}, \quad DB = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix},$$

whence rank BD = rank D and RDB is symmetric for any R. But since

$$RD = \begin{bmatrix} r_{11} + r_{12} & 0 \\ r_{12} + r_{22} & 0 \end{bmatrix} \quad \text{and} \quad B^TB^{\ddagger T}RD = \begin{bmatrix} 0 & 0 \\ r_{12} + r_{22} & 0 \end{bmatrix},$$

the compatibility condition holds only if $r_{11} + r_{12} = 0$. Nevertheless, the following theorem holds.

THEOREM 3.2. *If the rank condition* (2.4) *on BD and the eigenvector condition* (2.7) *hold, there exist* $R = R^T > 0$ *such that both the compatibility condition* (2.2) *and the symmetry condition* (2.3) *on RDB are satisfied, and all such R can be constructed by suitably choosing* Γ *in* (3.5).

Proof. This is shown in the Appendix, where B is first transformed to canonical form (2.16). For a set of eigenvectors of B^TD^T, all Γ in (3.5) are found such that the compatibility condition is satisfied.

Remark 3.1. The rule for structuring R in (3.3), or (3.5), so as to satisfy the compatibility condition (see (A.12)–(A.15) in the Appendix) rarely need be invoked because the compatibility condition always holds in the most common case when rank $B = m$ (see § 6). The compatibility condition also always holds if rank DB = rank D, as needed for the usual case $P \geqq 0$ (see Lemmas 2.3 and 2.4). Thus, in most cases of interest, the formula (3.5) for R provides all the requisite matrices R of the inverse problem.

Given a real $R = R^T > 0$ satisfying the compatibility condition (2.2) and the symmetry condition (2.3) on RDB, we next solve (2.1) for a real symmetric P. Let U be any real $n \times m$ matrix such that

$$B^T U^T RD = RD; \tag{3.6}$$

in view of (2.2), one such matrix is $U = B^{\ddagger}$. By inspection of (3.6), $-U^T RD$ is a solution of (2.1) for P, which, however, is not necessarily symmetric. To obtain a symmetric solution set

$$P_0 = -U^T RD - D^T RU + U^T RDBU. \tag{3.7}$$

Now by (3.6) and the symmetry of RDB,

$$B^T P_0 = -RD - B^T D^T RU + RDBU = -RD.$$

Further, if P is any real symmetric solution of (2.1), then

$$B^T(P - P_0) = 0,$$

whence the general solution of (2.1) for a real symmetric P is

$$P = -U^T RD - D^T RU + U^T RDBU + Y, \tag{3.8}$$

where Y is any real matrix such that

$$B^T Y = 0, \qquad Y = Y^T. \tag{3.9}$$

In summary, we have the following theorem.

THEOREM 3.3. *Let R be a real, symmetric and positive definite matrix satisfying the compatibility condition* (2.2) *and the symmetry condition* (2.3). *Then a real symmetric P satisfying* (2.1) *exists, and all such P are represented by* (3.8), *where U is any real matrix satisfying* (3.6) *and Y is any real matrix satisfying* (3.9).

Theorems 3.1, 3.2 and 3.3 provide the sufficiency part of Theorem 2.1.

Under the rank condition (2.8) on DB, additional representations of P are available. Furthermore, by Lemma 2.4, the compatibility condition (2.2) is then redundant. We then have the following.

THEOREM 3.4. *Let R be a real, symmetric and positive definite matrix such that RDB is symmetric. If*

$$\operatorname{rank} DB = \operatorname{rank} D, \tag{3.10}$$

then all real symmetric P satisfying (2.1) *are represented in terms of the given R by*

$$P = -D^T R(RDB)^{\dagger} RD + Y, \tag{3.11}$$

where † denotes the Penrose generalized inverse; in terms of the eigenvectors of $B^T D^T$, P is given by

$$P = -D^T V \Gamma \Lambda^{\dagger} V^T D + Y, \tag{3.12}$$

where Y are all real matrices that satisfy (3.9), *V and Λ are defined by* (3.1), *and Γ is defined by* (3.4).

Proof. Consider (3.6) and (3.7), and let

$$U = (RDB)^{\dagger} RD. \tag{3.13}$$

If DB is nonsingular, this U satisfies (3.6). Consider the case where DB is singular. By the identity $XX^{\dagger}X = X$,

$$[B^TD^TR(RDB)^{\dagger}R - R]DB = 0. \tag{3.14}$$

We note that (3.10) implies that $MDB = 0$ only if $MD = 0$. Hence (3.14) gives

$$B^TD^TR(RDB)^{\dagger}RD = RD.$$

Thus U given by (3.13) is seen to satisfy (3.6). Substituting U into the last term of (3.7), we have

$$\begin{aligned} U^TRDBU &= D^TR(RDB)^{\dagger}RDB(RDB)^{\dagger}RD \\ &= D^TR(RDB)^{\dagger}RD = D^TRU, \end{aligned}$$

so that (3.8) yields (3.11). To prove (3.12), we note that $\Lambda^{\dagger}$ is a diagonal matrix with elements $1/\lambda_i$ if $\lambda_i \neq 0$, and 0 if $\lambda_j = 0$. Thus $\Gamma\Lambda^{\dagger} = \Lambda^{\dagger}\Gamma$, and (3.12) is seen to be symmetric. Also, (3.10) implies that all eigenvectors v of B^TD^T that correspond to zero eigenvalues are such that $D^Tv = 0$. Considering all such eigenvectors we have that

$$D^TV = D^TV\Lambda^{\dagger}\Lambda,$$

whence P given by (3.12) is seen to satisfy (2.1):

$$B^TP = -B^TD^TV\Gamma\Lambda^{\dagger}V^TD = -V\Lambda\Gamma\Lambda^{\dagger}V^TD = -V\Gamma V^TD = -RD.$$

Q.E.D.

The rank condition (3.10) is, by Lemma 2.3, necessary for $P \geqq 0$ and $P > 0$. Formulas (3.11) and (3.12) are therefore useful for, but by no means restricted to, these cases. A general representation of P, equivalent to (3.8), or under the rank condition (3.10) equivalent to (3.12), and based on transformation of B to canonical form (2.16), is given in the Appendix (see (A.17)–(A.19)).

Remark 3.2. We used the generalized inverse in the sense of Moore–Penrose because it is unique and perhaps best known of the various pseudoinverses. However, of the identities

$$XX^{\dagger}X = X, \qquad X^{\dagger}XX^{\dagger} = X^{\dagger}, \qquad (XX^{\dagger})^T = XX^{\dagger}, \qquad (X^{\dagger}X)^T = X^{\dagger}X,$$

defining the Penrose inverse, we need only the first two. We may define $X^{\#}$ by $XX^{\#}X = X$ and $X^{\#}XX^{\#} = X^{\#}$, a pseudoinverse that is no longer unique. Let $(RDB)^{\#}$ be

$$(RDB)^{\#} = (V\Gamma\Lambda V^T)^{\#} \triangleq V^{T^{-1}}\Lambda^{\#}\Gamma^{-1}V^{-1}; \tag{3.15}$$

this is a symmetric matrix that satisfies the two identities for $^{\#}$. Then, replacing † in (3.11) by $^{\#}$, using (3.15), and $\Lambda^{\#} = \Lambda^{\dagger}$ and $\Lambda^{\dagger}\Gamma = \Gamma\Lambda^{\dagger}$, (3.11) reduces to (3.12). We remark that in passing from one general representation for P to another, the matrix Y in the second representation, while still satisfying $B^TY = 0$, may not be the same as in the first representation.

Since the symmetry of RDB and the compatibility condition (2.2) are necessary, R given by Theorems 3.1 and 3.2 is the general, real, symmetric, positive definite solution of (2.1). Given R, Theorem 3.3 provides the general, real, symmetric solution P. We thus have all solutions $R = R^T > 0$ and $P = P^T$ of (2.1); in the next section we obtain all such solutions where in addition P is nonnegative definite and positive definite.

4. Conditions for $P \geqq 0$ and for $P > 0$. In view of (1.6), conditions for $P \geqq 0$ (for $P > 0$) are necessary and sufficient for I_* to be nonnegative (positive) for all x_0 and all $t_0 < t_1$. They are important also for the solution of the inverse problem with positiveness conditions on Q in (1.3), as will be discussed in the sequel to this paper.

Recalling Lemma 2.3, we use the representations (3.11) and (3.12) for P, and seek conditions on Y for P to be nonnegative definite and positive definite. We have the following theorem.

THEOREM 4.1. *A real symmetric P given by* (3.11) *or* (3.12) *satisfies* (2.1) *with $R = R^T > 0$ and is nonnegative definite if and only if the following conditions hold*:

$$\text{rank } DB = \text{rank } D, \tag{4.1}$$

the eigenvalues λ of DB (*which by* (2.7) *must be real*) *are nonpositive*:

$$\lambda_i \leqq 0, \quad i = 1, 2, \cdots, m, \tag{4.2}$$

and in (3.11) *and* (3.12),

$$Y = Y^T \geqq 0; \tag{4.3}$$

P is positive definite if and only if

$$\text{rank } DB = \text{rank } D = \text{rank } B, \tag{4.4}$$

(4.2) *holds, and Y is of the form*

$$Y = WW^T, \tag{4.5}$$

where W has $n - r_B$ linearly independent columns ($r_B =$ rank B) *and*

$$B^TW = 0. \tag{4.6}$$

Proof. The necessity of (4.1), (4.2) and (4.4) is established in Lemma 2.3 and we now show the necessity of the conditions on Y. Consider formula (3.11) for P. Let

$$x = (I - B(RDB)^\dagger RD)\eta, \tag{4.7}$$

where η is any nonzero vector. Then

$$\begin{aligned} x^TPx = &-\eta^T(I - D^TR(RDB)^\dagger B^T)D^TR(RDB)^\dagger RD(I - B(RDB)^\dagger RD)\eta \\ &+\eta^TY\eta = \eta^TY\eta, \end{aligned} \tag{4.8}$$

whence (4.3) is seen to be necessary. To prove the necessity of (4.5), assume that the columns of W do not span $B^\perp$, the $(n - r_B)$-dimensional orthogonal complement of the range space of B. There is then a nonzero η in $B^\perp$ such that $\eta^TY\eta = 0$, and x given by (4.7) is nonzero (being the sum of a vector in $B^\perp$ and a vector in the range space of B). Then (4.8) shows that $x^TPx = 0, x \neq 0$, proving the necessity of (4.5). To establish sufficiency, we first note that as in the proof of Lemma 2.3, (4.2) implies that $RDB \leqq 0$; whence $(RDB)^\dagger \leqq 0$ (since RDB is symmetric, $RDB = H\Omega H^T, HH^T = I$, and $(RDB)^\dagger = H\Omega^\dagger H^T$, where $\Omega = \operatorname{diag}(\omega_i)$ and ω_i are eigenvalues of RDB). Thus the first term in (3.11) is nonnegative and $P \geqq 0$ under (4.3). For the case $P > 0$, let $x = \eta + B\rho$, where η is in the orthogonal complement $B^\perp$ of the range space of B. Then

$$x^TPx = -x^TD^TR(RDB)^\dagger RDx + \eta^TWW^T\eta,$$

where the first term is nonnegative and the second is, under (4.5), positive for all nonzero η. Thus $x^TPx > 0$ for $\eta \neq 0$. If $\eta = 0$, then $x = B\rho$, and for $x = B\rho \neq 0$ we have

$$x^TPx = -\rho^TB^TD^TR(RDB)^\dagger RDB\rho = -\rho^TRDB\rho.$$

Since $-RDB$ is nonnegative, $-RDB = Z^TZ$ for some real Z and then $DB\rho = -R^{-1}Z^TZ\rho$. But since $B\rho \neq 0$, then under (4.4), $DB\rho \neq 0$ and $Z^TZ\rho \neq 0$. Thus $x^TPx = -\rho^TRDB\rho = \rho^TZ^TZ\rho > 0$. Hence $P > 0$. For the representation (3.12) of P we can proceed in a similar manner, or simply note that (3.12) follows from (3.11) if we replace the Penrose inverse † by the pseudoinverse $^\#$ defined in Remark 3.2. Q.E.D.

Theorem 4.1 together with the theorems of § 3 establish the sufficiency part of Theorem 2.2. The equivalent of Theorem 4.1, for a partitioned P that results when B is in canonical form (2.16), is given in the Appendix (see (A.21)–(A.25)).

Since (3.10) is necessary for $P \geqq 0$, it is clear that (3.11) or (3.12), together with the conditions (4.3) and (4.5) on Y, are the general, real, symmetric solutions $P \geqq 0$ and $P > 0$ of (2.1). We thus have all the desired solutions R and P of (2.1), needed for solving the inverse problem.

5. The inverse problem. If matrices B and D of (1.1), (1.2) satisfy the conditions of Theorem 2.1, we can construct an $R = R^T > 0$ as in Theorems 3.1 and 3.2, and a $P = P^T$ as in Theorem 3.3, so that (1.4) is satisfied. To construct a bounded Q from (1.10) we require that P be differentiable. We therefore make the following assumption.

ASSUMPTION 5.1. In (1.1), (1.2), $B(t)$ and $D(t)$ are differentiable on $[t_0, t_1]$ and are of constant rank.

The latter part prevents an apparent discontinuity in P due to a change in the rank of $Y = WW^T$, mandated by Theorem 4.1 for $P > 0$, at the instant B changes rank. However, since the direct problem does not require Assumption 5.1, and our representations for P are general, it is clear that the assumption is for convenience only.

We now show that a performance index (1.3) with the weighting matrices R, Q and $F = P(t_1)$ constructed in §§ 3 and 4 is minimized by the control (1.2).

LEMMA 5.1. *Consider a closed-loop linear system* (1.1), (1.2). *Let* $R > 0$, Q, P, *and* $F = P(t_1)$ *be arbitrary uniformly bounded symmetric matrices satisfying* (1.4) *and the Riccati equation* (1.5). *Then the performance index* (1.3) *attains its absolute minimum* I_* *given by* (1.6), *over all square-integrable controls, for all* x_0 *and all* $t_0 < t_1 \leqq \infty$. *The optimal control is uniquely given by* (1.2).

Proof. Substituting (1.4) into (1.5) we have

$$-\dot{P} = PA + A^TP - D^TRD + Q. \tag{5.1}$$

Multiplying both sides by x and using $Ax = \dot{x} - Bu$, we have

$$-\frac{d}{dt}(x^TPx) = -(u - Dx)^TR(u - Dx) + x^TQx + u^TRu. \tag{5.2}$$

By integrating (5.2), setting $P(t_1) = F$, and multiplying by $\frac{1}{2}$, we have

$$\begin{aligned} &\frac{1}{2}x_0^TP(t_0)x_0 + \frac{1}{2}\int_{t_0}^{t_1} (u - Dx)^TR(u - Dx)\,dt \\ &\qquad = \frac{1}{2}x(t_1)^TFx(t_1) + \frac{1}{2}\int_{t_0}^{t_1} (x^TQx + u^TRu)\,dt. \end{aligned} \tag{5.3}$$

The right side of (5.3) is the performance index I of (1.3). Since $R > 0$, the integral on the left side is nonnegative. Thus I attains its absolute minimum if and only if $u = Dx$, as stated. Q.E.D.

When $t_1 = \infty$ in (1.3), the lemma is valid even if the closed-loop system (1.1), (1.2) is not asymptotically stable, since the integral on the left side of (5.3) must approach $+\infty$ if it is not finite. The time-invariant case, however, must be treated distinctly. We then arrive at (5.3) by starting with the constant quadratic matrix equation (1.9) rather than with the matrix Riccati equation (1.5), and since the performance index (1.7) does not have the terminal term $\frac{1}{2}x^T(t_1)Fx(t_1)$, (5.3) is, for $u = Dx$, replaced by

$$\tfrac{1}{2}x^T(0)P(0)x(0) - \tfrac{1}{2}x^T(\infty)P(\infty)x(\infty) = I.$$

The term $\frac{1}{2}x^T(\infty)P(\infty)x(\infty)$ can be positive and finite, raising the possibility, pointed out to us by B. P. Molinari, of an optimal control law that is unstable. Thus, to draw conclusions from (5.3), we restrict consideration to *stabilizing controls*, i.e., such that $x(t_1) \to 0$ as $t_1 \to \infty$. We have the following lemma.

LEMMA 5.2. *Consider a time-invariant asymptotically stable system* (1.1), (1.2), *and symmetric constant matrices* $R > 0$, P_∞, *and* Q *satisfying* (1.4) *and the quadratic matrix equation* (1.9). *Then the performance index* (1.7) *attains its absolute minimum over all square-integrable stabilizing controls for all* x_0, (1.2) *is the unique minimizing control, and* P_∞ *is the unique asymptotically stable equilibrium point of the Riccati equation* (1.8).

Proof. Only the last assertion remains to be proved. To prove it, we shift the origin of the Riccati equation (1.8) to P_∞ by considering

$$\bar{P}(t) = P(t) - P_\infty. \tag{5.4}$$

We find that

$$\dot{\bar{P}} = \bar{P}A_c + A_c^T\bar{P} - \bar{P}BR^{-1}B^T\bar{P}, \tag{5.5}$$

where

$$A_c = A + BD. \tag{5.6}$$

Since $\operatorname{Re}\{\lambda_i\} < 0$ for any eigenvalue λ_i of A_c, we have

$$\operatorname{Re}\{\lambda_i + \lambda_j\} < 0, \quad \text{all } i, j. \tag{5.7}$$

By (5.7), the linear part of (5.5) is asymptotically stable and hence, by Lyapunov's first method, $\bar{P}_\infty = 0$ is a locally asymptotically stable equilibrium point of (5.5); the same holds for P_∞ with respect to the Riccati equation (1.8). It remains to show that P_∞ is the only asymptotically stable equilibrium point of (1.8). First we rewrite (1.9) as

$$PA_c + A_c^TP = -D^TRD - Q. \tag{5.8}$$

For given A, B, D, R and Q, this is a linear matrix equation in P, and it has a unique solution P_∞ because by (5.7), $\lambda_i + \lambda_j \neq 0$ for all i, j. Thus any other solution of (1.9), say P'_∞, must yield D':

$$D' = -R^{-1}B^TP'_\infty \neq D = -R^{-1}B^TP_\infty.$$

Now suppose P'_∞ is an asymptotically stable equilibrium point of (1.8). Then, by reversing the previous arguments,

$$A'_c = A + BD'$$

is asymptotically stable, and $u = D'x$ provides the minimizing control for I. Hence

$$I_* = \tfrac{1}{2}x_0^TP_\infty x_0 = \tfrac{1}{2}x_0^TP'_\infty x_0, \quad \text{all } x_0,$$

whence $P'_\infty = P_\infty$ and $D' = D$.

Remark 5.1. Lemma 5.2 extends, to the case where Q is not necessarily nonnegative definite, the well-known facts (for $Q \geqq 0$) (i) that there is a one-to-one relation between the stability of the Riccati equation and that of the corresponding closed-loop optimal system, and (ii) that the Riccati equation (1.8) has at most one asymptotically stable equilibrium point. In contrast with the case $Q \geqq 0$, however, the equilibrium point P_∞ may not be positive definite and its domain of attraction is not generally known. See also recent results in [6] and [7].

We now have all the elements needed for solution of the inverse problem. Theorems 2.1, 2.2, 3.1, 3.2, 3.4, 3.5 and 4.1, and Lemmas 5.1 and 5.2, together imply the next theorem.

THEOREM 5.1. *Consider a closed-loop linear system* (1.1), (1.2) *satisfying Assumption* 5.1. *It is possible to construct a performance index* (1.3) *with*

$$F = F^T, \quad Q = Q^T, \quad R = R^T > 0, \tag{5.9}$$

that attains its absolute minimum I_* *over all square-integrable controls, for all* x_0 *and all* $t_0 < t_1 \leqq \infty$, *if and only if for all* $t, t_0 \leqq t \leqq t_1$, *the following conditions hold*:

(5.10) DB *has* m *linearly independent real eigenvectors*

and

$$\operatorname{rank} BD = \operatorname{rank} D. \tag{5.11}$$

The minimal value I_* *can be negative. An index* (1.3) *such that* $I_* \geqq 0$ *for all* x_0 *and all* $t_0 < t_1 \leqq \infty$ *can be constructed if and only if in addition to* (5.10), *for all* t, $t_0 \leqq t \leqq t_1$:

(5.12) *all eigenvalues of* DB *are nonpositive*, *and* (5.11) *is strengthened to*

$$\text{rank}\, DB = \text{rank}\, D. \tag{5.13}$$

An index (1.3), *such that* $I_* > 0$ *for all* x_0 *and all* $t_0 < t_1 \leqq \infty$, *can be constructed if and only if, in addition to* (5.10) *and* (5.12), *the rank condition* (5.13) *is strengthened to*

$$\text{rank}\, DB = \text{rank}\, D = \text{rank}\, B. \tag{5.14}$$

If the system (1.1), (1.2) *is constant and asymptotically stable, then it is possible to construct a performance index* (1.7) *with constant symmetric* Q *and* $R > 0$, *that attains its absolute minimum* I_* *over all square-integrable stabilizing controls, for all* x_0, *if and only if the above conditions hold. All performance indices corresponding to these conditions can be constructed by the general formulas of* § 3.

We observe that all conditions for the inverse problem are on the system matrices B and D, while A is arbitrary (aside from stability in the constant case). This is so because the inverse problem obviates the stability and existence problems of the direct problem. Conditions on A, as well as on B and D emerge when $Q \geqq 0$ is desired (see [1], [2], and the sequel to this paper); the conditions on B, D, BD, and DB discovered here remain of course necessary properties of a linear optimal system.

6. Consequences of B having full rank. Normally the $n \times m$ system matrix B has full rank and $m \leqq n$; in particular, this is so in a single-input system. It is therefore of interest to record the resulting simplifications in our previous results.

Case 1. rank $B = m, m < n$. We observe that the rank condition (5.11) always holds, because by Sylvester's inequality,

$$\text{rank}\, BD \geqq \text{rank}\, B + \text{rank}\, D - m = \text{rank}\, D,$$

which implies (5.11). Further, the compatibility condition (2.2), which somewhat complicates the construction of R (see Theorem 3.2 and Remark 3.1), is always satisfied because now a $B^{\ddagger}$ such that $B^T B^{\ddagger T} = I$ always exists (e.g., $B^{\ddagger} = B^{\dagger} = (B^T B)^{-1} B^T$).

If DB is nonsingular, then formula (3.11) for P reduces to

$$P = -D^T(B^T D^T)^{-1} R D + Y, \tag{6.1}$$

the rank conditions (5.13) and (5.14) always hold, and the eigenvalue condition (5.12) becomes simply

$$DB < 0. \tag{6.2}$$

Case 2. rank $B = m = 1$. Here R reduces to a scalar $r > 0$, B to a column vector b, and D to a row vector d^T. All the conditions of Lemma 2.1 are now satisfied and P can always be represented by (3.8), where now U is a row vector, say, $U = b^{\dagger} = b^T/b^T b$. In particular, if b is in canonical form, then since

$$b = \begin{bmatrix} 0 \\ 1 \end{bmatrix}, \quad b^{\dagger T} = \begin{bmatrix} 0 \\ 1 \end{bmatrix}, \quad d = \begin{bmatrix} d_1 \\ d_2 \end{bmatrix}, \quad (3.9) \Rightarrow Y = \begin{bmatrix} Y_{11} & 0 \\ 0 & 0 \end{bmatrix},$$

we find that (3.8) becomes

$$P = \begin{bmatrix} Y_{11} & -rd_1 \\ -rd_1^T & -rd_2 \end{bmatrix}. \tag{6.3}$$

For any b, rank DB = rank D reduces to

$$d^Tb \neq 0; \tag{6.4}$$

then (6.1) becomes

$$P = -(r/d^Tb)dd^T + Y, \tag{6.5}$$

and (6.2) reduces to

$$d^Tb < 0. \tag{6.6}$$

Since r can be any positive scalar and P can always be constructed as in (3.8) or (6.3), we have a corollary.

COROLLARY 6.1. *Every single-input linear feedback system* (1.1), (1.2), *such that Assumption* 5.1 *holds, minimizes a performance index of the type* (1.3). *A performance index* (1.3) *such that* $I_* > 0$ *for all* x_0 *and all* $t_0 < t_1$ *exists if and only if for all* t, $t_0 \leqq t \leqq t_1$, $d^Tb < 0$ *holds.*

Case 3. rank $B = m = n$. The case where B is nonsingular is rather trivial: P is simply $-B^{T^{-1}}RD$ and rank conditions (2.4) and (2.8) are automatically satisfied.

The case of $m > n$ is unusual, but will be discussed for completeness.

Case 4. rank $B = n < m$. In contrast with the case of $m < n$, the compatibility condition is not automatically satisfied, nor is the rank condition (5.11). However, (5.11) is implied by the symmetry condition

$$RDB = B^TD^TR$$

because, since BB^T is now nonsingular, it yields

$$(BB^T)^{-1}BRDBD = D^TRD,$$

whence, by (2.6),

$$\text{rank}\, BD \geqq \text{rank}\, (BB^T)^{-1}BRDBD = \text{rank}\, D^TRD = \text{rank}\, D,$$

which implies (5.11). Also in contrast with the case $m < n$, the rank condition (5.13) is now automatically satisfied as can be verified by Sylvester's inequality. Finally, (2.1) is now readily solvable for P, yielding

$$P = -(BB^T)^{-1}BRD, \tag{6.7}$$

which by postmultiplying by $BB^T(BB^T)^{-1}$ is seen to be symmetric under the symmetry of RDB,

$$P = -(BB^T)^{-1}BRD = (BB^T)^{-1}B(RDB)B^T(BB^T)^{-1}.$$

Appendix: Proof of Theorem 3.2. We first reduce B to canonical form by means of an equivalence transformation

$$\bar{B} = NBM = \begin{bmatrix} 0 & 0 \\ 0 & I_{r_B} \end{bmatrix}, \tag{A.1}$$

where N and M are suitable nonsingular matrices and r_B is the rank of B. If we define

$$\bar{R} = M^T R M, \qquad \bar{P} = (N^{-1})^T P N^{-1}, \qquad \bar{D} = M^{-1} D N^{-1}, \tag{A.2}$$

we find that (2.1)–(2.7) remain valid in terms of the new matrices. We may therefore assume with no loss of generality that B is initially in canonical form.

According to the hypotheses of Theorem 3.2, the rank condition (2.4) and the eigenvector condition (2.7) hold. We have

$$BD = \begin{bmatrix} 0 & 0 \\ 0 & I_{r_B} \end{bmatrix} \begin{bmatrix} D_{11} & D_{12} \\ D_{21} & D_{22} \end{bmatrix} = \begin{bmatrix} 0 & 0 \\ D_{21} & D_{22} \end{bmatrix},$$

whence rank BD = rank D requires

$$D = \begin{bmatrix} KD_{21} & KD_{22} \\ D_{21} & D_{22} \end{bmatrix} \tag{A.3}$$

for some matrix K. Now

$$B^T D^T = \begin{bmatrix} 0 & 0 \\ 0 & I_{r_B} \end{bmatrix} \begin{bmatrix} D_{21}^T K^T & D_{21}^T \\ D_{22}^T K^T & D_{22}^T \end{bmatrix} = \begin{bmatrix} 0 & 0 \\ D_{22}^T K^T & D_{22}^T \end{bmatrix}.$$

Thus the eigenvector equation $B^T D^T v = \lambda v$ has $m - r_B$ solutions

$$v = \begin{bmatrix} v_1 \\ -K^T v_1 \end{bmatrix}, \qquad \lambda = 0,$$

where the $m - r_B$ vectors v_1 are any set of linearly independent real $(m - r_B)$-vectors. The remaining r_B solutions are

$$v = \begin{bmatrix} 0 \\ v_2 \end{bmatrix},$$

where the v_2 are eigenvectors of D_{22}^T, and by the eigenvector condition (2.7) they are real and linearly independent. Thus in (3.1),

$$V = \begin{bmatrix} V_{11} & 0 \\ -K^T V_{11} & V_{22} \end{bmatrix}, \quad \Lambda = \begin{bmatrix} 0 & 0 \\ 0 & \Lambda_2 \end{bmatrix}, \tag{A.4}$$

where V_{11} is any nonsingular real matrix, and V_{22}, which is given by

$$D_{22}^T V_{22} = V_{22} \Lambda_2, \tag{A.5}$$

is nonsingular and real. By Theorem 3.1, all R given by

$$R = \begin{bmatrix} V_{11} & 0 \\ -K^T V_{11} & V_{22} \end{bmatrix} \begin{bmatrix} \Gamma_{11} & \Gamma_{12} \\ \Gamma_{12}^T & \Gamma_{22} \end{bmatrix} \begin{bmatrix} V_{11}^T & -V_{11}^T K \\ 0 & V_{22}^T \end{bmatrix} \tag{A.6}$$

satisfy the symmetry condition on RDB and we only have to satisfy the compatibility condition (2.2) by choice of Γ. Expanding (A.6) gives

$$\begin{aligned} R_{11} &= V_{11}\Gamma_{11}V_{11}^T, \qquad R_{12} = -R_{11}K + V_{11}\Gamma_{12}V_{22}^T, \\ R_{22} &= K^T R_{11} K + V_{22}\Gamma_{22}V_{22}^T - V_{22}\Gamma_{12}^T V_{11} K - K^T V_{11}\Gamma_{12}V_{22}^T. \end{aligned} \tag{A.7}$$

Letting $B^{\ddagger}$ be the $m \times n$ matrix:

$$B^{\ddagger} = \begin{bmatrix} 0 & 0 \\ 0 & I_{r_B} \end{bmatrix},$$

and using (A.3), (2.2) yields the conditions

$$\text{(A.8)} \qquad (R_{11}K + R_{12})D_{21} = 0, \qquad (R_{11}K + R_{12})D_{22} = 0.$$

From the expression for R_{12} in (A.7), $R_{11}K + R_{12} = V_{11}\Gamma_{12}V_{22}^T$, and (A.8) becomes

$$\text{(A.9)} \qquad V_{11}\Gamma_{12}V_{22}^T D_{21} = 0, \qquad V_{11}\Gamma_{12}V_{22}^T D_{22} = 0.$$

In view of (A.5) and the nonsingularity of V_{11} and V_{22}, these conditions reduce to

$$\text{(A.10)} \qquad \Gamma_{12}V_{22}^T D_{21} = 0$$

and

$$\text{(A.11)} \qquad \Gamma_{12}\Lambda_2 = 0,$$

both of which are satisfied by the choice $\Gamma_{12} = 0$.

Thus, *all $R = R^T > 0$ such that both the symmetry and compatibility conditions hold, are given by all Γ in* (A.6) *such that $\Gamma = \Gamma^T > 0$, $\Gamma\Lambda = \Lambda\Gamma$ and Γ_{12} satisfies* (A.10) *and* (A.11). *One such Γ_{12} is $\Gamma_{12} = 0$ which is also necessary when D_{22} (or equivalently Λ_2) is nonsingular.*

This proves Theorem 3.2. The rule for Γ can be broken down further:

Γ_{22} is any real $r_B \times r_B$ matrix such that

$$\text{(A.12)} \qquad \Gamma_{22} = \Gamma_{22}^T > 0, \qquad \Gamma_{22}\Lambda_2 = \Lambda_2\Gamma_{22};$$

$\Gamma_{12} = \Gamma_{21}^T$ is any real $(m - r_B) \times r_B$ matrix such that

$$\text{(A.13)} \qquad \Gamma_{12}V_{22}^T D_{21} = 0, \qquad \Gamma_{12}\Lambda_2 = 0;$$

Γ_{11} is any real $(m - r_B) \times (m - r_B)$ matrix such that

$$\text{(A.14)} \qquad \Gamma_{11} = \Gamma_{11}^T > \Gamma_{12}\Gamma_{22}^{-1}\Gamma_{12}^T.$$

Analysis of (A.13) shows that

$$\text{(A.15)} \qquad \operatorname{rank} \Gamma_{12} \leqq \operatorname{rank} B - \operatorname{rank} D.$$

We conclude this Appendix by deriving a general formula for P when B is in canonical form (A.1). Using B given by (A.1), D given by (A.3), and $U = B^{\dagger}$, (3.7) yields

$$\text{(A.16)} \qquad P_0 = -\begin{bmatrix} 0 & D_{21}^T(K^T R_{12} + R_{22}) \\ (R_{12}^T K + R_{22})D_{21} & (R_{12}^T K + R_{22})D_{22} \end{bmatrix},$$

and (3.9) yields

$$\text{(A.17)} \qquad Y = \begin{bmatrix} Y_{11} & 0 \\ 0 & 0 \end{bmatrix},$$

where Y_{11} is any real symmetric matrix. From the expression for R_{12} in (A.7),

$$K = -R_{11}^{-1}R_{12} + R_{11}^{-1}V_{11}\Gamma_{12}V_{22}^T.$$

Postmultiplying K by D_{21} and D_{22}, and using (A.9) gives

$$KD_{21} = -R_{11}^{-1}R_{12}D_{21}, \qquad KD_{22} = R_{11}^{-1}R_{12}D_{22}.$$

Thus (A.16) becomes

$$P_0 = -\begin{bmatrix} 0 & D_{21}^T(R_{22} - R_{12}^TR_{11}^{-1}R_{12}) \\ (R_{22} - R_{12}^TR_{11}^{-1}R_{12})D_{21} & (R_{22} - R_{12}^TR_{11}^{-1}R_{12})D_{22} \end{bmatrix}.$$

By defining R_0 as

(A.18) $$R_0 = R_{22} - R_{12}^TR_{11}^{-1}R_{12},$$

the general solution $P = P_0 + Y$, with Y given by (A.17), is

(A.19) $$P = \begin{bmatrix} P_{11} & -D_{21}^TR_0 \\ -R_0D_{21} & -R_0D_{22} \end{bmatrix},$$

where P_{11} is any real symmetric matrix. The term $-R_0D_{22}$ in (A.19) is symmetric (as expected), because we find that

(A.20) $$R_0 = V_{22}(\Gamma_{22} - \Gamma_{12}^T\Gamma_{11}^{-1}\Gamma_{12})V_{22}^T,$$

whence, using (A.13), we have

$$R_0D_{22} = V_{22}\Gamma_{22}V_{22}^TD_{22} = V_{22}\Gamma_{22}\Lambda_2V_{22}^T.$$

The conditions for $P \geqq 0$ and $P > 0$ can be obtained in terms of the partitioned blocks of P, from Theorem 4.1 or directly from (A.19) by the results in [5]. Corresponding to conditions (4.1), (4.2) and (4.3), we find that *P given by* (A.19) *is nonnegative definite if and only if*

(A.21) $$\operatorname{rank}[D_{21} \quad D_{22}] = \operatorname{rank} D_{22},$$

(A.22) *all eigenvalues of D_{22} are nonpositive,*

and

(A.23) $$P_{11} \geqq -D_{21}^TR_0D_{22}^{\#}D_{21},$$

where $D_{22}^{\#}$ is any matrix such that $D_{22}D_{22}^{\#}D_{22} = D_{22}$ and $D_{22}^{\#}D_{22}D_{22}^{\#} = D_{22}^{\#}$. For $P > 0$ it is necessary and sufficient that:

(A.24) *all eigenvalues of D_{22} are negative,*

and

(A.25) $$P_{11} > -D_{21}^TR_0D_{22}^{-1}D_{21}.$$

REFERENCES

[1] R. E. KALMAN, *When is a linear control system optimal?*, Trans. ASME Ser. D, J. Basic Engrg., 86 (1964), pp. 51–60.

[2] B. D. O. ANDERSON, *The inverse problem of optimal control*, Rep. SEL-66-038 (Tr. No. 6560–3), Stanford Electronics Laboratories, Stanford, Calif., 1966. See also *Linear Optimal Control*, B. D. O. Anderson and J. B. Moor, Prentice-Hall, Englewood Cliffs, N.J., 1971.

[3] E. KREINDLER AND J. K. HEDRICK, *On equivalence of quadratic loss functions*, Internat. J. Control, 11 (1970), pp. 213–222.

[4] E. KREINDLER AND A. JAMESON, *Optimality of linear control systems*, IEEE Trans. Automatic Control, AC-17 (1972), pp. 349–351.

[5] ———, *Conditions for nonnegativeness of partitioned matrices*, Ibid., AC-17 (1972), pp. 147–148.

[6] J. C. WILLEMS, *Least squares stationary optimal control and the algebraic Riccati equation*, Ibid., AC-16 (1971), pp. 621–634.

[7] B. P. MOLINARI, *Aspects of the constrained regulator problem*, IEEE trans. Automatic Control, to appear.

SIAM J. CONTROL
Vol. 11, No. 1, February 1973

OPEN-LOOP APPROXIMATION OF DIFFERENTIAL GAMES*

RONALD J. STERN†

Abstract. A linear differential game of fixed duration where the players are restricted to compact control sets is approximated in value by a game in which the players may choose any square-integrable functions as admissible controls. This is done by appending penalty terms to the payoff. The general approximation result leads to a computational method of estimating the value of a certain class of games in which the approximating games have open-loop solutions.

1. Introduction. The approach to differential game theory employed in this paper is that of A. Friedman [2]. The purpose is to give a procedure for computing the value of a certain class of linear differential games of fixed duration.

Consider the following system of differential equations:

$$\dot{x} = A(t)x(t) + B(t)y(t) + C(t)z(t), \qquad t_0 \leqq t \leqq T_0, \tag{1.1}$$

with initial condition

$$x(t_0) = x_0. \tag{1.2}$$

Here $y(t)$, which is the control of the player y, is a measurable function valued almost everywhere in Y, a given compact subset of the Euclidean space R^s. Similarly, $z(t)$, the control of the player z, is a measurable function valued almost everywhere in Z, a given compact subset of the Euclidean space R^q.

We shall require the following assumptions on the dynamics (1.1), (1.2):

(D) $A(t)$, $B(t)$ and $C(t)$ are continuous $m \times m$, $m \times s$ and $m \times q$ matrices, respectively, on the real line.

Condition (D) implies that the system (1.1), (1.2) has a unique solution for each pair of controls $(y(t), z(t))$.

The following payoff is now introduced:

$$\begin{aligned} P(y, z) = \alpha|x(T_0)|^2 &+ \int_{t_0}^{T_0} p(t)|x(t) - \tilde{x}(t)|^2\,dt \\ &- \int_{t_0}^{T_0} |y(t)|^2\,dt + \int_{t_0}^{T_0} |z(t)|^2\,dt. \end{aligned} \tag{1.3}$$

The goal of y is to maximize $P(y, z)$, whereas z tries to minimize it. The following assumptions are made regarding the payoff:

(P) $p(t) \in C^{0,1}(R^1)$ and $\tilde{x}(t) \in C^{0,m}(R^1)$.

Conditions (D) and (P) imply $P(y, z)$ is well-defined for each pair $(y = y(t), z = z(t))$.

* Received by the editors December 21, 1971.

† Department of Business Administration, University of Illinois, Urbana, Illinois 61801. This research was part of the author's doctoral dissertation in Applied Mathematics at Northwestern University.

The differential game associated with (1.1), (1.2) and (1.3) is denoted by $G(0,0)$. If (D) and (P) hold, then it is known [2] that $G(0,0)$ has value, which is denoted by $V(0,0)$.

Let $\rho(m, M)$ denote the Euclidean distance of a point $m \in R^p$ to a subset M of R^p. Let ε and ξ be positive real numbers, and introduce the following payoff functional:

$$(1.4)\qquad P_{\varepsilon,\xi}(y,z) = P(y,z) - \frac{1}{\varepsilon}\int_{t_0}^{T_0} \rho^2(y(t), Y)\,dt + \frac{1}{\xi}\int_{t_0}^{T_0} \rho^2(z(t), Z)\,dt.$$

Consider the differential game associated with (1.1), (1.2) and (1.4), where the players are no longer restricted to compact control sets but may choose any square-integrable functions as admissible controls. This game is denoted $G(\varepsilon, \xi)$. Let $V^\delta(\varepsilon, \xi)$ denote the upper δ-value of $G(\varepsilon, \xi)$ as defined in [2]. In § 2 it is proven that $V^\delta(\varepsilon, \xi) \to V^\delta(0,0)$ as $\varepsilon \to 0$, $\xi \to 0$, at a rate uniform in δ, where $V^\delta(0,0)$ is the upper δ-value of $G(0,0)$. Thus $V^\delta(\varepsilon, \xi)$ provides an estimate of $V(0,0)$ for small ε, ξ and δ. This approximation result extends to differential games with more general payoffs than (1.3), as is indicated.

In § 3 we take Y and Z to be unit balls. Instead of (1.4) we use the following payoff in the approximating game:

$$(1.5)\qquad P_k(y,z) = P(y,z) - \int_{t_0}^{T_0} |y(t)|^{2k}\,dt + \int_{t_0}^{T_0} |z(t)|^{2k}\,dt,$$

where k is a positive integer. The differential game associated with (1.1), (1.2) and (1.5), where the players are free to choose any square-integrable control functions, is denoted by G_k. The upper δ-value of G_k is denoted by V_k^δ. We prove that $V_k^\delta \to V^\delta(0,0)$ as $k \to \infty$ at a rate uniform in δ, by extending the result of § 2.

In § 4 we prove that G_k has a unique open-loop solution for each k. In § 5 a computational procedure is given for estimating the value of G_k, which in turn is an estimate of $V(0,0)$, for large k.

2. The general approximation theorem. We shall require the following lemma.

LEMMA 2.1. *Let* (D) *and* (P) *hold. Then there is a positive constant* η *such that* $T_0 - t_0 \leqq \eta$ *implies all of the following*:

(i) $\sup_{y \in L^{2,s}(t_0,T_0)} P(y,z)$ *exists for each* $z \in L^{2,q}(t_0, T_0)$;

(ii) $\inf_{z \in L^{2,q}(t_0,T_0)} P(y,z)$ *exists for each* $y \in L^{2,s}(t_0, T_0)$;

(iii) *there exists an* $N > 0$ *such that* $P(y,z) < N$ *for every* $y \in L^{2,s}(t_0, T_0)$ *and measurable* z *valued almost everywhere in* Z;

(iv) $P(y,z) > -N$ *for every* $z \in L^{2,q}(t_0, T_0)$ *and measurable* y *valued almost everywhere in* Y.

The proof of Lemma 2.1 is a computation which makes use of the form of the solutions to (1.1), (1.2) (see [1] and [2]), and repeated applications of Hölder's inequality.

In [2] Friedman defines the concept of a δ-game both for the case of compact control sets, and also for square-integrable function space control sets. Here $\delta = (T_0 - t_0)/n$, where n is a positive integer. Due to Lemma 2.1, an important property of δ-games of the former type generalizes to the latter; namely,

$$(2.1)\qquad V^\delta(\varepsilon, \xi) = \inf_{\Delta_\delta}\sup_{\Gamma^\delta} P_{\varepsilon,\xi}(\Delta_\delta, \Gamma^\delta) = \sup_{\Gamma^\delta}\inf_{\Delta_\delta} P_{\varepsilon,\xi}(\Delta_\delta, \Gamma^\delta).$$

Here Δ_δ denotes a lower δ-strategy for the player z, and Γ^δ denotes an upper δ-strategy for y, as defined in [2]. The outcome of a pair $(\Delta_\delta, \Gamma^\delta)$ is denoted (y^δ, z_δ), and the resultant path is denoted x^δ.

Consider the following payoff:

$$P_{\varepsilon,0}(y, z) = P(y, z) - \frac{1}{\varepsilon}\int_{t_0}^{T_0} \rho^2(y(t), Y)\,dt. \tag{2.2}$$

$G(\varepsilon, 0)$ denotes the differential game associated with (1.1), (1.2) and (2.2). An admissible control for y is as in $G(\varepsilon, \xi)$, but z remains restricted to Z as in $G(0, 0)$. A result similar to (2.1) also holds for $G(\varepsilon, 0)$.

Lemma 2.1 and (2.1) yield the following fact.

LEMMA 2.2. *Let* $T_0 - t_0 \leqq \eta$ *and assume* (D) *and* (P) *hold. Then for every* $\varepsilon > 0$, $\xi > 0$ *and each* δ *we have*

$$-N < V^\delta(\varepsilon, \xi) \leqq V^\delta(\varepsilon, 0) < N, \tag{2.3}$$

where $V^\delta(\varepsilon, 0)$ *denotes the upper* δ*-value of* $G(\varepsilon, 0)$ *and* N *is as in Lemma* 2.1.

We now shall define certain classes of δ-strategies. Let R be a nonnegative real number.

$C_y^\delta(R)$ is the class of upper δ-strategies Γ^δ for y such that if Δ_δ is any lower δ-strategy for z then the y-outcome of the pair $(\Delta_\delta, \Gamma^\delta)$ satisfies

$$\int_{t_0}^{T_0} \rho^2(y^\delta(t), Y)\,dt \leqq R.$$

Note that the y-outcomes of members of $C_y^\delta(0)$ are measurable functions valued almost everywhere in Y.

Analogously to $C_y^\delta(R)$, we define $C_\delta^z(R)$ as the class of lower δ-strategies Δ_δ for z such that if Γ^δ is any upper δ-strategy for y then the z-outcome of $(\Delta_\delta, \Gamma^\delta)$ satisfies

$$\int_{t_0}^{T_0} \rho^2(z_\delta(t), Z)\,dt \leqq R.$$

We shall require the following two lemmas.

LEMMA 2.3. *Let* $T_0 - t_0 \leqq \eta$ *and assume* (D) *and* (P) *hold. Then for each* $\varepsilon > 0$ *and each* δ *the following holds*:

$$V^\delta(\varepsilon, 0) = \sup_{\Gamma^\delta \in C_y^\delta(2N\varepsilon)} \inf_{\Delta_\delta \in C_\delta^z(0)} P_{\varepsilon,0}(\Delta_\delta, \Gamma^\delta). \tag{2.4}$$

LEMMA 2.4. *Let* $T_0 - t_0 \leqq \eta$ *and assume* (D) *and* (P) *hold. There exists a positive constant* M *such that for each* $\xi > 0$, $\varepsilon \in (0, 1]$ *and each* δ *the following holds*:

$$\inf_{\Delta_\delta} \sup_{\Gamma^\delta \in C_y^\delta(2N\varepsilon)} P_{\varepsilon,\xi}(\Delta_\delta, \Gamma^\delta) = \inf_{\Delta_\delta \in C_\delta^z(M\varepsilon)} \sup_{\Gamma^\delta \in C_y^\delta(2N\varepsilon)} P_{\varepsilon,\xi}(\Delta_\delta, \Gamma^\delta). \tag{2.5}$$

Proof of Lemma 2.3. Let $\gamma > 0$ be given, and let $\tilde{\Gamma}^\delta$ be such that

$$V^\delta(\varepsilon, 0) \leqq \inf_{\Delta_\delta \varepsilon C_\delta^z(0)} P_{\varepsilon,0}(\Delta_\delta, \tilde{\Gamma}^\delta) + \gamma.$$

Lemma 2.1 implies that if γ is sufficiently small the y-outcomes of $(\Delta_\delta, \tilde{\Gamma}^\delta)$ satisfy

$$\int_{t_0}^{T_0} \rho^2(\tilde{y}^\delta(t), Y)\,dt < 2N\varepsilon. \tag{2.6}$$

Now $\tilde{\Gamma}^\delta$ will be modified into a member of $C_y^\delta(2N\varepsilon)$, which we shall denote by $\hat{\Gamma}^\delta$. Let $z \in L^{2,q}(t_0, T_0)$. Denote the y-outcome of $(z, \tilde{\Gamma}^\delta)$ as $\tilde{\Gamma}^\delta(z)$. Suppose there is a $\hat{t} \in (t_0, T_0)$ such that

$$\int_{t_0}^{\hat{t}} \rho^2(\tilde{\Gamma}^\delta(z), Y)\,dt = 2N\varepsilon.$$

Then define $\hat{\Gamma}^\delta(z) = \tilde{\Gamma}^\delta(z)$ for $t_0 \leqq t \leqq \hat{t}$, and let $\hat{\Gamma}^\delta(z)$ be any point in Y for $\hat{t} < t \leqq T_0$. Upon varying z over the space $L^{2,q}(t_0, T_0)$, we see that $\hat{\Gamma}^\delta$ is indeed a member of $C_y^\delta(2N\varepsilon)$. Furthermore, the y-outcome of $(\Delta_\delta, \tilde{\Gamma}^\delta)$ coincides with the y-outcome of $(\Delta_\delta, \hat{\Gamma}^\delta)$ for each $\Delta_\delta \in C_\delta^z(0)$. This proves (2.4), since γ was arbitrary.

Proof of Lemma 2.4. Let $Q = \sup\{|y| : y \in Y\}$. All y-outcomes of pairs $(\Delta_\delta, \Gamma^\delta)$ where $\Gamma^\delta \in C_y^\delta(2N\varepsilon)$ satisfy

$$(2.7) \qquad \int_{t_0}^{T_0} |y^\delta(t)|^2\,dt \leqq 2N + Q^2(T_0 - t_0) + 2Q(2N)^{1/2}(T_0 - t_0)^{1/2}.$$

Similarly to Lemma 2.1 we have a $W > 0$ such that

$$(2.8) \qquad -W < \inf_{z \in L^{2,q}(t_0, T_0)} P(y, z) \quad \text{for every} \quad y \in L^{2,s}(t_0, T_0)$$

such that (2.7) holds.

Given $\gamma > 0$ sufficiently small there exists $\tilde{\Delta}_\delta$ such that

$$(2.9) \qquad \sup_{\Gamma^\delta \in C_y^\delta(2N\varepsilon)} P_{\varepsilon,\xi}(\tilde{\Delta}_\delta, \Gamma^\delta) \leqq \inf_{\Delta_\delta} \sup_{\Gamma^\delta \in C_y^\delta(2N\varepsilon)} P_{\varepsilon,\xi}(\Delta_\delta, \Gamma^\delta) + \gamma < N.$$

The remainder of the proof resembles the proof of Lemma 2.3. We can take $M = 3N + W$.

From Lemma 2.4 and (2.1) we obtain the following inequality:

$$(2.10) \qquad \inf_{\Delta_\delta \in C_\delta^z(M\xi)} \sup_{\Gamma^\delta \in C_y^\delta(2N\varepsilon)} P_{\varepsilon,0}(\Delta_\delta, \Gamma^\delta) \leqq V^\delta(\varepsilon, \xi)$$

provided $T_0 - t_0 \leqq \eta$. (Note that the z-outcomes here are not necessarily valued in Z almost everywhere. Nevertheless, $P_{\varepsilon,0}$ by its definition implies that z receives no penalty for leaving Z.)

We can now prove the following lemma.

LEMMA 2.5. *Let* $T_0 - t_0 \leqq \eta$ *and assume* (D) *and* (P) *hold. Then there is a positive constant F such that for each* $\delta = (T_0 - t_0)/n$, $\varepsilon \in (0, 1]$, $\xi \in (0, 1]$,

$$(2.11) \qquad V^\delta(\varepsilon, \xi) \geqq V^\delta(\varepsilon, 0) - F\xi^{1/2}.$$

Proof. By Lemma 2.3 and (2.10) it follows that it suffices to prove

$$(2.12) \qquad \inf_{\Delta_\delta \in C_\delta^z(M\xi)} \sup_{\Gamma^\delta \in C_y^\delta(2N\varepsilon)} P_{\varepsilon,0}(\Delta_\delta, \Gamma^\delta) \geqq \sup_{\Gamma^\delta \in C_y^\delta(2N\varepsilon)} \inf_{\Delta_\delta \in C_\delta^z(0)} P_{\varepsilon,0}(\Delta_\delta, \Gamma^\delta) - F\xi^{1/2}.$$

Let p be any point in R^q. Let $Z(p)$ denote the subset of Z such that

$$\rho(p, Z) = \rho(p, Z(p)).$$

We now shall define a lexicographic ordering on R^q. We say

$$(\hat{u}_1, \hat{u}_2, \cdots, \hat{u}_q) \ll (u_1, u_2, \cdots, u_q)$$

if either $\hat{u}_1 < u_1$, or $\hat{u}_1 = u_1, \cdots, \hat{u}_{s-1} = u_{s-1}$, $\hat{u}_s < u_s$ for some $s = 2, \cdots, q$.

Via this ordering it follows that there is a unique point $\hat{p} \in Z(p)$ such that $\hat{p} \ll w$ for all $w \in Z(p)$. Thus we have defined a single-valued map "$\hat{}$" from R^q onto Z, taking p into $\hat{p}$.

A slight modification in the proof of Fillipov's lemma [2] yields the following result.

LEMMA 2.6. *Let $z(t)$ be any measurable function on $[t_0, T_0]$ valued in R^q. Then $\hat{z}(t)$ is also measurable.*

We are now in the position to define a certain mapping L from the space of lower δ-strategies for z onto $C^z_\delta(0)$. Let $y_1 = (y_1, y_2, \cdots, y_n)$ be any member of $L^{2,s}(t_0, T_0)$, and let Δ_δ be any lower δ-strategy for z. $L(\Delta_\delta) = \hat{\Delta}_\delta$ is defined as follows:

$$\hat{\Delta}_{\delta,1} = \hat{z}_1, \quad \text{where} \quad z_1 = \Delta_{\delta,1},$$

$$\hat{\Delta}_{\delta,2}(\hat{z}_1, y_1) = \hat{z}_2, \quad \text{where} \quad z_2 = \Delta_{\delta,2}(z_1, y_1),$$

$$\hat{\Delta}_{\delta,j}(\hat{z}_1, y_1, \hat{z}_2, y_2, \cdots, \hat{z}_{j-1}, y_{j-1}) = \hat{z}_j,$$

where $z_j = \Delta_{\delta,j}(z_1, y_1, z_2, y_2, \cdots, z_{j-1}, y_{j-1})$, $2 < j \leqq n$, and

$$\hat{\Delta}_\delta = (\hat{\Delta}_{\delta,1}, \hat{\Delta}_{\delta,2}, \cdots, \hat{\Delta}_{\delta,n}).$$

Let Δ_δ be any lower δ-strategy for z, and y any control for player y. Denote the z-outcome of (Δ_δ, y) by z_δ, and denote the z-outcome of $(\hat{\Delta}_\delta, y)$ by $\hat{\Delta}_\delta(y)$. By the way in which $\hat{\Delta}_\delta$ is defined we have that

$$\hat{z}_\delta = \hat{\Delta}_\delta(y).$$

Let $\gamma > 0$ be given. There exists $\tilde{\Delta}_\delta \in C^z_\delta(M\xi)$ such that

$$\sup_{\Gamma^\delta \in C^\delta_y(2N\varepsilon)} P_{\varepsilon,0}(\tilde{\Delta}_\delta, \Gamma^\delta) \leqq \inf_{\Delta_\delta \in C^z_\delta(M\xi)} \sup_{\Gamma^\delta \in C^\delta_y(2N\varepsilon)} P_{\varepsilon,0}(\Delta_\delta, \Gamma^\delta) + \gamma, \tag{2.13}$$

and there exists $\tilde{\Gamma}^\delta \in C^\delta_y(2N\varepsilon)$ such that

$$\inf_{\Delta_\delta \in C^z_\delta(0)} P_{\varepsilon,0}(\Delta_\delta, \tilde{\Gamma}^\delta) \geqq \sup_{\Gamma^\delta \in C^\delta_y(2N\varepsilon)} \inf_{\Delta_\delta \in C^z_\delta(0)} P_{\varepsilon,0}(\Delta_\delta, \Gamma^\delta) - \gamma. \tag{2.14}$$

In view of (2.12)–(2.14) and the arbitrariness of γ, the proof of Lemma 2.5 will be completed if we can show that

$$P_{\varepsilon,0}(\tilde{\Delta}_\delta, \tilde{\Gamma}^\delta(\hat{\tilde{\Delta}}_\delta)) \geqq P_{\varepsilon,0}(\hat{\tilde{\Delta}}_\delta, \tilde{\Gamma}^\delta) - F\xi^{1/2}. \tag{2.15}$$

Here $\hat{\tilde{\Delta}}_\delta = L(\tilde{\Delta}_\delta)$, and $\tilde{\Gamma}^\delta(\hat{\tilde{\Delta}}_\delta)$ denotes the y-outcome of the pair $(\hat{\tilde{\Delta}}_\delta, \tilde{\Gamma}^\delta)$.

Let us denote the outcome of $(\tilde{\Delta}_\delta, \tilde{\Gamma}^\delta(\hat{\tilde{\Delta}}_\delta))$ by $(\tilde{z}_\delta, \tilde{y}_\delta)$, and the outcome of $(\hat{\tilde{\Delta}}_\delta, \tilde{\Gamma}^\delta)$ by $(\hat{\tilde{z}}_\delta, \tilde{y}_\delta)$. The corresponding trajectories for these pairs will be denoted by $\tilde{x}^\delta$ and $\hat{\tilde{x}}^\delta$ respectively. Thus (2.15) will hold if a constant F can be found such that

$$\left| |\tilde{x}^\delta(T_0)|^2 - |\hat{\tilde{x}}^\delta(T_0)|^2 \right| \leqq F\xi^{1/2}, \tag{2.16}$$

$$\left| \int_{t_0}^{T_0} p(t)|\tilde{x}^\delta(t) - \tilde{x}(t)|^2 \, dt - \int_{t_0}^{T_0} p(t)|\hat{\tilde{x}}^\delta(t) - \tilde{x}(t)|^2 \, dt \right| \leqq F\xi^{1/2} \tag{2.17}$$

and

$$\left| \int_{t_0}^{T_0} |\tilde{z}_\delta(t)|^2 \, dt - \int_{t_0}^{T_0} |\hat{\tilde{z}}_\delta(t)|^2 \, dt \right| \leqq F\xi^{1/2}. \tag{2.18}$$

Let $X(\varepsilon, \xi)$ denote the space of trajectories generated by pairs $(\Delta_\delta, \Gamma^\delta)$ as Δ_δ varies in $C_\delta^z(M\xi)$ and Γ^δ varies in $C_y^\delta(2N\varepsilon)$. If ε and ξ are both in $(0, 1]$, then one can show that $X(\varepsilon, \xi)$ has a uniform bound in the sense of $C^{0,m}[t_0, T_0]$. Relations (2.16) and (2.17) follow from this fact and from the following:

$$\int_{t_0}^{T_0} |\tilde{z}_\delta(t) - \hat{\tilde{z}}_\delta(t)|\, dt = \int_{t_0}^{T_0} \rho(\tilde{z}_\delta(t), Z)\, dt \leqq (M\xi)^{1/2}(T_0 - t_0)^{1/2}. \tag{2.19}$$

Relation (2.18) follows from (2.19) alone. This completes the proof.

LEMMA 2.7. *Let $T_0 - t_0 \leqq \eta$ and assume* (D) *and* (P) *hold. Then there is a constant H such that for each $\delta = (T_0 - t_0)/n$ and $\varepsilon \in (0, 1]$ we have*

$$|V^\delta(\varepsilon, 0) - V^\delta(0, 0)| \leqq H\varepsilon^{1/2}. \tag{2.20}$$

Proof. The proof is quite similar to the proof of Lemma 2.5. Details will not be given. Here one verifies that

$$\sup_{\Gamma^\delta \in C_y^\delta(2N\varepsilon)} \inf_{\Delta_\delta \in C_\delta^z(0)} P(\Delta_\delta, \Gamma^\delta) \leqq \inf_{\Delta_\delta \in C_\delta^z(0)} \sup_{\Gamma^\delta \in C_y^\delta(0)} P(\Delta_\delta, \Gamma^\delta) + H\varepsilon^{1/2} \tag{2.21}$$

in a fashion similar to the proof of (2.12).

We can now state the following general approximation theorem.

THEOREM 2.1. *Let $T_0 - t_0 \leqq \eta$ and assume* (D) *and* (P) *hold. Then there is a positive constant D such that for each δ and each $\varepsilon \in (0, 1]$, $\xi \in (0, 1]$ the following holds*:

$$|V^\delta(\varepsilon, \xi) - V(0, 0)| \leqq D(\varepsilon^{1/2} + \xi^{1/2} + \delta). \tag{2.22}$$

Proof. From Friedman's proof of existence of value for games like $G(0, 0)$ it follows that for a positive constant C we have $|V^\delta(0, 0) - V(0, 0)| \leqq C\delta$. Lemmas 2.5 and 2.7 then give (2.22).

Remark 1. In $G(\varepsilon, \xi)$ replace ρ by any function q which is 0 on the control sets Y and Z but which dominates $c\rho$ everywhere, $c > 0$. It follows that Theorem 2.1 generalizes to this case provided that q is measurable.

Remark 2. Consider the payoff

$$P(y, z) = g(x(T_0)) + \int_{t_0}^{T_0} h(t, x, y, z)\, dt, \tag{2.23}$$

where h is continuous on $[t_0, T_0] \times R^m \times Y \times Z$ and g is continuous on R^m. The general approximation theorem can be extended to the case where the payoff for $G(0, 0)$ is given by (2.23) and the dynamics are given by (1.1), (1.2) provided the following assumptions hold:

(i) g is uniformly Lipschitz continuous on compact subsets of R^m;

(ii) for any compact subset X of R^m the function h is uniformly Lipschitz continuous in x on $[t_0, T_0] \times X \times R^2 \times R^q$;

(iii) for any compact subset X of R^m the function h is uniformly Lipschitz continuous in $y(z)$ on $[t_0, T_0] \times X \times R^s \times Z$ $([t_0, T_0] \times X \times Y \times R^q)$;

(iv) for each $y \in L^{2,s}(t_0, T_0)$ and $z \in L^{2,q}(t_0, T_0)$ the numbers $\inf_{z \in L^{2,q}(t_0, T_0)} P(y, z)$ and $\sup_{y \in L^{2,s}(t_0, T_0)} P(y, z)$ are finite.

3. Polynomial penalty functions. In this section we shall specialize the control sets Y and Z, and use an approximating game different from $G(\varepsilon, \xi)$, namely G_k, which was introduced in § 1.

The control sets for $G(0, 0)$ are now given by

$$Y = \{y \in R^s : |y| \leqq 1\}, \quad Z = \{z \in R^q : |z| \leqq 1\}.$$

THEOREM 3.1. *Let* $T_0 - t_0 \leqq \eta$ *and let* (D) *and* (P) *hold. Then there is a positive constant* $\bar{D}$ *such that for each* δ *and each positive integer* k *the following holds*:

$$|V_k^\delta - V(0,0)| \leqq \bar{D}(1 - k^{-1/2k} + k^{-1/2} + \delta). \tag{3.1}$$

Proof. Define the following sets:

$$Y_k = \{y \in R^s : |y| \leqq k^{-1/2k}\}, \qquad Z_k = \{z \in R^q : |z| \leqq k^{-1/2k}\}.$$

Let $G_k(0, 0)$ denote the differential game associated with (1.1)–(1.3), where y and z choose controls valued almost everywhere in Y_k and Z_k respectively. $V_k^\delta(0, 0)$ and $V_k(0, 0)$ denote the upper δ-value and value of $G_k(0, 0)$ respectively.

Now define the following functions:

$$\phi_k(y) = \begin{cases} 0 & \text{if } |y| \leqq k^{-1/2k}, \\ |y|^{2k} & \text{otherwise,} \end{cases}$$

$$\psi_k(z) = \begin{cases} 0 & \text{if } |z| \leqq k^{-1/2k}, \\ |z|^{2k} & \text{otherwise.} \end{cases}$$

Define the following payoff functional:

$$P_{\phi_k, \psi_k}(y, z) = P(y, z) - \int_{t_0}^{T_0} \phi_k(y(t))\, dt + \int_{t_0}^{T_0} \psi_k(z(t))\, dt. \tag{3.2}$$

The differential game associated with (1.1), (1.2), (3.2) where both players choose their controls from square-integrable function spaces is denoted by $G(\phi_k, \psi_k)$. Its upper δ-value is denoted by $V^\delta(\phi_k, \psi_k)$. Even though the integrand of (3.2) is not continuous in the controls, a result similar to (2.1) holds for this game also, since all the results on δ-games in [2] extend to this case.

Define another payoff:

$$P_{\phi_k, 0}(y, z) = P(y, z) - \int_{t_0}^{T_0} \phi_k(y(t))\, dt. \tag{3.3}$$

The differential game associated with (1.1), (1.2), (3.3) where y chooses square-integrable controls but z chooses controls valued almost everywhere in Z_k is denoted by $G(\phi_k, 0)$. Its upper δ-value is denoted $V^\delta(\phi_k, 0)$. A result analogous to (2.1) holds also for this game. Similar to Lemma 2.2 we have, for each positive integer k, the following:

$$-N < V^\delta(\phi_k, \psi_k) \leqq V^\delta(\phi_k, 0) < N. \tag{3.4}$$

Theorem 3.1 will be proven upon verification of the following three inequalities:

$$|V^\delta(\phi_k, \psi_k) - V_k(0,0)| \leqq \bar{\bar{D}}(k^{-1/2} + \delta) \quad \text{for some} \quad D > 0, \tag{3.5}$$

$$|V_k^\delta(\phi_k, \psi_k) - V_k^\delta| \leqq \frac{2(T_0 - t_0)}{k}, \tag{3.6}$$

$$|V_k^\delta(0,0) - V^\delta(0,0)| \leqq S(1 - k^{-1/2k}) \quad \text{for some} \quad S > 0. \tag{3.7}$$

Relation (3.5) follows from the fact that $k^{1/2}\rho(y, Y_k) \leqq \phi_k(y)$ and $k^{1/2}\rho(z, Z_k) \leqq \psi_k(z)$ for all $y \in R^s$, $z \in R^q$ and Remark 1 of § 2.

Relation (3.6) follows easily from the fact that for each $y \in L^{2,s}(t_0, T_0)$ and $z \in L^{2,q}(t_0, T_0)$ we have $|P_{\phi_k,\phi_k}(y, z) - P_k(y, z)| \leqq 2(T_0 - t_0)/k$.

We prove (3.7) as follows: To each pair of controls $(y(t), z(t))$ in $G^\delta(0,0)$ we uniquely correspond a pair of controls for $G_k^\delta(0,0)$, namely, $(k^{-1/2k}y(t), k^{-1/2k}z(t))$. A constant S can be computed such that $|P(y, z) - P(k^{-1/2k}y, k^{-1/2k}z)| \leqq S(1 - k^{-1/2k})$. Using arguments in [2] (where continuity of the value for games of fixed duration is proven) we have that (3.7) holds. This completes the proof of Theorem 3.1.

4. Open-loop solution of G_k.

We shall require the following theorem, adapted from [2].

THEOREM 4.1. *Suppose* $\bar{y}_k \in L^{2,s}(t_0, T_0)$ *and* $\bar{z}_k \in L^{2,q}(t_0, T_0)$ *satisfy*

$$P_k(y, \bar{z}_k) \leqq P_k(\bar{y}_k, \bar{z}_k) \leqq P_k(\bar{y}_k, z) \tag{4.1}$$

for any $y \in L^{2,s}(t_0, T_0)$ *and* $z \in L^{2,q}(t_0, T_0)$. *Then* G_k *has value* V_k *and*

$$V_k = P_k(\bar{y}_k, \bar{z}_k). \tag{4.2}$$

The pair $(\bar{y}_k, \bar{z}_k)$ is referred to as an *open-loop solution* of G_k.

In [2] a variational method is given for computing the (unique) open-loop solution to G_0. We now shall extend the method to G_k, k a positive integer.

A sufficient condition for $(\bar{y}_k, \bar{z}_k)$ to be an open-loop solution of G_k is that the following four conditions hold:

$$\frac{d}{d\gamma} P_k(\bar{y}_k + \gamma w, \bar{z}_k)|_{\gamma=0} = 0 \quad \text{for all} \quad w \in L^{2,s}(t_0, T_0), \tag{4.3}$$

$$\frac{d}{d\gamma} P_k(\bar{y}_k, \bar{z}_k + \gamma v)|_{\gamma=0} = 0 \quad \text{for all} \quad v \in L^{2,q}(t_0, T_0), \tag{4.4}$$

(4.5) $P_k(\bar{y}_k, z)$ is a convex functional on $L^{2,q}(t_0, T_0)$ which is bounded below,

(4.6) $P_k(y, \hat{z})$ is a concave functional on $L^{2,s}(t_0, T_0)$ which is bounded above.

The fact that (4.3)–(4.6) constitute a set of sufficient conditions is due to well-known theorems on global optima.

Conditions (4.3) and (4.4) yield the following system of integral equations:

$$\begin{aligned} 2\bar{y}_k(t) + 2k|\bar{y}_k(t)|^{2k-2}\bar{y}_k(t) &= \alpha B^*(t)S^*(T_0, t)\bar{x}_k(T_0) \\ &\quad + \int_t^{T_0} p(\tau)B^*(t)S^*(t, \tau)[\bar{x}_k(\tau) - \tilde{x}(\tau)]\, d\tau, \end{aligned} \tag{4.7}$$

$$\begin{aligned} 2\bar{z}_k(t) + 2k|\bar{z}_k(t)|^{2k-2}\bar{z}_k(t) &= -\alpha C^*(t)S^*(T_0, t)\bar{x}_k(T_0) \\ &\quad - \int_t^{T_0} p(\tau)C^*(t)S^*(t, \tau)[\bar{x}_k(\tau) - \tilde{x}(\tau)]\, d\tau, \end{aligned} \tag{4.8}$$

where $\bar{x}_k(t)$ is the solution to (1.1)–(1.2).

We now define a certain map of R^s into itself:

$$\mathscr{Y}_k(y) = 2y + 2k|y|^{2k-2}y. \tag{4.9}$$

A map of R^q into itself is similarly given by

$$\mathscr{Z}_k(z) = 2z + 2k|z|^{2k-2}z. \tag{4.10}$$

Notice that the maps given by (4.9) and (4.10) have continuous inverses. The inverse of $\mathscr{Y}_k$ is given by

$$\mathscr{Y}_k^{-1}(w) = \frac{w}{2 + 2k[r_k(|w|)]^{2k-2}},$$

where $r_k(|w|)$ is the unique real root of the polynomial $2kx^{2k-1} + 2x - |w| = 0$. Similarly, the inverse of $\mathscr{Z}_k$ is given by

$$\mathscr{Z}_k^{-1}(v) = \frac{v}{2 + 2k[r_k(|w|)]^{2k-2}}.$$

Using the fact that $\mathscr{Y}_k$ and $\mathscr{Z}_k$ have inverses, we can rewrite the system (4.7)–(4.8) as follows:

$$(\mathscr{Y}_k(\bar{y}_k(t)), \mathscr{Z}_k(\bar{z}_k(t))) = T_k(\mathscr{Y}_k(\bar{y}_k), \ \mathscr{Z}(\bar{z}_k)), \tag{4.11}$$

where T_k is a mapping of $C^{0,s}[t_0, T_0] \times C^{0,q}[t_0, T_0]$ into itself. From this point on we shall denote this space simply by $C^{0,s} \times C^{0,q}$.

THEOREM 4.2. *Let* (D) *and* (P) *hold. If* $T_0 - t_0$ *is sufficiently small, then for each positive integer* k *there exists a unique pair* $(\bar{y}_k, \bar{z}_k)$ *such that conditions* (4.3)–(4.6) *hold.*

Proof. From the solution of G_0 found in [2] we have that if $T_0 - t_0$ is sufficiently small, say $T_0 - t_0 \leqq \bar{\eta}$, the following holds for each positive integer k:

$$\begin{aligned} &\| T_k(\mathscr{Y}_k(y_2), \mathscr{Z}_k(z_2)) - T_k(\mathscr{Y}_k(y_1), \mathscr{Z}_k(z_1))\|_{C^{0,s} \times C^{0,q}} \\ &\qquad \leqq \tfrac{1}{2}\|(y_2, z_2) - (y_1, z_1)\|_{C^{0,s} \times C^{0,q}} \end{aligned} \tag{4.12}$$

for every (y_2, z_2) and (y_1, z_1) in $C^{0,s} \times C^{0,q}$.

We claim that if $T_0 - t_0 \leqq \bar{\eta}$, then each T_k in (4.11) is a contraction. To this end it suffices to prove that

$$\begin{aligned} &\|(y_2, z_2) - (y_1, z_1)\|_{C^{0,s} \times C^{0,q}} \\ &\qquad \leqq \|(\mathscr{Y}_k(y_2), \mathscr{Z}_k(z_2)) - (\mathscr{Y}_k(y_1), \mathscr{Z}_k(z_1))\|_{C^{0,s} \times C^{0,q}} \end{aligned} \tag{4.13}$$

for every (y_2, z_2) and (y_1, z_1) in $C^{0,s} \times C^{0,q}$, and every positive integer k. Relation (4.13) follows from the definitions of the maps $\mathscr{Y}_k$ and $\mathscr{Z}_k$, and the following geometry: If v_1 and v_2 are members of the Euclidean space R^p such that $|v_2| \geqq |v_1|$, and $c_1 \in R^1$, $c_2 \in R^1$ are such that $c_2 \geqq c_1 \geqq 1$, then $|v_2 - v_1| \leqq |c_2v_2 - c_1v_1|$.

Thus, if $T_0 - t_0 \leqq \bar{\eta}$, each T_k has a unique fixed point $(\mathscr{Y}_k(\bar{y}_k), \mathscr{Z}_k(\bar{z}_k))$, the pair $(\bar{y}_k, \bar{z}_k)$ satisfying (4.3), (4.4).

From the solution of G_0 in [2], we have $\bar{\bar{\eta}} > 0$ such that if $T_0 - t_0 \leqq \bar{\bar{\eta}}$, then the following inequalities hold:

$$\frac{d^2}{d\gamma^2}P(y, z + \gamma v)|_{\gamma=0} \geqq 0 \quad \text{for every } z, v \text{ in} \quad L^{2,q}(t_0, T_0) \quad \text{and} \quad y \in L^{2,s}(t_0, T_0), \tag{4.14}$$

$$\frac{d^2}{d\gamma^2}P(y + \gamma w, z)|_{\gamma=0} \leqq 0 \quad \text{for every } y, w \text{ in} \quad L^{2,s}(t_0, T_0) \quad \text{and} \quad z \in L^{2,q}(t_0, T_0). \tag{4.15}$$

Thus $P(\bar{y}_k, z)$ is convex over $L^{2,q}(t_0, T_0)$ and $P(y, \bar{z}_k)$ is concave over $L^{2,s}(t_0, T_0)$. It easily follows that $P_k(\bar{y}_k, z)$ and $P_k(y, \bar{z}_k)$ are respectively convex and concave over these spaces for each positive integer k when $T_0 - t_0 \leqq \bar{\bar{\eta}}$.

The boundedness requirements in (4.5) and (4.6) follow from Lemma 3.1. Upon taking $T_0 - t_0 \leqq \min\{\bar{\bar{\eta}}, \eta\}$ the proof is completed.

Remark. In [4] it is proven that when Y and Z are compact convex sets (as they are in § 4), then $G(0, 0)$ has an open-loop solution $(\bar{y}, \bar{z})$. It is not known, however, how to compute $P(\bar{y}, \bar{z}) = V(0, 0)$. In the next section a computational procedure is given for estimating $V(0, 0)$.

5. Computational procedure. Letting $\delta \to 0$ in (3.1) we have

$$|V_k - V(0,0)| \leqq \bar{D}(1 - k^{-1/2k} + k^{-1/2}) \tag{5.1}$$

when $T_0 - t_0 \leqq \eta$, where η is as in Lemma 2.1. The right side of (5.1) tends to 0 as $k \to \infty$. Thus, upon determining $\bar{D}$ we can pick k sufficiently large to suit the tolerance we have set for our estimate of $V(0, 0)$.

Let (y_0, z_0) be any element of $C^{0,s} \times C^{0,q}$. Consider the following sequence of successive approximations:

$$(\mathscr{Y}_k(y_n), \mathscr{Z}_k(z_n)) = T_k(\mathscr{Y}_k(y_{n-1}), \mathscr{Z}_k(z_{n-1})) = (Q(y^{n-1}), S(z^{n-1})), \quad n = 1, 2, \cdots. \tag{5.2}$$

Assume $T_0 - t_0 \leqq \min\{\eta, \bar{\eta}, \bar{\bar{\eta}}\}$. Then $(\mathscr{Y}_k(y_n), \mathscr{Z}_k(z_n))$ converges in $C^{0,s} \times C^{0,q}$ to $(\mathscr{Y}_k(\bar{y}_k), \mathscr{Z}_k(\bar{z}_k))$, and $P_k(\bar{y}_k, \bar{z}_k) = V_k$ by Theorem 4.1. Note that the maps Q and S which are defined by the equality on the right in (5.2) are independent of k.

Notice that in (5.2) the inverses $\mathscr{Y}_k^{-1}$ and $\mathscr{Z}_k^{-1}$ are employed to generate the successive terms. Unfortunately, these maps do not have explicit representations, due to the fact that the root function r_k, introduced in the previous section, does not have an explicit form for $k > 2$. Our computational procedure circumvents this problem.

Let $\gamma > 0$ be given. Then there is an n_γ such that $n \geqq n_\gamma$ implies

$$\|(\mathscr{Y}_k(y_n), \mathscr{Z}_k(z_n)) - (\mathscr{Y}_k(\bar{y}), \mathscr{Z}_k(\bar{z}))\|_{C^{0,s} \times C^{0,q}} \leqq \gamma. \tag{5.3}$$

In particular, n_γ can be taken to be any n such that

$$(\tfrac{1}{2})^{n-1}\|(\mathscr{Y}_k(y_0), \mathscr{Z}_k(z_0)) - (Q(y_0), S(z_0))\|_{C^{0,s} \times C^{0,q}} \leqq \gamma, \tag{5.4}$$

due to (4.12), (4.13), and the fact that T_k is a contraction mapping.

Also notice that for each $n \geqq 0$ we have

$$\begin{aligned}\|(\mathscr{Y}_k(y_n), \mathscr{Z}_k(z_n)) - (Q(y_0), S(z_0))\|_{C^{0,s}\times C^{0,q}} \\ \leqq 2\|(\mathscr{Y}_k(y_0), \mathscr{Z}_k(z_0)) - (Q(y_0), S(z_0))\|_{C^{0,s}\times C^{0,q}}.\end{aligned} \tag{5.5}$$

Let B denote the closed ball of radius $\|(Q(y_0), S(z_0))\|_{C^{0,s}\times C^{0,q}} + 2\|(\mathscr{Y}_k(y_0), \mathscr{Z}_k(z_0)) - (Q(y_0), S(z_0))\|_{C^{0,s}\times C^{0,q}}$ in the space $C^{0,s} \times C^{0,q}$. In view of (5.5), the sequence (5.2) remains in B for all $n = 0, 1, 2, \cdots$.

We now shall give a computational procedure for estimating the terms of the sequence (5.2), up to stage n_γ.

Let $\varepsilon > 0$ be given. We shall find a pair of functions, $(\tilde{y}_1, \tilde{z}_1)$, such that

$$\|(\tilde{y}_1, \tilde{z}_1) - (y_1, z_1)\|_{C^{0,s}\times C^{0,q}} \leqq \varepsilon. \tag{5.6}$$

Let β be a positive integer. Define

$$t_j = t_0 + j\frac{(T_0 - t_0)}{\beta}, \qquad j = 0, 1, \cdots, \beta.$$

At each t_j compute the values $r_k(|Q(y_0(t_j))|)$ and $r_k(|S(z_0(t_j))|)$. Let $q_1(t)$ be a polynomial which agrees with $r_k(|Q(y_0(t_j))|)$ at each t_j. Similarly, let $s_1(t)$ be a polynomial which agrees with $r_k(|S(z_0(t_j))|)$ at each t_j. (See [3] for a method of accomplishing this.) Since the derivative of r_k is bounded on $[0, \infty)$ by $\frac{1}{2}$, the polynomials $q_1(t)$ and $s_1(t)$ approximate $r_k(|Q(y_0(t))|)$ and $r_k(|S(z_0(t))|)$ uniformly on $[t_0, T_0]$ to an arbitrary tolerance, depending on how large β is chosen to be.

Now define

$$\tilde{y}_1(t) = \frac{Q(y_0(t))}{2 + 2k[q_1(t)]^{2k-2}} \tag{5.7}$$

and

$$\tilde{z}_1(t) = \frac{S(z_0(t))}{2 + 2k[s_1(t)]^{2k-2}}. \tag{5.8}$$

Thus, given the pair $(y_0(t), z_0(t))$, β may be chosen to be sufficiently large so as to guarantee (5.6).

Now define a pair $(\tilde{y}_2', \tilde{z}_2')$ by the relation

$$(\mathscr{Y}_k(\tilde{y}_2'), \mathscr{Z}_k(\tilde{z}_2')) = (Q(\tilde{y}_2), S(\tilde{z}_2)). \tag{5.9}$$

In the same way that we constructed $(\tilde{y}_1, \tilde{z}_1)$, we construct $(\tilde{y}_2, \tilde{z}_2)$ such that

$$\|(\tilde{y}_2, \tilde{z}_2) - (\tilde{y}_2', \tilde{z}_2')\|_{C^{0,s}\times C^{0,q}} \leqq \varepsilon. \tag{5.10}$$

Let $\bar{B}$ be a compact subset of $C^{0,s} \times C^{0,q}$ such that $B \subset \bar{B}$. If ε is sufficiently small, then both $(\tilde{y}_2', \tilde{z}_2')$ and $(\tilde{y}_2, \tilde{z}_2)$ are in $\bar{B}$. Let $O_{\bar{B}}(\cdot)$ denote the modulus of continuity of $\mathscr{Y}_k$ on $\bar{B}$, and let $\bar{O}_{\bar{B}}(\cdot)$ be the modulus of continuity of $\mathscr{Z}_k$ on $\bar{B}$. Let $\bar{\bar{O}}_{\bar{B}}(\cdot) = O_{\bar{B}}(\cdot) + \bar{O}_{\bar{B}}(\cdot)$. Then (5.10) implies

$$\|(\mathscr{Y}_k(\tilde{y}_2), \mathscr{Z}_k(\tilde{z}_2)) - (\mathscr{Y}_k(\tilde{y}_2'), \mathscr{Z}_k(\tilde{z}_2'))\|_{C^{0,s}\times C^{0,q}} \leqq \bar{\bar{O}}_{\bar{B}}(\varepsilon). \tag{5.11}$$

In view of the fact that the maps Q and S are Lipschitz continuous on $C^{0,s}$ and $C^{0,q}$ respectively, we have a positive constant C such that

$$\|(\mathscr{Y}_k(y_2), \mathscr{Z}_k(z_2)) - (\mathscr{Y}_k(\tilde{y}_2), \mathscr{Z}_k(\tilde{z}_2))\|_{C^{0,s} \times C^{0,q}} \leqq \bar{\bar{O}}_B(\varepsilon) + C\varepsilon \tag{5.12}$$

due to (5.6), (5.9) and (5.11).

We continue in this way step by step up to stage n_γ. If ε is sufficiently small (which means β is sufficiently large at each stage), then $(\mathscr{Y}_k(\tilde{y}_n), \mathscr{Z}_k(\tilde{z}_n)) \in \bar{B}$ for each $n \leqq n_\gamma$.

Define the following function:

$$g(\alpha) = C\alpha + C\bar{\bar{O}}_{\bar{B}}(\alpha).$$

We then have

$$\|(\mathscr{Y}_k(y_{n_\gamma}), \mathscr{Z}_k(z_{n_\gamma})) - (\mathscr{Y}_k(\tilde{y}_{n_\gamma}), \mathscr{Z}_k(\tilde{z}_{n_\gamma}))\|_{C^{0,s} \times C^{0,q}} \leqq g^{(n_\gamma - 2)}(\varepsilon), \tag{5.13}$$

where the function on the right is the $(n_\gamma - 2)$-fold composition of g with itself, evaluated at ε. It is clear that $g^{(n_\gamma - 2)}(\varepsilon) \to 0$ as $\varepsilon \to 0$.

Thus, (5.1) and (5.13) yield

$$\begin{aligned} &|P_k(\mathscr{Y}_k(\tilde{y}_{n_\gamma}), \mathscr{Z}_k(\tilde{z}_{n_\gamma})) - V(0,0)| \\ &\qquad \leqq \bar{D}(1 - k^{-1/2k} + k^{-1/2}) + \omega_{\bar{B}}(\gamma + g^{(n_\gamma - 2)}(\varepsilon)), \end{aligned} \tag{5.14}$$

where $\omega_{\bar{B}}$ is the modulus of continuity of P_k on $\bar{B}$.

Remark 1. The method presented in §§4 and 5 extends to more general convex control sets Y and Z than the unit balls. For example, the procedure will generalize directly if

$$Y = \left\{ y \in R^s : \sum_{i=1}^{s} |y_i|^4 \leqq 1 \right\}, \quad Z = \left\{ z \in R^q : \sum_{i=1}^{q} |z_i|^4 \leqq 1 \right\}.$$

In this case one chooses polynomial penalty terms in G_k which are of degree $4k$.

Remark 2. Suppose that we defined the payoff in G_k as follows:

$$P_k(y, z) = P(y, z) - \int_{t_0}^{T_0} e^{k(|y|^2 - 1)}\,dt + \int_{t_0}^{T_0} e^{k(|z|^2 - 1)}\,dt. \tag{5.15}$$

An approximation result similar to Theorem 4.2 may be proven for this case. Upon generalizing the computational method of § 5, one encounters a variant of the difficulty presented by the fact that r_k has no explicit form; that is, the inverse of the function $d(x) = 2x + 2k \times e^{k(x^2 - 1)}$ when $x \in R^1_+$ has no explicit form.

Acknowledgment. The author is indebted to Prof. A. Friedman for his many useful suggestions, including that this area of study be undertaken.

REFERENCES

[1] E. CODDINGTON AND N. LEVINSON, *Theory of Ordinary Differential Equations*, McGraw-Hill, New York, 1955.

[2] A. FRIEDMAN, *Differential Games*, Interscience, New York, 1971.

[3] F. B. HILDEBRAND, *Introduction to Numerical Analysis*, McGraw-Hill, New York, 1956.

[4] D. J. WILSON, *Differential games with no information*, Doctoral dissertation, University of Adelaide, Australia, 1971.

SIAM J. CONTROL
Vol. 11, No. 1, February 1973

UNCONSTRAINED CONTROL PROBLEMS WITH QUADRATIC COST*

RICHARD DATKO†

Abstract. This paper considers a class of control problems where the cost is quadratic, the dynamics are linear and the controls are unbounded. The optimal control is obtained by computing the Fréchet derivative of the cost and setting it equal to the zero vector. In the case of linear autonomous differential-difference equations conditions are found for optimization of the cost over an infinite interval. These lead to feedback controls which stabilize the system.

Introduction. In § 1 and § 2 of this paper we show that a class of linear optimal control problems with quadratic cost functionals can be treated by a reasonably simple and straightforward procedure which reduces the problem to the solution of a Fredholm integral equation. The type of problem studied here was first considered by Kalman [13] for finite-dimensional systems and an excellent treatment of it can be found in the book of Lee and Markus [17]. Extensions of Kalman's work either to a Hilbert space setting or involving linear differential-difference equations can be found in [6], [7], [16], [18], [19] and [20].

The basic idea in this paper is to treat the problem as a variational one and to compute the Fréchet derivative of the quadratic cost functional in a Hilbert space. It is shown that the optimal control is unique and can be obtained by setting the Fréchet derivative equal to the zero vector. The optimal control can then be described as either a linear feedback or as a time-dependent function of the initial point in the phase space. From either viewpoint this leads to the solution of a Fredholm integral equation. Kushner and Barnea [16] have considered a problem similar to ours for differential-difference equations and have reduced it to solving a system of partial differential equations. We believe that our approach is more satisfactory for two reasons: first, we can deal with a more general class of problems and second, the solution of many systems of linear partial differential equations are obtained by solving integral equations.

Using results from § 2 we are able, in the case of linear autonomous differential-difference equations with quadratic cost functionals, to consider optimal control over an infinite interval. We show for these problems that if we can optimize them over the positive half-line, then the optimal control is a linear feedback control which gives rise to an asymptotically stable linear differential-difference equation. This result extends the work of Krasovskii [14] and Ross and Flügge-Lotz [22] since in both these papers the admissible controls are assumed to satisfy conditions which we show must hold a posteriori. In [14] and [22] it is assumed that the admissible controls are feedback controls which stabilize the given system.

Section 4 is essentially an appendix to § 3. Here we obtain a necessary and sufficient condition for the problem in § 3 to be optimized over the positive half-line.

Since § 1 is an extension of the work in [7] to a more general setting and the proofs of the results in this section are essentially the same as in that paper, they will thus only be sketched. The proofs of the results in the remaining sections will

* Received by the editors June 4, 1971, and in revised form February 15, 1972.

† Department of Mathematics, Georgetown University, Washington, D.C. 20007. This work was supported in part by the National Science Foundation under Grant GP 32325.

be given in more detail; however, in some instances routine but technically complicated proofs will be omitted.

Delfour and Mitter [24] have recently obtained results similar to those of § 3 for linear autonomous hereditary differential equations in a Hilbert space setting.

1. We first introduce some preliminary notation.

1. X will denote a Banach space over the real numbers and X^* its topological dual. The norms on both spaces will be denoted by $|\cdot|$ and the continuous linear functionals by the notation $\langle x^*, x\rangle$, where $x^* \in X$ and $x \in X$.

2. H will denote a real Hilbert space with inner product $(\cdot, \cdot)$ and norm $\|\cdot\|$.

3. Let $[0, T] = I$ be a finite closed interval. Then $L_2(I, H)$ will denote the equivalence classes of measurable mappings from I into H which are square integrable. The norm in $L_2(I, H)$ is given by $\|u\| = (\int_0^T \|u(t)\|^2\, dt)^{1/2}$. It is shown in [10] that $L_2(I, H)$ with the inner product $(u, v) = \int_0^T (u(t)v(t))\, dt$ is a Hilbert space.

4. Let X_1 and X_2 be any two real Banach spaces. The Banach space of continuous linear mappings from X_1 into X_2 will be denoted by $L(X_1, X_2)$. In the special case where X_2 is the real line, the notation $X_1^* = L(X_1, R)$ will be used.

The following operators will be used to define the control problem given by (1.1) and (1.2) below.

5. (a) $S(t, s)$ will denote a family of mappings in $L(X, X)$ which are strongly continuous in the triangle $\Delta = \{(t, x): 0 \leqq s \leqq t \leqq T\}$. The adjoint system $S^*(t, s)$ will also be assumed to be strongly continuous on Δ.

(b) The mapping $B: I \to L(H, X)$ will be assumed to be strongly measurable as well as the adjoint mapping $B^*: I \to L(X^*, H)$.

(c) The mapping $W: I \to L(X, X^*)$ will be assumed to be strongly continuous. In addition W will be assumed to be nonnegative and symmetric on I in the sense that for all pairs x, y in X and t in I, $\langle W(t)x, x\rangle \geqq 0$ and $\langle W(t)y, x\rangle = \langle W(t)x, y\rangle$.

(d) The mapping $U: I \to L(H, H)$ will be assumed strongly continuous and symmetric. Moreover, there exist two positive constants m_1 and m_2 such that for all $u \in H$ and t in I, $m_1\|u\|^2 \leqq (U(t)u, u) \leqq m_2\|u\|^2$. The last condition ensures the existence of $U^{-1}(t)$ as a strongly continuous mapping from I into $L(H, H)$.

Statement of the problem. Let x_0 in X be given. The problem is to minimize the functional defined on $L_2(I, H)$ by the equation

$$C(u, x_0) = \int_0^T \langle W(t)x(t, u, x_0), x(t, u, x_0)\rangle\, dt + \int_0^T (U(t)u(t), u(t))\, dt, \tag{1.1}$$

where

$$x(t, u, x_0) = S(t, 0)x_0 + \int_0^t S(t, s)B(s)u(s)\, ds. \tag{1.2}$$

DEFINITION 1.1. If the infimum of the functional $C(\cdot, x_0)$ exists, it will be denoted by $m(x_0)$. A u in $L_2(I, H)$ for which the infimum is attained will be denoted by u^m and the corresponding solution of (1.2) by x^m. The mappings u^m and x^m will be called, respectively, the *optimal control* and *optimal trajectory* for $C(\cdot, x_0)$.

LEMMA 1.1. *For each x_0 in X the infimum $m(x_0)$ of the problem* (1.1)–(1.2) *is finite.*

Proof. Observe that $C(u, x_0) \geqq 0$ for all $u \in L_2(I, H)$. Let $u(t) \equiv 0$ on I. Then since $W(t)$ is strongly continuous, $0 \leqq m(x_0) \leqq C(0, x_0) < +\infty$, which proves the lemma.

LEMMA 1.2. *There exists a unique u^m in $L_2(I, H)$ such that $m(x_0) = C(u^m, x_0)$.*

Proof. The proof is entirely similar to the proof of existence and uniqueness given in Theorem 1 in [7] and is therefore omitted.

THEOREM 1.1. *The optimal control for the problem* (1.1)–(1.2) *is given by the equation*

$$u^m(t) = -U^{-1}(t)B^*(t)\int_t^T S^*(s, t)W(s)x^m(s)\,ds. \tag{1.3}$$

Proof. Let $u \in L_2(I, H)$ be fixed. The Fréchet derivative of $C(u, x_0)$ can be computed as in [7, (10)] to be

$$C'(u, x_0)(t) = 2\left[B^*(t)\int_t^T S^*(s, t)W(t)x(s, u, x_0)\,ds + U(t)u(t)\right]. \tag{1.4}$$

Since u^m is the unique minimal point of the functional $C(\cdot, x_0)$, it follows that $C'(u^m, x_0)(t) = 0$ a.e. on I (see, for example, [4] or [8]). But $U^{-1}(t)$ exists for all $t \in I$. Hence (1.3) follows from (1.4).

The following corollaries are consequences of Theorem 1.1. Their proofs are omitted since they are similar to some numbered corollaries in [7, § 1].

COROLLARY 1.1. *The second Fréchet derivative of the functional $C(\cdot, x_0)$ exists and is a positive definite bilinear functional in $L_2(I, H)$.*

COROLLARY 1.2. *There exists only one point in $L_2(I, H)$ such that $C'(u, x_0) = 0$. Hence the integral equation* (1.4) *and the optimal control have unique solutions.*

COROLLARY 1.3. *Let a and b be scalars. If u^m and v^m are optimal controls for the functionals $C(\cdot, x_0)$ and $C(\cdot, x_1)$ respectively, then $au^m + bv^m$ is the optimal control for the functional $C(\cdot, ax_0 + bx_1)$.*

We define the mapping $(t, x) \to K(t)x$ from $I \times X$ into X^* given by the equation

$$K(t)x_0 = \int_t^T S^*(s, t)W(s)x^m(s, u^m, x_0)\,ds. \tag{1.5}$$

THEOREM 1.2. *The mapping from $I \times X$ into X^* defined by* (1.5) *is for each fixed t in I a linear mapping from X into X^* and $K(0)$ is symmetric in the sense that $\langle K(0)x, y\rangle = \langle K(0)y, x\rangle$ for all x, y in X. Moreover, for each x_0 in X, $\langle K(0)x_0, x_0\rangle = m(x_0)$.*

Proof. The linearity of $K(t)x$ is a consequence of Corollary 1.3. To prove the rest of the theorem, let x_1 and x_2 be fixed in X and let $u(t)$ and $v(t)$ be the optimal controls for $C(\cdot, x_1)$ and $C(\cdot, x_2)$ respectively. By (1.5) and (1.2),

$$\begin{aligned}\langle K(0)x_1, x_2\rangle &= \int_0^T \langle S^*(s,0)W(s)x^m(s,u,x_1), x_2\rangle\,ds \\ &= \int_0^T \langle W(s)x^m(s,u,x_1), S(s,0)x_2\rangle\,ds \\ &= \int_0^T \langle W(s)x^m(s,u,x_1), x^m(s,v,x_2)\rangle\,ds \\ &\quad - \int_0^T\int_0^s \langle W(s)x(s,u,x_1), S(s,\alpha)B(\alpha)v(\alpha)\rangle\,d\alpha\,ds.\end{aligned}\tag{1.6}$$

We apply Fubini's theorem to the multiple integral in (1.6), interchange the order of integration and use (1.3) to obtain

$$\begin{aligned}&\int_0^T\int_0^s \langle W(s)x^m(s,u,x_1), S(s,\alpha)B(\alpha)v(\alpha)\rangle\,d\alpha\,ds \\ &\qquad = \int_0^T d\alpha \int_\alpha^T \langle B^*(\alpha)S^*(s,\alpha)x^m(s,u,x_1), v(\alpha)\rangle\,ds \\ &\qquad = -\int_0^T \langle U(\alpha)u(\alpha), v(\alpha)\rangle\,d\alpha.\end{aligned}\tag{1.7}$$

Substitution of (1.7) into (1.6) leads to the equation

$$\begin{aligned}\langle K(0)x_1, x_2\rangle &= \int_0^T \langle W(s)x^m(s,u,x_1), x^m(s,v,x_2)\rangle\,ds \\ &\quad + \int_0^T (U(s)u(s), v(s))\,ds \\ &= \langle K(0)x_2, x_1\rangle.\end{aligned}\tag{1.8}$$

The last statement of the theorem is a consequence of equation (1.8) when $x_1 = x_2$. This completes the proof of the theorem.

Remark 1.1. The mapping $x_0 \to \langle K(0)x_0, x_0\rangle$ of Theorem 1.2 induces an inner product on X.

Utilizing the form of the optimal control described by (1.3) the optimal trajectory for the problem (1.1)–(1.2) has the form

$$\begin{aligned}x^m(t,u^m,x_0) &= S(t,0)x_0 \\ &\quad - \int_0^t\int_s^T [S(t,s)B(s)U^{-1}(s)B^*(s)S^*(\alpha,s)W(\alpha)x^m(\alpha,u^m,x_0)]\,d\alpha\,ds \\ &= S(t,0)x_0 \\ &\quad - \int_0^t\int_s^t [S(t,s)B(s)U^{-1}(s)B^*(s)S^*(\alpha,s)W(\alpha)x^m(\alpha,u^m,x_0)]\,d\alpha\,ds \\ &\quad - \int_0^t\int_s^T [S(t,s)B(s)U^{-1}(s)B^*(s)S^*(\alpha,s)W(\alpha)x^m(\alpha,u^m,x_0)]\,d\alpha\,ds.\end{aligned}\tag{1.9}$$

Applying Fubini's theorem to both of the multiple integrals in (1.9) we obtain

$$
\begin{aligned}
x^m(t, u^m, x_0) &= S(t, 0)x_0 \\
&\quad - \int_0^t d\alpha \int_0^\alpha [S(t, s)B(s)U^{-1}(s)B^*(s)S^*(\alpha, s)W(\alpha)x^m(\alpha, u^m, x_0)]\, ds \\
&\quad - \int_t^T d\alpha \int_0^t [S(t, s)B(s)U^{-1}(s)B^*(s)S^*(\alpha, s)W(\alpha)x^m(\alpha, u^m, x_0)]\, ds.
\end{aligned}
\tag{1.10}
$$

Let

$$
Q(t, \alpha) = \begin{cases} -\displaystyle\int_0^\alpha S(t, s)B(s)U^{-1}(s)B^*(s)S^*(\alpha, s)W(\alpha)\, ds & \text{if } 0 \leqq \alpha \leqq t, \\[2ex] -\displaystyle\int_0^t S(t, s)B(s)U^{-1}(s)B^*(s)S^*(\alpha, s)W(\alpha)\, ds & \text{if } t \leqq \alpha \leqq T. \end{cases}
\tag{1.11}
$$

Using (1.11) we can state the following theorem.

THEOREM 1.3. *The optimal trajectory for the problem* (1.1)–(1.2) *is the unique solution of the Fredholm integral equation*

$$
x^m(t, x_0) = S(t, 0)x_0 + \int_0^T Q(t, \alpha)x^m(\alpha, x_0)\, d\alpha,
\tag{1.12}
$$

where $Q(t, \alpha)$ *is given by* (1.11).

2. In this section we shall consider an analogue for functional differential equations of the control problem posed in § 1. As in § 1, we shall first introduce some conventions.

1. We shall use vector matrix notation. All vectors and matrices are real. Vectors will be denoted by lower-case letters and matrices by upper-case letters. The transpose of a vector will not be distinguished when it is clear what is meant from the context. In this regard xy will denote the usual inner product of two vectors of the same dimension. The transpose of a matrix A will be denoted by A^*. Thus if A is an $n \times n$ matrix and x and y are n-vectors, $xAy = yA^*x$. In general, unless otherwise specified, the notation will conform to that used in [2] or [11].

2. $A(t, s)$ will denote an $n \times n$ matrix with entries $a_{ij}(t, s)$ such that:

(i) $A(t, s)$ is Lebesgue measurable in (t, s) and for fixed t is of bounded variation in s.

(ii) There is a constant $\tau > 0$ such that $A(t, s) = A(t, -\tau)$ if $s \leqq -\tau$ and $A(t, s) = 0$ if $s \geqq 0$.

(iii) There exists a Lebesgue integrable function $m(t)$ defined on the finite interval $[0, T]$ such that $|a_{ij}(t, s)| \leqq m(t)$ for all s and the variation with respect to s satisfies the inequality

$$
\bigvee_{s=-\tau}^{0} a_{ij}(t, s) \leqq m(t)
$$

for all i, j and t.

3. $B(t)$ will denote an $n \times m$ matrix whose entries are uniformly bounded and measurable on $[0, T]$.

4. $L_2(R^+, R^m)$ will denote the equivalence classes of all measurable square integrable mappings from $[0, \infty)$ into R^m. This is a Hilbert space with inner product $(v, u) = \int_0^\infty v(t)u(t)\,dt$.

5. $W_0(t, s)$ will denote a symmetric $n \times n$ matrix which is measurable and uniformly bounded on $[0, T] \times [-\tau, 0]$ and such that $xW_0(t, s)x \geqq 0$ for all x in R^n and (t, s). $W_i(t)$, $1 \leqq i \leqq k$, will denote $n \times n$ positive semidefinite matrices which are measurable and bounded in $[0, T]$.

6. $U(t)$ will denote an $m \times m$ matrix which is symmetric and measurable and such that for all $u \in R^m$ and $t \in [0, T]$ the inequality $m_1 uu \leqq uU(t)u \leqq m_2 uu$ holds, where m_1 and m_2 are positive constants.

7. $BV[-\tau, 0]$ will denote the Banach space of all mappings $\phi : [-\tau, 0] \to R^n$ which are of bounded variation, and $C[-\tau, 0]$ will denote the Banach space of continuous mappings from $[-\tau, 0]$ into R^n. The norm on $C[-\tau, 0]$ will be given by $|\phi| = \sup_{-\tau \leqq s \leqq 0}(\sum_{i=1}^n (\phi_i(s))^2)^{1/2}$.

Statement of the problem. Let $\phi \in BV[-\tau, 0]$ be given. The problem is to minimize the functional $C(\cdot, \phi)$ defined on $L_2(R^+, R^m)$ by the equation

$$C(u, \phi) = \int_0^T \sum_{i=1}^k [x(t - \tau_i)W_i(t)x(t - \tau_i) + \int_{-\tau}^0 x(t + s)W_0(t, s)x(t + s)\,ds + u(t)U(t)u(t)]\,dt, \tag{2.1}$$

where $0 \leqq \tau_1 \leqq \tau_2 \leqq \cdots \leqq \tau_k \leqq \tau$, and $x(t)$ satisfies the differential-difference equation

$$\mathring{x}(t) = \int_{-\tau}^0 x(t + s)d_s A(t, s) + u(t)B(t) \tag{2.2}$$

with

$$x(t) = \phi(t) \quad \text{on} \quad [-\tau, 0]. \tag{2.3}$$

It has been shown in [1] and [2] that given $u \in L_2(R^+, R^m)$ and $\phi \in BV[-\tau, 0]$ there exists a unique absolutely continuous vector function $x(t)$ which satisfies (2.3) on $[-\tau, 0]$ and (2.2) a.e. on $[0, T]$. Moreover, $x(t)$ has the following analytic representation for $t \geqq 0$:

$$x(t) = \phi(0)S(t, 0) + \int_0^t u(\alpha)B(\alpha)S(t, \alpha)\,d\alpha + \int_{-\tau}^0 \phi(s)\,d_s\left[\int_0^\tau A(\alpha, s - \alpha)S(t, \alpha)\,d\alpha\right], \tag{2.4}$$

where $S(t, \alpha)$ is a unique $n \times n$ matrix which is measurable in (t, α) and of bounded variation in α for t fixed, and conversely. $S(t, \alpha)$ also satisfies the relation (see, for example, [2])

$$\begin{aligned}
&S(t,t) = I && \text{(the identity matrix)},\\
(2.5)\qquad &S(t,\alpha) = 0 \quad \text{if} \quad \alpha > t, &&\\
&S(t,\alpha) + \int_\alpha^t A(\beta, \alpha - \beta)S(t,\beta)\,d\beta = I, && \alpha \in [0,t].
\end{aligned}$$

Remark 2.1. Although we consider the control problem in this section to be over the interval $[0,T]$, this is not restrictive. The main reason for not considering a more general interval of the type $[t_0, T]$ is to keep notational complexity to a minimum. However, all results obtained in § 2 apply equally well to the same problem considered over intervals of the form $[t_0, T]$. This fact will be needed in § 3.

Definition 2.1. If the infimum of the functional $C(\cdot, \phi)$ exists, it will be denoted by $m(\phi)$. A u in $L_2(I)$ for which the infimum is attained will be denoted by u^m and the corresponding solution of (2.2) by x^m. The functions u^m and x^m will be called, respectively, the *optimal control* and *optimal trajectory* for $C(\cdot, \phi)$.

The following lemma is similar to Lemma 1.1, and therefore its proof is omitted.

Lemma 2.1. *For each $\phi \in BV[-\tau, 0]$ the infimum of the problem* (2.1)–(2.2) *is finite.*

Lemma 2.2. *For each $\phi \in BV[-\tau, 0]$ and $u \in L_2(R^+, R^m)$ the Fréchet derivative with respect to u exists and is given a.e. on* $[0, T)$ *by the equation*

$$(2.6)\qquad \begin{aligned}
C'(u,\phi)(t) = 2\Bigg[u(t)U(t) + \Bigg\{\sum_{i=1}^k \int_t^T x(\alpha - \tau_i)W_i(\alpha)S^*(\alpha - \tau_i, t)\,d\alpha\\
+ \int_t^T \int_{-\tau}^0 x(\alpha + s)W_0(\alpha, s)S^*(\alpha + s, t)\,ds\,d\alpha\Bigg\}B^*(t)\Bigg].
\end{aligned}$$

Proof. The existence of the Fréchet derivative is a consequence of the fact that $C(\cdot, \phi)$ is a quadratic functional on $L_2(R^+, R^m)$. Using the fact that $S(t,\alpha) = 0$ if $\alpha > 0$, it is easy to compute for each $h \in L_2(R^+, R^m)$ the identity

$$(2.7)\qquad \begin{aligned}
(C'(u,\phi), h) = 2\Bigg[\sum_{i=1}^k \int_0^T \int_0^t x(t - \tau_i)W_i(t)h(\alpha)B(\alpha)S(t - \tau_i, \alpha)\,d\alpha\,dt\\
+ \int_0^T \int_{-\tau}^0 \int_0^t x(t + s)W_0(t,s)h(\alpha)B(\alpha)S(t + s, \alpha)\,d\alpha\,ds\,dt\\
+ \int_0^T h(t)U(t)u(t)\,dt\Bigg].
\end{aligned}$$

Applying Fubini's theorem to the multiple integrals in (2.6) we obtain

$$(2.8)\qquad \begin{aligned}
(C'(u,\phi), h) = 2\Bigg[\sum_{i=1}^k \int_0^T \int_\alpha^T x(t - \tau_i)W_i(t)S^*(t - \tau_i, \alpha)B^*(\alpha)h(\alpha)\,dt\,d\alpha\\
+ \int_0^T \int_\alpha^T \int_{-\tau}^0 x(t + s)W_0(t,s)S^*(t + s, \alpha)B^*(\alpha)h(\alpha)\,ds\,dt\,d\alpha\\
+ \int_0^T u(\alpha)U(\alpha)h(\alpha)\,d\alpha\Bigg].
\end{aligned}$$

Since (2.8) holds for all $h \in L_2(I)$ it follows that $C'(u, \alpha)$ must be given by (2.6). This completes the proof.

THEOREM 2.1. *The optimal control for the problem described by equations* (2.1)–(2.3) *is given by*

$$\begin{aligned} u^m(t) = -\Bigg[\sum_{i=1}^{k} \int_t^T x^m(\alpha - \tau_i) W_i(\alpha) S^*(\alpha - \tau_i, t)\, d\alpha \\ + \int_t^T \int_{-\tau}^0 x^m(\alpha + s) W_0(\alpha, s) S^*(\alpha + s, t)\, ds\, d\alpha \Bigg] B^*(t) U^{-1}(t). \end{aligned} \tag{2.9}$$

Proof. The optimal control satisfies the condition $C'(u, \phi)(t) = 0$ a.e. on $[0, T]$. Thus (2.9) is an immediate consequence of (2.6).

The following corollaries are stated without proof since the proofs are similar to those given in [7, § 1].

COROLLARY 2.1. *The second Fréchet derivative of* $C(\cdot, \phi)$ *exists and is a positive definite bilinear functional in* $L_2(R^+, R^m)$.

COROLLARY 2.2. *There exists only one point in* $L_2(R^+, R^m)$ *such that* $C'(u, \phi) = 0$. *Hence* (2.9) *has a unique solution.*

COROLLARY 2.3. *Let a and b be scalars. If* u^m *and* v^m *are optimal controls for* $C(\cdot, \phi)$ *and* $C(\cdot, \psi)$ *respectively, then* $au^m + bv^m$ *is the optimal control for* $C(\cdot, a\phi + b\psi)$.

COROLLARY 2.4. *The solution of the control problem* (2.1)–(2.3) *is given by the linear integral equation*

$$\begin{aligned} x^m(t) = \phi(0)S(t, 0) - \int_0^t \Bigg[\sum_{i=1}^{k} \int_\alpha^T x^m(\beta - \tau_i) W_i(\beta) S^*(\beta - \tau_i, \alpha)\, d\alpha \\ + \int_\alpha^T \int_{-\tau}^0 x^m(\beta + s) W_0(\beta, s) S^*(\beta + s, \alpha)\, ds\, d\beta \Bigg] B^*(\alpha) U^{-1}(\alpha) B(\alpha) S(t, \alpha)\, d\alpha \\ + \int_{-\tau}^0 \phi(s)\, d_s \left[\int_0^\tau A(\alpha, s - \alpha) S(t, \alpha)\, d\alpha \right]. \end{aligned}$$

DEFINITION 2.2. Let ϕ_1 and ϕ_2 in $BV[-\tau, 0]$ be given. Let the pairs $(x_1(t), u_1(t))$, $(x_2(t), u_2(t))$ be the optimal control and optimal trajectory for the functionals $C(\cdot, \phi_1)$ and $C(\cdot, \phi_2)$ respectively. Define the expression

$$\begin{aligned} Q(\phi_1, \phi_2) = \int_0^T \Bigg[\sum_{i=1}^{k} x_1(t - \tau_i) W_i(t) x_2(t - \tau_i) \\ + \int_{-\tau}^0 x_1(t + s) W_0(t, s) x_2(t + s)\, ds + u_1(t) U(t) u_2(t) \Bigg] dt. \end{aligned} \tag{2.10}$$

Notice that $Q(\phi_1, \phi_2) = Q(\phi_2, \phi_1)$ and $Q(\phi_1, \phi_1) = m(\phi_1)$.

DEFINITION 2.3. Let ϕ be in $BV[-\tau, 0]$. Let x^m denote the optimal trajectory for $C(\cdot, \phi)$. For each $t \in [0, T]$ define the column vector

$$\begin{aligned} L(t)\phi = \sum_{i=1}^{k} \int_t^T S(\alpha - \tau_i, t) W_i(\alpha) x^m(\alpha - \tau_i)\, d\alpha \\ + \int_t^T \int_{-\tau}^0 [S(\alpha + s, t) W_0(\alpha, s) x^m(\alpha + s)]\, ds\, d\alpha. \end{aligned} \tag{2.11}$$

Remark 2.2. Notice that $(t, \phi) \to L(t)\phi$ is a continuous mapping from $[0, T] \times BV[-\tau, 0]$ into the space of continuous functions defined on R^n. Moreover, because of Corollary 2.3, it is linear on $BV[-\tau, 0]$ and for t fixed $L(t)\phi$ is a linear mapping from $BV[-\tau, 0] \to R^n$.

Consider the bilinear form on $BV[-\tau, 0]$ given by

$$\langle\!\langle \phi, \psi \rangle\!\rangle = \psi(0)L(0)\phi + \int_{-\tau}^{0} \psi(s)\, d_s \int_0^{\tau} A(\alpha, s - \alpha)L(\alpha)\phi\, d\alpha. \tag{2.12}$$

THEOREM 2.2. *The bilinear form defined by* (2.12) *is symmetric and equal to the form Q of Definition* 2.2. *That is, for all ϕ and ψ in $BV[-\tau, 0]$,*

$$\langle\!\langle \phi, \psi \rangle\!\rangle = \langle\!\langle \psi, \phi \rangle\!\rangle = Q(\phi, \psi) \quad \textit{and} \quad \langle\!\langle \phi, \phi \rangle\!\rangle = m(\phi).$$

Proof. Let ϕ_1 and ϕ_2 in $BV[-\tau, 0]$ be given and let the pairs $(x_i(t), u_i(t))$ denote the optimal trajectories and optimal controls for the functionals $C(\cdot, \phi_i)$, $i = 1, 2$. Using (2.11) we can write

$$\begin{aligned} \phi_2(0)L(0)\phi_1 = &\sum_{i=1}^{k} \int_0^T \phi_2(0)S(\alpha - \tau_i, 0)W_i(\alpha)x_1(\alpha - \tau_i)\, d\alpha \\ &+ \int_0^T \int_{-\tau}^{0} \phi_2(0)S(\alpha + s, 0)W_0(\alpha, s)x_1(\alpha + s)\, ds\, d\alpha. \end{aligned} \tag{2.13}$$

Observe that by (2.4), (2.5) and the fact that $-\tau \leqq s \leqq 0$,

$$\begin{aligned} \phi_2(0)S(\alpha - \tau_i, 0) = x_2(\alpha - \tau_i) &- \int_0^{\alpha} u_2(\sigma)B(\sigma)S(\alpha - \tau_i, \sigma)\, d\sigma \\ &- \int_{-\tau}^{0} \phi_2(\sigma)\, d_\sigma \int_0^{\alpha} A(\beta, \sigma - \beta)S(\alpha - \tau_i, \beta)\, d\beta \end{aligned} \tag{2.14}$$

and

$$\begin{aligned} \phi_2(0)S(\alpha + s, 0) = x_2(\alpha + s) &- \int_0^{\alpha} u_2(\sigma)B(\sigma)S(\alpha + s, \sigma)\, d\sigma \\ &- \int_{-\tau}^{0} \phi_2(\sigma)\, d_\sigma \int_0^{\tau} A(\beta, \sigma - \beta)S(\alpha + s, \beta)\, d\beta. \end{aligned} \tag{2.15}$$

Substitution of (2.14) and (2.15) into (2.13) yields

$$\begin{aligned} \phi_2(0)L(0)\phi_1 = &\sum_{i=1}^{k} \int_0^T x_2(\alpha - \tau_i)W_i(\alpha)x_1(\alpha - \tau_i)\, d\alpha \\ &+ \int_0^T \int_{-\tau}^{0} [x_2(\alpha + s)W_0(\alpha, s)x_2(\alpha + s)]\, ds\, d\alpha \\ &- \sum_{i=1}^{k} \int_0^T \int_0^{\alpha} [u_2(\sigma)B(\sigma)S(\alpha - \tau_i, \sigma)W_1(\alpha)x_1(\alpha - \tau_i)]\, d\sigma\, d\alpha \\ &- \sum_{i=1}^{k} \int_0^T \left[\int_{-\tau}^{0} \phi_2(\sigma)\, d_\sigma \int_0^{\tau} A(\beta, \sigma - \beta)S(\alpha - \tau_i, \beta)\, d\beta\right] \\ &\cdot W_i(\alpha)x_1(\alpha - \tau_i)\, d\alpha \qquad \text{(cont.)} \end{aligned} \tag{2.16}$$

$$-\int_0^T \int_{-\tau}^0 \int_0^\alpha [u_2(\sigma)B(\sigma)S(\alpha + s, \sigma)W_0(\alpha, s)x_1(\alpha + s)]\, d\sigma\, ds\, s\alpha$$

$$-\int_0^T \int_{-\tau}^0 \left[\int_{-\tau}^0 \phi_2(\sigma)\, d_\sigma \int_0^\tau A(\beta, \sigma - \beta)S(\alpha + s, \beta)\, d\beta \right]$$

$$\cdot\, W_0(\alpha, s)x_1(\alpha + s)\, ds\, d\alpha.$$

If we examine the multiple integrals in (2.16) which contain Stieltjes integrals and apply the unsymmetric Fubini's theorem (see, for example, [3] or [2]), Fubini's theorem, the fact that $S(\alpha, \beta) = 0$ if $\beta > \alpha$ and equation (2.11), we can express the integrals as follows:

$$\sum_{i=1}^k \int_0^T \left[\int_{-\tau}^0 \phi_2(\sigma)\, d_\sigma \int_0^\tau A(\beta, \sigma - \beta)S(\alpha - \tau_i, \beta)\, d\beta \right] W_i(\alpha)x_1(\alpha - \tau_i)\, d\alpha$$

$$+ \int_0^T \int_{-\tau}^0 \left[\int_{-\tau}^0 \phi_2(\sigma)\, d_\sigma \int_0^\tau A(\beta, \sigma - \beta)S(\alpha + s, \beta)\, d\beta \right]$$

$$\cdot\, W_0(\alpha, s)x_1(\alpha + s)\, ds\, d\alpha$$

$$(2.17) \qquad = \sum_{i=1}^k \int_{-\tau}^0 \phi_2(\sigma)\, d_\sigma \int_0^\tau \int_\beta^T A(\beta, \sigma - \beta)S(\alpha - \tau_i, \beta)W_i(\alpha)x_1(\alpha - \tau_i)\, d\alpha\, d\beta$$

$$+ \int_{-\tau}^0 \phi_2(\sigma)\, d_\sigma \int_0^\tau \int_\beta^T \int_{-\tau}^0 A(\beta, \sigma - \beta)S(\alpha + s, \beta)$$

$$\cdot\, W_0(\alpha, s)x_1(\alpha + s)\, ds\, d\alpha\, d\beta$$

$$= \int_{-\tau}^0 \phi_2(\sigma)\, d_\sigma \int_0^\tau A(\beta, \sigma - \beta)L(\beta)\phi_1\, d\beta.$$

We substitute the right side of (2.17) into (2.16) and apply Fubini's theorem in a judicious manner to obtain

$$\phi_2(0)L(0)\phi_1 = \sum_{i=1}^k \int_0^T x_2(\alpha - \tau_i)W_i(\alpha)x_1(\alpha - \tau_i)\, d\alpha$$

$$+ \int_0^T \int_{-\tau}^0 x_2(\alpha + s)W_0(\alpha, s)x_1(\alpha + s)\, ds\, d\alpha$$

$$(2.18) \qquad - \int_0^T \left[\sum_{i=1}^k \int_\sigma^T u_2(\sigma)B(\sigma)S(\alpha - \tau_i, \sigma)W_i(\alpha)x_1(\alpha - \tau_i)\, d\alpha \right.$$

$$\left. - \int_\sigma^T \int_{-\tau}^0 u_2(\sigma)B(\sigma)S(\alpha + s, \sigma)W_0(\alpha, s)x_1(\alpha + s)\, ds\, d\alpha \right] ds$$

$$- \int_{-\tau}^0 \phi_2(\sigma)\, d_\sigma \left[\int_0^\tau A(\beta, \sigma - \beta)L\beta\phi_1\, d\beta \right].$$

But because of (2.9),

$$\int_0^T \left[\sum_{i=1}^k \int_\sigma^T u_2(\sigma)B(\sigma)S(\alpha - \tau_i, \sigma)W_i(\alpha)x_1(\alpha - \tau_i)\, d\alpha \right.$$

$$\left. + \int_\sigma^T \int_{-\tau}^0 u_2(\sigma)B(\sigma)S(\alpha + s, \sigma)W_0(\alpha, s)x_1(\alpha + s)\, ds\, d\alpha \right] d\sigma$$

$$= -u_2(\sigma)U(\sigma)u_1(\sigma). \tag{2.19}$$

If we substitute (2.19) into (2.18) and apply (2.10) we obtain

$$\phi_2(0)L(0)\phi_1 = Q(\phi_1, \phi_2) - \int_{-\tau}^0 \phi_2(\sigma)\, d_\sigma \int_0^\tau A(\beta, \sigma - \beta)L(\beta)\phi_1\, d\beta, \tag{2.20}$$

which proves the theorem.

Example. The following example, a system considered by Kushner and Barnea [16], will perhaps motivate the reason for selecting the bilinear form defined by (2.12).

Consider the problem

$$\dot{x}(t) = A_1(t)x(t) + A_2(t)x(t - \tau) + \int_{-\tau}^0 C(t, s)x(t + s)\, ds + D(t)u(t), \tag{2.21}$$

$$C(u, \phi) = \int_0^T [x(t)W(t)x(t) + u(t)U(t)u(t)]\, dt. \tag{2.22}$$

Here $A_i(t)$, $i = 1, 2$, and $C(t, s)$ are continuous $n \times n$ matrices, $x(t)$ is an n-vector, $D(t)$ a continuous $n \times m$ matrix, $u(t)$ a measurable m-vector whose norm is square integrable, $W(t)$ a continuous $n \times n$ positive semidefinite matrix, $U(t)$ a continuous $m \times m$ positive definite matrix and ϕ is fixed in $BV[-\tau, 0]$. Using Theorem 2.1, we compute the optimal control to be

$$u(t) = -U^{-1}(t)B^*(t) \int_t^T S^*(\alpha, t)W(\alpha)x(\alpha)\, d\alpha \tag{2.23}$$

and

$$L(t)\phi = y(t) = \int_t^T S^*(\alpha, t)W(\alpha)x(\alpha)\, d\alpha. \tag{2.24}$$

Observing that $S^*(\alpha, t)$ satisfies the differential equation (see, for example, [11])

$$\begin{aligned} \frac{d}{dt}[S^*(\alpha, t)] = {} & -A_1^*(t)S^*(\alpha, t) - A_2^*(t + \tau)S^*(\alpha, t + \tau) \\ & - \int_{-\tau}^0 C^*(t - s, s)S^*(\alpha, t - s)\, ds, \end{aligned} \tag{2.25}$$

we find that

$$\begin{aligned} \frac{d}{dt}(L(t)\phi) = \frac{dy(t)}{dt} = {} & -W(t)x(t) - A_1^*(t)y(t) - A_2^*(t + \tau)y(t + \tau) \\ & - \int_{-\tau}^0 C^*(t - s, s)y(t - s)\, ds. \end{aligned} \tag{2.26}$$

Thus $L(t)\phi$ can be looked upon as a solution of a differential lead equation, and the optimal control and optimal trajectory are then obtained by looking at a particular solution of the coupled system (2.21) and (2.26). This is the procedure used by Nosov [20].

Define the system

$$(2.27)\quad \mathring{z}(t) = -A_1^*(t)z(t) - A_2^*(t+\tau)z(t+\tau) - \int_{-\tau}^{0} C^*(t-s,s)z(t-s)\,ds$$

and let $x(t)$ be a solution of (2.21) with $u(t) = 0$. The "inner product" of any solution of (2.27) with $x(t)$ is defined to be

$$(2.28)\quad \begin{aligned}\langle z(t), x(t)\rangle = z(t)x(t) &+ \int_{t-\tau}^{t} z(\alpha+\tau)A_2(\alpha+\tau)x(\alpha)\,d\alpha \\ &+ \int_{t-\tau}^{t}\int_{t}^{\alpha+\tau} z(s)C(s,\alpha-s)x(\alpha)\,ds\,d\alpha.\end{aligned}$$

It is shown in [2] or [11] that $\langle z(t), x(t)\rangle$ is constant for all t. Equations (2.26) and (2.28) are the motivation for defining the bilinear form on $BV[-\tau,0]$ given by the equation

$$(2.29)\quad \begin{aligned}\langle\!\langle \phi,\psi\rangle\!\rangle = \langle L(0)\phi,\psi(0)\rangle = \psi(0)L(0)\phi &+ \int_{-\tau}^{0} L(\alpha+\tau)\phi A_2(\alpha+\tau)\psi(\alpha)\,d\alpha \\ &+ \int_{-\tau}^{0}\int_{0}^{\alpha+\tau} L(s)\phi C(s,\alpha-s)\psi(\alpha)\,ds\,d\alpha.\end{aligned}$$

Using the fact that $S(\alpha, s) = 0$ if $\alpha < s$, it is not difficult to prove directly that $\langle\!\langle \phi,\psi\rangle\!\rangle = \langle\!\langle \psi,\phi\rangle\!\rangle$ and that $\langle\!\langle \phi,\phi\rangle\!\rangle = m(\phi)$. However, these are also consequences of Theorem 2.2.

3. In this section we shall apply the results of § 2 to an autonomous control problem involving differential-difference equations. Our main result in this section is to prove that for certain linear autonomous systems with quadratic cost functionals the optimal control is a feedback control which gives rise to an asymptotically stable linear differential-difference equation. It will become apparent that the results of this section remain valid for more complex linear systems. The reason for restricting the problem is to limit notational difficulties.

We shall first introduce some additional notation.

1. All vectors considered in this section will be column vectors. The inner product of two column vectors x and y will henceforth be denoted by $x \cdot y$.

2. Let $t_0 \geqq 0$; by the notation $\phi_{t_0}(\cdot)$ we shall mean a continuous mapping from $[-\tau, 0]$ into R^n which is given by the expression $\phi_{t_0}(s)$, $t_0 - \tau \leqq s \leqq t_0$.

We shall first consider the control problem given by the equations:

$$(3.1)\qquad \mathring{x}_u(t) = A_0 x_u(t) + A_1 x(t-\tau) + Bu(t)$$

for $t \geqq t_0$ and $u \in L_2(R^+, R^m)$,

$$(3.2)\qquad x_u(t) = \phi_{t_0}(t) \quad \text{if} \quad t_0 - \tau \leqq t \leqq t_0,$$

and cost functional

$$C(u, \phi_{t_0}, t_0, T) = \int_{t_0}^{T} [Wx_u(t) \cdot x_u(t) + Uu(t) \cdot u(t)]\, dt, \qquad t_0 \leqq T < \infty. \tag{3.3}$$

Here we assume A_0 and A_1 are $n \times n$ constant real matrices, B is an $n \times m$ constant real matrix, W is a positive definite real $n \times n$ matrix and U is a positive definite real $m \times m$ matrix.

DEFINITION 3.1. We define

$$m(\phi_{t_0}, t_0, T) = \inf_{u \in L_2(R^+, R^m)} C(u, \phi_{t_0}, t_0, T), \tag{3.4}$$

and denote by

$$x^m(t, t_0, T, \phi_{t_0}) \tag{3.5}$$

and

$$u^m(t, t_0, T, \phi_{t_0}), \tag{3.6}$$

respectively, the optimal control and optimal trajectory for the problem (3.1)–(3.3). In addition the notation

$$x_s^m(\cdot, t_0, T, \phi_{t_0}), \qquad t_0 \leqq s \leqq T, \tag{3.7}$$

will designate the point in $C[-\tau, 0]$ given by (3.5) for $s - \tau \leqq t \leqq s$.

Observe that, since (3.1) is autonomous in the variables involving x, the matrix $S(t - \alpha)$, $\alpha \leqq t$, of equation (2.5) can be written as in the form $S(t - \alpha)$ (see, for example, [12, p. 84]). Thus for each $\phi_{t_0} \in C[-\tau, 0]$ and triple of nonnegative real numbers (t_0, t, T) with $t_0 \leqq t \leqq T$, we define

$$L(t, t_0, T)\phi_{t_0} = \int_t^T S^*(\alpha - t)Wx^m(\alpha, t_0, T, \phi_{t_0})\, d\alpha. \tag{3.8}$$

Applying Theorem 2.1 we see that

$$u^m(t, t_0, T, \phi_{t_0}) = -U^{-1}B^*L(t, t_0, T)\phi_{t_0}. \tag{3.9}$$

(Theorem 2.1 is phrased in terms of row vectors, but by transposing (2.9) and using (3.8) we obtain (3.9).) In Remark 2.2 it was observed that $L(t, t_0, T)$ is a linear mapping from $BV[-\tau, 0]$ into R^n. In the present case it is clear that $L(t, t_0, T)$ is a linear mapping from $C[-\tau, 0]$ into R^n.

LEMMA 3.1. *If $\{\phi_{t_0}^n\} \subset C[-\tau, 0]$ tends in norm to $\phi_{t_0} \in C[-\tau, 0]$, then there exists a subsequence $\{q\}$ of the natural numbers such that for each $t \in [t_0, T]$ the subsequence $\{x_t^m(\cdot, t_0, T, \phi_{t_0}^q)\}$ tends in norm to $x_t^m(\cdot, t_0, T, \phi_{t_0})$ and the subsequence $\{u^m(\cdot, t_0, T, \phi_{t_0}^q)\}$ tends in $L_2(R^+, R^m)$ to $u^m(\cdot, t_0, T, \phi_{t_0})$.*

Proof. There exist positive constants M_1 and α such that $|\phi_{t_0}^n| \leqq M_1$ for all n and $|S(t)| \leqq M_1 e^{at}$ for all $t \in R^+$. Thus if we set $u \equiv 0$ in (2.4) and substitute the corresponding solutions $\{x_m\}$, with $x_n(t) = \phi_{t_0}^n(t)$ on $[t_0 - \tau, t_0]$, into the cost functional (3.3) we see that for some positive constant M_2, $m(\phi_{t_0}^n, t_0, T) \leqq M_2 \|\phi_{t_0}^n\|^2$ for all n. Since U is positive definite, the form of the cost functional (3.3) implies that the optimal controls $\{u^m(\cdot, t_0, T, \phi_{t_0}^n)\}$ are uniformly bounded in $L_2(R^+, R^m)$.

However, $L_2(R^+, R^m)$ is a Hilbert space and hence reflexive. Consequently there exists a subsequence $\{q\}$ of the natural numbers such that $\{u^m(\cdot, t_0, T, \phi_{t0}^q)\}$ converges weakly to some point u_0 in $L_2(R^+, R^m)$. Thus for each $t \in [t_0, T]$ the subsequence $\{x^m(t, t_0, T, \phi_{t0}^q)\}$ converges pointwise to $x_{u_0}(t, t_0, T, \phi_{t_0})$. Due to the assumption that both W and U are positive definite matrices we can infer that $C(u_0, \phi_{t_0}, t_0, T) \leqq \liminf m(\phi_{t_0}^q, t_0, T)$. By an argument similar to the proof of (c) in Theorem 4 in [7, p. 357], it can be shown that $C(u_0, \phi_{t_0}, t_0, T) = m(\phi_{t_0}, t_0, T)$. But then it easily follows that

$$\lim_{q\to\infty} \int_{t_0}^{T} Uu^m(t, t_0, T, \phi_{t0}^q) \cdot u^m(t, t_0, T, \phi_{t0}^q)\, dt = \int_{t_0}^{T} Uu_0(t) \cdot u_0(t)\, dt.$$

Again, since U is positive definite this implies that

$$\lim_{q\to\infty} |u(\cdot, t_0, T, \phi_{t0}^q)| = |u_0|.$$

In a Hilbert space weak convergence and convergence of the norms implies strong convergence (see, for example, [21]). Hence $\{u(\cdot, t_0, T, \phi_{t0}^q)\}$ converges in norm to u_0.

Lastly, observe that the operator $\int_{t_0}^{t} S(t-s)Bu(s)\, ds$ is a continuous linear mapping from $L_2(R^+, R^m)$ into $C[-\tau, 0]$. Thus it follows that $\{x_t^m(\cdot, t_0, T, \phi_{t0}^q)\}$ converges strongly to $x_t^m(\cdot, t_0, T, \phi_{t0})$ in $C[-\tau, 0]$.

Some properties of the mapping $L(t, t_0, T)$.

PROPERTY 1. *$L(t, t_0, T)$ is continuous.*

Proof. The proof is an immediate consequence of Lemma 3.1 and the closed graph theorem.

PROPERTY 2. *For ϕ_{t_0} fixed in $C[-\tau, 0]$ the function $L(t, t_0, T)\phi_{t_0}$ is continuous in the variables (t, T).*

Proof. The proof is by contradiction, but uses standard arguments and is therefore omitted.

PROPERTY 3. *For each $\phi_{t_0} \in C[-\tau, 0]$ the identity*

$$L(t, t_0, T)\phi_{t_0} = L(t, t, T)x_t^m(\cdot, t_0, T, \phi_{t0})$$

holds.

Proof. By the uniqueness property of Corollary 2.2 it follows that if $x_t^m(\cdot, t_0, T, \phi_{t0})$ is optimal over the interval $[t_0, T]$, then

$$x_s^m(\cdot, t, T, x_t^m(\cdot, t_0, T, \phi_{t0})) = x_s^m(\cdot, t_0, T, \phi_{t0})$$

is optimal over the interval $[t, T]$. Substitution of these mappings into (3.8) establishes the identity.

PROPERTY 4. *If $t_0 > 0$ and $\phi_{t0}(t_0 + s) = \psi(s)$ for $s \in [-\tau, 0]$, then*

$$L(t + t_0, t_0, T + t_0)\phi_{t0} = L(t, 0, T)\psi \quad and \quad m(\phi_{t_0}, t_0, T + t_0) = m(\psi, 0, T).$$

Proof. This property is a consequence of the fact that (3.1) is autonomous, W and U are independent of t and the definition of $L(t, t_0, T)$.

Necessary and sufficient conditions under which the following hypothesis is valid will be given in § 4.

HYPOTHESIS H. For all $t_0 \geqq 0$ and for each $\phi_{t_0} \in C[-\tau, 0]$ it will be assumed that $\lim_{T\to\infty} m(\phi_{t_0}, t_0, T) < \infty$.

DEFINITION 3.2. Let $t_0 \leqq T$ and ϕ^1 and ϕ^2 be in $C[-\tau, 0]$. Let the pairs (x^1, x^2) and (u^1, u^2) denote the respective optimal trajectories and optimal controls for the control problem (3.1)–(3.3) with initial values ϕ^1 and ϕ^2 over the interval $[t_0, T]$. We define the bilinear form $Q(t_0, T)$ on $C[-\tau, 0] \times C[-\tau, 0]$ by the expression

$$Q(t_0, T)(\phi^1, \phi^2) = \int_{t_0}^{T} [Wx^1(t) \cdot x^2(t) + Uu^1(t) \cdot u^2(t)]\, dt. \tag{3.10}$$

Notice that for each $\phi_{t_0} \in C[-\tau, 0]$,

$$Q(t_0, T)(\phi_{t_0}, \phi_{t_0}) = m(t_0, T, \phi_{t_0}). \tag{3.11}$$

LEMMA 3.2. *Let $0 \leqq t_0 \leqq T$. Let $R(t_0, T)$ be the bilinear form defined on $C[-\tau, 0] \times C[-\tau, 0]$ by the expression*

$$\begin{aligned} R(t_0, T)(\phi_{t_0}, \psi_{t_0}) &= L(t_0, t_0, T)\phi_{t_0}(t_0) \cdot \psi_{t_0}(t_0) \\ &\quad + \int_{t_0 - \tau}^{t_0} A_1^* L(s + \tau, t_0, T)\phi_{t_0} \cdot \psi_{t_0}(s)\, ds. \end{aligned} \tag{3.12}$$

Then $R(t_0, T)$ is symmetric and equal to the form $Q(t_0, T)$ defined by Definition 3.2.

Proof. This lemma is a special case of Theorem 2.2 in which column vectors replace row vectors.

The following lemma will be used in the sequel. It is given in an abstract setting since it is of interest in itself and extends a result which is well known for Hilbert spaces (see, for example, [21]).

LEMMA 3.3. *Let $\{R(t)\}$, $t \in R^+$, be a family of continuous symmetric bilinear forms defined on a real Banach space X such that for each x in X,*

$$\begin{gathered} 0 \leqq R(t)(x, x) \leqq R(\bar{t})(x, x) \quad \textit{if} \quad t \leqq \bar{t} \\ \textit{and} \\ \lim_{t \to \infty} R(t)(x, x) < \infty. \end{gathered} \tag{3.13}$$

To each $R(t)$ let there correspond the continuous linear mapping $g(t)$ from $X \to X^*$ which is given by the relation

$$\langle g(t)x, y \rangle = \langle g(t)y, x \rangle = R(t)(x, y) \tag{3.14}$$

(*see, for example,* [8, *p.* 104]). *Then there exists a continuous linear mapping $g: X \to X^*$ such that for all x, y in X,*

$$\langle gx, y \rangle = \langle gy, x \rangle = \lim_{t \to \infty} R(t)(x, y) \tag{3.15}$$

and such that for all $x \in X$,

$$\lim_{t \to \infty} |g(t)x - gx| = 0. \tag{3.16}$$

Proof. We shall first show that $|R(t)|$ is uniformly bounded on R^+. To this end define for each $t \in R^+$ the seminorm

$$\Phi(t)(x) = [R(t)(x, x)]^{1/2}.$$

By the principle of uniform boundedness (see, for example, [23, p. 68]) it follows that $\lim_{|x|\to 0} \Phi(t)(x) = 0$ uniformly on R^+. Hence given $\varepsilon > 0$, there exists $\delta(\varepsilon) > 0$ such that for all $t \in R^+$,

$$R(t)(x, x) < \varepsilon \quad \text{if} \quad |x| < \delta(\varepsilon).$$

This means

$$R(t)(x, x) \leqq M_4|x|^2 \tag{3.17}$$

for all $t \in R^+$, where $M_4 = \varepsilon/\delta^2$. This proves uniform boundedness since $R(t)$ is symmetric.

Since for all x, y in X and $t \in R^+$,

$$|\langle g(t)y, x\rangle| = \tfrac{1}{4}|R(t)(x + y, x + y) - R(t)(x - y, x - y)|,$$

it follows for $|x| = |y| = 1$ that

$$\langle g(t)y, x\rangle \leqq \tfrac{1}{4}M_4[|x + y|^2 + |x - y|^2] \leqq 2M_4. \tag{3.18}$$

This proves that $|g(t)| \leqq 2M_4$ for all $t \in R^+$.

Because $R(t)$ is symmetric monotone increasing and uniformly bounded on R^+ it follows that there exists a bilinear form R defined on X and a continuous mapping $g: X \to X^*$ such that for each pair x, y in X,

$$\begin{aligned}\lim_{t\to\infty} R(t)(x, y) &= \lim_{t\to\infty} \langle g(t)y, x\rangle = R(x, y)\\ &= \langle gy, x\rangle = \langle gx, y\rangle.\end{aligned} \tag{3.19}$$

Because of (3.18), $|g| \leqq 2M_4$. Moreover, $\langle g - g(t)x, x\rangle \geqq 0$ for all $t \in R^+$. Thus we can apply the Cauchy–Schwarz inequality to obtain

$$\begin{aligned}[\langle g - g(t)y, x\rangle]^2 &= [\langle (g - g(t))x, y\rangle]^2\\ &\leqq \langle (g - g(t))y, y\rangle\langle (g - g(t))x, x\rangle\\ &\leqq 2M_4\langle (g - g(t))y, y\rangle|x|^2.\end{aligned} \tag{3.20}$$

Since $\lim_{t\to\infty} \langle (g - g(t))y, y\rangle = 0$ for all $y \in X$, (3.20) shows that $g(t)y \to gy$ strongly in X^*. This completes the proof of the lemma.

THEOREM 3.1. *Assume Hypothesis* H *holds: then there exist a continuous bilinear form* $\hat{Q}$ *on* $C[-\tau, 0]$ *and a continuous linear mapping*

$$q: C[-\tau, 0] \to BV[-\tau, 0]$$

such that for ϕ_{t_0}, ψ_{t_0} *in* $C[-\tau, 0]$,

$$\begin{aligned}\lim_{T\to\infty} Q(t_0, T)(\psi_{t_0}, \phi_{t_0}) &= \hat{Q}(\psi_{t_0}, \phi_{t_0})\\ &= \langle q\phi_{t_0}, \psi_{t_0}\rangle = \langle q\psi_{t_0}, \phi_{t_0}\rangle.\end{aligned} \tag{3.21}$$

Moreover, for each $\phi_{t_0} \in C[-\tau, 0]$,

$$\lim_{T\to\infty} L(t_0, t_0, T)\phi_{t_0} = \lim_{s\to 0^-} [(q\phi_{t_0})(0) - (q\phi_{t_0})(s)]. \tag{3.22}$$

Proof. The first conclusion of the theorem is a special case of Lemma 3.3 in which $R(t)$ is replaced by $Q(t_0, t)$ and it is observed that the topological dual of

$C[-\tau, 0]$ is isometric to the subspace of $BV[-\tau, 0]$ which consists of those mappings ϕ of bounded variation from $[-\tau, 0] \to R^n$ for which $\phi(-\tau) = 0$ (see, for example, [9]). The fact that both $\hat{Q}$ and q are independent of t_0 is a consequence of Property 4.

To prove the second conclusion of the theorem, observe that by Lemma 3.2 for (t_0, T) fixed and ϕ_{t_0}, ψ_{t_0} in $C[-\tau, 0]$,

$$Q(t_0, T)(\phi_{t_0}, \psi_{t_0}) = L(t_0, t_0, T)\phi \cdot \psi_{t_0}(t_0) + \int_{t_0-\tau}^{t_0} A_1^* L(\alpha + \tau, t_0, T)\phi_{t_0} \cdot \psi_{t_0}(\alpha)\, d\alpha.$$

The above representation of $Q(t_0, T)$ gives rise to a linear mapping

$$q(t_0, T): C[-\tau, 0] \to BV[-\tau, 0]$$

given by the equations

$$\begin{aligned} &(q(t_0, T)\phi_{t_0})(s) = \int_{t_0-\tau}^{t_0+s} A_1^* L(\alpha + \tau, t_0, T)\phi_{t_0}\, d\alpha, \qquad -\tau \leqq s < 0, \\ (3.23)\qquad & \\ &(q(t_0, T)\phi_{t_0})(0) = L(t_0, t_0, T)\phi_{t_0} + \int_{t_0-\tau}^{t_0} A_1^* L(\alpha + \tau, t_0, T)\phi_{t_0}\, d\alpha. \end{aligned}$$

Thus

$$\lim_{s \to 0^-} [(q(t_0, T)\phi_{t_0})(0) - (q(t_0, T)\phi_{t_0})(s)] = L(t_0, t_0, T)\phi_{t_0}.$$

From (3.16) of Lemma 3.3 it follows that for each $\phi_{t_0} \in C[-\tau, 0]$,

$$(3.24) \qquad \lim_{T \to \infty} q(t_0, T)\phi_{t_0} = q\phi_{t_0}$$

in the normed topology of $BV[-\tau, 0]$, which implies that

$$\lim_{T \to \infty} L(t_0, t_0, T)\phi_{t_0} = \lim_{s \to 0^-} [(q\phi_{t_0})(0) - (q\phi_{t_0})(s)],$$

and establishes (3.22).

Definition 3.3. We shall denote by L the mapping from $C[-\tau, 0] \to R^n$ given by (3.22).

Corollary. *L is a continuous linear mapping from $C[-\tau, 0] \to R^n$.*

Proof. The proof is an immediate consequence of Theorem 3.1 since $q: C[-\tau, 0] \to BV[-\tau, 0]$ is continuous.

We now want to consider the problem posed by (3.1)–(3.3) when $T = \infty$. To this end we shall allow T in (3.3) to be infinity if for each $\phi_{t_0} \in C[-\tau, 0]$ the cost functional in (3.3) can be made finite over R^+ for some choice of $u \in L_2(R^+, R^m)$. If this is the case we shall denote the cost by $C(u, \phi_{t_0}, t_0, \infty)$.

Theorem 3.2. *Assume Hypothesis* H *holds and let L be as in Definition* 3.3 *and $\hat{Q}$ be defined by* (3.21). *Consider the linear differential-difference equation given by*

$$\begin{aligned} &\mathring{x}(t) = A_0 x(t) + A_1 x(t - \tau) - BU^{-1}B^* L x_t, \\ (3.25)\qquad & \\ &x(t) = \phi_{t_0}(t), \qquad\qquad t_0 - \tau \leqq t \leqq t_0, \end{aligned}$$

and let $x(t, \phi_{t_0})$ *denote the solution of* (3.25) *and* $x_t(\phi_{t_0})$ *the corresponding point in* $C[-\tau, 0]$. *Then for any sequence of points* $\{T_n\} \subset R^+$ *such that* $\lim_{n\to\infty} T_n = \infty$ *the following holds*:

(i) $\lim_{n\to\infty} x^m(\cdot, \phi_{t_0}, t_0, T_n) = x_t(\phi_{t_0})$ *for each* $t \in [t_0, \infty)$,

(ii) $\lim_{n\to\infty} m(\phi_{t_0}, t_0, T_n) = \hat{Q}(\phi_{t_0}, \phi_{t_0})$ *and*

(iii) *the point* $u \in L_2(R^+, R^m)$ *given by* $u(t) = -U^{-1}B^*Lx(t, \phi_{t_0})$ *optimizes the cost functional* $C(\cdot, \phi_{t_0}, t_0, \infty)$.

Proof of (i). Notice that by using the equation for the optimal control, (3.9), and the Property 3 we can form for any n and $t \geqq t_0$ the expression

$$
\begin{aligned}
&x^m(t, \phi_{t_0}, t_0, T_n) - x(t, \phi_{t_0}) \\
(3.26)\qquad &= -\int_{t_0}^{t} S(t-s)BU^{-1}B^*[L(s, t_0, T_n)\phi_{t_0} - Lx_s(\phi_{t_0})]\,ds \\
&= -\int_{t_0}^{t} S(t-s)BU^{-1}B^*[L(s, s, T_n)x_s^m(\cdot, t_0, T_n, \phi_{t_0}) - Lx_s(\phi_{t_0})]\,ds.
\end{aligned}
$$

Hence,

$$
\begin{aligned}
&|x^m(t, \phi_{t_0}, t_0, T_n) - x(t, \phi_{t_0})| \\
(3.27)\qquad &\leqq \left|\int_{t_0}^{t} S(t-s)BU^{-1}B^*L(s, s, T_n)[x_s^m(\cdot, t_0, T_n, \phi_{t_0}) - x_s(\phi_{t_0})]\,ds\right| \\
&\quad + \left|\int_{t_0}^{t} S(t-s)BU^{-1}B^*[L - L(s, s, T_n)]x_s(\phi_{t_0})\,ds\right|.
\end{aligned}
$$

Let

$$
(3.28)\qquad r_n(t) = \int_{t_0}^{t} M_1 e^{\alpha(t-s)}|BU^{-1}B^*|\,|[L - L(s, s, T_n)]x_s(\phi_{t_0})]|\,ds.
$$

There exist constants M_5 and $\alpha > 0$ such that

$$
(3.29)\qquad |S(t-s)BU^{-1}B^*L(s, s, T_n)| \leqq M_5 e^{\alpha(t-s)}.
$$

Hence using (3.27), (3.28) and (3.29) we can make the estimate

$$
\begin{aligned}
&|x^m(t, \phi_{t_0}, t_0, T_n) - x(t, \phi_{t_0})| \\
(3.30)\qquad &\leqq r_n(t) + M_5 \int_{t_0}^{t} |x_s^m(\cdot, t_0, T_n, \phi_{t_0}) - x_s(\phi_{t_0})| e^{\alpha(t-s)}\,ds.
\end{aligned}
$$

Observe that for t fixed the left-hand side of (3.30) also majorizes the expression

$$
|x^m(t + \alpha, \phi_{t_0}, t_0, T_n) - x(t + \alpha, \phi_{t_0})|
$$

as long as $t_0 \leqq t - \tau \leqq t + \alpha \leqq t$. Thus

$$
\begin{aligned}
&|x_t^m(\cdot, \phi_{t_0}, t_0, T_n) - x_t(\phi_{t_0})| \\
(3.31)\qquad &\leqq r_n(t) + M_5 \int_{t_0}^{t} |x_s^m(\cdot, t_0, T_n, \phi_{t_0}) - x_s(\phi_{t_0})| e^{\alpha(t-s)}\,ds.
\end{aligned}
$$

Using Theorem 3.1 and the Lebesgue dominated convergence theorem we can show that $r_n(t) \to 0$ as $n \to \infty$. Thus by Gronwall's inequality (see, for example, [5, p. 37, Prob. 1]) it follows that $\lim_{n\to\infty} x_t^m(\cdot, \phi_{t_0}, t_0, T_n) = x_t(\phi_{t_0})$ for each $t \in [t_0, \infty)$.

Proof of (ii). This is a consequence of Theorem 3.1 and the fact that

$$Q(t_0, T_n)(\phi_{t_0}, \phi_{t_0}) = m(\phi_{t_0}, t_0, T_n).$$

Proof of (iii). Let

$$x_n(t) = x(t, \phi_{t_0}, t_0, T_n), \tag{3.32}$$

$$u(t) = -U^{-1}B^*Lx_t(\phi_{t_0}) \tag{3.33}$$

and

$$u_n(t) = -U^{-1}B^*L(t, t, T_n)x_t(\cdot, \phi_{t_0}, t_0, T_n). \tag{3.34}$$

Since $|L(t, t, T_n)| \leqq M_4$ for all n, $L(t, t, T_n)\phi \to L(t, t, \phi)$ for all $\phi \in C[-\tau, 0]$ and by part (i) above $x_t^m(\cdot, \phi_{t_0}, t_0, T_n) \to x_t(\phi_{t_0})$ as $n \to \infty$, we can show that

$$\lim_{n\to\infty} L(t, t, T_n)x_t(\cdot, \phi_{t_0}, t_0, T_n) = Lx_t(\phi).$$

Hence using the notation in (3.32)–(3.34) it is easy to prove that for each $t \in [t_0, \infty)$,

$$\begin{aligned} &\lim_{n\to\infty} \int_{t_0}^{t} [Wx_n(s) \cdot x_n(s) + Uu_n(s) \cdot u_n(s)]\, ds \\ &\qquad = \int_{t_0}^{t} [Wx(s, \phi_{t_0}) \cdot x(s, \phi_{t_0}) + Uu(s) \cdot u(s)]\, ds. \end{aligned} \tag{3.35}$$

From this it follows, since Hypothesis H holds, that

$$\lim_{n\to\infty} m(\phi_{t_0}, t_0, T_n) \geqq \lim_{t\to\infty} \int_{t_0}^{t} [Wx(s, \phi_{t_0}) \cdot x(s, \phi_{t_0}) + Uu(s) \cdot (s)]\, ds. \tag{3.36}$$

Suppose strict inequality holds in (3.36). Then for some $\bar{n}$,

$$m(\phi_{t_0}, t_0, T_{\bar{n}}) > \int_{t_0}^{T_{\bar{n}}} [Wx(s, \phi_{t_0}) \cdot x(s, \phi_{t_0}) + Uu(s) \cdot u(s)]\, ds,$$

which is impossible. Hence equality holds in (3.36).

On the other hand, suppose that the optimal cost is attained by $\bar{u} \neq u$ and $C(\bar{u}, \phi_{t_0}, t_0, \infty) < \hat{Q}(\phi_{t_0}, \phi_{t_0})$. Then for some n it must follow that $C(\bar{u}, \phi_{t_0}, t_0, T_n) \leqq C(\bar{u}, \phi_{t_0}, t_0, \infty) < m(\phi_{t_0}, t_0, T_n)$ which is also impossible. Thus the optimal control is given by (3.33) and the optimal cost is $\hat{Q}(\phi_{t_0}, \phi_{t_0})$.

COROLLARY 1. *If Hypothesis* H *is satisfied, then there exist positive constants* β *and* M_6 *such that solutions of* (3.25) *satisfy the inequality*

$$|x_t(\phi_{t_0})| \leqq M_6 e^{-\beta(t-t_0)}|\phi_{t_0}|, \qquad t \geqq t_0, \tag{3.37}$$

that is, (3.25) *is exponentially stable.*

Proof. We complexify the spaces $C[-\tau, 0]$, R^n and R^m in the usual way and observe that because of Theorem 3.2,

$$\int_{t_0}^{\infty} Wx(t, \phi_{t_0}) \cdot \bar{x}(t, \phi_{t_0})\, dt < \infty$$

for all $\phi_{t_0} \in C[-\tau, 0]$. Hence (3.25) is asymptotically stable (see, for example, [12] or [15]). But this implies that all the eigenvalues associated with the semigroup generated by (3.25) lie in the right half-plane $\operatorname{Re} Z < 0$. This is sufficient to guarantee the conclusion of the corollary (see, for example, [12]).

4. In this section we give a necessary and sufficient condition for Hypothesis H to be valid.

First observe that if the uncontrolled system

$$\dot{x}(t) = A_0 x(t) + A_1 x(t - \tau) \tag{4.1}$$

is asymptotically stable in the sense that all solutions tend to the origin as t tends to infinity, then Hypothesis H is valid. This is because asymptotic stability for (4.1) implies exponential decay (see, for example, [12]). It is known (again see [12]) that solutions of (4.1) with values in the complex version of $C[-\tau, 0]$ generate a semigroup of operators, $T(t)$, which has an infinitesimal generator A, and A can have at most a finite number of eigenvalues lying in any right half-plane. Moreover, the complex version of $C[-\tau, 0]$ can be decomposed into a direct sum of two closed subspaces P and Q, with P finite-dimensional, which enjoy the properties that $T(t)P \subseteq P$, $T(t)Q \subseteq Q$, $AP \subseteq P$, the spectrum of A restricted to P lies in the half-plane $\operatorname{Re} Z \geqq 0$ and $T(t)|_Q$ is exponentially stable. Hale [12] has also shown that on P solutions of (4.1) are determined by an ordinary differential equation of the form

$$\dot{y}(t) = Fy(t), \tag{4.2}$$

and that the projections of solutions of (3.1)–(3.2) on P are described by solutions of an ordinary differential equation of the form

$$\dot{y}(t) = Fy(t) + \tilde{B}u(t), \tag{4.3}$$

where $\tilde{B}$ is a linear mapping from the complex m-space into a finite-dimensional complex linear space which has the dimensions of P (see, for example, [12, p. 122]). Thus we can state the following theorem.

THEOREM 4.1. *A necessary and sufficient condition for Hypothesis* H *to hold is that the system* (4.3) *be completely controllable.*

Proof. Assume that (4.3) is completely controllable; then given any $\phi \in C[-\tau, 0]$ let ϕ_p denote its projection on P. There exist $u \in L_2(R^+, R^m)$ and $T > 0$ such that $u(t) \equiv 0$ on $[T, \infty)$ and the solution $x_u(\cdot, \phi)$ of (3.1) has its projection, $x_u^p(\cdot, \phi)$, on P reach the origin of P, O_p, in time T. But then since $u(t) \equiv 0$ on $[T, \infty)$, the solution of (3.1) on $[T, \infty)$ is really a solution of (4.1) which lies in Q. However, all solutions of (4.1) with values in Q are exponentially stable.

On the other hand, if (4.3) is not completely controllable, then because (4.2) has its entire spectrum in $\operatorname{Re} Z \geqq 0$ there exists $\phi \in P$ such that all solutions of (3.1)–(3.2) with ϕ as initial value are uniformly bounded away from O_p (see, for example, [17]). Hence Hypothesis H cannot hold since

$$\int_0^t Wx_u(t, \phi) \cdot x_u(t, \phi)\, dt$$

diverges for all $u \in L_2(R^+, R^m)$.

REFERENCES

[1] H. T. Banks, *Variational problems involving functional differential equations*, this Journal, 7 (1969), pp. 1–17.

[2] ———, *Representations for solutions of linear functional differential equations*, J. Differential Equations, 5 (1969), pp. 399–409.

[3] R. H. Cameron and W. T. Martin, *An unsymmetric Fubini theorem*, Bull. Amer. Math. Soc., 47 (1941), pp. 121–125.

[4] H. Cartan, *Calcul différentiel*, Hermann, Paris, 1967.

[5] E. A. Coddington and N. Levinson, *Theory of Ordinary Differential Equations*, McGraw-Hill, New York, 1955.

[6] D. H. Chyung and E. B. Lee, *Linear optimal systems with time delays*, this Journal, 4 (1966), pp. 548–575.

[7] R. Datko, *A linear control problem in an abstract Hilbert space*, J. Differential Equations, 9 (1971), pp. 346–359.

[8] J. Dieudonné, *Foundations of Modern Analysis*, Academic Press, New York, 1960.

[9] N. Dunford and J. T. Schwartz, *Linear Operators*, Part I, John Wiley, New York, 1958.

[10] R. E. Edwards, *Functional Analysis, Theory and Applications*, Holt, Rinehart and Winston, New York, 1965.

[11] A. Halanay, *Differential Equations: Stability Oscillations, Time Lags*, Academic Press, New York, 1966.

[12] J. Hale, *Functional Differential Equations*, Springer-Verlag, New York, 1971.

[13] R. E. Kalman, *Contributions to the theory of optimal control*, Bol. Soc. Mat. Mexicana, 5 (1960), pp. 102–119.

[14] N. N. Krasovskii, *On the analytic construction of an optimum control in a system with time lags*, J. Appl. Math. Mech., 26 (1962), pp. 50–67.

[15] ———, *Stability of Motion*, Stanford University Press, Stanford, Calif., 1963.

[16] H. J. Kushner and D. I. Barnea, *On the control of a linear differential equation with quadratic cost*, this Journal, 8 (1970), pp. 257–272.

[17] E. B. Lee and L. Markus, *Foundations of Optimal Control Theory*, John Wiley, New York, 1967.

[18] J. L. Lions, *Contrôle optimal de systèmes gouvernés par des équations aux dérivées partielles*, Dunod, Paris, 1968.

[19] D. L. Lukes and D. L. Russell, *The quadratic criterion for distributed systems*, this Journal, 7 (1969), pp. 101–121.

[20] V. R. Nosov, *On a problem arising in the theory of optimal control with after effect*, Prikl. Mat. Mekh., 2 (1966), pp. 399–403.

[21] F. Riesz and B. Sz.-Nagy, *Functional Analysis*, Fredrick Ungar, New York, 1955.

[22] D. W. Ross and I. Flügge-Lotz, *An optimal control problem for systems with differential-difference equation dynamics*, this Journal, 7 (1969), pp. 609–623.

[23] K. Yosida, *Functional Analysis*, Springer-Verlag, New York, 1968.

[24] M. C. Delfour and S. K. Mitter, *L^2-stability, stabilizability and operator Riccati equations for linear autonomous hereditary differential systems*, to appear.

SIAM J. CONTROL
Vol. 11, No. 1, February 1973

THE RITZ–GALERKIN METHOD FOR ABSTRACT OPTIMAL CONTROL PROBLEMS*

JAMES W. DANIEL†

Abstract. This paper studies the Ritz–Galerkin procedure and abstract computational implementations of it as applied to solve abstract optimal control problems of the form: Minimize $f(x, u)$ over the set of (x, u) satisfying $G(x, u) = 0$ and possibly additional convex constraints $(x, u) \in B$. The specific model problem has $f(x, u) = \int_0^1 c(t, x(t), u(t))\,dt$, $G(x, u) = \dot{x}(t) + s(t, x(t), u(t))$, and, say, $B = \{(x, u); |x| \leqq 1, |u| \leqq 1\}$. Included in the computational procedures are generalized finite difference methods.

Consider the following basic optimal control problem: Minimize

$$f(x, u) = \int_0^1 c(t, x(t), u(t))\,dt \tag{1.1}$$

over the set of functions (x, u) satisfying the differential equation

$$\dot{x}(t) + s(t, x(t), u(t)) = 0 \quad \text{for almost all } t \text{ in } [0, 1], \qquad x(0) = 0, \tag{1.2}$$

and the additional control constraints

$$|u(t)| \leqq 1 \quad \text{for almost all } t \text{ in } [0, 1]. \tag{1.3}$$

We suppose that this problem has a unique solution (x^*, u^*) in some space $X \times U$. The simplest direct Rayleigh–Ritz method for treating this problem is well known and has been analyzed rather carefully [9], [7]. We simply let U_n be a finite-dimensional subspace of U and solve the above problem again with the additional restriction $u_n \in U_n$. Since x_n is uniquely determined from u_n via (1.2), $f(x_n, u_n)$ becomes a function of only finitely many variables u_n and can "easily" be minimized to yield an approximate solution to the original problem that can turn out to be quite accurate [1], [2]. In practice however one cannot solve (1.2) exactly for x from u except in special cases, for example when $s(t, x, u) = Ax + Bu$ with A and B constant; instead a numerical method must be used. If we approximately solve (1.2) by the Galerkin method in which x_n is chosen as an element of a finite-dimensional space $X_n \subset X$ satisfying, say,

$$\int_0^1 [\dot{x}_n(t) + s(t, x_n(t), u_n(t))]z(t)\,dt = 0 \tag{1.4}$$

for some finite set of functions z, the overall procedure of keeping $u_n \in U_n$, $x_n \in X_n$, and approximating (1.2) by (1.4) is called a Ritz–Galerkin method. Since (x_n, u_n) as determined by (1.4) no longer in general satisfies (1.2), the usual simple convergence proofs for ordinary Ritz methods do not apply; we give a convergence proof in §2.

* Received by the editors June 15, 1971, and in revised form December 27, 1971.

† Departments of Mathematics and Computer Sciences and Center for Numerical Analysis, The University of Texas at Austin, Austin, Texas 78712. This work was supported in part by the Office of Naval Research under Contract N0014-67-A-0128-0004 through the University of Wisconsin.

Example. Suppose for U_n we take the piecewise constant functions with joints at the equally spaced points $t_i = t_{n,i} = i/n$ for $0 \leqq i \leqq n$, for X_n we take the piecewise linear functions with the same joints, and for the functions z in (1.4) we choose the n functions

$$z_i(t) = \begin{cases} 1 & \text{if } t \in [t_i, t_{i+1}), \\ 0 & \text{elsewhere}, \end{cases} \quad \text{for} \quad 0 \leqq i \leqq n - 1.$$

Then the Galerkin equations (1.4) become, in terms of $x_{n,i} = x_n(t_i)$ and $u_{n,i} = u_n(t_i)$,

$$\text{(1.5)} \qquad x_{n,i+1} - x_{n,i} + \int_{t_i}^{t_{i+1}} s(t, x_n(t), u_{n,i})\, dt = 0 \quad \text{for } 0 \leqq i \leqq n - 1.$$

We immediately observe that (1.5) will be difficult to solve unless $s(t, x, u)$ is simple; in addition, evaluation of $f(x_n, u_n) = \int_0^1 c(t, x_n(t), u_n(t))\, dt$ will be difficult as well. Thus we need a further computational modification of the Ritz–Galerkin method in order to implement it. One obvious approach would be to replace (1.5) by

$$\text{(1.6)} \qquad x_{n,i+1} - x_{n,i} + \frac{1}{n} s(t_i, x_{n,i}, u_{n,i}) = 0 \quad \text{for } 0 \leqq i \leqq n - 1$$

and to evaluate $f(x_n, u_n)$ approximately as

$$\text{(1.7)} \qquad \frac{1}{n} \sum_{i=0}^{n-1} c(t_i, x_{n,i}, u_{n,i}).$$

Roughly speaking (1.6) amounts to replacing (1.4) by

$$\int_0^1 [\dot{x}_n(t) z_1(t) + s(t, x_n(t), u_n(t)) z_2(t)]\, dt = 0$$

for the set of pairs (z_1, z_2), where z_1 is the indicator step function used above and z_2 is a Dirac "function" at t_i. In § 3 we examine abstract models of this computational process.

Thus far we have not considered additional state and control constraints. If we have control constraints of the simple form $|u(t)| \leqq 1$ almost everywhere—or more generally $u(t) \in C$ almost everywhere for a fixed set C, the constraint $|u_n(t)| \leqq 1$ almost everywhere is easily implemented in terms of the finitely many variables $u_{n,i}$ by means of $|u_{n,i}| \leqq 1$ for $i = 0, 1, \cdots, n$. Thus there is usually no difficulty implementing discrete versions of control constraints; state constraints however can cause some problems. It is quite possible to construct examples—see our § 4—in which there are no points satisfying the discretized control, system and *state* constraints but in which there are points satisfying all the continuous constraints. For this reason, we shall have to examine in § 4 how one can conveniently treat complicated constraints in addition to the systems dynamics constraints (1.2).

2. The basic Ritz–Galerkin method. For simplicity, we describe our abstract theory in a Hilbert space setting, although essentially all that is needed is a space that is the dual of another and that has weak*-sequentially compact spheres such as does $L^\infty(0, 1)$, any reflexive space, or the dual of a separable space. For notational convenience, we use the same symbol $\|\cdot\|$ for norms in all spaces, where the context will make the meaning clear.

Let X and U be real Hilbert spaces and let $G\colon X \times U \to Z$, a real normed linear space, be continuous from the product norm topology to the norm topology. Let f be a norm continuous and weakly lower semicontinuous functional on $X \times U$ and let $B \subset U$ be a closed convex set. We consider the problem of minimizing $f(x, u)$ over the set

$$C = \{(x, u) \in X \times U;\ G(x, u) = 0,\ u \in B\}. \tag{2.1}$$

Although we could explicitly assume the existence of a solution to our minimization problem, some hypotheses that we shall need later serve to guarantee such existence; therefore we henceforth assume the following hypotheses.

H1. For each $u \in U$, there is a unique $x = x_u \in X$ solving $G(x, u) = 0$, and if u ranges over a bounded set then the resulting set of x_u is bounded.

H2. $G(x, u)$ is weakly sequentially closed at zero in Z, that is, if (x_n, u_n) and $G(x_n, u_n)$ converge weakly to (x, u) and zero respectively, then $G(x, u) = 0$.

H3. $f(x, u)$ tends to infinity with $\|u\|$ and $u \in U$.

By the usual arguments using minimizing sequences [7], it is easy to show that under hypotheses H1–H3 there exists at least one solution (x^*, u^*) to our minimization problem.

We now describe the Ritz–Galerkin discretization. Let X_n, U_n, Z'_n be finite-dimensional subspaces of X, U and Z' (the dual space of Z) respectively for each n. Let π_n denote orthogonal projection in X onto X_n, and let b_n denote orthogonal projection in U onto $B \cap U_n$ which is assumed to be nonempty. In many cases of interest, when $B = \{u; |u(t)| \leqq 1\}$ for example in $L^2(0, 1)$, if $u \in B$, then $b_n(u)$ is just the orthogonal projection of u onto U_n. Let $\{z'_{n,i}\}$ for $1 \leqq i \leqq m_n$ be a basis for Z'_n, and let $s_n\colon Z \to \mathbb{R}^{m_n}$ be defined as $s_n(z) = ((z'_{n,i}(z)))_{i=1}^{m_n}$. Our approximate problem now is that of minimizing $f(x_n, u_n)$ over

$$C_n = \{(x_n, u_n) \in X_n \times U_n;\ s_nG(x_n, u_n) = 0,\ u_n \in B\}, \tag{2.2}$$

which we assume is nonempty. We assume also the following hypothesis.

H4. For each $u_n \in U_n$, there is a unique $x_n \in X_n$ such that $s_nG(x_n, u_n) = 0$, and if there is a constant c_1 such that $\|u_n\| \leqq c_1$ for all n with $u_n \in B$, then there is a constant c_2 such that the generated x_n satisfy $\|x_n\| \leqq c_2$.

First we wish to show that there are some points in C_n near to (x^*, u^*).

LEMMA 2.1. *In addition to the preceding hypotheses, assume the following ones.*

H5. *There exists a constant*[1] $c > 0$ *such that* $\|s_n(z)\| \leqq c\|z\|$ *for all* $z \in Z$ *and such that* $\|s_nG(x_n, u_n) - s_nG(y_n, u_n)\| \geqq c\|x_n - y_n\|$ *for all* $x_n, y_n \in X_n$ *and* $u_n \in B \cap U_n$.

H6. $\|b_n(u^*) - u^*\| + \|\pi_n x^* - x^*\| \to 0$ *as* $n \to \infty$.

Define $r_n(x^*, u^*) = (r_nx^*, r_nu^*) \in C_n$ *via* $r_nu^* = b_n(u^*)$ *and* $r_nx^* = w_n$, *where* $s_nG(w_n, r_nu^*) = 0$. *Then* $\|r_n(x^*, u^*) - (x^*, u^*)\| \to 0$ *as* $n \to \infty$.

Proof. We need only show $\|x^* - r_nx^*\| \to 0$. We have

$$\begin{aligned}\|x^* - r_nx^*\| &\leqq \|x^* - \pi_nx^*\| + \|\pi_nx^* - r_nx^*\| \\ &\leqq \|x^* - \pi_nx^*\| + c\|s_nG(\pi_nx^*, r_nu^*) - s_nG(r_nx^*, r_nu^*)\|\end{aligned}$$

(cont.)

[1] We use c as a generic constant, perhaps different in each occurrence.

$$\leqq \|x^* - \pi_n x^*\| + c\|s_n G(\pi_n x^*, b_n(u^*))\|$$

$$\leqq \|x^* - \pi_n x^*\| + c\|G(\pi_n x^*, b_n(u^*)) - G(x^*, u^*)\|$$

which tends to zero. This completes the proof.

Let $(x_n^*, u_n^*) \in C_n$ satisfy $f(x_n^*, u_n^*) \leqq \inf_{C_n} f(x_n, u_n) + \varepsilon_n$ where $\varepsilon_n \geqq 0$ tends to zero. We know that $f(x_n^*, u_n^*) \leqq f(r_n x^*, r_n u^*) + \varepsilon_n$ which converges to $f(x^*, u^*)$, and hence by H3 and H4 there is a constant c_0 such that $\|x^*\| + \|u^*\| + \|x_n^*\| + \|u_n^*\| \leqq c_0$. Later we will need to know that limit points of $\{(x_n^*, u_n^*)\}$ lie in C. To this end we define

$$C^n = (C \cup C_n) \cap \{(x, u);\ \|x\| + \|u\| \leqq c_0\}. \tag{2.3}$$

LEMMA 2.2. *In addition to the above hypotheses, assume the following hypotheses.*

H7. *$\{G(x, u);\ x \in D,\ u \in E\}$ is a bounded set whenever D and E are bounded sets.*

H8. *For every $z' \in Z'$ and $\varepsilon > 0$, there exists N such that $n \geqq N$ implies $\varepsilon > d(z', Z'_n) = \inf\{\|z' - z'_n\|;\ z'_n \in Z'_n\}$.*

If $(x_{n_i}, u_{n_i}) \in C^{n_i}$ converges weakly to (x, u), then $(x, u) \in C$. Moreover, if $\gamma_n \equiv f(x^*, u^*) - \inf_{C^n} f(x, u)$, then $\gamma_n \geqq 0$ converges to zero as $n \to \infty$.

Proof. Let (x_{n_i}, u_{n_i}) satisfy the above assumptions; we claim that $G(x_{n_i}, u_{n_i})$ converges weakly to zero. To show this, let $z' \in Z'$, $\varepsilon > 0$, and choose N according to H8 and $z'_n \in Z'_n$ for $n \geqq N$ such that $\|z' - z'_n\| < \varepsilon$. Then, for large i,

$$|\langle z', G(x_{n_i}, u_{n_i})\rangle| \leqq |\langle z' - z'_{n_i}, G(x_{n_i}, u_{n_i})\rangle| + |\langle z'_{n_i}, G(x_{n_i}, u_{n_i})\rangle \leqq \varepsilon\|G(x_{n_i}, u_{n_i})\|$$

which, because of H7, can be made arbitrarily small; thus $G(x_{n_i}, u_{n_i})$ converges weakly to zero and so by H2 we have $G(x, u) = 0$. Since B is closed and convex, $u \in B$ and thus $(x, u) \in C$ as required. For the second part, clearly $\gamma_n \geqq 0$. Choose $(x_n, u_n) \in C^n$ such that $f(x_n, u_n) \leqq \inf_{C^n} f(x, u) + 1/n$; since the C^n are uniformly bounded, from any subsequence of $\{(x_n, u_n)\}$ we can choose a further subsequence $\{(x_{n_i}, u_{n_i})\}$ weakly converging to some point (x, u), which we now know must lie in C. Thus

$$0 \leqq f(x, u) - f(x^*, u^*) \leqq \liminf_{i \to \infty} f(x_{n_i}, u_{n_i}) - f(x^*, u^*)$$

$$\leqq \liminf_{i \to \infty} [\inf_{C^{n_i}} f(x, u)] - f(x^*, u^*) \leqq \liminf_{i \to \infty} (-\gamma_{n_i}).$$

Since $\gamma_n \geqq 0$, we have

$$0 \leqq \liminf_{i \to \infty} \gamma_{n_i} \leqq \limsup_{i \to \infty} \gamma_{n_i} = -\liminf_{i \to \infty} (-\gamma_{n_i}) \leqq 0,$$

and thus γ_{n_i} converges to zero.

With these basic lemmas in hand, it is now straightforward to prove our first theorem on the convergence of the Ritz–Galerkin method, following the style of the arguments in [6], [7].

THEOREM 2.3. *Under all the hypotheses of this section, let $(x_n^*, u_n^*) \in C_n$ satisfy $f(x_n^*, u_n^*) \leqq \inf_{C_n} f(x_n, u_n) + \varepsilon_n$, where $\varepsilon_n \geqq 0$ converges to zero. Then $\{(x_n^*, u_n^*)\}$ is a minimizing sequence, that is, $f(x_n^*, u_n^*)$ converges to $f(x^*, u^*) = \inf_C f(x, u)$, and $\{(x_n^*, u_n^*)\}$ has weak limit points all of which lie in C and minimize f over C; if f has a unique minimizer over C, then (x_n^*, u_n^*) converges weakly to it.*

Proof. By the definitions we can write $f(x^*, u^*) = \inf_{C^n} f(x, u) + \gamma_n \leqq f(x_n^*, u_n^*) + \gamma_n \leqq f(r_n x^*, r_n u^*) + \varepsilon_n + \gamma_n$, the latter of which converges to $f(x^*, u^*)$ by

Lemma 2.1 and the norm continuity of f. Therefore $\{(x_n^*, u_n^*)]$ is a minimizing sequence and, by the remarks following Lemma 2.1, it is a bounded sequence. It therefore has weak limit points (x_0, u_0) which, by Lemma 2.2, must lie in C. The weak lower semicontinuity of f then implies that $f(x_0, u_0) = f(x^*, u^*)$. The final statement is obvious.

In this paper we shall not discuss conditions under which the weak convergence concluded above can be replaced by norm convergence; once one has a minimizing sequence, there are quite general hypotheses which allow one to deduce norm convergence in [8], [10], [7], to which we refer the reader. The Ritz method has been successful in unconstrained minimization problems largely because of the excellent (essentially optimal) norm error bounds that one can deduce; it would be good to have similar results for our constrained problem. By using Lagrange multipliers and the Lagrangian $L(x, u, z') = f(x, u) + \langle z', G(x, u)\rangle$ when $B = U$, for the special case of (1.1) and (1.2), error bounds have been derived [1], [3] under the extremely strong and a priori unverifiable (except in special cases) hypothesis that $L(x, u, z')$ is uniformly convex in (x, u) for (x, u, z') near the optimal point (x^*, u^*, z'^*); these bounds carry over in a straightforward fashion to our general setting with $B = U$, but because of the extremely restrictive hypothesis mentioned above we examine this no further here.

Example. For clarity we present a simple example of the above results in a simple special case of the problem (1.1), (1.2), (1.3). In this format, let

$$s(t, x, u) = a(t, x) + b(t)u, \tag{2.4}$$

where $a(t, x)$ is Lipschitz continuous in x uniformly for t in $[0, 1]$ and $b(t)$ is continuous. Let

$$c(t, x, u) = q(t, x) + ru^2, \tag{2.5}$$

where $r > 0$ and $q(t, x)$ is continuous in (t, x) and convex and bounded below in x. More complicated dependence on u than that described here can be treated similarly, but the technical details tend to obscure the ideas and are therefore not presented here. To put this in our abstract setting, we let $U = Z = L^2(0, 1)$ and $X = \{x; x \in W^{1,2}(0, 1), x(0) = 0\}$, where $W^{1,2}(0, 1)$ is the usual Sobolev space of absolutely continuous functions having derivatives in $L^2(0, 1)$; the operator $G(x, u) = \dot{x} + a(t, x) + b(t)u$ satisfies H1, H2 and H7 as is easily shown. We have $B = \{u; |u(t)| \leqq 1 \text{ almost everywhere}\}$.

Suppose that for U_n and Z_n' we take the piecewise constant functions with equally spaced joints at $t_i = t_{n,i} = i/n$ for $0 \leqq i \leqq n$, with $\{z_{n,i}'\}_{i=0}^{n-1}$ chosen as the obvious orthogonal basis, while for X_n we take the piecewise linear functions with the same joints, exactly as in the example of § 1. Thus we seek to minimize

$$\sum_{i=0}^{n-1} \int_{t_i}^{t_{i+1}} \left\{ q\left(t, x_{n,i} + \frac{x_{n,i+1} - x_{n,i}}{h}(t - t_i) \right) + ru_{n,i}^2 \right\} dt,$$

where $h = 1/n$, over the (x_n, u_n) satisfying

$$x_{n,i+1} - x_{n,i} = \int_{t_i}^{t_{i+1}} \left\{ a\left(t, x_{n,i} + \frac{x_{n,i+1} - x_{n,i}}{h}(t - t_i) \right) + b(t)u_{n,i} \right\} dt,$$

$$x_{n,0} = 0 \quad \text{and} \quad |u_{n,i}| \leqq 1$$

for all i. By essentially the usual stability analysis in the numerical solution of ordinary differential equations, it is easy to see that H4 and H5 are satisfied, where the norm in $\mathbb{R}^n$ is given by $\|(w_1, \cdots, w_n)\| = \{h \sum_{i=1}^n w_i^2\}^{1/2}$. Clearly H3, H6 and H8 are satisfied. The weak convergence asserted by Theorem 2.3 in this case refers to $L^2(0, 1)$-weak convergence of $\dot{x}_n$ and u_n and to uniform (i.e., $C[0, 1]$) convergence of x_n.

3. Computational Ritz–Galerkin methods. In actual computations with control problems, as we saw in § 1, the pure Ritz–Galerkin method can be rather difficult to implement; one would rather use discrete constraints in the form (1.6) instead of (1.5) and evaluate costs via (1.7) instead of (1.1), for example. These discretizations represent the result of replacing $s(t, x(t), u(t))$ and $c(t, x(t), u(t))$ by interpolating piecewise constant functions; we now show how these ideas can be treated abstractly.

Let X, U, Z, X_n, U_n, Z_n', $z_{n,i}'$, s_n, B, b_n and π_n be as defined in § 2. Let $G: X \times U \to Z$ have the special structure

$$G(x, u) = G_1(x, u) + G_2(x, u), \tag{3.1}$$

where we shall treat in a different way $G_1(x, u)$, playing the role of $\dot{x}$ in applications, and $G_2(x, u)$, playing the role of $s(t, x, u)$. We suppose that we have operators p_n mapping $G_2(X_n, U_n)$ into Z. Suppose that $f: X \times U \to \mathbb{R}$ is norm continuous and weakly lower semicontinuous, with the special form

$$f(x, u) = g(D(x, u)), \tag{3.2}$$

where D maps $X \times U$ into a normed space W and g maps W into $\mathbb{R}$. We suppose that we have operators q_n mapping $D(X_n, U_n)$ into W; the operators p_n and q_n in applications will be, for example, piecewise constant interpolation mappings. Instead of minimizing $f(x, u)$ over $Q = \{(x, u) \in X \times U;\ G(x, u) = 0,\ u \in B\}$ we now minimize

$$f_n(x_n, u_n) \equiv g(q_n D(x_n, u_n)) \tag{3.3}$$

over

$$Q_n = \{(x_n, u_n) \in X_n \times U_n; s_n[G_1(x_n, u_n) + p_n G_2(x_n, u_n)] = 0, u_n \in B\}. \tag{3.4}$$

We wish to make the same kind of arguments as in § 2, with obviously necessary modifications.

Assume that H1 and H3 from § 2 hold.

We now need something somewhat different from H2, so we now assume the following

H2′. If $G(x_n, u_n) = 0$ for $(x_n, u_n) \in X \times U$ and if (x_n, u_n) converges weakly to (x, u), then $G(x, u) = 0$. If $G_1(x_n, u_n) + p_n G_2(x_n, u_n)$ converges weakly to zero with $(x_n, u_n) \in X_n \times U_n$ converging weakly to (x, u), then $G(x, u) = 0$.

Again it is easy to show that at least one (x^*, u^*) minimizing f over Q exists. To be sure that the Ritz–Galerkin method can be used at all, we assume the following.

H4′. For each $u_n \in U_n$, there is a unique $x_n \in X_n$ such that $s_n[G_1(x_n, u_n) + p_n G_2(x_n, u_n)] = 0$, and if there is a constant c_1 such that $\|u_n\| \leqq c_1$ and $u_n \in B$ for $u_n \in U_n$, then there is a constant c_2 such that the generated x_n satisfy $\|x_n\| \leqq c_2$.

As in Lemma 2.1, we now show there are points in Q_n near (x^*, u^*).

LEMMA 3.1. *In addition to the preceding hypotheses of this section, assume the following ones.*

H5′. *There exists a constant $c > 0$ such that $\|s_n(z)\| \leqq c\|z\|$ for all $z \in Z$ and such that*

$$\|s_n[G_1(x_n, u_n) + p_n G_2(x_n, u_n)] - s_n[G_1(y_n, u_n) + p_n G_2(y_n, u_n)]\| \geqq c\|x_n - y_n\|$$

for all $x_n, y_n \in X_n$ and $u_n \in B \cap U_n$.

H6′.

$$\|b_n(u^*) - u^*\| + \|\pi_n x^* - x^*\| + \|G_1(\pi_n x^*, b_n(u^*)) + p_n G_2(\pi_n x^*, b_n(u^*))\| \to 0 \quad \text{as } n \to \infty.$$

Define $r_n(x^*, u^*) = (r_n x^*, r_n u^*) \in Q_n$ via $r_n u^* = b_n(u^*)$ and $r_n x^* = w_n$, where $s_n[G_1(w_n, r_n u^*) + p_n G_2(w_n, r_n u^*)] = 0$. Then $\|r_n(x^*, u^*) - (x^*, u^*)\| \to 0$ as $n \to \infty$.

Proof. The proof is an obvious and trivial modification of that of Lemma 2.1.

Now let $(x_n^*, u_n^*) \in Q_n$ satisfy $f_n(x_n^*, u_n^*) \leqq \inf_{Q_n} f_n(x_n, u_n) + \varepsilon_n$, where $\varepsilon_n \geqq 0$ tends to zero. We know now that

$$f_n(x_n^*, u_n^*) \leqq f_n(r_n x^*, r_u u^*) + \varepsilon_n \tag{3.5}$$

and we want to be able to conclude that $\{(x_n^*, u_n^*)\}$ is uniformly bounded. This requires more work then in § 2. We here assume the following hypotheses.

H9. If $(x_n, u_n) \in X_n \times U_n, u_n \in B$, and if there is a constant c such that $\|x_n\| + \|u_n\| \leqq c$, then $|g(D(x_n, u_n)) - g(q_n D(x_n, u_n))| \to 0$ as $n \to \infty$.

H10. If $u_n \in U_n \cap B$ and if $\|u_n\| \to \infty$ as $n \to \infty$, then $g(q_n D(x_n, u_n)) \to \infty$.

From the hypotheses H9, H6′, Lemma 3.1, and the norm continuity of f, it follows that $f_n(r_n x^*, r_n u^*) = [f_n(r_n x^*, r_n u^*) - f(r_n x^*, r_n u^*)] + [f(r_n x^*, r_n u^*) - f(x^*, u^*)] + f(x^*, u^*)$ must converge to $f(x^*, u^*)$; thus from (3.5) we see that $f_n(x_n^*, u_n^*)$ is bounded which implies the boundedness of (x_n^*, u_n^*) via H10 and H4′. As in § 2 we can therefore choose c_0 such that $\|x_n^*\| + \|u_n^*\| + \|x^*\| + \|u^*\| \leqq c_0$ and define

$$Q^n = (Q \cup Q_n) \cap \{(x, u); \|x\| + \|u\| \leqq c_0\}. \tag{3.6}$$

Using the same kinds of arguments as in Lemma 2.2, we can easily prove the following lemma.

LEMMA 3.2. *In addition to the preceding hypotheses of this section, assume the following one.*

H7′. *If $(x_n, u_n) \in X_n \times (U_n \cap B)$ for each n and if there exists a c_1 such that $\|x_n\| + \|u_n\| \leqq c_1$, then there exists a c_2 such that $\|G_1(x_n, u_n) + p_n G_2(x_n, u_n)\| \leqq c_2$.*

Assume that H8 *from* § 2 *holds.*

If $(x_{n_i}, u_{n_i}) \in Q^{n_i}$ converges weakly to (x, u), then $(x, u) \in Q$. Moreover, if $\gamma_n \equiv f(x^*, u^*) - \inf_{Q_n} f(x, u)$, then $\gamma_n \geqq 0$ converges to zero as $n \to \infty$.

Proof. See the proof of Lemma 2.2.

With these lemmas, we now can prove convergence for the computationally feasible Ritz–Galerkin method.

THEOREM 3.1. *Under all the hypotheses of this section, let* $(x_n^*, u_n^*) \in Q_n$ *satisfy* $f_n(x_n^*, u_n^*) \leqq \inf f_n(x_n, u_n) + \varepsilon_n$, *where* $\varepsilon_n \geqq 0$ *tends to zero as* $n \to \infty$. *Then* $\{(x_n^*, u_n^*)\}$ *is a minimizing sequence, that is,* $f(x_n^*, u_n^*)$ *and* $f_n(x_n^*, u_n^*)$ *converge to* $f(x^*, u^*) = \inf_Q f(x, u)$, *and* $\{(x_n^*, u_n^*)\}$ *has weak limit points all of which lie in* Q *and minimize* f *over* Q; *if* f *has a unique minimizer over* Q, *then* (x_n^*, u_n^*) *converges weakly to it.*

Proof. Once that we prove $\{(x_n^*, u_n^*)\}$ to be a minimizing sequence, the remainder of the argument follows precisely as in the proof of Theorem 2.1. To this end, we consider the inequality

$$\begin{aligned} f(x^*, u^*) = \inf_{Q^n} f(x, u) + \gamma_n &\leqq f(x_n^*, u_n^*) + \gamma_n \\ &= f_n(x_n^*, u_n^*) + \gamma_n + \delta_n \leqq f_n(r_n x^*, r_n u^*) + \gamma_n + \delta_n + \varepsilon_n \\ &= f(x^*, u^*) + \gamma_n + \delta_n + \varepsilon_n + \eta_n, \end{aligned}$$

where γ_n and ε_n are as defined before while $\delta_n \equiv f(x_n^*, u_n^*) - f_n(x_n^*, u_n^*)$ and $\eta_n \equiv f_n(r_n x^*, r_n u^*) - f(x^*, u^*)$. We already know that γ_n and $\varepsilon_n \to 0$ by Lemma 3.2 and the hypothesis of this theorem; $\delta_n \to 0$ by H9 and the previously deduced boundedness of (x_n^*, u_n^*). We have $\eta_n = f_n(r_n x^*, r_n u^*) - f(r_n x^*, r_n u^*) + f(r_n x^*, r_n u^*) - f(x^*, u^*)$ which tends to zero by H6′, H9, Lemma 3.1, and the norm continuity of f. Therefore, shortening the above inequality to $f(x^*, u^*) \leqq f(x_n^*, u_n^*) + \gamma_n = f_n(x_n^*, u_n^*) + \gamma_n + \delta_n \leqq f(x^*, u^*) + \gamma_n + \delta_n + \varepsilon_n + \eta_n$, we conclude that $\{(x_n, u_n^*)\}$ is a minimizing sequence. The rest of the proof follows as in that of Theorem 2.1.

Again we remark that general conditions can be used to deduce norm convergence; error bounds could also be found under the restrictive hypothesis mentioned in § 2. We pursue these topics no further here.

Example. For clarity we look at the same examples as at the end of § 2, namely (1.1), (1.2), (1.3) in the special case of (2.4) and (2.5), although more complicated nonlinearities can be treated as well. We let $U, Z, X, X_n, U_n, Z_n', B$ and G be as in the example of § 2. We now let

$$G_1(x, u) = \dot{x}, \tag{3.7}$$

$$G_2(x, u) = a(t, x) + bu, \tag{3.8}$$

and let $p_n z$ and $q_n z$ both be, for any piecewise continuous function z, the piecewise constant interpolation of z at the joints $t_i = t_{n,i} = i/n = ih$ for $0 \leqq i \leqq n - 1$, the piecewise constant function chosen to be continuous from the right; that is, $(p_n z)(t_i) = z(t_i)$ for $t_i \leqq t < t_{i+1}$. We let $D(x, u) = c(t, x, u)$ and $g(w) = \int_0^1 w(t)\,dt$. The computational Ritz–Galerkin procedure now is to minimize $h \sum_{i=0}^{n-1} [q(t_{n,i}, x_{n,i}) + r u_{n,i}^2]$ subject to $|u_{n,i}| \leqq 1$ for $0 \leqq i \leqq n$, $x_{n,0} = 0$, and $x_{n,i+1} - x_{n,i} + h[a(t_{n,i} x_{n,i}) + b(t_{n,i}) u_{n,i}] = 0$ for $0 \leqq i \leqq n - 1$. As in § 2, it is a straightforward exercise to verify that the hypotheses in Theorem 3.1 are satisfied, allowing us to conclude the same kind of convergence of (x_n^*, u_n^*) as described at the end of § 2.

4. Complicated constraints. Thus far we have considered problems for which the constraints, except for the system laws $G(x, u) = 0$, were rather simple, namely $u \in B$ where $B \cap U_n$ was nonempty and presumably the discrete constraint $u_n \in U_n \cap B$ was easy to implement in terms of the finitely many parameters determining u_n; there were no state constraints such as $x \in E$. In the model control problem with $B = \{u; |u(t)| \leqq 1 \text{ almost everywhere}\}$ and U_n being the usual space of piecewise constant functions with joints at $\{i/n\}_{i=0}^{n}$, the constraint $u_n \in U_n \cap B$ is implemented simply as $|u_{n,i}| \leqq 1$ for $0 \leqq i \leqq n$. The same approach can of course be used on state constraints $x \in E$, such as $x(1) = 3$ and $|x(t)| \leqq 3$ everywhere; unfortunately this can lead to empty discrete constraint sets.

For example [6], [7], if the set

$$C = \{(x, u); \dot{x} = u, t^2 \leqq x(t) \leqq t^2 + t, 0 \leqq u(t) \leqq 3t \text{ on } [0, 1]\}$$

is discretized via

$$C_n = \{(x_n, u_n); x_{n,i+1} - x_{n,i} = (1/n)u_{n,i}, t_{n,i}^2 \leqq x_{n,i} \leqq t_{n,i}^2 + t_{n,i} \text{ and } 0 \leqq u_{n,i} \leqq 3t_{n,i} \quad \text{for } i = 0, 1, \cdots, n\},$$

then C is *nonempty* while C_n is *empty* for all n. However it is clear that there are points (x_n, u_n) with u_n satisfying the control constraints, (x_n, u_n) satisfying the discrete system dynamic constraints, and x_n being very *near* the state constraint set; we are led to allowing the discretized state constraints to be relaxed and expanded slightly. In this section, therefore, we wish to examine the Ritz–Galerkin procedure when complicated constraints of the form $x \in E$, $u \in B$ are present, for which the requirements $x_n \in E \cap X_n$ are difficult to implement, leading us to expand the discretized constraints. The question arises first as to how to discretize the constraint $x \in E$ say, other than by $x_n \in E \cap X_n$. Consider the case when $E = \{x; |x(t)| \leqq 1 \text{ everywhere}\}$ is discretized by requiring $|x_{n,i}| \leqq 1$ for $i = 0, 1, \cdots, n$. If one lets $j_{n,i}$ for $i = 1, \cdots, n + 1 = l_n$ be the linear functional $j_{n,i}(x) = x((i-1)/n)$, then this discretization merely asks $j_{n,i}(x_n) \in j_{n,i}(E) = \{r; r = j_{n,i}(e) \text{ for some } e \in E\}$ for $i = 1, \cdots, \ell_n$. This approach generalizes.

For each n, let $J_n = \{j_{n,i}; i = 1, \cdots, \ell_n\}$ be a set of linear functionals of norm one over X; let E be a closed convex subset of X. In the setting of § 3, we seek to minimize $f(x, u) = g(D(x, u))$ over

$$T = \{(x, u) \in X \times U; G(x, u) = 0, u \in B, x \in E\}. \tag{4.1}$$

We instead approximately minimize $f_n(x_n, u_n) = g(q_n D(x_n, u_n))$ over

$$T_n = \{(x_n, u_n) \in X_n \times U_n; s_n[G_1(x_n, u_n) + p_n G_2(x_n, u_n)] = 0, \quad d(J_n(x_n), J_n(E)) \leqq d_n, u_n \in B\}, \tag{4.2}$$

where

$$d(J_n(x), J_n(E)\} = \max_{1 \leqq i \leqq \ell_n} \inf_{e \in E} |j_{n,i}(x) - j_{n,i}(e)|. \tag{4.3}$$

Note that the requirements $d(J_n(x_n), J_n(E)) \leqq d_n$ merely replace $x_n \in E$ by the requirement that $j_{n,i}(x_n)$ lie within d_n of $j_{n,i}(E)$ for all i. Here d_n describes a small expansion of the constraint sets; its magnitude is not yet determined. We wish to choose d_n so that T_n will not be empty.

Assume H1, H2′ and H3 of §§ 2 and 3.

Then we have the existence of (x^*, u^*) as usual, if T is nonempty.

Assume H4′, H5′ and H6′ of § 3. Then the mapping r_n constructed in Lemma 3.1 satisfies $\|r_n(x^*, u^*) - (x^*, u^*)\| \to 0$. Let us define

$$d_n^* \geqq \|r_n x^* - x^*\|. \tag{4.4}$$

Then we can choose d_n^* so as to tend to zero, and $r_n(x^*, u^*) \in T_n$ if we define $T_n = T_n^*$ with $d_n = d_n^*$.

Assume H4′, H9, H10, H7′ and H8 of § 3.

Under these hypotheses, all of the arguments leading to Theorem 3.1 can be duplicated in our new setting, once we know that weak limits (x, u) of $(x_{n_i}, u_{n_i}) \in T^{n_i} = (T \cup T_{n_i}^*) \cap \{(x, u); \|x\| + \|u\| \leqq c_0\}$ must lie in T, as in Lemma 3.2. The arguments of Lemma 2.2 show easily that $G(x, u) = 0$ and $u \in B$; the only problem is to demonstrate that $x \in E$. To do this, clearly we need to know that J_n accurately represents E. By the construction of T^{n_i} from $T_{n_i}^*$ using $d_n - d_n^* \to 0$, we know that $d_{n_i}(J_{n_i}(x_{n_i}), J_{n_i}(E)) \to 0$. We can conclude $(x, u) \in E \times B$ from the following assumption.

H11. If (x_{n_i}, u_{n_i}) converges weakly to (x, u) and if $d_{n_i}(J_{n_i}(x_{n_i}), J_{n_i}(E))$ converges to zero as i tends to infinity, then $x \in E$.

We have thus outlined the development of the following result, the details of whose proof are easy to provide.

THEOREM 4.1. *Under all the hypotheses of this section, let* $(x_n^*, u_n^*) \in T_n$ *satisfy* $f_n(x_n^*, u_n^*) \leqq \inf_{T_n^*} f_n(x_n, u_n) + \varepsilon_n$, *where* $\varepsilon_n \geqq 0$ *tends to zero as* $n \to \infty$. *Then* $\{(x_n^*, u_n^*)\}$ *is a minimizing sequence, that is,* $f(x_n^*, u_n^*)$ *and* $f_n(x_n^*, u_n^*)$ *converge to* $f(x^*, u^*) = \inf_T f(x, u)$, $\{(x_n^*, u_n^*)\}$ *has weak limit points all of which lie in T and minimize f over T; if f has a unique minimizer over T, then* (x_n^*, u_n^*) *converges weakly to it.*

In order to implement the above algorithm, we require bounds d_n^* satisfying $d_n^* \geqq \|r_n x^* - x^*\|$, so that we know by how much to expand the discrete constraints. The problem of by how much to expand constraints in the special case of (1.1), (1.2) with complicated state and control constraints has been well studied. In [4] it was shown that there exist amounts by which we can expand and prove convergence, much as in Theorem 3.1; in [6], [7], under regularity hypotheses on u^*, it was shown essentially that $\|r_n x^* - x^*\|$ was $(1/n)$, so that $d_n^* = \sqrt{1/n}$, say, would work for large n. Most recently in [5], d_n was determined as the minimal expansion under which T_n is nonempty, and it was proved, under additional hypotheses, that d_n tends to zero and yields a theorem like Theorem 4.1.

The hypothesis H11 is interesting in itself and leads to the following question. Given a closed convex set E in a Hilbert space X, how can one choose $J_n = \{j_{n,i}; i = 1, \cdots, \ell_n\}$ such that the weak convergence of x_n to x and the convergence of $d_n(J_n(x_n), J_n(E))$ to zero implies $x \in E$? We are not aware of general results in this direction. In our model control problems, one often has $E = \{x; m(t) \leqq x(t) \leqq M(t)$ for all t, $x(\xi_i) \in C_i$ for $i = 1, 2, \cdots, q\}$, where m and M are continuous functions, the $\{\xi_i\}$ are given real numbers, and the C_i are given closed sets; we have $X = W^{1,2}(0, 1)$ as usual. Using our standard X_n of piecewise linear functions with joints at $t_{n,i} = i/n$ for $i = 0, 1, \cdots, n$, we can take $\ell_n = n + 1 + q$ and $j_{n,i}(x) = x(t_{n,i-1})$ for $i = 1, \cdots, n + 1$, $j_{n,i}(x) = x(\xi_{i-n-1})$ for $i = n + 2, \cdots, n + 1 + q$. If x_n converges weakly to x—so that $x_n(t)$ converges to

$x(t)$ uniformly on $[0, 1]$—and if $d_n(J_n(x_n), J_n(E))$ converges to zero, then it is easy to see that $x \in E$.

It is also possible in some cases that one have rather complex control constraints for which it would be difficult to implement the discrete constraint $u_n \in U_n \cap B$; this would be true for example if $B = \{u; 0 \leqq u(t) \leqq t \sin(1/t) + 1$ almost everywhere$\}$. If one has a constraint set $B = \{u; u(t) \leqq M(t)$ almost everywhere$\}$ with M continuous, it is tempting to discretize this by requiring

$$u_{n,i} \leqq M(t_{n,i}) \quad \text{for all } i. \tag{4.5}$$

This is the same as

$$\frac{1}{h}\int_{t_{n,i}}^{t_{n,i+1}} u_n(t)\,dt \leqq \frac{1}{h}\int_{t_{n,i}}^{t_{n,i+1}} [M(t_{n,i}) - M(t) + M(t)]\,dt = \frac{1}{h}\int_{t_{n,i}}^{t_{n,i+1}} M(t)\,dt + \varepsilon_n,$$

where ε_n is given by the modulus of continuity of M over intervals of length $h = 1/n$. Writing

$$j_{n,i}(u) = \frac{1}{h^{1/2}}\int_{t_{n,i}}^{t_{n,i+1}} u(t)\,dt,$$

we have $j_{n,i}(u_n) \leqq j_{n,i}(M) + \varepsilon_n h^{1/2}$, a format we can easily generalize much as in (4.2). Thus it is straightforward to generalize the results leading to Theorem 4.1 to allow expansions of the control constraints so as to include the simple discretization of (4.5). We leave the simple details for the reader.

REFERENCES

[1] W. E. Bosarge, Jr. and O. G. Johnson, *Direct method approximation to the state regulator problem using a Ritz-Trefftz suboptimal control*, IEEE Trans. Automatic Control, AC-15 (1970).

[2] ———, *Error bounds of high order accuracy for the state regulator problem via piecewise polynomial approximations*, this Journal, 9 (1971), pp. 15–28.

[3] W. E. Bosarge, Jr., O. G. Johnson, R. S. McKnight and W. P. Timlake, *The Ritz-Galerkin procedure for nonlinear control problems*, SIAM J. Numer. Anal., 10 (1973).

[4] J. Cullum, *Discrete approximation to continuous optimal control problems*, this Journal, 7 (1969), pp. 32–49.

[5] ———, *An explicit procedure for discretizing continuous optimal control problems*, J. Optimization Theory Appl., to appear.

[6] J. W. Daniel, *On the convergence of a numerical method for optimal control problems*, Ibid., 4 (1969), pp. 330–342.

[7] ———, *The Approximate Minimization of Functionals*, Prentice-Hall, Englewood Cliffs, N.J., 1971.

[8] E. S. Levitin and B. T. Poljak, *Convergence of minimizing sequences in conditional extremum problems*, Soviet. Math. Dokl., 7 (1966), pp. 764–767.

[9] S. C. Mikhlin and K. L. Smolitskiy, *Appropriate Methods for Solution of Differential and Integral Equations*, American Elsevier, New York, 1967.

[10] B. T. Poljak, *Existence theorems and convergence of minimizing sequences in extremum problems with restrictions*, Soviet Math. Dokl., 7 (1966), pp. 72–75.

SIAM J. CONTROL
Vol. 11, No. 1, February 1973

ON THE MINIMIZATION OF A QUADRATIC FUNCTIONAL SUBJECT TO A CONTINUOUS FAMILY OF LINEAR INEQUALITY CONSTRAINTS*

GRACE WAHBA†

Abstract. The problem of minimizing a positive definite quadratic functional subject to a continuous family of linear inequality constraints is studied. Upper and lower bounds are given for the value of the functional at the minimum. In certain cases, the given bounds coincide, and an explicit formula for the solution is given. Convergence rates for a sequence of (computable) approximate solutions obtained by discretizing the constraint set are established.

1. Introduction. Let $\mathscr{H}$ be a separable Hilbert space of real-valued functions and let $J(f)$ be a quadratic functional over $\mathscr{H}$ with $m\|f\|^2 \leqq J(f) \leqq M\|f\|^2$ for some $0 < m \leqq M < \infty$. Then, without further loss of generality, we may take as a norm

$$\|f\|_{\mathscr{H}}^2 = J(f) \tag{1.1}$$

and $2\langle f_1, f_2\rangle_{\mathscr{H}} = \|f_1 + f_2\|_{\mathscr{H}}^2 - \|f_1\|_{\mathscr{H}}^2 - \|f_2\|_{\mathscr{H}}^2$. Let $T = D \cup E$, where D is a finite set of points not in E and $E = [a, b]$ is a closed, bounded interval. Suppose $\{\eta_t, t \in T\}$ is a family of elements in $\mathscr{H}$, and $\alpha(t)$ a real-valued function defined on T such that

(i) every finite subfamily of $\{\eta_t, t \in T\}$ is linearly independent,

(ii) $0 < \alpha(t), t \in T$.

We consider the basic problem:

$$\text{Minimize} \quad J(f)\,(= \|f\|_{\mathscr{H}}^2) \tag{1.2}$$

subject to

$$\alpha(t) \leqq \langle \eta_t, f\rangle_{\mathscr{H}}, \qquad t \in T. \tag{1.3}$$

A solution $f^* \in \mathscr{H}$ of this minimization problem always exists and is unique. This follows since the feasible set $\mathscr{S}$ of functions satisfying (1.3) is closed and convex in $\mathscr{H}$, being the intersection of a family of closed half-spaces,

$$\mathscr{S} = \bigcap_t \mathscr{F}_t,$$

where $\mathscr{F}_t = \{f : \alpha(t) \leqq \langle \eta_t, f\rangle_{\mathscr{H}}\}$. Thus, $\mathscr{S}$ has a unique element of minimal norm.

The main result of this paper is the establishment of convergence rates for approximate solutions f_n^* obtained by discretizing the constraints of (1.3). Here f_n^* is the unique solution of the problem (1.2) subject to

$$\alpha(t) \leqq \langle \eta_t, f\rangle_{\mathscr{H}}, \qquad t \in T_n, \tag{1.4}$$

where $T_n = D \cup E_n$ and $E_n = \{a = t_1 < t_2 < \cdots < t_n = b\}$.

* Received by the editors September 28, 1971, and in revised form January 27, 1972.

† Department of Statistics, University of Wisconsin, Madison, Wisconsin 53706. This work was supported by the Wisconsin Alumni Research Foundation, and by the U.S. Air Force Office of Scientific Research under Grant AF-AFOSR 69-1803.

Let $Q(s, t)$ be the symmetric kernel on $T \times T$ defined by

$$Q(s,t) = \frac{1}{\alpha(s)\alpha(t)}\langle \eta_s, \eta_t\rangle_{\mathscr{H}}, \qquad s, t \in T. \tag{1.5}$$

Since the $\{\eta_t, t \in T\}$ are linearly independent, we always have that Q is positive definite in the sense that

$$\sum_{i,j=1}^{k} c_i c_j Q(s_i, s_j) > 0 \tag{1.6}$$

for any finite constants $c_1, c_2, \cdots, c_k$ not all 0, and distinct $s_1, s_2, \cdots, s_k \in T$. In this note we reduce the study of the solution f^* and its approximants f_n^* to the study of the properties of the kernel Q. In § 2 we show how this reduction is carried out. In § 3 we find upper and lower bounds on $\|f^*\|_{\mathscr{H}}$ in terms of Q. When the upper and lower bounds coincide we may determine f^* by inspection. This easy result is summarized in Theorem 1. Rates of convergence of f_n^* to f^* in terms of the continuity properties of Q on $E = [a, b]$ are given in § 4, Theorem 2.

The convergence results are germane to questions raised by works of Daniel [2] and Mangasarian and Schumaker [7]. In § 5 we discuss these applications and collect a number of examples, other applications, and remarks.

2. Transformation of the problem to canonical form. We now introduce a Hilbert space associated with Q which will be convenient in the sequel. Let Q_t be, for each fixed t, that function defined on T by

$$Q_t(s) = Q(s, t), \qquad s \in T.$$

Let $\mathscr{H}_Q^0$ be the vector space of real-valued functions defined on T of the form

$$g^0 = \sum_{i=1}^{l} c_i Q_{s_i} \tag{2.1}$$

for some finite l, finite constants $c_1, c_2, \cdots, c_l$ and any $s_1, s_2, \cdots, s_l$ points in T, with the inner product induced by

$$\langle Q_s, Q_t\rangle_Q = Q(s, t). \tag{2.2}$$

Equation (2.2) induces an inner product on $\mathscr{H}_Q^0$ because of (1.6). Let $\mathscr{H}_Q$ be the completion of $\mathscr{H}_Q^0$ with this inner product. $\mathscr{H}_Q$ is the reproducing kernel Hilbert space with reproducing kernel Q (see, for example, [9, § 6.2]). Q_t has the usual property for reproducing kernel spaces, namely,

$$\langle Q_t, g\rangle_Q = g(t), \qquad g \in \mathscr{H}_Q, \quad t \in T. \tag{2.3}$$

The reader may verify (2.3) by noting that if

$$g_k = \sum_{i=1}^{k} c_{ik} Q_{s_{ik}} \in \mathscr{H}_Q^0$$

and

$$\|g_k - g\|_Q \to 0,$$

then

$$\langle Q_t, g_k\rangle_Q = \left\langle Q_t, \sum_{i=1}^{k} c_{ik} Q_{s_{ik}} \right\rangle_Q = \sum_{i=1}^{k} c_{ik} Q_{s_{ik}}(t) = g_k(t) \tag{2.4}$$

and

$$|g_k(t) - \langle Q_t, g\rangle_Q| = |\langle Q_t, g_k - g\rangle_Q| \leqq \|Q_t\|_Q \|g_k - g\|_Q \to 0 \tag{2.5}$$

so that $g_k(t)$ converges for each fixed t to $\langle Q_t, g\rangle_Q$, giving (2.3). The span of the family $\{Q_t, t \in T\}$ is dense in $\mathscr{H}_Q$, since $\langle Q_t, g\rangle_Q = 0$, $t \in T$, implies $g = 0$.

Let V be the closure of the span of $\{\eta_t, t \in T\}$ in $\mathscr{H}$. For any $f \in \mathscr{H}$, let $f = f_1 + f_2$, where $f_1 \in V$ and $f_2 \in V^{\perp}$. Since $\langle \eta_t, f_2\rangle_{\mathscr{H}} = 0$, $t \in T$, then f satisfies (1.3) if f_1 satisfies (1.3); and it is obvious that the solution f^* must be in V. There is an isometric isomorphism between V and $\mathscr{H}_Q$ generated by the correspondence

$$\frac{1}{\alpha(t)}\eta_t \in V \sim Q_t \in \mathscr{H}_Q. \tag{2.6}$$

This follows since $\{\eta_t, t \in T\}$ span V, and

$$\frac{1}{\alpha(s)\alpha(t)}\langle \eta_s, \eta_t\rangle_{\mathscr{H}} = Q(s, t) = \langle Q_s, Q_t\rangle_Q. \tag{2.7}$$

Furthermore, $f \in V \sim g \in \mathscr{H}_Q$ if and only if

$$\frac{1}{\alpha(t)}\langle \eta_t, f\rangle_{\mathscr{H}} = g(t) = \langle Q_t, g\rangle_Q, \qquad t \in T. \tag{2.8}$$

If $f \in V \sim g \in \mathscr{H}_Q$, then

$$\|f\|_{\mathscr{H}} = \|g\|_Q. \tag{2.9}$$

Thus the minimization problem may now be reformulated as: Find $g \in \mathscr{H}_Q$ to

$$\text{Minimize} \quad \|g\|_Q^2 \tag{2.10}$$

subject to

$$1 \leqq \langle Q_t, g\rangle_Q = g(t), \qquad t \in T. \tag{2.11}$$

Then, if g^* is the solution to this problem, the solution f^* to the problem of (1.2) and (1.3) is given by finding f^* which corresponds to g^* under the isomorphism "~" of (2.6) and (2.8),

$$\frac{1}{\alpha(t)}\langle \eta_t, f^*\rangle_{\mathscr{H}} = g^*(t), \qquad t \in T. \tag{2.12}$$

In many cases (2.12) may be solved analytically for f^*, given g^*, by noting that if

$$g^*(t) = \sum_{j=1}^{l} c_j Q_{t_j}(t) \tag{2.13}$$

for some constants $\{c_j\}_{j=1}^l$, then

$$f^*(t) = \sum_{j=1}^{l} c_j\eta_{t_j}(t)/\alpha(t_j), \tag{2.14}$$

and if

$$g^*(t) = \lim_{l\to\infty} \sum_{j=1}^{l} c_{jl}Q_{t_{jl}}(t), \tag{2.15}$$

then

$$f^*(t) = \lim_{l\to\infty} \sum_{j=1}^{l} c_{jl}\eta_{t_{jl}}(t)/\alpha(t_{jl}). \tag{2.16}$$

3. Determination of the exact solution in certain cases. In certain cases, the solution f^* may be found by inspection. In any case, Theorem 1 gives upper and lower bounds for $\|f^*\|_{\mathscr{H}}$.

THEOREM 1. *Suppose* $\{\eta_t, t \in T\}$ *and* α *satisfy* (i) *and* (ii) *of* § 1 *and* Q *is defined by* (1.5). *Let* $T' = \{s : s \in T, \inf_{t\in T} Q_s(t) > 0\}$. *Then*

(i)

$$\inf_{s\in T'} \sup_{t\in T} \frac{\sqrt{Q(s,s)}}{Q(s,t)} \geqq \|f^*\|_{\mathscr{H}} \geqq \sup_{t\in T} \frac{1}{\sqrt{Q(t,t)}}, \tag{3.1}$$

where the left-hand side is interpreted as ∞ *if* T' *is empty.*

(ii) *If there exists a* (*necessarily unique*) s_* *for which*

$$\sup_{s\in T'} \frac{1}{\sqrt{Q(s,s)}} = \frac{1}{\sqrt{Q(s_*,s_*)}} \tag{3.2}$$

and

$$Q(s_*,s_*) = \inf_{t\in T} Q(s_*,t), \tag{3.3}$$

then

$$f^* = \frac{1}{Q(s_*,s_*)\alpha(s_*)}\eta_{s_*}. \tag{3.4}$$

Proof. The right-hand inequality in (i) follows from the Cauchy–Schwarz inequality:

$$\alpha(t) \leqq \langle \eta_t, f^*\rangle_{\mathscr{H}} \leqq \|\eta_t\|_{\mathscr{H}}\|f^*\|_{\mathscr{H}} = \alpha(t)\sqrt{Q(t,t)}\|f^*\|_{\mathscr{H}}. \tag{3.5}$$

The left-hand inequality in (i) follows by considering elements of the form g_s given by

$$g_s = \frac{Q_s}{\inf_{t\in T} Q_s(t)} \tag{3.6}$$

for $s \in T'$. Each such g_s satisfies the constraints (2.11) so that

$$\inf_{s\in T'} \|g_s\|_Q = \inf_{s\in T'} \frac{\|Q_s\|_Q}{\inf_{t\in T} Q_s(t)} = \inf_{s\in T'} \sup_{t\in T} \frac{\sqrt{Q(s,s)}}{Q(s,t)} \geqq \|g^*\|_Q = \|f^*\|_{\mathscr{H}}. \tag{3.7}$$

If there exists an s_* satisfying (3.2) and (3.3) it must be unique, since, by (1.6), if $s_* \neq s_{**}$,

$$Q(s_*, s_*)Q(s_{**}, s_{**}) > Q^2(s_*, s_{**}). \tag{3.8}$$

To prove (ii), note that if (3.2) and (3.3) hold, $s_* \in T'$, and

$$\frac{1}{\sqrt{Q(s_*, s_*)}} = \sqrt{Q(s_*, s_*)} \sup_{t \in T} \frac{1}{Q(s_*, t)} \geqq \inf_{s \in T'} \sup_{t \in T} \frac{\sqrt{Q(s, s)}}{Q(s, t)} \geqq \frac{1}{\sqrt{Q(s_*, s_*)}} \tag{3.9}$$

so that

$$\|g_{s_*}\|_Q = \frac{1}{\sqrt{Q(s_*, s_*)}} = \|f^*\|_{\mathscr{H}} = \|g^*\|_Q. \tag{3.10}$$

Hence, by the uniqueness of the solution,

$$g^* = g_{s_*} = \frac{Q_{s_*}}{Q(s_*, s_*)} \tag{3.11}$$

and, by (2.13) and (2.14),

$$f^* = \frac{1}{Q(s_*, s_*)\alpha(s_*)} \eta_{s_*}. \tag{3.12}$$

4. Properties of the approximate solution. Let f_n^* be the (unique) solution to the problem:

$$\text{Minimize} \quad \|f\|_{\mathscr{H}} \tag{4.1}$$

subject to

$$\alpha(t) \leqq \langle \eta_t, f \rangle_{\mathscr{H}}, \qquad t \in T_n, \tag{4.2}$$

where $T_n = \{s_1, s_2, \cdots, s_n, s_i \in T\}$.

f_n^* is obtained as follows: For any $f \in \mathscr{H}$, we may write

$$f = \sum_{i=1}^{n} c_i \eta_{s_i} / \alpha(s_i) + \rho,$$

where $\langle \eta_t, \rho \rangle_{\mathscr{H}} = 0$, $t \in T_n$. Then (4.1) and (4.2) become:

$$\text{Minimize} \quad cQ_n c' + \langle \rho, \rho \rangle_{\mathscr{H}} \tag{4.3}$$

subject to

$$\begin{pmatrix} 1 \\ 1 \\ \cdot \\ \cdot \\ \cdot \\ 1 \end{pmatrix} \leqq Q_n c', \tag{4.4}$$

where $c = (c_1, c_2, \cdots, c_n)$ and Q_n is the $n \times n$ matrix (of full rank) with (i, j)th element

$$[Q_n]_{ij} = \frac{1}{\alpha(s_i)\alpha(s_j)}\langle \eta_{s_i}, \eta_{s_j}\rangle_{\mathscr{H}} = Q(s_i, s_j).$$

It is obvious that ρ must be 0 in (4.3). Thus the problem is reduced to finding c to minimize cQ_nc' subject to (4.4), a standard quadratic programming problem. Let $c^* = (c_1^*, c_2^*, \cdots, c_n^*)$ be the (unique) solution to this problem.

Thus, $f_n^* \in V$ is given by

$$f_n^* = \sum_{i=1}^n c_i^* \eta_{s_i}/\alpha(s_i) \tag{4.5}$$

and

$$g_n^* = \sum_{i=1}^n c_i^* Q_{s_i} \tag{4.6}$$

is that element in $\mathscr{H}_Q$ corresponding to f_n^* under the isomorphism of (2.6). g_n^* is the solution to the problem:

$$\text{Minimize} \quad \|g\|_Q \tag{4.7}$$

subject to

$$1 \leqq \langle Q_t, g\rangle = g(t), \qquad t \in T_n. \tag{4.8}$$

It is our purpose to study the behavior of $\|f_n^* - f^*\|_{\mathscr{H}}$. Since $\|f_n^* - f^*\|_{\mathscr{H}} = \|g_n^* - g^*\|_Q$, we may restrict our attention to $\|g_n^* - g^*\|_Q$. For notational simplicity we let $T = \{\xi_i\}_{i=1}^l \cup [a, b]$, where $\{\xi_i\}_{i=1}^l = D$ consists of l isolated points not in $[a, b] = E$. The argument below may be carried through for $[a, b]$ replaced by any finite union of closed bounded intervals. The study of the behavior of $\|g_n^* - g^*\|_Q$ now proceeds by studying Q. We have the following lemma.

LEMMA 1. *Let $T = D \cup E$, $T_n = D \cup E_n$, where E is a closed, bounded interval $[a, b]$ and $E_n = \{a = t_1 < t_2 < \cdots < t_n = b\}$, with $\Delta = \max_i |t_{i+1} - t_i|$. Let Q satisfy* (1.6) *and have continuous mixed partial derivatives of all orders to $2p - 2$, and bounded mixed partial derivatives to order $2p - 1$ on $E \times E$. Then there exists $k = k(Q)$ depending only on Q such that*

$$g_n^*(t) \geqq 1 - k\|g_n^*\|_Q \Delta^{\min(p-1/2,2)} \tag{4.9a}$$

$$\geqq 1 - k\|g^*\|_Q \Delta^{\min(p-1/2,2)}, \qquad t \in T. \tag{4.9b}$$

Proof. Relation (4.9b) follows from (4.9a) since $\|g_n^*\|_Q \leqq \|g^*\|_Q$. Since $D \subset T_n$, it is only necessary to consider $t \in E$. Let $\{d_i\}_{i=1}^n$ be any set of real numbers with $d_i \geqq 0$, $\sum_{i=1}^n d_i = 1$. Then since $g_n^*(t_i) \geqq 1$, $\sum_{i=1}^n d_i g_n^*(t_i) \geqq 1$, and

$$\begin{aligned} \left| g_n^*(t) - \sum_{i=1}^n d_i g_n^*(t_i) \right| &= \left| \left\langle g_n^*, Q_t - \sum_{i=1}^n d_i Q_{t_i} \right\rangle_Q \right| \\ &\leqq \|g_n^*\|_Q \left\| Q_t - \sum_{i=1}^n d_i Q_{t_i} \right\|_Q . \end{aligned} \tag{4.10}$$

Thus

$$g_n^*(t) \geqq 1 - \|g_n^*\|_Q \left\| Q_t - \sum_{i=1}^n d_i Q_{t_i} \right\|_Q . \tag{4.11}$$

Now, for $p = 1, 2$, we find, for each t, a set of $\{d_i\}$ for which

$$\left\| Q_t - \sum_{i=1}^n d_i Q_{t_i} \right\|_Q \leqq k(Q)\Delta^{p-1/2}. \tag{4.12}$$

For $t \in [t_j, t_{j+1}]$, let $d_i = 0$, $i \neq j, j+1$, and

$$d_j = \frac{(t_{j+1} - t)}{(t_{j+1} - t_j)}, \qquad d_{j+1} = \frac{(t - t_j)}{(t_{j+1} - t_j)}. \tag{4.13}$$

Then, for $t \in [t_j, t_{j+1}]$,

$$\begin{aligned} \|Q_t - d_jQ_{t_j} - d_{j+1}Q_{t_{j+1}}\|^2 &= Q(t, t) - 2d_jQ(t_j, t) - 2d_{j+1}Q(t_{j+1}, t) \\ &\quad + d_j^2 Q(t_j, t_j) + 2d_jd_{j+1}Q(t_j, t_{j+1}) \\ &\quad + d_{j+1}^2 Q(t_{j+1}, t_{j+1}). \end{aligned} \tag{4.14}$$

For $p = 1$, Q has a bounded first derivative in each variable and

$$Q(u, v) = Q(t_j, t_j) + (u - t_j)k_1 + (v - t_j)k_2, \tag{4.15}$$

where k_1 and k_2 are bounded in absolute value by

$$\max_{t,x} \left| \frac{\partial}{\partial x} Q(t, x) \right| . \tag{4.16}$$

Substituting (4.15) into (4.14) with u, v replaced by t_j, t and t_{j+1} as appropriate, the zeroth order terms all drop out, leaving only terms involving $(u - t_j)$ and $(v - t_j)$. So, for $t \in [t_j, t_{j+1}]$, there exists $k = k(Q)$ such that

$$\|Q_t - d_jQ_{t_j} - d_{j+1}Q_{t_{j+1}}\|_Q^2 \leqq k^2\Delta, \qquad p = 1. \tag{4.17}$$

For $p = 2$, Q has continuous mixed partial derivatives of order 2 with the third order mixed partial derivatives bounded. Thus we may write

$$\begin{aligned} Q(u, v) &= Q(t_j, t_j) + [(u - t_j) + (v - t_j)]\alpha_1 + \left[\frac{(u - t_j)^2}{2!} + \frac{(v - t_j)^2}{2!}\right]\alpha_2 \\ &\quad + \frac{(u - t_j)(v - t_j)}{2!}\beta + \sum_{i=0}^{3} \frac{(u - t_j)^i}{i!}\frac{(v - t_j)^{3-i}}{(3 - i)!}k_{i+3}, \end{aligned} \tag{4.18}$$

where

$$\begin{aligned} \alpha_1 &= \frac{\partial}{\partial x} Q(t_j, x)|_{x=t_j}, \\ \alpha_2 &= \frac{\partial^2}{\partial x^2} Q(t_j, x)|_{x=t_j}, \\ \beta &= \frac{\partial^2}{\partial x \partial y} Q(x, y)|_{x=y=t_j}, \end{aligned} \tag{4.19}$$

and the k_i, $i = 3, 4, 5, 6$, are all bounded in absolute value by the bound on the absolute third mixed partial derivatives of Q.

Substituting (4.18) into (4.14), some tedious calculations, partly reproduced in the Appendix, show that all but the third order terms cancel out, giving the existence of a $k = k(Q)$ such that

$$\|Q_t - d_j Q_{t_j} - d_{j+1} Q_{t_{j+1}}\|_Q^2 \leqq k^2 \Delta^3, \qquad p = 2. \tag{4.20}$$

To begin the study of the case $p \geqq 3$, we next show that, if $g \in \mathscr{H}_Q$ with $p \geqq 3$, then $g''(t)$ exists and is continuous for $t \in E$. First, note that, for any δ,

$$\frac{g(t + 2\delta) - 2g(t + \delta) + g(t)}{\delta^2} = \langle g, (Q_{t+2\delta} - 2Q_{t+\delta} + Q_t)/\delta^2 \rangle_Q. \tag{4.21}$$

Next, letting $Q''_{t,\delta} = (Q_{t+2\delta} - 2Q_{t+\delta} + Q_t)/\delta^2$, it can be shown that Q''_{t,δ_m}, $m = 1, 2, \cdots$, is a Cauchy sequence in $\mathscr{H}_Q$ as $\delta_m \uparrow 0$ and as $\delta_m \downarrow 0$ through any sequence of real numbers, and that the two limit elements are the same. This demonstration proceeds by expanding out $\|Q''_{t,\delta_m} - Q''_{t,\delta_n}\|_Q^2$, using (2.2) and then using the hypothesis that $\partial^4 Q(u, v)/\partial^2 u \partial^2 v$ exists and is continuous on $E \times E$. The limit element, $\lim_{\delta_m \to 0} Q''_{t,\delta_m} = Q''_t$, say, is given by

$$Q''_t(t') = \frac{\partial^2}{\partial u^2} Q(u, t')|_{u=t}$$

and

$$\|Q''_t\|_Q^2 = \lim_{\delta \to 0} \|Q''_{t,\delta}\|_Q^2 = \left.\frac{\partial^4}{\partial^2 u \partial^2 v} Q(u, v)\right|_{u=v=t}. \tag{4.22}$$

Thus

$$\begin{aligned} \lim_{\delta \to 0} \frac{g(t + 2\delta) - 2g(t + \delta) + g(t)}{\delta^2} &= \langle g, \lim_{\delta \to 0} Q''_{t,\delta} \rangle_Q \\ &= \langle g, Q''_t \rangle_Q = h(t), \quad \text{say.} \end{aligned} \tag{4.23}$$

h is continuous, since $|h(t_1) - h(t_2)| = |\langle g, Q''_{t_1} - Q''_{t_2} \rangle_Q| \leqq \|g\|_Q \|Q''_{t_1} - Q''_{t_2}\|_Q \to 0$ as $t_1 \to t_2$, again by the continuity of $\partial^4 Q(u, v)/\partial^2 u \partial^2 v$. By a similar argument, we prove that g and g' are continuous, so that $h(t)$ of (4.23) equals $g''(t)$. Finally, then,

$$|g''(t)| \leqq \|g\|_Q \|Q''_t\|_Q \leqq \|g\|_Q \sup_t \left[\left.\frac{\partial^4}{\partial^2 u \partial^2 v} Q(u, v)\right|_{u=v=t} \right]^{1/2}. \tag{4.24}$$

We may now expand g in a Taylor series as

$$g(t_i) = g(t) + (t_i - t)g'(t) + \int_t^{t_i} (t_i - u) g''(u)\, du. \tag{4.25}$$

For $t \in [t_j, t_{j+1}]$, choose $\{d_i\}$ as in (4.13). Then, with the aid of (4.24),

$$\begin{aligned} |g(t) - d_j g(t_j) - d_{j+1} g(t_{j+1})| &= d_j \int_t^{t_j} (t_j - u) g''(u)\, du \\ &\quad + d_{j+1} \int_t^{t_{j+1}} (t_{j+1} - u) g''(u)\, du \\ &\leqq k \|g\|_Q \frac{(t_{j+1} - t_j)^2}{2} \leqq \frac{k}{2} \|g\|_Q \Delta^2, \end{aligned} \tag{4.26}$$

where

$$k = \sup_t \left[\frac{\partial^4}{\partial^2 u \partial^2 v} Q(u, v) \bigg|_{u=v=t} \right]^{1/2}. \tag{4.27}$$

Thus we have proved (4.9a) for $p \geqq 3$.

We may ask if these rates may be improved upon. Suppose

(i) $\partial^l Q(t, t')/\partial t^l$ exists and is continuous on $E \times E$ for $t \neq t', l = 0, 1, 2, \cdots, 2p$, $\partial^l Q(t, t')/\partial t^l$ exists and is continuous on $E \times E$ for $l = 0, 1, 2, \cdots, 2p - 2$.

(ii) $\lim_{t \uparrow t'} \partial^{2p-1} Q(t, t')/\partial t^{2p-1}$ and $\lim_{t \downarrow t'} \partial^{2p-1} Q(t, t')/\partial t^{2p-1}$ exist and are bounded for all $t' \in E$.

It may be shown (see [11] and [12]) that if (i) and (ii) hold, there exist constants $\{e_i\}$ for which

$$\left\| Q_t - \sum_{i=1}^n e_i Q_{t_i} \right\|_Q = O(\Delta^{p-1/2}). \tag{4.28}$$

Examples may be found in [10] for which it can be shown that this rate is exact.

We now give an example, with $p = 3$, which demonstrates that this method of proof cannot be used to show that $g_n^*(t) \geqq 1 - k\|g_n^*\|_Q \Delta^{2+\varepsilon}$ for $\varepsilon > 0$. The method of proof depended on finding $\{d_i\}$ such that $d_i \geqq 0$, $\sum_{i=1}^n d_i = 1$ so that

$$\left| g(t) - \sum_{i=1}^n d_i g(t_i) \right| \leqq k\|g\|_Q \Delta^{\min(p-1/2,2)} \tag{4.29}$$

for all $g \in \mathscr{H}_Q$ and all t, where k depends only on Q.

Let

$$Q(s, t) = \sum_{j=0}^2 \frac{s^j t^j}{(j!)^2} + \int_0^1 \frac{(s-u)_+^2}{2!} \frac{(t-u)_+^2}{2!} du,$$

where

$$(x)_+ = \begin{cases} x & \text{if } x > 0, \\ 0 & \text{otherwise.} \end{cases}$$

Q satisfies the hypothesis of Theorem 2 with $p = 3$. Then $\mathscr{H}_Q = \{g : g''$ absolutely continuous, $g''' \in \mathscr{L}_2[0, 1]\}$ and

$$\langle g_1, g_2 \rangle_Q = \sum_{j=0}^2 \frac{g_1^{(j)}(0) g_2^{(j)}(0)}{j!} + \int_0^1 g_1'''(u) g_2'''(u)\, du.$$

Here we may expand g in a Taylor series with remainder to order 2 and have

$$\begin{aligned} \left| g(t) - \sum_{i=1}^n g(t_i) d_i(t) \right| = \bigg| g(t) - \sum_{i=1}^n \bigg\{ & g(t) + (t_i - t) g'(t) + \frac{(t_i - t)^2}{2!} g''(t) \\ & + \int_t^{t_i} \frac{(t_i - u)^2}{2!} g'''(u)\, du \bigg\} d_i(t) \bigg|. \end{aligned} \tag{4.30}$$

Here

$$\left|\int_t^{t_i}\frac{(t_i-u)^2}{2!}g'''(u)\,du\right| \leqq \left[\int_t^{t_i}\frac{(t_i-u)^4}{(2!)^2}\,du\right]^{1/2}\left[\int_t^{t_i}[g'''(u)]^2\,du\right]^{1/2}$$
$$\leqq \frac{(t_i-t)^{5/2}}{\sqrt{5}\cdot 2!}\|g\|_Q.$$

Thus, in order that the right-hand side of (4.30) is $o(\Delta^2)$ independently of the choice of g we need that

$$\sum_{i=1}^{n}(t_i-t)\,d_i = o(\Delta^2), \tag{4.31}$$

$$\sum_{i=1}^{n}(t_i-t)^2\,d_i = o(\Delta^2). \tag{4.32}$$

For (4.32) to hold with $d_i > 0$, $\sum d_i = 1$, we need for each t that there be an i such that $(t_i - t)^2 = o(\Delta^2)$. But suppose $t = \frac{1}{2}(t_j + t_{j+1})$, where $(t_{j+1} - t_j) = \Delta$; then $(t_i - t)^2 \geqq (\frac{1}{2}\Delta)^2$ for all i, so that (4.32) is impossible.

We are now able to prove the following theorem.

THEOREM 2. *Let T, T_n, Q and Δ be as in Lemma* 1. *Then*

$$\|f_n^* - f^*\|_{\mathscr{H}} = O(\Delta^{\min(p-1/2,2)/2}). \tag{4.33}$$

Proof. We may replace $\|f_n^* - f^*\|_{\mathscr{H}}$ by $\|g_n^* - g^*\|_Q$. Let k be as in (4.9a) and let $\gamma_n = k\|g^*\|_Q\Delta^{\min(p-1/2,2)}$. Then by Lemma 1 we have

$$\inf_t g_n^*(t) \geqq 1 - \gamma_n. \tag{4.34}$$

Since

$$\frac{g_n^*(t)}{1-\gamma_n} \geqq 1 \quad\text{and}\quad \|g_n^*\|_Q \leqq \|g^*\|_Q, \tag{4.35}$$

we must have

$$\left\|\frac{g_n^*}{(1-\gamma_n)}\right\|_Q = \frac{1}{(1-\gamma_n)}\|g_n^*\|_Q \geqq \|g^*\|_Q \geqq \|g_n^*\|_Q \geqq \|g^*\|_Q(1-\gamma_n) \tag{4.36}$$

and so

$$\|g_n^*\|_Q = \|g^*\|_Q(1-\theta\gamma_n) \tag{4.37}$$

for some $\theta \in [0, 1]$, or

$$\|g_n^*\|_Q^2 = \|g^*\|_Q^2(1 - 2\theta\gamma_n + \theta^2\gamma_n^2), \tag{4.38}$$

$$2\|g^*\|_Q^2\gamma_n \geqq \|g^*\|_Q^2 - \|g_n^*\|_Q^2 = \langle g^* - g_n^*, 2g_n^*\rangle_Q + \langle g^* - g_n^*, g^* - g_n^*\rangle_Q. \tag{4.39}$$

Now, $\langle g^*, g_n^*\rangle_Q \geqq \langle g_n^*, g_n^*\rangle_Q$ since g^* is in the closed convex set of functions satisfying the constraints for $t \in T_n$, and g_n^* is the element of minimal norm in this convex set. (Draw a picture.) Thus, we have $\langle g^* - g_n^*, 2g_n^*\rangle_Q \geqq 0$ and

$$2k\|g^*\|_Q^3\Delta^{\min(p-1/2,2)} = 2\|g^*\|_Q^2\gamma_n \geqq \|g^* - g_n^*\|_Q^2 = \|f^* - f_n^*\|_{\mathscr{H}}^2. \tag{4.40}$$

5. Remarks and applications. 1. As an example of a Hilbert space and quadratic functional, let $\mathscr{H}$ be the Sobolev space $W^{m,2}$,

$$W^{m,2} = \{f: f^{(m-1)} \text{ absolutely continuous}, f^{(m)} \in \mathscr{L}_2[0, 1]\}$$

and

$$J(f) = \sum_{j=0}^{m-1} \lambda_j (f^{(j)}(0))^2 + \int_0^1 (L_m f(t))^2 \, dt, \tag{5.1}$$

where $\lambda_i > 0$, $i = 0, 1, \cdots, m-1$, and

$$L_m f(t) = \sum_{j=0}^{m} a_j(t) f^{(j)}(t), \qquad a_m(t) \neq 0 \text{ in } [0, 1], \tag{5.2}$$

$$a_j(t) \in C^{2m}, \quad j = 0, 1, 2, \cdots, m.$$

This Hilbert space is such that all the linear functionals N_t^ν defined by $N_t^\nu f \to f^{(\nu)}(t)$, $t \in [0, 1]$, are continuous for $\nu = 0, 1, \cdots, m-1$.

Whenever the evaluation functionals $N_s f \to f(s)$ and $N_s^\nu f \to f^{(\nu)}(s)$, $\nu = 0, 1, \cdots, m-1$, are all continuous in $\mathscr{H}$, then strong convergence entails pointwise convergence of $f_n^{*(\nu)}(s)$ to $f^{*(\nu)}(s)$ for $\nu = 0, 1, \cdots, m-1$, as follows: If R_s^ν is the representer of N_s^ν, that is,

$$N_s^\nu f = \langle R_s^\nu, f \rangle_{\mathscr{H}} = f^{(\nu)}(s), \tag{5.3}$$

$$f \in \mathscr{H}, \quad \nu = 0, 1, \cdots, m-1,$$

then

$$|f_n^{*(\nu)}(s) - f^{*(\nu)}(s)| = |\langle f_n^* - f^*, R_s^\nu \rangle_{\mathscr{H}}| \leqq \|R_s^\nu\|_{\mathscr{H}} \|f_n^* - f^*\|_{\mathscr{H}}. \tag{5.4}$$

2. Atteia [1] and Ritter [8] considered minimizing

$$J(f) = \int_0^1 (L_m f(t))^2 \, dt \tag{5.5}$$

in $W^{m,2}$ subject to the constraints

$$\alpha(t) \leqq \sum_{j=0}^{m-1} c_j(t) f^{(j)}(t) \leqq \beta(t), \qquad t \in T_n, \tag{5.6}$$

for some real numbers $\{c_j(t), j = 0, 1, \cdots, m-1, t \in T_n\}$, where T_n is a finite set, $T_n = \{t_1, t_2, \cdots, t_n\}$. For $\beta(t) = \infty$, this problem is of the type considered here, with $\{\eta_t, t \in T_n\}$ being the representers of the continuous linear functionals $\sum_{j=0}^{m-1} c_j(t) N_t^j$, $t \in T_n$, appearing in (5.6). However, $J(f)$ of (5.5) is only a seminorm on $W^{m,2}$, and the solution may not be unique. It was shown in [8] that the solution of this problem can be reduced to the solution of a (finite) standard quadratic programming problem. This problem was also discussed by Jerome and Schumaker [3], who also considered continuous equality constraints.

3. Mangasarian and Schumaker [6] and Laurent [5] considered generalizations of the problem in the above remark, obtaining them by enlarging the constraint set by replacing T_n by $T = D \cup E$, where D is a finite set of points and E is a finite union of closed, bounded intervals. An example of such a set of

constraints is:

$$\begin{aligned}(5.7)\qquad &\alpha_D(t) \leqq \sum_{j=0}^{m-1} c_j(t)f^{(j)}(t) \leqq \beta_D(t), \\ &\qquad\qquad\qquad t \in D = \{\xi_1, \xi_2, \cdots, \xi_l\}, \\ &\alpha_{E_i}(t) \leqq \sum_{j=0}^{m-1} b_{ij}(t)f^{(j)}(t) \leqq \beta_{E_i}(t), \\ &\qquad t \in E_i = [0, 1], \quad E = \bigcup_i E_i, \quad i = 1, 2, \cdots, k.\end{aligned}$$

Various characterizations of the (possibly nonunique) solutions are discussed.

4. Mangasarian and Schumaker [7] discussed a completely discrete analogue of the problem of Remark 3. That is, $W^{m,2}$ is replaced by a Euclidean space, and differential operators and derivatives are replaced by difference operators and divided differences.

5. Daniel [2] considered an approximate method of solving the minimization problem of Remark 3 with J and the constraints as in (5.5) and (5.7). Daniel's procedure consists of replacing E_i in (5.7) by the discrete set $E_{in} \subset E_i$,

$$E_{in} = \left\{\frac{j}{n}, j = 0, 1, 2, \cdots, n\right\}.$$

A solution f_n^* to the problem: minimize $J(f)$ of (5.5) subject to

$$\begin{aligned}(5.8)\qquad &\alpha_D(t) \leqq \sum_{j=0}^{m-1} c_j(t)f^{(j)}(t) \leqq \beta_D(t), \qquad t \in D, \\ &\alpha_{E_i}(t) \leqq \sum_{j=0}^{m-1} b_{ij}(t)f^{(j)}(t) \leqq \beta_{E_i}(t), \qquad t \in E_{in},\end{aligned}$$

may be found by the methods described in [8], and is taken by Daniel as the nth approximation to the solution of the original minimization problem with constraints (5.7). He discusses the convergence properties of f_n^* as $n \to \infty$. No convergence rates are given however.

6. If $\mathscr{H}$ itself possesses a (known) reproducing kernel, we may always find $\langle \eta_s, \eta_t \rangle_{\mathscr{H}}$ of (1.5), where η_s is the representer of an arbitrary continuous linear functional M_s on $\mathscr{H}$. The necessary and sufficient condition that $\mathscr{H}$ possess a (real-valued) reproducing kernel $R(u, v)$, $u, v \in U$, is that $\mathscr{H}$ be a Hilbert space of real-valued functions defined on U and all the evaluation functionals N_u, $u \in U$, $N_u f \to f(u)$ be continuous. In this case,

$$(5.9)\qquad \langle \eta_s, \eta_t \rangle_{\mathscr{H}} = M_{s(u)}M_{t(v)}R(u, v),$$

where $M_{s(u)}$ means the linear functional M_s applied to R considered as a function of u. (For further discussion of Hilbert spaces possessing a reproducing kernel, see [4], [9], [11] and references cited therein.)

7. We now use (5.9) and Theorem 2 to obtain convergence rates for approximate solutions to a problem similar to that considered by Daniel. Consider the

problem of minimizing $J(f)$ of (5.1) subject to

$$\alpha(t) \leqq M_t f \overset{\text{def}}{=} \sum_{j=0}^{q} b_j(t) f^{(j)}(t), \qquad t \in [a, b] \subset [0, 1], \tag{5.10}$$

where $q \leqq m - 1$. The linear functionals M_t, $t \in [a, b]$, are all continuous on $W^{m,2}$. It is well known (see, for example, Kimeldorf and Wahba [4]) that $W^{m,2}$ with the norm defined by $\| f \|_{\mathscr{H}} = J(f)$ of (5.1) possesses the reproducing kernel $R(s, t)$ given by

$$R(s, t) = \sum_{j=0}^{m-1} \frac{\phi_j(s)\phi_j(t)}{\lambda_j} + \int_0^{\min(s,t)} G(s, u)G(t, u)\, du, \qquad s, t \in [0, 1], \tag{5.11}$$

where

$$\begin{aligned} L_m \phi_j &= 0, & j &= 0, 1, 2, \cdots, m - 1, \\ \phi_j^{(v)}(0) &= \delta_{vj}, & v, j &= 0, 1, 2, \cdots, m - 1, \end{aligned}$$

and $G(s, u)$ is the Green's function for the problem

$$\begin{aligned} L_m f &= g, & & \\ f^{(v)}(0) &= 0, & v &= 0, 1, 2, \cdots, m - 1. \end{aligned}$$

Then

$$\begin{aligned} Q(s, t) &= \frac{1}{\alpha(s)\alpha(t)} \langle \eta_s, \eta_t \rangle_{\mathscr{H}} = \frac{1}{\alpha(s)\alpha(t)} M_{s(u)} M_{t(v)} R(u, v) \\ &= \frac{1}{\alpha(s)\alpha(t)} \sum_{j,k=0}^{q} b_j(s) b_k(t) \frac{\partial^{j+k}}{\partial s^j \partial t^k} R(s, t), \qquad s, t \in [a, b]. \end{aligned} \tag{5.12}$$

By recalling the properties of Green's functions, we see that

$$\frac{\partial^{j+k}}{\partial s^j \partial t^k} R(s, t), \quad j, k = 0, 1, \cdots, q, \quad s, t \in [a, b], \tag{5.13}$$

has continuous mixed partial derivatives to order at least $2(m - q) - 2$ and bounded (left and right) derivatives of order $2(m - q) - 1$. Thus if $\alpha, b_j \in C^{2(m-q)-2}$, $\alpha^{2(m-q)-1}$, $b_j^{2(m-q)-1}$ bounded, then Q of (5.12) satisfies the hypotheses of Theorem 2 with $p = m - q$. Thus, if f_n^* are the approximate solutions to this problem, as described in § 1, then Theorem 2 and (5.4) give

$$| f_n^{*(v)}(t) - f^{*(v)}(t)| = O(\Delta^{(\min(m-q-1/2, 2))/2}), \qquad v = 0, 1, \cdots, m - 1. \tag{5.14}$$

The approximate solutions f_n^* to this problem are Lg-spline functions in the sense of [3]. To see the connection with [3], note that f_n^* is a linear combination of the functions $\{\eta_t, t \in T_n\}$, where η_t is the representer of M_t of (5.10). We may find $\eta_t(s)$ as follows: Letting R_s be the representer of the evaluation functional in $W^{m,2}$ with norm given by $J(f)$ of (5.1), then $R_s(s') = R(s, s')$ with R as in (5.11). Then

$$\eta_t(s) = \langle \eta_t, R_s \rangle_{\mathscr{H}} = M_t R_s = \sum_{j=0}^{q} b_j(t) \frac{\partial^j}{\partial t^j} R(s, t).$$

When R is given by (5.11), then, for each fixed t, $\partial^j R(s,t)/\partial t^j$, considered as a function of s, is an Lg-spline with knot at t. Thus f_n^* is an Lg-spline with knots for $t \in T_n$.

8. We give an example of the application of part (ii) of Theorem 1. We seek the solution to the problem: Find $f^* \in W^{2,2}$ to minimize

$$J(f) = \lambda_0 f^2(0) + \lambda_1 (f'(0))^2 + \int_0^1 (w(t) f''(t))^2 \, dt \tag{5.15}$$

subject to

$$\alpha(t) \leqq f'(t), \qquad t \in T = [\tfrac{1}{2}, 1], \tag{5.16}$$

where $\lambda_0, \lambda_1 > 0$, and α and w are given positive functions. The reproducing kernel for $W^{2,2}$ with $\|f\|^2_{\mathscr{H}} = J(f)$ is

$$R(s,t) = \frac{1}{\lambda_0} + \frac{st}{\lambda_1} + \int_0^{\min(s,t)} \frac{(s-u)(t-u)}{w^2(u)} du. \tag{5.17}$$

The representer η_t of the continuous linear functional N'_t defined by $N'_t f \to f'(t)$ is given by

$$\eta_t(s) = \langle \eta_t, R_s \rangle_{\mathscr{H}} = N'_t R_s = \frac{\partial}{\partial t} R(s,t) = \frac{s}{\lambda_1} + \int_0^{\min(s,t)} \frac{(s-u)}{w^2(u)} du. \tag{5.18}$$

Here

$$\begin{aligned} Q(s,t) &= \frac{1}{\alpha(s)\alpha(t)} \langle \eta_s, \eta_t \rangle_{\mathscr{H}} = \frac{1}{\alpha(s)\alpha(t)} \frac{\partial^2}{\partial s \partial t} R(s,t) \\ &= \frac{1}{\alpha(s)\alpha(t)} \left[\frac{1}{\lambda_1} + q(\min(s,t)) \right], \qquad s, t \in T, \end{aligned} \tag{5.19}$$

where

$$q(s) = \int_0^s \frac{1}{w^2(u)} du.$$

If $\alpha(t)$ is nonincreasing on $T = [\frac{1}{2}, 1]$, then with $s_* = \frac{1}{2}$,

$$\begin{gathered} \sup_{t \in T} \frac{1}{\sqrt{Q(t,t)}} = \frac{1}{\sqrt{Q(s_*, s_*)}}, \\ Q(s_*, s_*) = \inf_{t \in T} Q(s_*, t). \end{gathered} \tag{5.20}$$

Thus Q with $s_* = \frac{1}{2}$ satisfies (3.2) and (3.3) and so

$$f^* = \frac{1}{\alpha(s_*) Q(s_*, s_*)} \eta_{s_*},$$

$$f^*(s) = \frac{\alpha(1/2)}{(1/\lambda_1 + q(1/2))} \left[\frac{s}{\lambda_1} + \int_0^{\min(s,1/2)} \frac{(s-u)}{w^2(u)} \right] du. \tag{5.21}$$

Similarly, if

$$\frac{(1+\lambda_1 q(t))}{(1+\lambda_1 q(1))} \geqq \frac{\alpha(t)}{\alpha(1)} \geqq \left(\frac{\alpha(t)}{\alpha(1)}\right)^2, \tag{5.22}$$

then Q with $s_* = 1$ satisfies (3.2) and (3.3) and so

$$f^* = \frac{1}{\alpha(1)Q(1,1)}\eta_1, \tag{5.23}$$

$$f^*(s) = \frac{\alpha(1)}{(1/\lambda_1 + q(1))}\left\{\frac{s}{\lambda_1} + \int_0^s \frac{(s-u)}{w^2(u)}du\right\}.$$

Appendix. Table for calculations, substitution of (4.18) into (4.14).

			Coefficients in (4.18)			
			$Q(t_j, t_j)$	α_1	$\frac{1}{2!}\alpha_2$	$\frac{1}{2!}\beta$
(u, v)	Coefficient of $Q(u, v)$ in (4.14)	$(u - t_j)(v - t_j)$		$(u - t_j) + (v - t_j)$	$(u - t_j)^2 + (v - t_j)^2$	$(u - t_j)(v - t_j)$
(t, t)	$+1$	δ_1 δ_1	1	$2\delta_1$	$2\delta_1^2$	δ_1^2
(t_j, t)	$-2\frac{\delta_2}{\delta}$	0 δ_1	1	δ_1	δ_1^2	0
(t_{j+1}, t)	$-2\frac{\delta_1}{\delta}$	δ δ_1	1	$\delta + \delta_1$	$\delta^2 + \delta_1^2$	$\delta\delta_1$
(t_j, t_j)	$+\frac{\delta_2^2}{\delta^2}$	0 0	1	0	0	0
(t_j, t_{j+1})	$+2\frac{\delta_1\delta_2}{\delta^2}$	0 δ	1	δ	δ^2	0
(t_{j+1}, t_{j+1})	$+\frac{\delta_1^2}{\delta^2}$	δ δ	1	2δ	$2\delta^2$	δ^2

$$\delta_1 = (t - t_j), \quad \delta_2 = (t_{j+1} - t), \quad \delta = \delta_1 + \delta_2 = (t_{j+1} - t_j)$$

To obtain the coefficient of $Q(t_j, t_j)$ in (4.14) with (4.18) substituted in, multiply entries in column 2 with the corresponding entries under the column headed by $Q(t_j, t_j)$, and add. To obtain the coefficient of α_1, multiply entries in column 2 with the corresponding entries under the column headed by α_1, and add, and similarly for α_2 and β. The results are all 0.

Acknowledgment. The author wishes to thank one of the referees for a simplification of the proof of Theorem 2.

REFERENCES

[1] M. Atteia, *Fonctions splines avec contraintes linéaires de type inégalité*, Congres de l'A.F.I.R.O., Nancy, 1967.

[2] JAMES W. DANIEL, *Convergence of a discretization for constrained spline function problems*, this Journal, 9 (1971), pp. 83–96.

[3] J. W. JEROME AND L. L. SCHUMAKER, *On Lg-splines*, J. Approximation Theory, 2 (1969), pp. 29–49.

[4] GEORGE KIMELDORF AND GRACE WAHBA, *Some results on Tchebycheffian spline functions*, J. Math. Anal. Appl., 33 (1971), pp. 82–95.

[5] P. J. LAURENT, *Construction of spline functions in a convex set*, Approximations with Special Emphasis on Spline Functions, I. J. Schoenberg, ed., Academic Press, London, 1969, pp. 415–446.

[6] O. L. MANGASARIAN AND L. L. SCHUMAKER, *Splines via optimal control*, Approximations with Special Emphasis on Spline Functions, I. J. Schoenberg, ed., Academic Press, London, 1969, pp. 119–156.

[7] ———, *Discrete splines via mathematical programming*, this Journal, 9 (1971), pp. 174–183.

[8] KLAUS RITTER, *Generalized spline interpolation and nonlinear programming*, Approximations with Special Emphasis on Spline Functions, I. J. Schoenberg, ed., Academic Press, London, 1969, pp. 75–118.

[9] HAROLD S. SHAPIRO, *Topics in Approximation Theory*, Springer-Verlag, Berlin, 1971.

[10] GRACE WAHBA, *On the regression design problem of Sacks and Ylvisaker*, Ann. Math. Statist., 42 (1971), pp. 1035–1053.

[11] ———, *Convergence rates for certain approximate solutions to Fredholm integral equations of the first kind*, J. Approximation Theory, to appear.

[12] ———, *A class of approximate solutions to linear operator equations*, Ibid., to appear.

SIAM J. CONTROL
Vol. 11, No. 1, February 1973

AN ADAPTIVE PRECISION GRADIENT METHOD FOR OPTIMAL CONTROL*

R. KLESSIG† AND E. POLAK‡

Abstract. This paper presents a gradient algorithm for unconstrained optimal control problems. The algorithm is stated in terms of numerical integration formulas, the precision of which is controlled adaptively by a test that ensures convergence. Empirical results show that this algorithm is considerably faster than its fixed precision counterpart.

1. Introduction. In certain unconstrained minimization problems, precise evaluation of the cost function and its derivative is very expensive. One well-known problem of this type is the continuous optimal control problem. In this problem, evaluation of the cost and its derivative requires the numerical solution of a system of differential equations. Thus, if a gradient-type algorithm is applied to this problem, most of the computational time will be taken up by numerical integration. If the numerical integration is of high accuracy, the algorithm will require a large amount of time to solve a problem. On the other hand, if the numerical integration is not accurate enough, the convergence of the algorithm to a solution may be lost, thus nullifying any time saving.

It is to this dilemma that we address ourselves. Specifically, we propose a method for adaptively varying the accuracy of the numerical integration, as the computation progresses in a gradient algorithm for solving unconstrained optimal control problems. We use low accuracy numerical integration while far from a solution and improve the accuracy as a solution is approached. This is achieved by means of a test in the algorithm which decides on the required integration precision. The resulting adaptive precision gradient method, for solving unconstrained continuous optimal control problems, is convergent and considerably more efficient than its fixed precision counterpart. Our adaptive precision method is based on the algorithm models described in [4] and [5]. Because of this, in § 2 we state the model on which our scheme is based. In § 3, we define the control problem, state our new algorithm, relate it to the model in § 2, and then apply the theory of § 2 to establish convergence of our new method. In § 4 we present some computational results that support our claim of efficiency for our method.

2. An abstract prototype. Let $\mathscr{X}$ be a linear space with a topology defined by the seminorm $\|\cdot\|_{\mathscr{X}}$ containing a set of *desirable* points denoted by Δ. The algorithm prototype we are about to present is designed to compute points in Δ. It is similar to (A.1.1) in [5].

Let $\{D_j\}_{j=0}^{\infty}$ be a sequence of subsets of $\mathscr{X}$ such that

$$D_j \subset D_{j+1}, \qquad j = 0, 1, \cdots . \tag{2.1}$$

* Received by the editors April 6, 1971, and in final revised form March 9, 1972. This research was supported in part by the National Aeronautics and Space Administration under Grant NGL-05-003-016, by the Joint Services Electronics Program under Grant AFOSR-68-1488 and by the National Science Foundation under Grant GK-10656X.

† Bell Telephone Laboratories, Holmdel, New Jersey 07733.

‡ Department of Electrical Engineering and Computer Sciences, and the Electronics Research Laboratory, University of California, Berkeley, California 94720.

Let

$$D = \bigcup_{j=0}^{\infty} D_j. \tag{2.2}$$

Next we introduce a sequence of search functions $\{A_j\}_{j=0}^{\infty}$, $A_j : D_j \to 2^{D_j}$, and a sequence of test functions $\{c_j\}_{j=0}^{\infty}$, $c_j : D_j \to \mathbb{R}^1$. We make the following assumptions.

2.3. *Assumptions.*

(i) $\bar{D} = \mathscr{X}$.

(ii) There exists a set $M \subset \mathscr{X}$, satisfying $M \cap \Delta \neq \emptyset$, such that for every $z \in M$, $z \notin \Delta$, there exist an $\varepsilon(z) > 0$, a $\delta(z) > 0$ and an integer $N(z) \geqq 0$ satisfying

$$c_j(z'') - c_j(z') \leqq -\delta(z) \tag{2.4}$$

for all $z' \in B(z, \varepsilon(z)) \cap M \cap D_j$, for all $z'' \in A_j(z')$ and for all $j \geqq N(z)$, where

$$B(z, \varepsilon(z)) \triangleq \{z' \in \mathscr{X} \mid \|z' - z\|_{\mathscr{X}} \leqq \varepsilon(z)\}. \tag{2.5}$$

(iii) There exists a continuous function $c : M \to \mathbb{R}^1$ and a sequence $\{\beta_s\}_{s=0}^{\infty} \subset \mathbb{R}^+$, possibly depending on M, such that

$$\sum_{s=0}^{\infty} \beta_s < \infty \tag{2.6}$$

and

$$|c_j(z) - c(z)| \leqq \beta_s \quad \text{for all } z \in D_j \cap M, \quad \text{for all } j \geqq s. \tag{2.7}$$

2.8. *Remark.* Since c_j is an approximation to c, Assumption 2.3 (iii) can be interpreted as a requirement on how fast $c_j(z)$ converges to $c(z)$. Note that if $c_j(z)$ converges to $c(z)$ linearly, Assumption 2.3 (iii) should not be difficult to satisfy.

2.9. Algorithm Prototype.

Step 0. Select an integer $j_0 \geqq 0$. Select $z_0 \in D_{j_0}$, $\varepsilon_0 > 0$ and $\alpha \in (0, 1)$. Set $i = 0, j = j_0$, and $\varepsilon = \varepsilon_0$.

Step 1. Compute a $y \in A_j(z_i)$.

Step 2. If $c_j(y) - c_j(z_i) \leqq -\varepsilon$, go to Step 3; else, set $j = j + 1$, set $\varepsilon = \alpha\varepsilon$, and go to Step 1.

Step 3. Set $z_{i+1} = y$.

Step 4. Set $i = i + 1$ and go to Step 1.

2.10. *Remark.* Because of (2.1), $A_j(z_i)$ and $c_j(z_i)$ are always well-defined in (2.9).

We now state the convergence properties of (2.9) in the following theorem. We omit a proof since the theorem is only a minor modification of Theorem (A.1.26) in [5].

2.11. Theorem. *Let* $\{z_i\}$ *be a sequence constructed by* (2.9). *Suppose that Assumptions* 2.3 *are satisfied and suppose that* $\{z_i\} \subset M$. *If* $\{z_i\}$ *is finite* (*because* (2.9) *is jammed up between Step* 1 *and Step* 2) *with last point* z_s, *then* $z_s \in \Delta$. *If* $\{z_i\}$ *is infinite, then every accumulation point of* $\{z_i\}$ *is in* Δ.

3. Adaptive optimal control algorithm. We now turn to the optimal control problem and relate it to the quantities in the previous section. This requires some notation.

3.1. *Notation.* (i) We denote the space of square integrable functions from $[0, T]$ into $\mathbb{R}^m$ by $L_2^m[0, T]$.[1] (ii) With $T > 0$, and $\|\cdot\|$ and $\langle \cdot, \cdot \rangle$ denoting the Euclidean norm and inner product respectively, we define, for u, $v \in L_2^m[0, T]$,

$$\langle u, v\rangle_2 \triangleq \int_0^T \langle u(t), v(t)\rangle \, dt, \tag{3.2}$$

$$\|u\|_2 \triangleq [\langle u, u\rangle_2]^{1/2}, \tag{3.3}$$

$$\|u\|_\infty \triangleq \operatorname*{ess\,sup}_{t \in [0, T]} \|u(t)\|. \tag{3.4}$$

3.5. *Optimal control problem.*
Minimize $g(x(T, u))$ subject to

$u \in L_2^m[0, T]$ and $x(\cdot, u)$ satisfying the differential equation

$$\dot{x}(t, u) = f(x(t, u), u(t), t) \quad \text{a.e. on } [0, T], \quad x(0, u) = x_0, \tag{3.6}$$

where $g: \mathbb{R}^n \to \mathbb{R}^1, f: \mathbb{R}^n \times \mathbb{R}^m \times \mathbb{R}^1 \to \mathbb{R}^n$, $x_0 \in \mathbb{R}^n$ is given, and $T > 0$ is given.

3.7. *Assumptions.* Throughout this section we assume that

(i) $g(\cdot)$ is continuously differentiable on $\mathbb{R}^n$;

(ii) $f(\cdot, \cdot, \cdot)$ is continuously differentiable on $\mathbb{R}^n \times \mathbb{R}^m \times \mathbb{R}^1$;

(iii) there exists an $\omega > 0$ such that (3.6) has a unique, absolutely continuous solution, $x(\cdot, u)$, with $\|x(T, u)\| \leqq K_1 < \infty$, for all $u \in M' \triangleq \{u' \in L_2^m[0, T] | \, \|u'\|_\infty < 2\omega\}$.

In applying the results of § 2, we shall take $\mathscr{X} = L_2^m[0, T]$, the seminorm $\|\cdot\|_{\mathscr{X}} = \|\cdot\|_2$, and $M = \{u' \in L_m^2[0, T] | \, \|u'\|_\infty < \omega\}$. Thus the convergence properties of our algorithm will be established with respect to this seminorm topology. We believe that Theorem 3.48 can be extended to hold also with respect to the weak topology, with $L_2^m[0, T]$ viewed as a space of equivalence classes. This appears to be possible because the results of § 2 can be stated in terms of a general topological space, and thus we could use the weak topology on $L_2^m[0, T]$. However, verifying Assumptions 2.3 for the weak topology on $L_2^m[0, T]$ seems to require stronger assumptions on the optimal control problem. For example, it seems necessary that (3.6) be such that $u_i \to u^*$ weakly implies $x(\cdot, u_i) \to x(\cdot, u^*)$ strongly (in the space of continuous functions). In any event, we shall concern ourselves solely with the seminorm topology on $L_2^m[0, T]$ in this paper and leave the weak topology case as a question for further research.

We should also explain the use of the sets M and M'. Since the algorithm to be presented generates a sequence of step functions, it is natural to work in a space in which the step functions are dense. The space $L_2^m[0, T]$ with the seminorm topology appears to be the natural choice. Unfortunately, the performance index cannot be shown to be differentiable in this topological space. However, in the Appendix, we show that the performance index is differentiable in an appropriate sense on M', and of course, the step functions are dense in M'. Finally, by requiring the sequence generated by our algorithm to lie in M, we insure that the performance index is "differentiable" at accumulation points of the sequence which allows us to apply the theory of § 2.

[1] Note that this is not the space of equivalence classes.

Next we define the subsets D_j.

3.8. DEFINITION. For any integer $j \geqq 0$, let

$$\eta_j = T/2^j \tag{3.9}$$

and let

$$t_{js} = s\eta_j, \qquad s = 0, 1, \cdots, 2^j. \tag{3.10}$$

Then we define $D_j \subset L_2^m[0, T]$ as the set of all functions $u:[0, T] \to \mathbb{R}^m$ such that the components $u^i(\cdot)$ of $u(\cdot)$ are real-valued step functions having mesh points at t_{js}, which are continuous from the right, and which satisfy $u^i(T) = u^i(T - \eta_j)$.

We see that as j increases, these step functions have smaller step sizes. It is these step sizes that we shall use to control the accuracy of the numerical integration.

3.11. DEFINITION. For any $u \in M'$, we define

$$c(u) = g(x(T, u)) \tag{3.12}$$

and

$$h(u)(t) = \frac{\partial}{\partial u} f(x(t, u), u(t), t)^T p(t, u), \quad t \in [0, T], \tag{3.13}$$

where $p(\cdot, u)$ is the solution of the differential equation

$$\dot{p}(t, u) = -\frac{\partial}{\partial x} f(x(t, u), u(t), t)^T p(t, u), \quad t \in [0, T], \tag{3.14}$$

$$p(T, u) = \frac{\partial}{\partial x} g(x(T, u))^T. \tag{3.15}$$

The following theorem is proven in the Appendix.

3.16. THEOREM. *For any $u \in M'$ and $\delta u \in \{u' \in L_2^m[0, T] |\ \|u'\|_\infty < \infty\}$,*

$$\lim_{\lambda \downarrow 0} \frac{c(u + \lambda \delta u) - c(u)}{\lambda} \triangleq c'(u, \delta u) = \langle h(u), \delta u \rangle_2 \tag{3.17}$$

and $c' : M' \times \{u' \in L_2^m[0, T] |\ \|u'\|_\infty < \infty\} \to \mathbb{R}^1$ is continuous.

Finally, to introduce c_j and A_j, we make use of numerical integration formulas which lead to approximations to c and h.

3.18. DEFINITION. Let $F_1 : \mathbb{R}^n \times \mathbb{R}^m \times \mathbb{R}^1 \times \mathbb{R}^1 \to \mathbb{R}^n$ and $F_2 : \mathbb{R}^n \times \mathbb{R}^n \times \mathbb{R}^m \times \mathbb{R}^1 \times \mathbb{R}^1 \to \mathbb{R}^{2n}$ be one-step integration functions (see (8.3) in [2]) for the differential equations

$$\dot{x}(t, u) = f(x(t, u), u(t), t), \quad t \in [0, T], \tag{3.19}$$

and

$$\begin{bmatrix} \dot{x}(t, u) \\ \hdashline \dot{p}(t, u) \end{bmatrix} = \begin{bmatrix} f(x(t, u), u(t), t) \\ \hdashline -\dfrac{\partial}{\partial x} f(x(t, u), u(t), t)^T p(t, u) \end{bmatrix}, \quad t \in [0, T]. \tag{3.20}$$

Then for any $u \in D_j$, we define the sequences $\{x_j(s,u)\}_{s=0}^{2^j}$, $\{x'_j(s,u)\}_{s=0}^{2^j}$ and $\{p_j(s,u)\}_{j=0}^{2^j}$ as follows:

$$x_j(s+1,u) = x_j(s,u) + \eta_j F_1(x_j(s,u), u(t_{js}), t_{js}, \eta_j), \quad s = 0, 1, \cdots, 2^j - 1, \tag{3.21}$$

$$x_j(0,u) = x_0, \tag{3.22}$$

$$\begin{bmatrix} x'_j(s,u) \\ \hline p_j(s,u) \end{bmatrix} = \begin{bmatrix} x'_j(s+1,u) \\ \hline p_j(s+1,u) \end{bmatrix} + \eta_j F_2(x'_j(s+1,u), p_j(s+1,u), u(t_{js}), t_{js}, \eta_j), \quad s = 0, 1, \cdots, 2^j - 1, \tag{3.23}$$

$$\begin{bmatrix} x'_j(2^j,u) \\ \hline p_j(2^j,u) \end{bmatrix} = \begin{bmatrix} x_j(2^j,u) \\ \hline \dfrac{\partial}{\partial x} g(x_j(2^j,u))^T \end{bmatrix}. \tag{3.24}$$

Finally, we use these sequences to construct step functions $\bar{x}_j:[0,T] \times D_j \to \mathbb{R}^n$, $\bar{x}'_j:[0,T] \times D_j \to \mathbb{R}^n$ and $\bar{p}_j:[0,T] \times D_j \to \mathbb{R}^n$ as follows:

$$\bar{x}_j(t,u) = \begin{cases} x_j(s,u) & \text{for } t \in [s\eta_j, (s+1)\eta_j), \quad s = 0, 1, \cdots, 2^j - 1, \\ x_j(2^j - 1, u) & \text{for } t = T; \end{cases} \tag{3.25}$$

$$\bar{x}'_j(t,u) = \begin{cases} x'_j(s+1,u) & \text{for } t \in [s\eta_j, (s+1)\eta_j), \quad s = 0, 1, \cdots, 2^j - 1, \\ x'_j(2^j, u) & \text{for } t = T; \end{cases} \tag{3.26}$$

$$\bar{p}_j(t,u) = \begin{cases} p_j(s+1,u) & \text{for } t \in [s\eta_j, (s+1)\eta_j), \quad s = 0, 1, \cdots, 2^j - 1, \\ p_j(2^j, u) & \text{for } t = T. \end{cases} \tag{3.27}$$

3.28. Definition. For $j = 0, 1, \cdots$, we define approximations to c and h, $c_j : D_j \to \mathbb{R}^1$ and $h_j : D_j \to D_j$, as follows:

$$c_j(u) \triangleq g(x_j(2^j, u)), \tag{3.29}$$

$$h_j(u)(t) \triangleq \frac{\partial}{\partial u} f(x'_j(s+1,u), u(s\eta_j), s\eta_j)^T p_j(s+1,u) \tag{3.30}$$

$$\text{for } t \in [s\eta_j, (s+1)\eta_j), \quad s = 0, 1, \cdots, 2^j - 1,$$

$$h_j(u)(T) \triangleq h_j(u)(T - \eta_j). \tag{3.31}$$

We now state our adaptive precision algorithm for solving (3.5). This algorithm defines the search function A_j.

3.32. Algorithm.

Step 0. Select an integer j_0. Select $u_0 \in D_{j_0}$, $\varepsilon_0 > 0$, $\alpha \in (0,1)$, $\beta \in (0,1)$ and $\lambda_{\min} \in (0,1]$. Set $i = 0$, $j = j_0$ and $\varepsilon = \varepsilon_0$.

Step 1. Compute $h_j(u_i)(\cdot)$ according to (3.18) and (3.28).

Step 2. Set $\lambda = 1$.

Step 3. Compute $\theta_j(u_i, \lambda) = c_j(u_i - \lambda h_j(u_i)) - c_j(u_i) + (\lambda/2)\|h_j(u_i)\|_2^2$ according to (3.18) and (3.28).[2]

Step 4. If $\theta_j(u_i, \lambda) \leqq 0$, set $y = u_i - \lambda h_j(u_i)$ and go to Step 6; else, set $\lambda = \beta\lambda$ and go to Step 5.

[2] Since $h_j(u_i)$ is a step function, precise evaluation of $\|h_j(u_i)\|_2$ offers no difficulty.

Step 5. If $\lambda \geqq \varepsilon\lambda_{\min}$, go to Step 3; else, set $y = u_i$ and go to Step 6.

Step 6. Compute $c_j(y)$ according to (3.18) and (3.28).

Step 7. If $c_j(y) - c_j(u_i) \leqq -\varepsilon$, go to Step 8; else, set $j = j + 1$, set $\varepsilon = \alpha\varepsilon$, and go to Step 1.

Step 8. Set $u_{i+1} = y$.

Step 9. Set $i = i + 1$ and go to Step 1.

3.33. *Remark.* We have not stated algorithm (3.32) in the form in which it should be programmed. For example, in the second part of Step 5, it would be more efficient to go directly to Step 8. The reason for this form is to make (3.32) similar to (2.9) so that the search function A_j, defined by (3.32), is more apparent. Thus, our way of presenting (3.32) facilitates analysis rather than computational efficiency. There should be no difficulty in identifying and removing these awkward points in programming the algorithm.

3.34. DEFINITION. We define the *search function* $A_j : D_j \to 2^{D_j}$ *for Algorithm* 3.32 as the function defined by Steps 1 through 6.

It should now be evident that (3.32) is of the same form as (2.9). Consequently, to obtain convergence results, we only need to verify Assumptions 2.3 since Theorem 2.11 will then apply. First, since $h(\hat{u})(\cdot) = 0$ is a necessary condition of optimality, we make the following definition.

3.35. DEFINITION. We define $\Delta \subset L_2^m[0, T]$ as

$$\Delta \triangleq \{u' \in L_2^m[0, T] | h(u')(\cdot) = 0\}. \tag{3.36}$$

Before proceeding to establish Assumptions 2.3, we must introduce additional assumptions.

3.37. *Assumption.* The $\omega > 0$ in Assumption 3.7 (iii) is such that

$$\{u' \in L_2^m[0, T] | \, \|u'\|_\infty < \omega\} \cap \Delta \neq \emptyset.$$

3.38. *Assumption.* The one-step integration formulas used in (3.18) are such that given any $\varepsilon > 0$, there exist an integer $J \geqq 0$ and a $K \in (0, \infty)$ satisfying

$$\|x(\cdot, u) - \bar{x}_j(\cdot, u)\|_\infty \leqq K/2^j, \tag{3.39}$$

$$\left\| \begin{bmatrix} x(\cdot, u) \\ p(\cdot, u) \end{bmatrix} - \begin{bmatrix} \bar{x}_j(\cdot, u) \\ \bar{p}_j(\cdot, u) \end{bmatrix} \right\|_\infty \leqq \varepsilon \tag{3.40}$$

for all $u \in M' \cap D_j$ and for all $j \geqq J$.

Assumption 3.38 does not appear to be unduly restrictive. For example, it is satisfied by the Euler–Cauchy method under a Lipschitz assumption on f and $(\partial/\partial x)f$. The proof of this fact closely follows that in (8.1) of [2] (the fact that $u(\cdot)$ is constant over $(s\eta_j, (s + 1)\eta_j)$ must be used). The proof of this fact can also be found in [3].

Before applying Theorem 2.11, we present some preliminary results. Recall that we are using Assumptions 3.7 and 3.38.

3.41. LEMMA. *Given any* $u \in M'$ *and any* $\gamma > 0$, *there exists an* $\varepsilon(u, \gamma) > 0$ *and an integer* $N(u, \gamma) > 0$ *such that*

$$|c_j(u') - c(u)| \leqq \gamma, \tag{3.42}$$

$$\|h_j(u') - h(u)\|_2 \leqq \gamma \tag{3.43}$$

for all $u' \in B(u, \varepsilon(u, \gamma)) \cap D_j \cap M'$, *for all* $j \geqq N(u, \gamma)$.

Proof. First we note that (A.1) and the continuity of g imply that c is continuous on M'. Thus (3.42) follows directly from (3.39).

Similarly, (A.9) and the continuous differentiability of f and g imply that $h : M' \to L_2^m[0, T]$ is continuous. Thus (3.43) follows directly from (3.40).

3.44. Lemma. *Suppose that $u \in M$ is such that $h(u) \neq 0$ and let $\beta \in (0, 1)$. Then there exists an $\varepsilon(u) > 0$, an integer $k(u) \geqq 0$, and an integer $N(u) \geqq 0$ such that*

$$(3.45) \qquad c_j(u' - \beta^{k(u)}h_j(u')) - c_j(u') + \frac{\beta^{k(u)}}{2}\|h_j(u')\|_2^2 \leqq 0$$

for all $u' \in B(u, \varepsilon(u)) \cap M \cap D_j$ for all $j \geqq N(u)$.

Proof. Since $u \in M$, it follows that $\|h(u)\|_\infty < \infty$. Therefore there exists an $\varepsilon'(u) > 0$ and an integer $k(u) \geqq 0$ such that the line segment $[u', u' - \beta^{k(u)}h(u')]$ is in M' for all $u' \in B(u, \varepsilon'(u)) \cap M$ and such that

$$(3.46) \qquad c(u - \beta^{k(u)}h(u)) - c(u) + \frac{\beta^{k(u)}}{2}\|h(u)\|_2^2 \leqq -\frac{\beta^{k(u)}}{4}\|h(u)\|_2^2$$

(recall that $c'(u, h(u)) = \|h(u)\|_2^2$ from (3.16)). But since $\|\cdot\|_2$ is continuous, we can choose the appropriate $\gamma > 0$ in (3.41) to find $\varepsilon(u) \in (0, \varepsilon'(u)]$ and $N(u) \geqq 0$ such that

$$(3.47) \qquad \left| \begin{array}{l} c(u - \beta^{k(u)}h(u)) - c(u) + \dfrac{\beta^{k(u)}}{2}\|h(u)\|_2^2 \\ -c_j(u' - \beta^{k(u)}h_j(u')) + c_j(u') - \dfrac{\beta^{k(u)}}{2}\|h_j(u')\|_2^2 \end{array} \right| \leqq \frac{\beta^{k(u)}}{4}\|h(u)\|_2^2$$

for all $u' \in B(u, \varepsilon(u)) \cap M$, for all $j \geqq N(u)$. Thus, (3.46) and (3.47) imply the desired result.

3.48. Theorem. *Suppose that Assumptions* 3.7, 3.37, *and* 3.38 *are satisfied. Consider any sequence $\{u_i\}$ constructed by* (3.32) *such that $\{u_i\} \subset M$.*[3] *If $\{u_i\}$ is finite because* (3.32) *jammed up at u_s, then $h(u_s) = 0$. If $\{u_i\}$ is infinite and u^* is an accumulation point of $\{u_i\}$, then $h(u^*) = 0$.*

Proof. All we need do is verify that various sets and functions introduced in this section satisfy Assumption 2.3.

(2.3) (i). Since the step functions are dense under the topology we are using, it is straightforward to show that $\bar{D} = L_2^m[0, T]$.

(2.3) (ii). First, (3.37) implies that $M \cap \Delta \neq \emptyset$. Now suppose $u \in M$ is such that $u \notin \Delta$ $(h(u) \neq 0)$. Then because of our definition of A_j, (3.34), and Lemma 3.44, there exists an $\varepsilon'(u) > 0$, an integer $k(u) \geqq 0$ and an integer $N'(u) \geqq 0$ such that $u'' \in A_j(u')$ is of the form $u'' = u' - \beta^{l(u')}h_j(u')$, with $l(u') \leqq k(u)$, for all $u' \in B(u, \varepsilon'(u)) \cap M \cap D_j$, for all $j \geqq N'(u)$. (In other words, the test in Step 4 of Algorithm 3.32 is satisfied.) Now, by the appropriate choice of $\gamma > 0$, we can apply Lemma 3.41 to find an $\varepsilon(u) \in (0, \varepsilon'(u)]$ and an integer $N(u) \geqq N'(u)$ such that $\|h_j(u')\|_2^2 \geqq \| \, |h(u)\|_2^2/2 > 0$ for all $u' \in B(u, \varepsilon(u)) \cap M \cap D_j$, for all $j \geqq N(u)$. Thus using the

[3] Whenever (3.6) has a solution for all $u \in L_2^m[0, T] \cap L_\infty^m[0, T]$, the condition $\{u_i\} \subset M$ can be replaced by $\sup_i\|u_i\|_\infty < \infty$ and $\sup_i\|x(\cdot, u_i)\|_\infty < \infty$.

inequality satisfied in Step 4 of Algorithm 3.32, we obtain

$$\begin{aligned} c_j(u'') - c_j(u') &= c_j(u' - \beta^{l(u')}h_j(u')) - c_j(u') \\ &\leqq -\frac{\beta^{l(u')}}{2}\|h_j(u')\|_2^2 \qquad (3.49) \\ &\leqq -\frac{\beta^{k(u)}}{4}\|h(u)\|_2^2 \triangleq -\delta(u) < 0 \end{aligned}$$

for all $u' \in B(u, \varepsilon(u)) \cap M \cap D_j$, for all $u'' \in A_j(u')$ for all $j \geqq N(u)$.

(2.3) (iii). Because of Assumption 3.7 (iii), there exists $K_1 \geqq 0$ such that $\|x(T, u')\| \leqq K_1$ for all $u' \in M$. Thus, because of (3.38), there exists $K_2 \geqq K_1$ such that $\|\bar{x}_j(T, u')\| \leqq K_2$ for all $u' \in M \cap D_j$, for $j = 0, 1, \cdots$. Define

$$G = \sup\left\{\left\|\frac{\partial}{\partial x}g(\xi)\right\| \,\middle|\, \|\xi\| \leqq K_2\right\}.$$

Then from the definition of c_j and (3.38) we have

$$\begin{aligned} |c_j(u) - c(u)| &= |g(\bar{x}_j(T, u)) - g(x(T, u))| \\ &\leqq G\|\bar{x}_j(T, u) - x(T, u)\| \leqq GK/2^j \end{aligned} \qquad (3.50)$$

for all $u \in M \cap D_j$, for all $j = 0, 1, \cdots$. It is now obvious that Assumption 2.3 (iii) is satisfied.

4. Computational results. In this section we present numerical results for two simple problems. On each problem we have used both the adaptive integration approach and a nonadaptive integration approach. We have used two types of numerical integration for each problem and each approach.

Because of memory limitations and roundoff effects, we modified (3.32) by placing an upper bound on j, called $j_{\max}$. When j reached $j_{\max}$, ε was set to zero thus insuring that j would not increase beyond $j_{\max}$. The nonadaptive algorithm was derived from the adaptive one by setting $\varepsilon_0 = 0$ and $j_0 = j_{\max}$. Thus the maximum accuracy was used for each iteration. In both algorithms, execution was stopped when the reduction in cost for an iteration was below .0001 % and $j = j_{\max}$. The two types of numerical integration used were the Euler–Cauchy method and the fourth order Runge–Kutta method.

4.1. *Example.* The first example is a simple linear plant, quadratic cost problem:

$$c(u) = x^0(1), \qquad (4.2)$$

$$\dot{x}^0(t) = [x^1(t)]^2 + [x^2(t)]^2 + 0.5[u(t)]^2, \qquad (4.3)$$

$$\dot{x}^1(t) = -0.1x^1(t) + x^2(t), \qquad (4.4)$$

$$\dot{x}^2(t) = -x^2(t) + u(t), \qquad (4.5)$$

$$x^0(0) = 0, \quad x^1(0) = 0, \quad x^2(0) = -1, \quad T = 1. \qquad (4.6)$$

The results are summarized in Table 4.7. The parameters used were $\varepsilon_0 = 0.01$ ($=0$ for the nonadaptive algorithm), $\alpha = 0.35$, $\beta = 0.4$, and $\lambda_{\min} = 10.0$. The time

TABLE 4.7

	Nonadaptive	Adaptive
Euler–Cauchy	$j_0 = j_{\max} = 10$ opt. cost = 0.20408 time = 65.620 sec.	$j_0 = 2, j_{\max} = 10$ opt. cost = 0.20408 time = 36.062 sec.
Fourth Order Runge–Kutta	$j_0 = j_{\max} = 8$ opt. cost = 0.20409 time = 43.870 sec.	$j_0 = 2, j_{\max} = 8$ opt. cost = 0.20409 time = 36.078 sec.

referred to is the execution time in seconds required to solve the problem on a CDC 6400.

4.8. *Example.* This example is a simple optimal control problem with a bang-bang solution converted to an unconstrained form by penalty functions. Such problems are generally considered to cause difficulty when solved by gradient methods.

$$c(u) = x^0(1), \tag{4.9}$$

$$\dot{x}^0(t) = [x^1(t)]^2 + [x^2(t)]^2 + 10[\max(0, u(t) - 1)]^2 + 10[\max(0, -u(t) - 1)]^2, \tag{4.10}$$

$$\dot{x}^1(t) = x^2(t), \tag{4.11}$$

$$\dot{x}^2(t) = -(0.01 + 4\pi^2)x^1(t) - 0.2x^2(t) + u(t), \tag{4.12}$$

$$x^0(0) = 0, \quad x^1(0) = 1, \quad x^2(0) = 0, \quad T = 1. \tag{4.13}$$

The parameters used were $\varepsilon_0 = 0.1$ for Euler–Cauchy, $\varepsilon_0 = 0.05$ for Runge–Kutta, $\varepsilon_0 = 0$ for the nonadaptive algorithm, $\alpha = 0.35$, $\beta = 0.4$, and $\lambda_{\min} = 0.1$. The results are given in Table 4.14.

TABLE 4.14

	Nonadaptive	Adaptive
Euler–Cauchy	$j_0 = j_{\max} = 10$ opt. cost = 16.436 time = 152.414 sec.	$j_0 = 2, j_{\max} = 10$ opt. cost = 16.438 time = 30.138 sec.
Fourth Order Runge–Kutta	$j_0 = j_{\max} = 8$ opt. cost = 16.435 time = 131.684 sec.	$j_0 = 2, j_{\max} = 8$ opt. cost = 16.435 time = 43.184 sec.

These two examples indicate that the use of adaptive accuracy numerical integration will result in significant improvement in computer time.

5. Conclusion. The method presented here is of both theoretical and practical interest. On the theoretical side, we have shown that it is possible to control to good advantage the effects of imprecise function evaluations when dealing with algorithms for solving minimization problems. In particular, we have considered the effects of numerical integration when solving optimal control problems. On the practical side, our method appears to provide significant savings in computer time as compared to its fixed precision counterpart.

Finally we note that Algorithm 3.32 is of the humble, first order variety. However, it is not difficult to see that the adaptive integration technique can also be used in conjunction with Newton–Raphson or conjugate gradient-type methods.

Appendix. Differentiability of c. In this Appendix, we closely follow many of the proofs in [1].

A.1. LEMMA. *Under the assumptions and definitions of* § 3, *given any* $u^* \in M'$, *there exist* $k > 0$ *and* $\gamma > 0$ *such that*

$$\|x(\cdot, u) - x(\cdot, u^*)\|_\infty \leqq k\|u - u^*\|_2 \quad \textit{for all } u \in B(u^*, \gamma) \cap M'. \tag{A.2}$$

Proof. Let $\omega' \in (0, \infty)$ be such that

$$\sup_{t \in [0,T]} \|x(t, u^*)\| < \omega'. \tag{A.3}$$

Because of Assumption 3.7 (ii) and the generalized mean value theorem, there exist $K_1, K_2 \in (0, \infty)$ such that

$$\|f(x_1, v_1, t) - f(x_2, v_2, t)\| \leqq K_1\|x_1 - x_2\| + K_2\|v_1 - v_2\| \tag{A.4}$$

for all $x_1, x_2 \in \{x' \in \mathbb{R}^n | \, \|x'\| \leqq 2\omega'\}$, for all $v_1, v_2 \in \{v' \in \mathbb{R}^n| \, \|v'\| \leqq 2\omega\}$ and for all $t \in [0, T]$ (ω was defined in Assumption 3.7 (iii)). Now set

$$\gamma = \frac{\omega'}{2}[K_2\sqrt{T}\, e^{K_1 T}]^{-1} \tag{A.5}$$

and choose any $u \in B(u^*, \gamma) \cap M'$. Since $x(\cdot, u)$ is continuous and $\|x(0, u)\| = \|x_0\| < \omega'$, there exists $\bar{t} \in (0, T]$ such that $\|x(t, u)\| \leqq 2\omega'$ for all $t \in [0, \bar{t}]$. Now, because of (A.4), we find that for any $t \in [0, \bar{t}]$,

$$\|x(t, u) - x(t, u^*)\| \leqq \int_0^t K_1\|x(t', u) - x(t', u^*)\|\, dt' + \int_0^t K_2\|u(t') - u^*(t')\|\, dt'. \tag{A.6}$$

Since by the Hölder inequality, for $t \in [0, T]$,

$$\int_0^t K_2\|u(t') - u^*(t')\|\, dt' \leqq K_2\sqrt{T}\|u - u^*\|_2,$$

the Bellman–Gronwall inequality together with (A.6) yields

$$\|x(t, u) - x(t, u^*)\| \leqq K_2\sqrt{T}\|u - u^*\|_2\, e^{K_1 T} \quad \text{for all } t \in [0, \bar{t}]. \tag{A.7}$$

We now show by contradiction that $\bar{t} = T$. If $\bar{t} < T$, we can assume without loss of generality that $\|x(\bar{t}, u)\| = 2\omega'$. Then from (A.5) and (A.7), we find that

$$\|x(\bar{t}, u^*)\| \geqq \|x(\bar{t}, u)\| - \frac{\omega'}{2} = \frac{3\omega'}{2}, \tag{A.8}$$

which contradicts (A.3), and hence we conclude that $\bar{t} = T$. Consequently, (A.7) and the fact that u was arbitrary in $B(u^*, \gamma) \cap M'$ imply the desired result.

A.9. LEMMA. *Under the assumptions and definitions of* § 3, *given any* $u^* \in M'$ *and any* $\tau > 0$, *there exists an* $\varepsilon > 0$ *such that*

$$\|p(\cdot, u) - p(\cdot, u^*)\|_\infty \leqq \tau \quad \textit{for all } u \in B(u^*, \varepsilon) \cap M'. \tag{A.10}$$

Proof. Let $K_3 = \|p(\cdot, u^*)\|_\infty$. Then by Assumption 3.7 (ii) and (A.1), there exist $K_4 \in (0, \infty)$ and $\varepsilon' > 0$ such that

$$\left\|\frac{\partial}{\partial x} f(x(\cdot, u), u(\cdot), \cdot)^T\right\|_\infty \leqq K_4, \tag{A.11}$$

$$\int_0^T \left\|\frac{\partial}{\partial x} f(x(t', u), u(t'), t')^T - \frac{\partial}{\partial x} f(x(t', u^*), u^*(t'), t')^T\right\| dt' \leqq \frac{\tau\, e^{-K_4 T}}{2K_3 T} \tag{A.12}$$

for all $u \in B(u^*, \varepsilon') \cap M'$. By Assumption 3.7 (i), (3.15), and (A.1), there exists $\varepsilon \in (0, \varepsilon']$ such that

$$\|p(T, u) - p(T, u^*)\| \leqq \frac{\tau}{2} e^{-K_4 T} \quad \text{for all } u \in B(u^*, \varepsilon) \cap M'. \tag{A.13}$$

Now using (3.14) and adding and subtracting appropriate terms, we obtain from (A.11), (A.12), and (A.13) that

$$\begin{aligned} \|p(T - s, u) - p(T - s, u^*)\| &\leqq \frac{\tau}{2} e^{-K_4 T} + \int_0^s K_4 \|p(T - s', u) - p(T - s', u^*)\|\, ds' \\ &\quad + \int_0^T \left\| \frac{\partial}{\partial x} f(x(T - s', u), u(T - s'), T - s') \right. \\ &\qquad \left. - \frac{\partial}{\partial x} f(x(T - s', u^*), u^*(T - s'), T - s') \right\| \\ &\qquad \cdot \|p(T - s', u^*)\|\, ds' \\ &\leqq \tau\, e^{-K_4 T} + \int_0^s K_4 \|p(T - s', u) - p(T - s', u^*)\|\, ds' \end{aligned} \tag{A.14}$$

for all $u \in B(u^*, \varepsilon) \cap M'$, for all $s \in [0, T]$. An application of the Bellman–Gronwall inequality to (A.14) completes the proof.

A.15. LEMMA. *Let the assumptions of* § 3 *hold. Then for any* $u^* \in M'$ *and* u *such that* $\|u\|_\infty < \infty$,

$$\lim_{\lambda \downarrow 0} \left\| \frac{x(\cdot, u^* + \lambda u) - x(\cdot, u^*) - \lambda v(\cdot)}{\lambda} \right\|_\infty = 0, \tag{A.16}$$

where $v(\cdot)$ is a solution of the differential equation

$$\text{(A.17)}\quad \begin{aligned} \dot{v}(t) &= \frac{\partial}{\partial x} f(x(t, u^*), u^*(t), t)v(t) \\ &\quad + \frac{\partial}{\partial u} f(x(t, u^*), u^*(t), t)u(t), \quad v(0) = 0, \quad t \in [0, T]. \end{aligned}$$

Proof. Let $\delta_\lambda x(\cdot) = x(\cdot, u^* + \lambda u) - x(\cdot, u^*)$. Then $\delta_\lambda x$ must satisfy the differential equation

$$\text{(A.18)}\quad \begin{aligned} \frac{d}{dt}\delta_\lambda x(t) &= f(x(t, u^* + \lambda u), u^*(t) + \lambda u(t), t) \\ &\quad - f(x(t, u^*), u^*(t), t), \quad \delta_\lambda x(0) = 0, \quad \text{a.e. on } [0, T]. \end{aligned}$$

By the mean value theorem, for any $t \in [0, T]$ there exist real numbers $v^i(t)$, $\mu^i(t) \in [0, 1]$, for $i = 1, 2, \cdots, n$, such that

$$\text{(A.19)}\quad \begin{aligned} \frac{d}{dt}\delta_\lambda x^i(t) &= \frac{\partial}{\partial x} f^i(x(t, u^*) + v^i(t)\delta_\lambda x(t), u^*(t) + \mu^i(t)\lambda u(t), t)\delta_\lambda x(t) \\ &\quad + \frac{\partial}{\partial u} f^i(x(t, u^*) + v^i(t)\delta_\lambda x(t), u^*(t) + \mu^i(t)\lambda u(t), t)\lambda u(t), \\ &\qquad i = 1, 2, \cdots, n, \quad \text{a.e. on } [0, T]. \end{aligned}$$

For simplicity, we write (A.19) in the following vector form:

$$\text{(A.20)}\quad \frac{d}{dt}\delta_\lambda x(t) = \frac{\partial}{\partial x} f_{v,\mu}\delta_\lambda x(t) + \frac{\partial}{\partial u} f_{v,\mu}\lambda u(t).$$

We now can combine (A.17) and (A.20) to obtain

$$\text{(A.21)}\quad \begin{aligned} \frac{d}{dt}[\delta_\lambda x(t) - \lambda v(t)] &= \frac{\partial}{\partial x} f \cdot [\delta_\lambda x(t) - \lambda v(t)] + \left(\frac{\partial}{\partial u} f_{v,\mu} - \frac{\partial}{\partial u} f\right)\lambda u(t) \\ &\quad + \left(\frac{\partial}{\partial x} f_{v,\mu} - \frac{\partial}{\partial x} f\right)\delta_\lambda x(t) \quad \text{a.e. on } [0, T]. \end{aligned}$$

It follows from (A.21) that

$$\text{(A.22)}\quad \begin{aligned} \|\delta_\lambda x(t) - \lambda v(t)\| &\leqq \int_0^t \left\|\frac{\partial}{\partial x} f\right\| \|\delta_\lambda x(t') - \lambda v(t')\|\, dt' \\ &\quad + \lambda\|u\|_\infty \int_0^T \left\|\frac{\partial}{\partial u} f_{v,\mu} - \frac{\partial}{\partial u} f\right\| dt' \\ &\quad + \|\delta_\lambda x\|_\infty \int_0^T \left\|\frac{\partial}{\partial x} f_{v,\mu} - \frac{\partial}{\partial x} f\right\| dt' \end{aligned}$$

for all $t \in [0, T]$. Applying the Bellman–Gronwall inequality to (A.22), we obtain

$$\text{(A.23)}\quad \|\delta_\lambda x(t) - \lambda v(t)\| \leqq (I_1 + I_2)\exp\left(\int_0^T \left\|\frac{\partial}{\partial x} f\right\| dt'\right)$$

for all $t \in [0, T]$, where

$$I_1 = \lambda \|u\|_\infty \int_0^T \left\| \frac{\partial}{\partial u} f_{\nu,\mu} - \frac{\partial}{\partial u} f \right\| dt', \tag{A.24}$$

$$I_2 = \|\delta_\lambda\|_\infty \int_0^T \left\| \frac{\partial}{\partial x} f_{\nu,\mu} - \frac{\partial}{\partial x} f \right\| dt'. \tag{A.25}$$

Since $\|u\|_\infty < \infty$, we conclude that

$$\left\| \frac{\partial}{\partial u} f_{\nu,\mu} - \frac{\partial}{\partial u} f \right\|_\infty \to 0 \quad \text{and} \quad \left\| \frac{\partial}{\partial x} f_{\nu,\mu} - \frac{\partial}{\partial x} f \right\|_\infty \to 0$$

as $\lambda \downarrow 0$. Thus the integrals in (A.24) and (A.25) converge to 0 as $\lambda \downarrow 0$. Consequently, making use of (A.1) and the fact that $\|u\|_\infty < \infty$, we obtain

$$\lim_{\lambda \downarrow 0} \frac{I_1}{\lambda} = \lim_{\lambda \downarrow 0} \|u\|_\infty \int_0^T \left\| \frac{\partial}{\partial u} f_{\nu,\mu} - \frac{\partial}{\partial u} f \right\| dt' = 0, \tag{A.26}$$

and since $\|u\|_2 \leqq \|u\|_\infty \sqrt{T}$,

$$\lim_{\lambda \downarrow 0} \frac{I_2}{\lambda} \leqq \lim_{\lambda \downarrow 0} k\|u\|_2 \int_0^T \left\| \frac{\partial}{\partial x} f_{\nu,\mu} - \frac{\partial}{\partial x} f \right\| dt' = 0. \tag{A.27}$$

But now, if we combine (A.23), (A.26), and (A.27), we see that $(\|\delta_\lambda x - \lambda v\|_\infty/\lambda) \to 0$ as $\lambda \downarrow 0$ and the proof is complete.

A.28. THEOREM. *Let the assumptions of* § 3 *hold. Then for any* $u^* \in M'$ *and* u *such that* $\|u\|_\infty < \infty$,

$$c'(u^*, u) \triangleq \lim_{\lambda \downarrow 0} \frac{c(u^* + \lambda u) - c(u^*)}{\lambda} = \langle h(u^*), u \rangle_2, \tag{A.29}$$

where

$$h(u^*)(t) = \frac{\partial}{\partial u} f(x(t, u^*), u^*(t), t)^T p(t, u^*), \tag{A.30}$$

with

$$\dot{p}(t, u) = -\frac{\partial}{\partial x} f(x(t, u^*), u^*(t), t)^T p(t, u^*) \quad \textit{a.e. on } [0, T], \tag{A.31}$$

$$p(T, u) = \frac{\partial}{\partial x} g(x(T, u^*))^T, \tag{A.32}$$

and $c' : M' \times \{u \in L_2^m[0, T] \mid \|u\|_\infty < \infty\} \to \mathbb{R}^1$ *is continuous.*

Proof. From (A.17), (A.31), and (A.32) there exists a continuous matrix-valued function $\Phi: \mathbb{R}^1 \times \mathbb{R}^1 \to \mathbb{R}^{n \times n}$ such that

$$v(t) = \int_0^t \Phi(t, s) \frac{\partial}{\partial u} f(x(s, u^*), u^*(s), s) u(s)\, ds, \tag{A.33}$$

$$p(t, u^*) = \Phi^T(T, t) \frac{\partial}{\partial x} g(x(T, u^*))^T \tag{A.34}$$

for all $t \in [0, T]$ (see [6]). Because of Lemma A.15,

$$c(u^* + \lambda u) - c(u^*) = g(x(T, u^* + \lambda u)) - g(x(T, u^*))$$

$$\text{(A.35)} \qquad = \frac{\partial}{\partial x} g(x(T, u^*))[\lambda v(T) + o_1(\lambda)] + o_2(x(T, u^* + \lambda u) - x(T, u^*)),$$

where $o_1(\lambda)/\lambda \to 0$ as $\lambda \downarrow 0$ and $o_2(\xi)/\|\xi\| \to 0$ as $\|\xi\| \to 0$. However, Lemma A.1 implies that

$$\text{(A.36)} \qquad \frac{o_2(x(T, u^* + \lambda u) - x(T, u^*))}{\lambda} \to 0 \quad \text{as } \lambda \downarrow 0.$$

Thus we have that

$$\lim_{\lambda \downarrow 0} \frac{c(u^* + \lambda u) - c(u^*)}{\lambda} = \frac{\partial}{\partial x} g(x, (T, u^*))v(T).$$

So if we apply (A.33) and (A.34) we obtain

$$c'(u^*, u) = \int_0^T \left\langle \frac{\partial}{\partial x} g(x(T, u^*))^T, \Phi(T, s) \frac{\partial}{\partial u} f(x(s, u^*), u^*(s), s) u(s) \right\rangle ds$$

$$\text{(A.37)} \qquad = \int_0^T \left\langle \frac{\partial}{\partial u} f(x(s, u^*), u^*(s), s)^T p(s, u^*), u(s) \right\rangle ds$$

$$= \langle h(u^*), u \rangle_2 .$$

Finally, by (A.1) and (A.9) we see that x and p, as maps from M' into the space of n-vector-valued continuous functions, are continuous. Therefore it follows that the map $u^* \in M' \to h(u^*) \in L_2^m[0, T]$ is continuous. But since $\langle \cdot , \cdot \rangle_2$ is continuous, $c'(\cdot , \cdot)$ is continuous and the proof is complete.

REFERENCES

[1] M. K. INAN, *On the perturbational sensitivity of solutions to nonlinear differential equations*, Memo. ERL-M270, Electronics Research Laboratory, Univ. of California, Berkeley, 1970.

[2] E. ISAACSON AND H. KELLER, *Analysis of Numerical Methods*, John Wiley, New York, 1966.

[3] R. KLESSIG AND E. POLAK, *An adaptive algorithm for unconstrained optimization with application to optimal control*, Memo. ERL-M297, Electronics Research Laboratory, Univ. of California, Berkeley, 1971.

[4] G. MEYER AND E. POLAK, *Abstract models for the synthesis of optimization algorithms*, Memo. ERL-268, Electronics Research Laboratory, Univ. of California, Berkeley, 1969.

[5] E. POLAK, *Computational Methods in Optimization: A Unified Approach*, Academic Press, New York, 1971.

[6] L. A. ZADEH AND C. A. DESOER, *Linear System Theory*, McGraw-Hill, New York, 1963.

SIAM J. Control
Vol. 11, No. 1, February 1973

CONSTRUCTION OF STABILITY MULTIPLIERS WITH PRESCRIBED PHASE CHARACTERISTICS: AN IMPROVED VALUE FOR σ_* *

S. RAMARAJAN AND M. A. L. THATHACHAR†

Abstract. For a feedback system consisting of a transfer function $G(s)$ in the forward path and a time-varying gain $n(t)$ $(0 \leqq n(t) \leqq k)$ in the feedback loop, a stability multiplier $Z(s)$ has been constructed (and used to prove stability) by Freedman [2] such that $Z(s)(G(s) + 1/K)$ and $Z(s - \sigma)$ $(0 < \sigma < \sigma_*)$ are strictly positive real, where σ_* can be computed from a knowledge of the phase-angle characteristic of $G(i\omega) + 1/k$ and the time-varying gain $n(t)$ is restricted by σ_* by means of an integral inequality. In this note it is shown that an improved value for σ_* is possible by making some modifications in his derivation.

Introduction. Considering the linear time-varying system consisting of a transfer function $G(s)$ in the forward path and a time-varying gain $n(t)$ $(0 \leqq n(t) \leqq k)$ in the feedback path, Gruber and Willems [4] proved stability under the existence of a multiplier $Z(s)$ such that $Z(s)(G(s) + 1/k)$ and $Z(s - \sigma)$ are strictly positive real (for some $0 < \sigma < \sigma_*$) and

$$\frac{dn(t)}{dt} \leqq 2\sigma n(t)\left(1 - \frac{n(t)}{k}\right) \quad \text{for } t \geqq 0.$$

Even though $Z(s)$ is allowed to be any general positive real function, it may be difficult under certain circumstances to find out such a $Z(s)$ satisfying the above conditions (since it need not be a rational function of s). A geometrical method of constructing such a positive real function, which avoids the abovementioned trial procedures, is given by Freedman [2]. He also obtained an approximate value of the upper bound σ_*, which in turn depended on the phase-angle characteristic of $G(i\omega) + 1/k$. Also Freedman [2] replaced Gruber and Willems' [4] time domain restriction by a more general integral criterion, namely, that the time-varying gain $n(t)$ satisfies the condition

$$\left|\frac{1}{t}\int_0^t \left|\frac{d}{d\tau}\left[\log\frac{n(\tau)}{(1 - n(\tau)/k)}\right] \pm 2\sigma\right| d\tau - 2\sigma\right| \leqq \frac{K}{t}$$

for some $K > 0$ and all $t > 0$. In this note we follow a similar procedure and obtain an improved value for the upper bound σ_*.

In the following we show that the value of σ_* can be improved to

$$\sigma_* = \frac{4(1 - \alpha/\pi)^2\alpha^2}{\int_{-W_\mu}^{W_\mu} |\Phi'(\omega)|^2\, d\omega}$$

from that given by Freedman [2], namely,

$$\sigma_* = \frac{(3\pi/16)^2\alpha^2}{\int_{-W_\nu}^{W_\nu} |\Phi'(\omega)|^2\, d\omega},$$

* Received by the editors September 14, 1971.

† Department of Electrical Engineering, Indian Institute of Science, Bangalore 12, India.

where

(i) $\Phi(\omega) \triangleq \arg[G(i\omega) + 1/k]$;

(ii) α is such that $|\Phi(\omega)| < \pi - \alpha$ for all ω;

(iii) $\mu = (\pi - \alpha)(\pi - 2\alpha)/(3\pi - 2\alpha)$ and $\nu = (\pi - \alpha)/3$ (we note that $\mu > 0$ for $\alpha < \pi/2$) and

(iv) $W_\mu \triangleq \min\{W \,|\, |\Phi(\omega)| \leqq \mu$ for $|\omega| \geqq W\}$ and $W_\nu \triangleq \min\{W \,|\, |\Phi(\omega)| \leqq \nu$ for $|\omega| \geqq W\}$.

For small α we see that $\mu \approx \nu$ and hence $W_\mu \approx W_\nu$ so that the value of σ_* obtained in this correspondence is approximately 12 times (i.e., $4(16/3\pi)^2$) that obtained by Freedman [2].

The detailed proofs and derivations are not given since they more or less follow that given in Freedman [2]. To start with we state a simple result which will be used later.

LEMMA 1.

$$\int_0^\infty \frac{\sqrt{\tau}}{1 + \tau^2}\, d\tau = \frac{\pi}{\sqrt{2}}.$$

Proof. Using the residue theorem in complex algebra we can show that [1, p. 256]

$$\int_0^\infty \frac{t^{z-1}}{1 + t}\, dt = \pi \operatorname{cosec} \pi z \quad \text{for } 0 < \operatorname{Re} z < 1.$$

Now taking $z = \frac{3}{4}$ and putting $t = \tau^2$ in the L.H.S. of the above equation we get the desired result.

For the sake of continuity we state without proof the lemma which ensures the existence of a causal multiplier with a prescribed phase characteristic. For a detailed proof one can refer to Freedman and Zames [3].

LEMMA 2 (Operators with prescribed phase). *If*:

(i) $\Phi_0(\omega)$ *is a real-valued continuous a.e. differentiable odd function of* ω *for* $\omega \in (-\infty, \infty)$;

(ii) $\Phi_0(\omega)$ *and* $\Phi_0'(\omega)$ *are in* $L_2(-\infty, \infty)$;

then:

(a) *there is a function* $\lambda(\cdot)$ *in* $L_1(-\infty, \infty)$ *with* $\lambda(t) = 0$ *for* $t < 0$ *and with a Laplace transform* $\Lambda(s)$ *satisfying* $\operatorname{Im} \Lambda(i\omega) = \Phi_0(\omega)$;

(b) *there is a* $y(\cdot)$ *in* $L_1(-\infty, \infty)$ *with* $y(t) = 0$ *for* $t < 0$ *and with a Laplace transform* $Y(s)$ *satisfying*

$$1 + Y(s) = \exp[\Lambda(s)] \quad \textit{for } \operatorname{Re} s \geqq 0;$$

(c) *if* $-\pi < \Phi_0(\omega) < \pi$, *there is a* $y(\cdot) \in L_1(-\infty, \infty)$ *with* $y(t) = 0$ *for* $t < 0$, $1 + Y(s) \neq 0$ *in* $\operatorname{Re} s \geqq 0$ *and*

$$\arg(1 + Y(i\omega)) = \Phi_0(\omega).$$

The following lemma is a slight modification of Lemma 2 in Freedman [2].

LEMMA 3. *Under the same notations and assumptions as in Lemma* 2, *it follows that for every* $\sigma > 0$,

$$\sup_{-\infty < \omega < \infty} |\operatorname{Im} \Lambda(i\omega + \sigma) - \Phi_0(\omega)| \leqq \sqrt{\sigma}\left(\int_{-\infty}^{\infty} |\Phi_0'(\omega)|^2\, d\omega\right)^{1/2}.$$

Thus if $-\pi < \Phi_0(\omega) < \pi$ *for all* ω *and if the principal value of the* arg $\{\cdot\}$ *is taken, then*

$$\sup_{-\infty<\omega<\infty} |\arg(1 + Y(i\omega + \sigma)) - \Phi_0(\omega)| \leqq \sqrt{\sigma}\left(\int_{-\infty}^{\infty} |\Phi_0'(\omega)|^2\, d\omega\right)^{1/2}.$$

Proof. If we denote Im $\Lambda(s) = V(\sigma, y)$ where $s = \sigma + iy$, it can be shown [2] that for $\sigma > 0$,

$$\text{(1)}\qquad \begin{aligned} V(\sigma, y) - \Phi_0(y) &= \frac{1}{\pi}\int_{-\infty}^{y} \frac{\sigma}{\sigma^2 + (y-\omega)^2}[\Phi_0(\omega) - \Phi_0(y)]\, d\omega \\ &\quad + \frac{1}{\pi}\int_{y}^{\infty} \frac{\sigma}{\sigma^2 + (y-\omega)^2}[\Phi_0(\omega) - \Phi_0(y)]\, d\omega. \end{aligned}$$

Now

$$\begin{aligned} &\left|\frac{1}{\pi}\int_{y}^{\infty} \frac{\sigma}{\sigma^2 + (y-\omega)^2}[\Phi_0(\omega) - \Phi_0(y)]\, d\omega\right| \\ &\leqq \frac{1}{\pi}\int_{y}^{\infty} \frac{\sigma}{\sigma^2 + (y-\omega)^2}\left(\int_{y}^{\omega} |\Phi_0'(t)|\, dt\right) d\omega \\ &\leqq \frac{1}{\pi}\int_{y}^{\infty} \frac{\sigma}{\sigma^2 + (y-\omega)^2}\left(\left[\int_{y}^{\omega} |\Phi_0'(t)|^2\, dt\right]^{1/2} (w - y)^{1/2}\right) d\omega \\ &\leqq \left(\int_{y}^{\infty} |\Phi_0'(t)|^2\, dt\right)^{1/2} \cdot \frac{1}{\pi}\left(\int_{y}^{\infty} \frac{\sigma\sqrt{\omega - y}}{\sigma^2 + (\omega - y)^2}\, d\omega\right) = A. \end{aligned}$$

(Taking $(\omega - y)/\sigma = \tau$, the second integral in the R.H.S. of the above inequality becomes

$$\sqrt{\sigma}\int_0^{\infty} \frac{\sqrt{\tau}}{1+\tau^2}\, d\tau = \frac{\pi}{\sqrt{2}}\sqrt{\sigma},$$

after using Lemma 1.) Now

$$\text{(2)}\qquad A = \frac{\sqrt{\sigma}}{\sqrt{2}}\left(\int_{y}^{\infty} |\Phi_0'(t)|^2\, dt\right)^{1/2}.$$

Similarly we can show that

$$\text{(3)}\qquad \left|\frac{1}{\pi}\int_{-\infty}^{y} \frac{\sigma}{\sigma^2 + (y-\omega)^2}[\Phi_0(\omega) - \Phi_0(y)]\, d\omega\right| \leqq \frac{\sqrt{\sigma}}{\sqrt{2}}\left(\int_{-\infty}^{y} \cdot|\Phi_0'(t)|^2\, dt\right)^{1/2}.$$

Combining (1), (2) and (3), we get

$$\begin{aligned} |V(\sigma, y) - \Phi_0(y)| &\leqq \frac{\sqrt{\sigma}}{\sqrt{2}}\left[\left(\int_{-\infty}^{y} |\Phi_0'(t)|^2\, dt\right)^{1/2} + \left(\int_{y}^{\infty} |\Phi_0'(t)|^2\, dt\right)^{1/2}\right] \\ &\leqq \sqrt{\sigma}\left(\int_{-\infty}^{\infty} |\Phi_0'(t)|^2\, dt\right)^{1/2} \end{aligned}$$

after using the inequality $\sqrt{a} + \sqrt{b} \leqq \sqrt{2}\sqrt{a+b}$ for $a, b \geqq 0$. Since the R.H.S.

is independent of y,

$$\sup_{-\infty<y<\infty} |V(\sigma, y) - \Phi_0(y)| \leqq \sqrt{\sigma}\left(\int_{-\infty}^{\infty} |\Phi_0'(t)|^2\, dt\right)^{1/2}$$

and hence the proof is complete.

Finally we state the main result.

LEMMA 4. *Under the usual notations, let*

$$\sigma_* \triangleq \frac{4(1 - \alpha/\pi)^2\alpha^2}{\int_{-W_\mu}^{W_\mu} |\Phi_0'(\omega)|^2\, d\omega}.$$

Then for any $\sigma \in [0, \sigma_*)$ *there is a* $y(\cdot)$ *in* $L_1[0, \infty)$ *with Laplace transform* $Y(s)$ *and a* $\delta > 0$, *such that*

(i) $\operatorname{Re}\{1 + Y(i\omega)\} \geqq \delta > 0$,

(ii) $\operatorname{Re}\{[1 + Y(i\omega + \sigma)][G(i\omega) + 1/k]\} \geqq \delta > 0$.

Remarks. Since $\arg[G(i\omega) + 1/k]$ lies in the interval $(-\pi - \alpha, \pi - \alpha)$, as an initial attempt we set

$$\arg[1 + Y(i\omega)] = -\frac{1}{2}\left(\arg\left[G(i\omega) + \frac{1}{k}\right]\right)\left(\frac{\pi}{\pi - \alpha}\right)$$

so that $\arg[1 + Y(i\omega)]$ lies in the interval $(-\pi/2, \pi/2)$. In the following we adopt a procedure similar to that of Freedman so that the phase function and its derivative satisfy the square integrability conditions stated in Lemma 2 for the existence of a $y(\cdot) \in L_1[0, \infty)$.

Proof. For each $\varepsilon > 0$, choose $l_\varepsilon(\omega)$ a continuous, a.e. differentiable real-valued function defined on $[W_\mu, \infty)$ and satisfying

(a) $l_\varepsilon(W_\mu) = -\dfrac{\Phi(W_\mu)}{2}\cdot\left(\dfrac{\pi}{\pi - \alpha}\right) = -\dfrac{\arg[G(i\omega) + 1/k]}{2}\left(\dfrac{\pi}{\pi - \alpha}\right)\Bigg|_{\omega = W_\mu}$;

(b) $|l_\varepsilon(\omega)| \leqq \dfrac{\mu}{2}\left(\dfrac{\pi}{\pi - \alpha}\right)$ for all $\omega \in (W_\mu, \infty)$;

(c) $\displaystyle\int_{W_\mu}^{\infty} |l_\varepsilon(\omega)|^2\, d\omega < \infty$ and $\displaystyle\int_{W_\mu}^{\infty} |l_\varepsilon'(\omega)|^2\, d\omega < \frac{\varepsilon^2}{2}$.

(As stated in Freedman [2] such functions $l_\varepsilon(\omega)$ are easily constructed.) For any $\varepsilon > 0$, define

$$\Phi_\varepsilon(\omega) = \begin{cases} -\dfrac{\Phi(\omega)}{2}\left(\dfrac{\pi}{\pi - \alpha}\right) = -\dfrac{1}{2}\left(\arg\left[G(i\omega) + \dfrac{1}{k}\right]\right)\left(\dfrac{\pi}{\pi - \alpha}\right) & \text{for } \omega \in [-W_\mu, W_\mu], \\ l_\varepsilon(\omega) & \text{for } \omega > W_\mu, \\ -l_\varepsilon(-\omega) & \text{for } \omega < -W_\mu. \end{cases}$$

Hence by the application of Lemma 2, it follows that there is a $y_\varepsilon(\cdot) \in L_1[0, \infty)$ with Laplace transform $Y_\varepsilon(s)$ and $\arg[1 + Y_\varepsilon(i\omega)] = \Phi_\varepsilon(\omega)$. Also by the above construction it is assured that $|\Phi_\varepsilon(\omega)| < \pi/2$ and hence

$$\operatorname{Re}[1 + Y_\varepsilon(i\omega)] \geqq \delta(\varepsilon) > 0$$

for some constant $\delta(\varepsilon)$. Thus (i) in Lemma 4 holds with $y(\cdot)$ equal to $y_\varepsilon(\cdot)$ chosen by the above procedure.

For further reference we see that the inequality

$$\int_{-\infty}^{\infty} |\Phi_\varepsilon'(\omega)|^2\, d\omega \leqq \frac{1}{4}\left(\frac{\pi}{\pi-\alpha}\right)^2 \left(\int_{-W_\mu}^{W_\mu} |\Phi'(\omega)|^2\, d\omega\right) + \varepsilon^2 \tag{4}$$

holds. Now, in order for (ii) in Lemma 4 to hold it is sufficient to show that

$$|\arg[1 + Y_\varepsilon(i\omega + \sigma)] + \Phi(\omega)| < \pi/2 \quad \text{for all } \omega. \tag{5}$$

Proceeding along lines similar to that of Freedman [2] and using the result of Lemma 3 and the definition of μ, one can show that (see Appendix) a sufficient condition for (5) to hold is that

$$\sigma\left(\int_{-\infty}^{\infty} |\Phi_\varepsilon'(\omega)|^2\, d\omega\right) < \alpha^2. \tag{6}$$

Now using (4) and making ε sufficiently small, we see that (6) is satisfied for any σ, $0 < \sigma < \sigma_*$, where

$$\sigma_* = \frac{4(1-\alpha/\pi)^2\alpha^2}{\int_{-W_\mu}^{W_\mu} |\Phi'(\omega)|^2\, d\omega}.$$

Thus $y_\varepsilon(\cdot)$ corresponding to such a choice of ε will be an acceptable choice for $y(\cdot)$ in the statement of this lemma.

Also by defining $Z(s) = 1 + Y(s + \sigma)$ we see that $Z(s)(G(s) + 1/k)$ and $Z(s - \sigma)$ $(0 < \sigma < \sigma_*)$ are strictly positive real with such a choice of σ_*.

Appendix. Inequality (5) can be written as

$$|\arg[1 + Y_\varepsilon(i\omega + \sigma)] - \Phi_\varepsilon(\omega) + \Phi_\varepsilon(\omega) + \Phi(\omega)| < \pi/2 \quad \text{for } -\infty < \omega < \infty. \tag{A.1}$$

From Lemma 3 it follows that

$$|\arg[1 + Y_\varepsilon(i\omega + \sigma)] - \Phi_\varepsilon(\omega)| \leqq \sqrt{\sigma}\left(\int_{-\infty}^{\infty} |\Phi_\varepsilon'(\omega)|^2\, d\omega\right)^{1/2}.$$

Hence a sufficient condition for (A.1) to hold is that

$$\sqrt{\sigma}\left(\int_{-\infty}^{\infty} |\Phi_\varepsilon'(\omega)|^2\, d\omega\right)^{1/2} + \sup_{-\infty<\omega<\infty} |\Phi_\varepsilon(\omega) + \Phi(\omega)| < \frac{\pi}{2}. \tag{A.2}$$

Now

$$\sup_{\omega, |\omega| \leqq W_\mu} |\Phi_\varepsilon(\omega) + \Phi(\omega)| = \sup_{\omega, |\omega| \leqq W_\mu} \left|\frac{\pi - 2\alpha}{2(\pi - \alpha)}\Phi(\omega)\right|$$

$$< \frac{\pi - 2\alpha}{2(\pi-\alpha)}(\pi - \alpha) = \frac{\pi}{2} - \alpha$$

and

$$\sup_{\omega, |\omega| > W_\mu} |\Phi_\varepsilon(\omega) + \Phi(\omega)| \leqq \sup_{\omega, |\omega| > W_\mu} |\Phi_\varepsilon(\omega)| + \sup_{\omega, |\omega| > W_\mu} |\Phi(\omega)|$$

$$\leqq \frac{\mu}{2}\left(\frac{\pi}{\pi - \alpha}\right) + \mu = \frac{\pi}{2} - \alpha.$$

Thus it follows that

$$\sup_{-\infty < \omega < \infty} |\Phi_\varepsilon(\omega) + \Phi(\omega)| \leqq \frac{\pi}{2} - \alpha.$$

Using this we see that a sufficient condition for (A.2) and hence for (5) to hold is that

$$\sqrt{\sigma}\left(\int_{-\infty}^{\infty} |\Phi_\varepsilon'(\omega)|^2 \, d\omega\right)^{1/2} < \alpha,$$

which is nothing but (6).

REFERENCES

[1] M. Abramowitz and I. A. Stegun, eds., *Handbook of Mathematical Functions*, Dover, New York, 1965.

[2] M. I. Freedman, *L_2-stability of time-varying systems—construction of multipliers with prescribed phase characteristics*, this Journal, 6 (1968), pp. 559–578.

[3] M. I. Freedman and G. Zames, *Logarithmic variation criteria for the stability of systems with time-varying gains*, this Journal, 6 (1968), pp. 487–507.

[4] M. Gruber and J. L. Willems, *On a generalization of the circle criterion*, Proc. 4th Allerton Conference on Circuit and System Theory, University of Illinois, Urbana, 1966, pp. 827–848.

SIAM J. CONTROL
Vol. 11, No. 1, February 1973

SADDLE POINTS FOR LINEAR DIFFERENTIAL GAMES*

R. J. ELLIOTT†, N. J. KALTON‡ AND L. MARKUS‡

Abstract. It is proved that there is a saddle point over the relaxed controls, and so over the strategies defined on the relaxed controls, for differential games in which the trajectory variable appears linearly in the dynamical equation and payoff. This is a strong saddle point property, but the example of Berkovitz [1], of a game that does not have a saddle point in pure strategies, does have a saddle point in this sense. Saddle points over the chattering controls are obtained for linear games in which the opposing control variables appear separated. The introduction of relaxed controls into differential games is analogous to the introduction by von Neumann of mixed strategies into two person, zero sum games.

1.1. Introduction. Motivated by work in the calculus of variations and control theory (see [6], [12], [13] and [14]), we introduce relaxed controls to study two person, zero sum differential games. For differential games which possess a certain linearity property in the trajectory variable (in the differential equation describing the game and in the payoff function), we prove there is a saddle point over the relaxed controls, and so over the strategies defined on the relaxed controls. As pointed out below this concept of a saddle point over the (relaxed) controls is a strong property because a player can use his saddle point control, and, independently of what the other player does, not loose in comparison with the saddle point payoff.

1.2. Outline. We first define a two person, zero sum differential game and describe the notions of strategy and saddle point. In § 3 and § 4 we introduce relaxed controls in a rigorous manner as elements of the dual of a Banach space of integrable vector-valued functions. The presentation follows Warga [13]. The relaxed controls are shown to be a compact convex set and we prove that the "piecewise constant" relaxed controls are dense.

The games considered have a payoff which, though bilinear, is only separately continuous on the product space of relaxed controls for the two players. However, a general theorem of Sion [10] then states there is a saddle point, so our principal result concerns the existence of a saddle point among the relaxed controls, for linear differential games. We observe that the well-known example of Berkovitz [1] (see also [3]) of a differential game that does not have a saddle point in pure strategies does have a saddle point in our sense in the form of constant strategies on the relaxed controls.

Finally, for linear games in which the control variables appear separated, saddle points are obtained over the particularly simple relaxed controls known as chattering controls.

* Received by the editors July 27, 1971, and in revised form April 26, 1972. The first two authors were supported by Research Fellowships from the Science Research Council and the third author was supported in part by the Office of Naval Research under Contract N00014-67-A-0113-0019.

† Department of Mathematics, Northwestern University, Evanston, Illinois 60201.

‡ Mathematics Institute, University of Warwick, Coventry, Warwickshire CV4 7AL, England.

2. Differential games. The situation to be discussed is a two person, zero sum differential game described as follows.

Notation 2.1. We have a dynamical system

$$\dot{x}(t) = f(t, x, u, v), \tag{2.1}$$

where the trajectory $x(t) \in R^m$, an initial position $x(0) = (x_1(0), \cdots, x_m(0)) = x_0$ is given, and the time t belongs to some closed bounded interval of T of R—say $T = [0, 1]$. There are two sets of control variables: $\{u\} \subset Y$, a compact subset of R^p, and $\{v\} \subset Z$, a compact subset of R^q.

To ensure integrability of the differential equation we shall assume that the vector function

$$f(t, x, u, v) = (f_1(t, x, u, v), \cdots, f_m(t, x, u, v))$$

is continuous in t, x, u and v and satisfies a Lipschitz condition in x of the form:

$$|f(t, x, u, v) - f(t, x', u, v)| \leqq \Psi(t)|x - x'| \quad \text{for } x, x' \in R^m$$

(or at least for x, x' in some compact subset of R^m wherein all trajectories $x(t)$ are known to lie), $t \in T$, $u \in Y$ and $v \in Z$. Here $\Psi(t)$ is an integrable real-valued function of t and $| \cdot |$ denotes the usual distance in R^m.

Together with the above dynamical system we have a payoff of the form:

$$P(u, v) = \mu(x(t)) + \int_0^1 h(t, x, u, v)\, dt. \tag{2.2}$$

Here μ is a (not necessarily linear) real-valued function on the Banach space $C([0, 1])$ of curves in R^m over $[0, 1]$, and h is a continuous real-valued function on $T \times R^m \times Y \times Z$.

DEFINITION 2.2. A *two player, zero sum differential game* is a dynamical system described by differential equations of the above form, together with the compact sets Y and Z and the payoff P.

The first player chooses $u \in Y$ at each time $t \in T$ in a measurable way, thus generating a function $u(t)$, so that the final payoff $P(u(t), v(t))$ is as large as possible. At the same time the second player chooses v at each time $t \in T$ so that $P(u(t), v(t))$ is as small as possible.

Remark 2.3. Denote by U (resp. V) the set of measurable functions from T to Y (resp. Z).

A (pure) strategy for the first player would ideally be defined as some rule which, for each time $t \in T$, determines for him his choice of $u(t)$ on the basis of what has happened in the game so far, that is, from the knowledge of $x(\tau)$, $u(\tau)$ and $v(\tau)$ for $0 \leqq \tau < t$. Similarly, there would be a notion of strategy for the second player.

On the grounds that a player knows his own previous choice of controls, Roxin [5] has defined a strategy for the first player as a function α from V to U which is nonanticipatory in the sense that:

if $v_1(t), v_2(t) \in V$ and $v_1(\tau) = v_2(\tau)$ for $0 \leqq \tau \leqq t_0$, then $\alpha(v_1)(\tau) = \alpha(v_2)(\tau)$ for $0 \leqq \tau \leqq t_0$.

The difficulty with this definition is that, in general, given a strategy α for the first player and a strategy β for the second player we do not know that there is an "outcome" for α and β, that is, a pair of control functions $u(t)$ and $v(t)$ such that

$$\alpha v = u \quad \text{and} \quad \beta u = v.$$

(We are looking for a fixed point of $\alpha \circ \beta : V \to V$, but α and β are not necessarily continuous.)

If a pair of strategies α, β do have an outcome $u(t)$, $v(t)$, we can write $P(\alpha, \beta)$ for the resulting payoff $P(u, v)$. A solution to the game is a pair of strategies α^*, β^*, which have an outcome which is "simultaneously best" for both players. This means that if α, β are other strategies such that α^*, β and α, β^* both have outcomes, then

$$P(\alpha^*, \beta) \geqq P(\alpha^*, \beta^*) \geqq P(\alpha, \beta^*). \tag{2.3}$$

DEFINITION 2.4. A pair of strategies α^*, β^* which satisfy (2.3) are said to form a *saddle point* for the differential game (*over the pure strategies*).

A much stronger notion of saddle point is now described.

DEFINITION 2.5. A pair of control functions $u^*(t) \in U$, $v^*(t) \in V$ are said to form a *saddle point* (*over the controls*) if for any other controls $u(t) \in U$, $v(t) \in V$,

$$P(u^*, v) \geqq P(u^*, v^*) \geqq P(u, v^*).$$

LEMMA 2.6. *If* (u^*, v^*) *are a saddle point over the controls, then there are identically constant strategies* (α^*, β^*) *which are a saddle point in the sense of Definition* 2.4.

Proof. For any $v(t) \in V$ and $u(t) \in U$ we define $\alpha^* v = u^*$, $\beta^* u = v^*$. That is, α^* is described by saying that, whatever the second player does, the first player continues to play his control $u^*(t)$. β^* is described similarly. If α, β are any other strategies, then α^*, β have an outcome u^*, v, say, and α, β^* have an outcome u, v^*. Thus, $P(\alpha^*, \beta) \geqq P(\alpha^*, \beta^*) \geqq P(\alpha, \beta^*)$.

Remarks 2.7. A saddle point over the controls is, therefore, a strong concept, because if one player uses his saddle point control then independently of what the other player does, he cannot lose. However, a saddle point over the strategies does not give us a saddle point over the controls.

If (u^*, v^*) form a saddle point over the controls and if u^* is played, then v must play to minimize $P(u^*, v)$. Hence $v^*(t)$ is an optimal controller subject to the maximum principle. Similarly, $u^*(t)$ is an optimal controller maximizing $P(u, v^*)$.

Also, note that any other saddle point $(\bar{u}, \bar{v})$ over the controls must give the same payoff, but nonsaddle plays might give the same value of P without satisfying the saddle point inequalities.

Background 2.8. We have already noted that there are difficulties surrounding the notion of a strategy for a player in a differential game. Friedman [3] circumvents these difficulties by considering "upper" and "lower" approximating games in which one player or the other has advance information. To obtain the convergence of the resulting "upper" and "lower" values of the game he requires in effect that the payoff $P(u, v)$ be jointly continuous in both control functions. This he ensures by considering only dynamical systems of the form

$$\dot{x} = f_1(t, x, u) + f_2(t, x, v)$$

and payoffs of the form

$$P(u, v) = \mu(x(t)) + \int_0^1 h_1(t, x, u) + h_2(t, x, v)\, dt,$$

so that u and v are "separated."

Somewhat earlier (cf. [1]), Berkovitz studied differential games by a "variational approach," introducing related Hamilton–Jacobi equations which are, unfortunately, hard to analyze.

2.1. A game with no pure strategy solutions. In [1] Berkovitz considers a differential game described by an equation

$$\dot{x}(t) = 4(u - v)^2, \qquad x \in R^1, \qquad t \in [0, 1], \qquad x(0) = 0,$$

and with payoff

$$P(u(t), v(t)) = \int_0^1 x(t)\, dt.$$

The control sets are

$$Y = \{u : 0 \leqq u \leqq 1\} \quad \text{and} \quad Z = \{v : 0 \leqq v \leqq 1\}.$$

In terms of his "variational approach," Berkovitz shows this game has no solution in pure strategies.

A. Friedman also cites this example in [3] and shows that for this game his upper value $V^+ = \frac{1}{2}$ while his lower value $V_- = 0$. Thus, as $V^+ \neq V_-$, this game does not have a value in his theory.

We shall see below, however, that by introducing the idea of relaxed controls, this game, for example, does have a saddle point over the relaxed controls.

3. Relaxed controls. The notion of relaxed, or generalized, curve was introduced into the calculus of variations by Young [14], and applied to control theory by Warga [13], McShane [6] and Young [15]. For further introduction of relaxed controls into differential games see the paper by Smoljakov [11]. The method and content of Smoljakov's paper, however, is quite different from the treatment below. In the discussion of relaxed controls described below we shall follow the setting described by Warga [13].

Suppose we have a dynamical system and payoff as described in Notation 2.1. We have already introduced U and V for the spaces of measurable functions from T to Y and Z respectively. U and V are, of course, just the spaces of (classical) control functions.

DEFINITION 3.1. Denote by (PY) and (PZ) the space of all regular probability measures defined on the Borel subsets of $Y \subset R^p$ and $Z \subset R^q$, respectively. A *relaxed control* for the first player is a function σ from $T = [0, 1]$ to PY. A relaxed control is *continuous* (resp. *measurable*) if $\int_Y f(u)\sigma(du; t)$ is a continuous (resp. measurable) function of $t \in T$, for every continuous real-valued function f on Y. A relaxed control for the second player: $\tau : T \to PZ$ is defined similarly. We shall identify relaxed controls which differ only on a set of measure zero.

Remarks 3.2. Here $\sigma(A, t)$ denotes the $\sigma(t)$-measure of any Borel set $A \subset Y$. By approximating the characteristic function of A with continuous functions on

Y, for example, it is easy to see that, if σ is a measurable relaxed control and A is a Borel subset of Y, then $\sigma(A, t)$ is measurable and integrable over T.

Notation 3.3. Denote by $\mathscr{S}(Y)$ the set of measurable relaxed controls on Y.

If $u(t) \in U$ is a classical control, then $\delta(u(t))$ can be thought of as an associated relaxed control giving, in effect, the same control. (By $\delta(u(t))$ we mean the probability measure on Y which at time t has total unit mass at $u(t)$—that is, a Dirac δ-function at $u(t)$.)

DEFINITION 3.4. If $u_1(t), \cdots, u_k(t)$ are classical controls and $\alpha_1(t) \geqq 0, \cdots, \alpha_k(t) \geqq 0$ are measurable functions on T such that $\sum \alpha_j(t) = 1$ almost everywhere, then we shall say that $\sum_{j=1}^k \alpha_j(t)\, \delta(u_j(t))$ is a *chattering control of degree k*. Such controls are special cases of relaxed controls and are discussed in Lee and Markus [4].

DEFINITIONS 3.5. $C(Y)$ will denote the Banach space of continuous real-valued functions f on Y with the usual norm: $\|f\| = \sup_{u \in y} |f(u)|$. $L^1_{[0,1]}(C(Y))$ will denote the Lebesgue space of integrable $C(Y)$-valued functions $\{\varphi\}$ defined on $[0, 1]$ with the norm

$$\|\varphi\| = \int_0^1 \sup_{u \in y} |\varphi(u, t)|\, dt.$$

For a discussion of Lebesgue spaces of Banach-space-valued functions, see Dunford and Schwartz [2] and Schwartz [9].

A real-valued function $\varphi(u, t)$ defined on $Y \times [0, 1]$ defines a function in $L^1_{[0,1]}(C(Y))$ if:

(i) $\varphi(u, t)$ is measurable in t for each $u \in Y$;

(ii) $\varphi(u, t)$ is continuous in u for each $t \in [0, 1]$ and

(iii) there exists an integrable real-valued function $\Phi(t)$ on $[0, 1]$ such that $|\varphi(u, t)| \leqq \Phi(t)$ on $Y \times [0, 1]$.

Conditions (i), (ii) and (iii) ensure that $\sup_{u \in Y} |\varphi(u, t)|$ is integrable and so $\|\varphi\|$ is finite.

Notation 3.6. Write B for the Banach space $L^1_{[0,1]}(C(Y))$.

DEFINITION 3.7. Denote by B^* the dual of B and by $\langle \varphi, \lambda \rangle$ the value of $\lambda \in B^*$ at $\varphi \in B$. We shall consider B^* to have the weak star topology. A sequence $\{\lambda_i\}$ converges to $\lambda \in B^*$ in this topology if

$$\lim_{i \to \infty} \langle \varphi, \lambda_i \rangle = \langle \varphi, \lambda \rangle \quad \text{for all } \varphi \in B.$$

Elements of B^* are described by the following lemma, which follows from results in [9, Exposé 4, p. 3].

LEMMA 3.8. *Suppose $\lambda \in B^*$. Then there is a measurable map μ from* $[0, 1]$ *to the class of regular signed Borel measures on Y such that*

$$\langle \varphi, \lambda \rangle = \int_0^1 \int_Y \varphi(u, t) \mu(du\,; t)\, dt \quad \textit{for all } \varphi \in B.$$

Furthermore, $|\mu|(Y, t) \in L^\infty([0, 1])$.

Note that the norm of $\lambda \in B^*$ is just $\operatorname{ess\,sup}_{t \in [0,1]} |\mu|(Y, t)$. Hence the norm of any relaxed controller $\sigma \in \mathscr{S}(Y)$ is just 1. From this lemma we have the following theorem.

THEOREM 3.9. *The set $\mathscr{S}(Y)$ of relaxed controls can be considered as a closed convex subset of the unit ball of B^*, and so with the weak star topology $\mathscr{S}(Y)$ is compact.*

For the proof again see [13].

For simplicity of exposition consider, instead of a differential game, a control system with just one control variable $u \in Y \subset R^p$; that is, a dynamical system

$$(3.1) \qquad \dot{x}(t) = f(t, x, u), \qquad x(t) \in R^m,$$

with initial condition $x(0) = x_0$ and f satisfying a Lipschitz condition

$$|f(t, x, u) - f(t, x', u)| \leqq \Psi(t)|x - x'|$$

with $\Psi(t)$ integrable.

Under these hypotheses we quote from [13].

THEOREM 3.10. *Suppose $\sigma \in \mathscr{S}(Y)$ is a measurable relaxed control. Then there is a unique absolutely continuous solution $x(t)$ of the differential equation*

$$(3.2) \qquad \dot{x}(t) = \int_Y f(t, x, u)\sigma(du; t)$$

differentiable and satisfying (3.2) *almost everywhere, with initial condition*

$$x(0) = x_0 \in R^m.$$

DEFINITION 3.11. Such a solution is called a *relaxed trajectory*.

From this Warga [13] proves the next result.

THEOREM 3.12. *Denote by $x(t; \sigma)$ the relaxed trajectory solution of* (3.2). *Suppose $\{\sigma_i\}$ is a sequence of measurable relaxed controls such that $\sigma_i \to \sigma$ in the weak star topology of $\mathscr{S}(Y)$. Then $x_i(t; \sigma_i)$ converges to $x(t; \sigma)$ in the uniform topology on* $[0, 1]$.

A corollary of this result is that the space of relaxed trajectories is compact in the uniform topology.

In preparation for our discussion of differential games let us return to our discussion of dynamical systems with two sets of control variables as described in Notation 2.1.

LEMMA 3.13. *If σ is a measurable relaxed control on Y and τ is a measurable relaxed control on Z, then $\sigma \times \tau$ is a measurable relaxed control on $Y \times Z$, and $\sigma \times \tau$ can be considered to belong to the unit sphere of the dual of $L^1(C(Y \times Z))$.*

Proof. The first statement is a simple consequence of results on product measures.

Given a function $f(t, u, v)$ in $L^1(C(Y \times Z))$ we have that for each $v' \in Z$, $f(t, u, v') \in L^1(C(Y))$. Therefore, $\int_Y f(t, u, v)\sigma(du; t)$ is continuous in v for each $t \in [0, 1]$ and is measurable and dominated by an integrable function in t (uniformly for all $v \in Z$). Thus we can consider

$$\int_0^1 \int_Y \int_Z f(t, u, v)\sigma(du; t)\tau(dv; t)\, dt = \langle f, \sigma \times \tau \rangle.$$

It is clear that $\sigma \times \tau$ is a probability measure on $Y \times Z$ for each $t \in [0, 1]$ and that $\sigma \times \tau$ has unit norm as a linear functional on $L^1(C(Y \times Z))$. Thus $\sigma \times \tau$ belongs to $\mathscr{S}(Y \times Z)$.

Remark 3.14. Note that by Fubini's theorem,

$$\int_Z \int_Y f(t,u,v)\sigma(du;t)\tau(dv;t) = \int_Y \int_Z f(t,u,v)\tau(dv;t)\sigma(du;t).$$

Discussion 3.15. Suppose $Y \subset R^p$, $Z \subset R^q$ are compact sets as above, and write

$$B_1 = L^1(C(Y)), \quad B_2 = L^1(C(Z)), \quad B_3 = L^1(C(Y \times Z)).$$

The dual spaces B_j^* will as usual be given the weak star topology. Denote by $\mathscr{S}(Y)$, $\mathscr{S}(Z)$, $\mathscr{S}(Y \times Z)$ the spaces of relaxed controls over $[0,1]$ on Y, Z and $Y \times Z$ respectively, so that we have

$$\mathscr{S}(Y) \subset B_1^*, \quad \mathscr{S}(Z) \subset B_2^*, \quad \mathscr{S}(Y \times Z) \subset B_3^*.$$

Lemma 3.13 above tells us that we have a natural mapping from $\mathscr{S}(Y) \times \mathscr{S}(Z)$ to $\mathscr{S}(Y \times Z)$. Of course this map is not surjective, but more surprising this (bilinear on convex combinations) mapping is *not* jointly continuous in both variables, as the following example shows.

Example 3.16. Suppose $Y = [0,1]$ and also $Z = [0,1]$. Consider a partition of the time interval $T = [0,1]$ into 2^n equal intervals $T_1 = [0, 1/2^n]$, $T_j = (j-1)/2^n, j/2^n]$, $j = 2, \cdots, 2^n$. Corresponding to the 2^n partition of T consider a relaxed control σ_n on Y and a relaxed control τ_n on Z which are piecewise constant on each T_j, $j = 1, \cdots, 2^n$, and which are such that

$$\sigma_n(\cdot\,;t) \quad \text{and} \quad \tau_n(\cdot\,;t)$$

are the unit mass at the point $1 \in Y$ (resp. $1 \in Z$) if $t \in T_1 \cup T_3 \cup T_5 \cup \cdots \cup T_{2^n-1}$, and $\sigma_n(\cdot\,;t)$ and $\tau_n(\cdot\,;t)$ are the unit mass at the point $0 \in Y$ (resp. $0 \in Z$) if $t \in T_2 \cup T_4 \cup T_6 \cup \cdots \cup T_{2^n}$.

Then it is easy to see that both σ_n and τ_n converge in B_1^* (resp. in B_2^*) to the constant relaxed control σ on Y (resp. τ on Z) which consists of a mass $\frac{1}{2}$ at 0 and mass $\frac{1}{2}$ at 1.

However, the product relaxed control $\sigma_n \times \tau_n$ on $Y \times Z = [0,1] \times [0,1]$ converges in B_3^* to the constant relaxed control π which consists of a mass $\frac{1}{2}$ at $(0,0) \in Y \times Z$ and mass $\frac{1}{2}$ at $(1,1) \in Y \times Z$.

Clearly $\pi \neq \sigma \times \tau$, so the map is not jointly continuous. (To check the above statements about the weak star convergence of σ_n, τ_n and $\sigma_n \times \tau_n$, it is sufficient to check how these relaxed controls act on products of functions $f(t)$, $\varphi(u)$, $\psi(v)$, where f is continuous on T, φ is continuous on Y and ψ is continuous on Z. This is because, for example, sums of products of the form $f(t)\,\varphi(u)\,\psi(v)$ are dense in B_3.)

DEFINITION 3.17. In Definition 3.1, we introduced the idea of measurable relaxed controls. Returning to a differential game described as in Notation 2.1, following Theorem 3.10, if the first player uses a relaxed control $\sigma(\cdot\,;t)$ and the second player uses a relaxed control $\tau(\cdot\,;t)$, then we define the dynamical equations to be given by the system

$$\dot{x}(t) = \int_Y \int_Z f(t,x,u,v)\sigma(du;t)\tau(dv;t). \tag{3.3}$$

Furthermore, the payoff corresponding to the relaxed controls σ and τ is defined to be

$$(3.4)\qquad P(\sigma, \tau) = \mu(x(t)) + \int_0^1 \int_Y \int_Z h(t, x, u, v)\sigma(du; t)\tau(du; t)\, dt.$$

Remarks 3.18. It is a consequence of Example 3.16 and Definition 3.17 that $P(\sigma, \tau)$ is not in general jointly continuous in σ and τ. However, in special situations, for example those discussed by Friedman [3] in which the control variables are separated, the payoff is jointly continuous and the definition of $P(\sigma, \tau)$ can be motivated by continuity because (cf. [13, Thm. 2.4]) the classical control functions are dense in the relaxed controls.

Similar to Definition 2.5 we have the strong concept of a saddle point over the relaxed controls.

DEFINITION 3.19. A pair of relaxed controls $\sigma^* \in \mathscr{S}(Y)$, $\tau^* \in \mathscr{S}(Z)$ is said to form a *saddle point over the relaxed controls* if for any other relaxed controls $\sigma \in \mathscr{S}(Y)$, $\tau \in \mathscr{S}(Z)$,

$$P(\sigma^*, \tau) \geqq P(\sigma^*, \tau^*) \geqq P(\sigma, \tau^*).$$

Remarks 3.20. Having made the above definitions we remark that, as is easily seen, one reason the Berkovitz game (cf. § 2.1) is difficult to analyze is that having introduced relaxed controls its payoff is not jointly continuous on $\mathscr{S}(Y) \times \mathscr{S}(Z)$.

4. Certain linear games. Suppose the system of equations describing the game has the form

$$(4.1)\qquad \dot{x}(t) = A(t)x(t) + f(t, u, v)$$

with initial condition $x(0) = x_0 \in R^m$. Here $A(t)$ is a continuous linear function of $t \in [0, 1]$, that is, $A(t)$ is an $m \times m$ matrix whose entries are continuous functions of time. $f(t, u, v)$ is a continuous function on $T \times Y \times Z$.

Furthermore, suppose the payoff has the form

$$(4.2)\qquad P(u, v) = \mu(x(t)) + \int_0^1 h(t, u, v)\, dt,$$

where μ is a continuous real-valued linear function on the Banach space $C([0, 1])$ of continuous R^m-valued functions on $[0, 1]$.

We are now in a position to prove our final result.

THEOREM 4.1. *Consider the differential game with dynamics and payoff given by equations of the above form* (4.1) *and* (4.2). *Then there is a pair of relaxed controls* $\sigma^*(\cdot, t)$, $\tau^*(\cdot, t)$ *which give a saddle point for the game when each player can play over his set of all relaxed controls. That is, if* $\sigma(\cdot, t)$ (*resp.* $\tau(\cdot, t)$) *is any other relaxed control for the first* (*resp. second*) *player,*

$$P(\sigma^*, \tau) \geqq P(\sigma^*, \tau^*) \geqq P(\sigma, \tau^*).$$

Proof. Since the system equations are linear in x, it follows that for fixed τ in $\mathscr{S}(Z)$ the mapping $\sigma \to \mu(x(\cdot))$ from $\mathscr{S}(Y)$ to the real numbers is continuous and linear on $\mathscr{S}(Y)$, where $\mathscr{S}(Y)$ is a subset of B_1^*, endowed with the weak star topology. Similarly, for fixed σ in $\mathscr{S}(Y)$ the mapping $\tau \rightarrow \mu(x(\cdot))$ from $\mathscr{S}(Z)$ to the real numbers is continuous and linear. Hence for fixed τ in $\mathscr{S}(Z)$ the mapping $\sigma \rightarrow P(\sigma, \tau)$ is continuous and linear on $\mathscr{S}(Y)$. Similarly for fixed σ in $\mathscr{S}(Y)$ the mapping $\tau \rightarrow P(\sigma, \tau)$ is continuous and linear on $\mathscr{S}(Z)$.

Since $\mathscr{S}(Y)$ and $\mathscr{S}(Z)$ are convex and compact, the existence of a saddle point follows from the general theorem of M. Sion [10]. The results in Sion give inf sup = sup inf, but since we have P continuous in each variable and $\mathscr{S}(Y)$ and $\mathscr{S}(Z)$ compact, it is easy to see that we also have max min = min max and the existence of a saddle point, that is, there are relaxed controls $\sigma^* \in \mathscr{S}(Y)$ and $\tau^* \in \mathscr{S}(Z)$ such that (4.3) is satisfied.

Remark 4.2. We note again that the example of Berkovitz [1] described in § 2.1 is a differential game described by an equation of the form (4.1) and with payoff of the form (4.2). This game, therefore, has a saddle point over the relaxed controls.

In fact, Smoljakov [11] proves by variational methods that a saddle point is obtained over the relaxed controls in the Berkovitz game if u (trying to maximize the payoff) "plays" a constant probability measure $\sigma^* =$ (mass $\frac{1}{2}$ at 0 and mass $\frac{1}{2}$ at 1) throughout the time interval, while v plays the constant control $v(t) = \frac{1}{2}$ throughout the interval.

5. Chattering control saddle points. In this section we examine several special cases of Theorem 4.1; in particular, we consider when the saddle point (σ^*, τ^*) over the relaxed controls can be reduced to a saddle point over classical controls or, perhaps, chattering controls (see Definition 3.4) of a specified degree.

THEOREM 5.1. *Consider a game with dynamics*

$$\dot{x} = A(t)x + B(u, t) + C(v, t),$$

$$x(0) = x_0 \in R^m,$$

and payoff

$$P(u, v) = \mu(x(t)) + \int_0^1 (F(u, t) + G(v, t))\, dt,$$

where

$$B: Y \times [0, 1] \rightarrow R^m, \quad C: Z \times [0, 1] \rightarrow R^m,$$

$$F: Y \times [0, 1] \rightarrow R, \quad G: Z \times [0, 1] \rightarrow R$$

are each continuous, and A and μ are as in (4.1). *Then if for each $t \in [0, 1]$ the sets*

$$L_t = \left\{\begin{pmatrix} B(u, t) \\ F(u, t) \end{pmatrix} u \in Y\right\}, \qquad M_t = \left\{\begin{pmatrix} C(v, t) \\ G(v, t) \end{pmatrix} v \in Z\right\}$$

in R^{m+1} are convex, there is a saddle point $(u(t), v(t))$ over the classical controls.

Proof. Let $(\sigma^*(t), \tau^*(t))$ be the saddle point over the relaxed controls obtained by Theorem 4.2. We determine $s(t) \in R^{m+1}$ by

$$s_i(t) = \begin{cases} \displaystyle\int_Y B_i(u, t)\, d\sigma^*(t, u), & 1 \leqq i \leqq m, \\ \displaystyle\int_Y F(u, t)\, d\sigma^*(t, u), & i = m + 1. \end{cases}$$

L_t is by assumption convex, and it is also compact as Y is compact and B and F are continuous. Hence it follows that $s(t) \in L_t$ and is a measurable function of $t \in [0, 1]$. By the Filippov implicit function theorem (cf. [5]) there is a measurable function $u^*:[0,1] \to Y$ such that

$$\begin{pmatrix} B(u^*(t), t) \\ F(u^*(t), t) \end{pmatrix} = s(t) = \int_Y \begin{pmatrix} B(u, t) \\ F(u, t) \end{pmatrix} d\sigma^*(t, u).$$

Similarly we determine $v^*(t)$ such that

$$\begin{pmatrix} C(v^*(t), t) \\ G(v^*(t), t) \end{pmatrix} = \int_Z \begin{pmatrix} C(v, t) \\ F(v, t) \end{pmatrix} d\tau^*(t, v).$$

It is clear that $u^*(t)$ has the same "effect" on the game as the relaxed control $\sigma^*(t)$, and similarly $v^*(t)$ has the same effect as $\tau^*(t)$. Hence it follows easily that

$$P(u^*, v) \geqq P(u^*, v^*) \geqq P(u, v^*)$$

for any other pair of control functions $u(t)$, $v(t)$.

COROLLARY 5.2. *The above result holds when Y and Z are compact and convex and*

$$B(t, u) = B'(t)u, \qquad C(t, v) = C'(t)v,$$

$$F(t, u) = F'(t)u, \qquad G(t, v) = G'(t)v,$$

where $B'(t)$, $C'(t)$, $F'(t)$, $G'(t)$ are each matrix-valued.

The assumption that L_t and $\mathcal{M}_t$ are convex for each t may be dropped if we are only interested in establishing a saddle point over the chattering controls of suitable degree.

THEOREM 5.3. *Consider a game with the same form as in Theorem 5.1 except that we do not assume that L_t and M_t are convex. Then there is a saddle point (σ_*, τ_*) over the chattering controls of a degree $m + 2$. If Y and Z are connected, we may take (σ_*, τ_*) of degree $m + 1$.*

Proof. Let $\Gamma(L_t)$ be the closed convex cover of L_t in R^{m+1}; then by a theorem of Carathéodory, if $\xi \in \Gamma(L_t)$,

$$\xi = \sum_{i=1}^{m+2} \alpha_i \xi_i,$$

where $\sum \alpha_i = 1$, $\alpha_i \geqq 0$ and $\xi_i \in L_t$. Now consider the set $\Delta^{m+2} \times Y^{m+2} \times [0, 1]$, where $\Delta^{m+2} \subset R^{m+2}$ is the set of all $\{\alpha_i\}_{i=1}^{m+2}$ such that $\sum \alpha_i = 1$ and $\alpha_i \geqq 0$. The map

$$\theta : \Delta^{m+2} \times Y^{m+2} \times [0, 1] \to R^{m+1}$$

given by

$$\theta(\alpha_1, \cdots, \alpha_{m+2}, u_1, \cdots, u_{m+2}, t) = \sum_{i=1}^{m+2} \alpha_i \begin{pmatrix} B(u_i, t) \\ F(u_i, t) \end{pmatrix}$$

is continuous; hence if (σ^*, τ^*) is the saddle point over relaxed controls, we may apply Fillipov's theorem to deduce the existence of measurable functions

$$\alpha_i:[0,1] \to R, \qquad i = 1, 2, \cdots, m+2,$$

$$v_i:[0,1] \to Y, \qquad i = 1, 2, \cdots, m+2,$$

such that $\alpha_i(t) \geqq 0$, $\sum \alpha_i(t) = 1$, and

$$\sum_{i=1}^{m+2} \alpha_i \begin{pmatrix} B(u_i, t) \\ F(u_j, t) \end{pmatrix} = \int_Y \begin{pmatrix} B(u, t) \\ F(u, t) \end{pmatrix} d\sigma^*(t, u)$$

for all t. The chattering control $\sigma_* = \sum_{i=1}^{m+2} \alpha_i(t)\delta_{u_i(t)}$ has the same "effect" as σ^*. A similar argument to that of Theorem 5.1 concludes the proof.

If Y (and then L_t) is connected, Carathéodory's result may be improved to expressing $\xi \in \Gamma(L_t)$ as

$$\xi = \sum_{i=1}^{m+1} \alpha_i \xi_i$$

and the proof proceeds as before.

This theorem may be extended to cases in which the u- and v-dependence in the dynamics does not split entirely, but becomes "polynomial-like" (in the terminology used in simple game theory). For simplicity we consider only the case where x is a real variable (i.e., we assume $m = 1$), with the dynamics of the game given by

$$\dot{x} = A(t)x + \sum_{i=0}^{p} \sum_{j=0}^{q} a_{ij}(t)\varphi_i(u, t)\psi_j(v, t) \tag{5.1}$$

(where we assume that $\varphi_0(u, t) \equiv 1$, $\psi_0(v, t) \equiv 1$), subject to the initial condition $x(0) = 0$; the payoff is also "polynomial-like":

$$P = \lambda(x(t)) + \int_0^1 \sum_{i=0}^{p} \sum_{j=0}^{q} b_{ij}(t)\varphi_i(u, t)\psi_j(\sigma, t)\, dt. \tag{5.2}$$

We assume that

$$\begin{aligned} &\varphi_i: Y \times [0,1] \to R, && i = 0, 1, 2, \cdots, p, \\ &\psi_j: Z \times [0,1] \to R, && j = 0, 1, 2, \cdots, q, \\ &\left.\begin{aligned} &a_{ij}:[0,1] \to R \\ &b_{ij}:[0,1] \to R \end{aligned}\right\} && \left\{\begin{aligned} &i = 0, 1, 2, \cdots, p, \\ &j = 0, 1, 2, \cdots, p, \end{aligned}\right. \end{aligned}$$

are all continuous. Then we can state the following theorem.

THEOREM 5.4. *The game described by* (5.1) *and* (5.2) *has a saddle point in chattering controls* (σ^*, τ^*) *of degree* $p+1$ *and* $q+1$ *respectively. If* Y *and* Z *are connected,* σ^* *and* τ^* *may be taken of degrees* p *and* q.

Proof. Consider the map $\Phi: Y \times [0,1] \to R^p$,

$$\Phi(u, t) = \begin{pmatrix} \varphi_1(u, t) \\ \vdots \\ \varphi_p(u, t) \end{pmatrix}.$$

Let $\Phi_t(Y) = \Phi(Y \times \{t\})$, $0 \leqq t \leqq 1$; then

$$\int_Y \Phi(u, t)\, d\sigma^*(t, u) \in \Gamma(\Phi_t(Y)).$$

Consider the map

$$\theta : \Delta^{p+1} \times Y^{p+1} \times [0, 1] \to R^p,$$

$$\theta(\alpha_1, \cdots, \alpha_{p+1}, u_1, \cdots, u_{p+1}, t) = \sum_{i=1}^{p+1} \alpha_i \Phi(u_i, t).$$

Then by Carathéodory's result,

$$\theta(\Delta^{p+1} \times Y^{p+1} \times \{t\}) = \Gamma(\Phi_t(Y))$$

and θ is continuous. Hence by the Filippov theorem we may determine measurable functions $\alpha_1(t), \cdots, \alpha_{p+1}(t)$, $v_1(t), \cdots, v_{p+1}(t)$ such that

$$\sum_{i=1}^{p+1} \alpha_i(t) \Phi(u_i(t), t) = \int_Y \Phi(u, t)\, d\sigma^*(t, u).$$

We show as before that the chattering control

$$\sigma_* = \sum_{i=1}^{p+1} \alpha_i(t) \delta_{v_i(t)}$$

has the same effect as σ^*. Let $\tau(t)$ be any relaxed control for the second player. Then the trajectory described by (σ_*, τ) is given by

$$\begin{aligned} \dot{x} &= A(t)x + \sum_{i=0}^{p} \sum_{j=0}^{q} a_{ij}(t) \int_Z \int_Y \varphi_i(u, t) \psi_j(v, t)\, d\sigma_*(t, u)\, d\tau(t, \sigma) \\ &= A(t)x + \sum_{i=0}^{p} \sum_{j=0}^{q} a_{ij}(t) \int_Z \int_Y \varphi_i(u, t) \psi_j(\sigma, t)\, d\sigma^*(t, u)\, d\tau(t, \sigma) \end{aligned}$$

and is therefore the same as the trajectory described by (σ^*, τ); a similar argument may be used on the payoff. We determine τ_* for the second player and the result follows. Once again if Y and Z are connected, we may reduce the $(p + 1)$-degree to p, as in Theorem 5.3.

A similar theorem may be stated in R^m, where for each coordinate the dynamical equation is "polynomial-like." We conclude by observing that the Berkovitz game (see § 2.1) may be analyzed by Theorem 5.4. Thus if

$$\dot{x} = (u - v)^2 = u^2 - 2uv + v^2,$$

we may take $\varphi_1(u, t) = u$, $\varphi_2(u, t) = u^2$, while $\psi_1(v, t) = v$, $\psi_2(v, t) = v^2$. As the payoff

$$p = \int_0^1 x(t)\, dt$$

does not depend on u and v, these are the only functions required. Thus $p = q = 2$, and Y and Z are connected. We may therefore expect a saddle point over chattering controls of order 2 (see Remarks 4.2).

Acknowledgment. We are indebted to the referees for pointing out reference [10] and the consequent simplification of Theorem 4.1.

REFERENCES

[1] L. D. BERKOVITZ, *A differential game with no pure strategy solution*, Annals of Mathematics Studies No. 52, Princeton University Press, Princeton, 1964, pp. 175–194.

[2] N. DUNFORD AND J. SCHWARTZ, *Linear Operators. I*, Interscience, New York, 1964.

[3] A. FRIEDMAN, *On the definition of differential games and the existence of Value and Saddle points*, J. Differential Equations, 7 (1970), pp. 69–91.

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

[5] E. J. MCSHANE, *Integration*, Princeton University Press, Princeton, 1944.

[6] ———, *Relaxed controls and variational problems*, this Journal, 5 (1967), pp. 438–485.

[7] G. OWEN, *Game Theory*, W. B. Saunders, Philadelphia, 1968.

[8] E. ROXIN, *Axiomatic approach in differential games*, J. Optimization Theory Appl., 3 (1969), pp. 153–163.

[9] L. SCHWARTZ, *Products tensoriels topologiques d'éspaces vectoriels topologiques. Espaces vectoriels topologiques nucléaires. Applications*, Sèminaire 1953/1954, Faculté des Sciences, Paris, 1954.

[10] M. SION, *On general minimax theorems*, Pacific J. Math., 8 (1958), pp. 171–176.

[11] E. R. SMOLJAKOV, *Differential games in mixed strategies*, Soviet Math. Dokl., 11 (1970), pp. 330–334.

[12] J. WARGA, *Relaxed variational problems*, J. Math. Anal. Appl., 4 (1962), pp. 111–128.

[13] ———, *Functions of relaxed controls*, this Journal, 5 (1967), pp. 628–641.

[14] L. C. YOUNG, *Generalized curves and the existence of an attained absolute minimum in the calculus of variations*, C.R. Soc. Sci. Lettres Varsovie, CL 111, 30 (1937), pp. 212–234.

[15] ———, *Lectures on the Calculus of Variations and Optimal Control Theory*, W. B. Saunders, Philadelphia, 1969.

SIAM J. CONTROL
Vol. 11, No. 1, February 1973

A DRIVABLE METHOD OF FEASIBLE DIRECTIONS*

GERARD G. L. MEYER†

Abstract. This paper presents a feasible directions algorithm for solving a general nonlinear programming problem. The algorithm is characterized by the fact that it is parametrized by functions. One may, through proper choice of the parametrizing maps, model the algorithm to optimize its behavior with respect to classes of problems.

Introduction. This paper presents a feasible directions algorithm for solving a general nonlinear programming problem. The algorithm is characterized by the fact that it is parametrized by functions. This allows for a great freedom in the application of the algorithm. One may, through proper choice of the parametrizing maps, model the algorithm so as to optimize its behavior with respect to a class of problems.

The paper's first part is devoted to preliminaries. The problem under consideration and the notations are defined. A necessary and sufficient condition of optimality for the problem is also given. The following part of the paper presents the drivable algorithm together with its proof of convergence. It is shown that the parametrizing maps need only to satisfy a simple assumption to ensure convergence of the algorithm. Finally, two examples of parametrizing maps are given and it is shown that the corresponding algorithms are the Zoutendijk's and Polak's method of feasible directions.

1. Preliminaries. The following nonlinear programming problem will be considered in this paper.

Problem 1. Given $m + 1$ continuously differentiable and convex maps $f^0(\cdot)$, $f^1(\cdot), \cdots, f^m(\cdot)$ from R^n into R^1, let T be the subset of R^n defined by $T = \{z | f^j(z) \leqq 0, j = 1, 2, \cdots, m\}$. Suppose that T is nonempty, compact and that for any z in T, the set of vectors $\{\nabla f^j(z) | f^j(z) = 0, j = 1, 2, \cdots, m\}$ is linearly independent. Find a point z^* in T, such that $f^0(z^*) \leqq f^0(z)$ for all z in T.

Notation 2.

(i) Let $\mathscr{I}_0$ be the set $\{1, 2, \cdots, m\}$, let $\mathscr{I}$ be the set $\mathscr{I}_0 \cup \{0\}$ and let $\mathscr{P}(\mathscr{I})$ be the family of all subsets of $\mathscr{I}$.

(ii) Let $\mathscr{F}$ be the family of all compact neighborhoods of the origin in R^n. A set in $\mathscr{F}$ is closed, bounded and contains an open set around the origin in R^n. An example of this type of set is the set $\{z \in R^n | |z^i| \leqq 1, i = 1, 2, \cdots, n\}$.

(iii) Let N be the positive integers, let R^1_+ be the positive real line and let $\bar{R}^1_+$ be the completed positive real line, i.e., $\bar{R}^1_+ = [0, \infty)$.

(iv) Given a sequence $\{z_i\}$ and a subset K of N, let $\{z_i\}_K$ be the subsequence defined by $\{z_i | i \in K\}$.

(v) Given a set J in $\mathscr{I}$, let $\bar{J}$ be its complement with respect to $\mathscr{I}$, i.e., let $\bar{J} = \{j \in \mathscr{I} | j \notin J\}$.

* Received by the editors January 4, 1972.

† Department of Electrical Engineering, The John Hopkins University, Baltimore, Maryland 21218.

DEFINITIONS 3. Let $J(\cdot,\cdot)$ be the map from $T \times \bar{R}^1_+$ into $\mathscr{P}(\mathscr{I})$ and $\phi(\cdot,\cdot,\cdot)$ be the map from $T \times \bar{R}^1_+ \times R^n$ into R^1 defined as follows:

$$J(z,\alpha) = \{j\in\mathscr{I}_0 | f^j(z) + 1/\alpha \geqq 0\} \cup \{0\},$$

$$\phi(z,\alpha,h) = \max_{j\in J(z,\alpha)} \langle\nabla f^j(z), h\rangle.$$

Necessary and sufficient conditions of optimality for Problem 1 are well known. One of their most convenient forms was given by Zoutendijk [10].

LEMMA 4. *A point z^* in T is a solution of Problem* 1 *if and only if*

$$\min_{h\in S} \phi(z^*, \infty, h) = 0$$

for any set S in $\mathscr{F}$.

The proofs of the two lemmas below are not difficult and have not been included.

LEMMA 5. *Given z in T, S in $\mathscr{F}$ and $\delta > 0$, there exist a neighborhood $N(z)$ of z and $k > 0$, both of which depend on z, S and δ, such that*

$$\phi(z',\alpha,h) \leqq \phi(z,\infty,h) + \delta,$$

for all z' in $N(z) \cap T$, for all $\alpha \geqq k$ and for all h in S.

LEMMA 6. *Let z in T, h in R^n, J in $\mathscr{P}(\mathscr{I})$ and $\delta > 0$ be such that $f^j(z) \leqq -\delta$ for all j in $\bar{J}$ and $\langle\nabla f^j(z), h\rangle \leqq -\delta$ for all j in J. Then there exist a neighborhood $N(z)$ of z, a neighborhood $N(h)$ of h, $\varepsilon_1 > 0$ and $\varepsilon_2 > 0$, all of which depend on z, h, J and δ, such that for all z' in $N(z) \cap T$, for all h' in $N(h)$ and for all j in $\mathscr{I}$,*

$$f^j(z' + \varepsilon_2 h') - \beta_j f^j(z') \leqq -\varepsilon_1,$$

with $\beta_j = 1$ if $j \in J$ and $\beta_j = 0$ otherwise.

2. A drivable method of feasible directions.

The algorithm proposed to solve Problem 1 uses a point y in T, a compact neighborhood S of the origin in R^n and two maps $\ell_1(\cdot)$ and $\ell_2(\cdot)$ from N into R^1_+.

ALGORITHM 7.

Step 0. Set $z_0 = y$, set $j_0 = 0$ and set $i = 0$.

Step 1. Compute h_i in S such that for all h in S,

$$\phi(z_i, \ell_1(j_i), h_i) \leqq \phi(z_i, \ell_1(j_i), h).$$

Step 2. If $\phi(z_i, \ell_1(j_i), h_i) \leqq -1/\ell_2(j_i)$, go to Step 3; else, set $z_{i+1} = z_i$, set $j_{i+1} = j_i + 1$, set $i = i + 1$ and go to Step 1.

Step 3. Compute λ_i in R^1 satisfying

$$\lambda_i = \max\{\lambda | f^k(z_i + \lambda h_i) \leqq 0, k = 1, 2, \cdots, m\}.$$

Step 4. Compute μ_i in $[0, \lambda_i]$ such that for all μ in $[0, \lambda_i]$,

$$f^0(z_i + \mu_i h_i) \leqq f^0(z_i + \mu h_i).$$

Step 5. Set $z_{i+1} = z_i + \mu_i h_i$, set $j_{i+1} = j_i$, set $i = i + 1$ and go to Step 1.

The assumption below ensures that Algorithm 7 can be used to solve Problem 1.

Hypothesis 8. Given any m in R^1_+ there exists a k in N depending on m such that $\ell_1(i) \geqq m$ and $\ell_2(i) \geqq m$ for all $i \geqq k$, i in N.

THEOREM 9. *Any sequence* $\{j_i\}$, *generated by Algorithm* 7 *when applied to Problem* 1, *has the following properties*:

(i) *The sequence* $\{j_i\}$ *is monotonically increasing.*

(ii) *The set* $M = \{i|\phi(z_i, \ell_1(j_i), h_i) > -1/\ell_2(j_i)\}$ *contains an infinite number of elements.*

(iii) *Given any m in N there exists n in N, depending on m, such that* $j_i \geqq m$, *for all* $i \geqq n$, *i in N.*

Proof. Let $\{z_i, h_i, j_i\}$ be a sequence generated by Algorithm 7 when applied to Problem 1. By construction, j_{i+1} is equal to either j_i or j_{i+1} and therefore $j_{i+1} \geqq j_i$ for all i in N.

Each set $J(z_i, \ell_1(j_i))$ is a subset of $\mathscr{I}$ which is finite. It follows that there exists an infinite subset L of N and a subset I of $\mathscr{I}$ such that $J(z_i, \ell_1(j_i)) = I$ for all i in L. The sets T and S are compact and therefore there exist an infinite subset K of L and two points z^* and h^* in T and S respectively, such that the subsequence $\{z_i\}_K$ converges to z^* and the subsequence $\{h_i\}_K$ converges to h^*.

Suppose that part (ii) of the theorem is false, i.e., suppose that the set

$$M = \{i|\phi(z_i, \ell_1(j_i), h_i) > -1/\ell_2(j_i)\}$$

is a finite subset of N. Then there exist p and m such that $j_i = m$ and $\phi(z_i, \ell_1(m), h_i) \leqq -1/\ell_2(m)$ for all $i \geqq p$. Let

$$\delta = \min\{1/\ell_1(m), 1/\ell_2(m)\}.$$

Then $f^k(z_i) \leqq -\delta$ for all k in $\bar{I}$ and $\langle \nabla f^k(z_i), h_i\rangle \leqq -\delta$ for all k in I, provided that $i \geqq p$.

The maps $f^k(\cdot)$ are continuously differentiable for k in $\mathscr{I}$ and therefore $f^k(z^*) \leqq -\delta$ for all k in $\bar{I}$ and $\langle \nabla f^k(z^*), h^*\rangle \leqq -\delta$ for all k in I. Lemma 6 implies that there exist $\tilde{p} \geqq p$, $\varepsilon_1 > 0$ and $\varepsilon_2 > 0$ so that $f^k(z_i + \varepsilon_2 h_i) - \beta_k f^k(z_i) \leqq -\varepsilon_1$ for all $i \geqq \tilde{p}$, i in K and for all k in $\mathscr{I}$. It follows immediately that ε_2 is in $[0, \lambda_i]$ for all $i \geqq \tilde{p}$, i in K. Step 4 of Algorithm 7 shows that

$$f^0(z_i + \mu_i h_i) \leqq f^0(z_i + \varepsilon_2 h_i) \leqq f^0(z_i) - \varepsilon_1$$

for all $i \geqq \tilde{p}$, i in K. The set K is infinite and by construction the sequence $\{f^0(z_i)\}$ is monotonically decreasing. It follows that $f^0(\cdot)$ is unbounded from below on T. This contradicts the assumptions that $f^0(\cdot)$ is continuous and that the set T is compact. Therefore the set M is an infinite subset of N.

In view of the determination of j_i in Step 2 of Algorithm 7, part (iii) is an immediate consequence of part (ii).

It can be noticed that Theorem 9 does not require that Hypothesis 8 be satisfied.

THEOREM 10. *If Hypothesis* 8 *is satisfied, any sequence* $\{z_i\}$, *generated by Algorithm* 7 *when applied to Problem* 1, *has the following properties*:

(i) *The sequence* $\{z_i\}$ *has at least one cluster point.*

(ii) *The sequence* $\{f^0(z_i)\}$ *is monotonically decreasing.*

(iii) *Any cluster point* z^* *of the sequence* $\{z_i\}$ *is a solution of Problem* 1.

Proof. Parts (i) and (ii) of the theorem are clear and only part (iii) will be proved.

Let $\{z_i\}$ be an infinite sequence constructed by Algorithm 7 when applied to Problem 1. Theorem 9 implies that $M = \{i|\phi(z_i, \ell_1(j_i), h_i) > -1/\ell_2(j_i)\}$ is an infinite subset of N. The sequence $\{z_i\}$ is in T which is compact and therefore there exist an infinite subset K of M and a point z^* in T, such that the subsequence $\{z_i\}_K$ converges to z^*.

Let $\delta > 0$ be given. Lemma 5 implies that there exists k_1 such that $\phi(z^*, \infty, h) \geqq \phi(z_i, \alpha, h) - \delta/2$ for all $\alpha \geqq k_1$, for all $i \geqq k_1$, i in K, and for all h in S. The maps $\ell_1(\cdot)$ and $\ell_2(\cdot)$ satisfy Hypothesis 8, and in view of part (iii) of Theorem 9 there exists k_2 such that $\ell_1(j_i) \geqq k_1$ and $\ell_2(j_i) \geqq 2/\delta$ for all $i \geqq k_2$. It follows that

$$\phi(z_i, \ell_1(j_i), h_i) \geqq -\delta/2$$

and

$$\min_{h \in S} \phi(z^*, \infty, h) \geqq \phi(z_i, \ell_1(j_i), h_i) - \delta/2$$

for all $i \geqq k$, i in K, with $k = \max(k_1, k_2)$. This immediately implies that $\min_{h \in S} \phi(z^*, \infty, h) \geqq -\delta$. The set S contains the origin of R^n and therefore

$$0 \geqq \min_{h \in S} \phi(z^*, \infty, h) \geqq -\delta \quad \text{for all } \delta > 0,$$

i.e.,

$$\min_{h \in S} \phi(z^*, \infty, h) = 0$$

and z^* is a solution to Problem 1.

The sequence $\{f^0(z_i)\}$ is monotonically decreasing and therefore if z^{**} is any cluster point of the sequence $\{z_i\}$, then $f^0(z^{**}) = f^0(z^*)$. It follows that any cluster point of a sequence $\{z_i\}$ generated by Algorithm 7 when applied to Problem 1 is a solution of Problem 1.

3. Classical methods of feasible directions. Algorithm 7 is parametrized by the two maps $\ell_1(\cdot)$ and $\ell_2(\cdot)$. These maps need only satisfy Hypothesis 8 to ensure that Algorithm 7 can be used to find solutions of Problem 1. There is therefore a great deal of freedom in the choice of these maps; thus the name drivable method of feasible directions given to Algorithm 7. One may use maps which not only satisfy Hypothesis 8 but also ensure that Algorithm 7 behaves in an efficient way when applied to a given class of problems.

The maps $\ell_1(\cdot)$ and $\ell_2(\cdot)$ may be chosen a priori or may be determined by Algorithm 7 itself. An example of each possibility is now presented.

DEFINITION 11. Given $\varepsilon > 0$, let $\ell_1(\cdot)$ and $\ell_2(\cdot)$ be the maps defined as follows:

$$\ell_1(j) = \ell_2(j) = \varepsilon(j + 1) \quad \text{for all } j \text{ in } N.$$

The maps $\ell_1(\cdot)$ and $\ell_2(\cdot)$ given by Definition 11 clearly satisfy Hypothesis 8. Therefore by employing these maps, Algorithm 7 can be used to find the solutions of Problem 1. One may notice that when the maps given in Definition 11 are used, Algorithm 7 is parametrized by a scalar.

DEFINITION 12. Given $\varepsilon > 0$ and $\beta > 1$, let $\ell_1(\cdot)$ and $\ell_2(\cdot)$ be the maps defined as follows:

(i) $\ell_1(0) = \varepsilon$.

(ii) For all $i \geqq 1$, i in N, let $\ell_1(j_i) = \varepsilon$ if $j_i = j_{i-1}$ and let $\ell_1(j_i) = \beta\ell_1(j_{i-1})$, otherwise.

(iii) $\ell_2(j_i) = \ell_1(j_i)$ for all i in N.

The maps $\ell_1(\cdot)$ and $\ell_2(\cdot)$ given by Definition 12 depend on Algorithm 7 and the problem under consideration. It is easy to show that the maps $\ell_1(\cdot)$ and $\ell_2(\cdot)$ generated by Algorithm 7 when applied to Problem 1 satisfy Hypothesis 8. Therefore, by utilizing these maps, Algorithm 7 can be used to find the solutions of Problem 1. One may notice that when the maps given in Definition 12 are used, Algorithm 7 is parametrized by two scalars.

There exist many ways of defining the maps $\ell_1(\cdot)$ and $\ell_2(\cdot)$ either explicitly as in Definition 11 or implicitly as in Definition 12. The two examples presented are interesting because they correspond to well-known algorithms. Algorithm 7 which employs the maps given by Definition 11 is called the *Zoutendijk's method of feasible directions.* Algorithm 7 which utilizes the maps given by Definition 12 is called the *Polak's method of feasible directions.*

4. Conclusion. The drivable method of feasible directions presented in this paper has a marked theoretical advantage over the classical methods of feasible directions. It is parametrized by two functions, $\ell_1(\cdot)$ and $\ell_2(\cdot)$, which need only satisfy Hypothesis 8 to ensure that the algorithm may be used to solve Problem 1. It follows that there is no limit on the complexity of the procedures used to obtain the maps $\ell_1(\cdot)$ and $\ell_2(\cdot)$. Although it is possible to use predetermined maps, it is the opinion of the author that maps determined implicitly, taking into account the behavior of the algorithm when solving a particular problem, are preferable. Definition 12 gives an example of such a map. In this case the determination of $\ell_1(\cdot)$ and $\ell_2(\cdot)$ during iteration i depends on the value of j_{i-1}. One may synthesize the maps $\ell_1(\cdot)$ and $\ell_2(\cdot)$ at iteration i by taking into account a greater quantity of information on the behavior of the algorithm at iteration $i - 1, i - 2, \cdots$ etc. For example, one may determine the maps $\ell_1(\cdot)$ and $\ell_2(\cdot)$ by taking into account the behavior of $f^0(z_i)$, $J(z_i, \ell_1(j_i))$, $\phi(z_i, \ell_1(j_i), h_i)$, j_i and h_i for $i - 1, i - 2, \cdots$ etc.

The computational efficiency of such schemes will most likely depend on the types of problems under consideration. These schemes may be better determined by direct experimentation on digital computers.

REFERENCES

[1] M. D. Canon, C. D. Cullum, Jr. and E. Polak, *Theory of Optimal Control and Mathematical Programming*, McGraw-Hill, New York, 1970.

[2] N. Dunford and J. T. Schwartz, *Linear Operators*, Interscience, New York, 1957.

[3] E. S. Levitin and B. T. Polyak, *Constrained minimization methods*, Zh. Vychisl. Mat. i Mat. Fiz., 6 (1966), pp. 787–823 = U.S.S.R. Comput. Math. and Math. Phys., 6 (1966), pp. 1–50.

[4] G. G. L. Meyer, *Abstract models for the synthesis of optimization algorithms*, Doctoral thesis, Department of Electrical Engineering, University of California, Berkeley, 1970.

[5] G. G. L. Meyer and E. Polak, *Abstract models for the synthesis of optimization algorithms*, Memo. ERL-286, Electronics Research Laboratory, College of Engineering, University of California, Berkeley, 1969.

[6] E. Polak, *On the convergence of optimization algorithms*, Rev. Française Informat. Recherche Opérationnelle, Série Rouge, 16 (1969), pp. 17–34.

[7] D. M. Topkis and A. Veinott, *On the convergence of some feasible directions algorithms for nonlinear programming*, this Journal, 5 (1967), pp. 286–379.

[8] W. I. Zangwill, *Convergence conditions for nonlinear programming algorithms*, Working Paper 196, Center for Research in Management Science, University of California, Berkeley, 1966.

[9] ———, *Nonlinear Programming: A Unified Approach*, Prentice-Hall, Englewood Cliffs, N.J., 1969.

[10] G. Zoutendijk, *Methods of Feasible Directions*, Elsevier, Amsterdam, 1960.

SIAM J. CONTROL
Vol. 11, No. 1, February 1973

MINIMUM-ENERGY TERMINAL STATE CONTROL OF FIRST ORDER LINEAR HYPERBOLIC SYSTEMS IN ONE SPATIAL VARIABLE USING THE METHOD OF CHARACTERISTICS*

TIMOTHY L. JOHNSON†

Abstract. The initial value problem for a hyperbolic system of (n) linear constant-coefficient first order partial differential equations is considered, on a finite time interval. Control over the evolution of the state variables is exercised by (n) forcing functions which are assumed to be distributed in space and time. These functions are to be determined subject to the *hard constraint* that the system state lie in a given target set at the terminal time, and to the *soft constraint* that a given scalar quadratic functional termed the "control energy" be minimized. For certain types of target sets, the method of characteristics is shown to be useful in determination of the optimal controls. Discontinuities in initial and terminal data are reflected in these controls.

1. Introduction. A first order linear constant-coefficient hyperbolic system of (n) equations in one spatial variable may be decoupled into (n) first order hyperbolic equations. The initial value problem for such a system is readily solved by applying the method of characteristics to the decoupled equations; (n) one-parameter families of ordinary differential equations are thus obtained. The system to be considered is nonhomogeneous, each equation being driven by an independent control function, distributed in space and time. These (n) control functions are to be determined subject to the *hard constraint* that the system state lie in a given set at the terminal time, and to the *soft constraint* that a given quadratic functional termed the "control energy" be minimized. Under the conditions described below, controllability, viz., the existence of one or more control functions satisfying the hard constraint, is guaranteed.

General theory for solving this type of optimization problem is available [3, pp. 272–347], [4], [5]. However, the optimal control will in general be specified by the solution of a two-point boundary value problem involving partial differential equations, which would in practice require a large amount of digital computation. Herein, the terminal target set and the cost functional are chosen in such a way that the optimal controls are easily computable. Furthermore, insight may be gained regarding the dependence of optimal controls upon the wave properties of the original system.

The method of characteristics and solutions of the system equations are reviewed in the preliminary section. In the following sections, optimal controls are derived for various combinations of target set and energy functional. Control to a linear integral manifold, a terminal state value at a single point, and a terminal state specified on a line segment are discussed. Generalizations and conclusions are examined in the final section.

* Received by the editors March 30, 1972.

† Department of Electrical Engineering, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139. This work was carried out at the Decision and Control Sciences Group of the M.I.T. Electronic Systems Laboratory and was supported in part by the National Science Foundation under Grant GK-25781.

2. Preliminaries. Systems of linear first order partial differential equations in one spatial variable are considered:

$$\partial \mathbf{x}(z,t)/\partial t = \mathbf{A}\partial \mathbf{x}(z,t)/\partial z + \mathbf{u}(z,t), \tag{2.1}$$

where the state $\mathbf{x}$ and the control $\mathbf{u}$ are n-vectors and the space-time domain of interest is $(z,t) \in D \subset (R \times [0,T])$, and $\mathbf{A} \in R^{n \times n}$. The initial and terminal states will be referred to as

$$\mathbf{x}(z,0) = \mathbf{x}_0(z), \tag{2.2}$$

$$\mathbf{x}(z,T) = \mathbf{x}_T(z), \tag{2.3}$$

where $\mathbf{x}_0(z) \in C^1(R)$ and $\mathbf{x}_T(z) \in C^1(Z_T)$ is specified only on a subset $Z_T = [z^-, z^+]$ of R. There are no other boundary conditions on $\mathbf{x}$ and no other explicit constraints on the control function $\mathbf{u}$. Initial data is actually only required on the domain of dependence of Z_T at time $t = 0$. For simplicity of exposition, boundary conditions will not be considered in this discussion; for well-posed initial boundary value problems, some of the methods below may be adapted by including control constraints involving the boundary data.

In order that the initial value problem (2.1)–(2.2) be well-posed, the system is assumed to be strictly hyperbolic, i.e., the constant matrix $\mathbf{A}$ is assumed to have real distinct roots and hence linearly independent eigenvectors:

$$\mathbf{VA} = \mathbf{\Lambda V}, \tag{2.4}$$

where $\mathbf{\Lambda} \doteq \operatorname{diag}(\lambda_1, \cdots, \lambda_n)$ and $\mathbf{V}$ is invertible.

By change of dependent variable, the above system may be decoupled into a set of (n) independent single-variable subsystems, each governed by a first order linear partial differential equation. Let

$$\mathbf{y}(z,t) = \mathbf{V}\mathbf{x}(z,t), \tag{2.5}$$

$$\mathbf{w}(z,t) = \mathbf{V}\mathbf{u}(z,t). \tag{2.6}$$

Then by (2.1), (2.4) derive the decoupled system

$$\partial \mathbf{y}(z,t)/\partial t = \mathbf{\Lambda}\partial y(z,t)/\partial z + \mathbf{w}(z,t) \tag{2.7}$$

subject to the transformed boundary conditions:

$$\mathbf{y}(z,0) = \mathbf{V}\mathbf{x}_0(z) \doteq \mathbf{y}_0(z), \tag{2.8}$$

$$\mathbf{y}(z,T) = \mathbf{V}\mathbf{x}_T(z) \doteq \mathbf{y}_T(z). \tag{2.9}$$

Consider the problem of solving the ith equation of (2.7)–(2.9):

$$\begin{aligned} \partial y_i(z,t)/\partial t &= \lambda_i \partial y_i(z,t)/\partial z + w_i(z,t), \\ y_i(z,0) &= y_{i0}(z), \\ y_i(z,T) &= y_{iT}(z). \end{aligned} \tag{2.10}$$

The solution may be found using the method of characteristics. Introduce a new set of (characteristic) coordinates

$$\varkappa_i(z,t) = (z - z^-) + \lambda_i(t - T), \tag{2.11}$$

$$\tau_i(z,t) = t - T. \tag{2.12}$$

Since this transformation is everywhere invertible, y_i and w_i may be considered as functions of $(\varkappa_i, \tau_i)$ rather than (z, t) via the association

$$\tilde{y}_i(\varkappa_i, \tau_i) = y_i(\varkappa_i(z,t), \tau_i(z,t)) = y_i(z,t) \tag{2.13}$$

and an ordinary differential equation for $\tilde{y}_i(\varkappa_i, \tau_i)$ in the characteristic coordinate system may be obtained via the chain rule:

$$\partial\tilde{y}_i/\partial t = \partial\tilde{y}_i/\partial\tau_i + \lambda_i\partial\tilde{y}_i/\partial\varkappa_i,$$
$$\partial\tilde{y}_i/\partial z = \partial\tilde{y}_i/\partial\varkappa_i.$$

Thus (2.10) is converted to a one-parameter family:

$$\begin{aligned} \partial\tilde{y}_i(\varkappa_i, \tau_i)/\partial\tau_i &= \tilde{w}_i(\varkappa_i, \tau_i), \\ \tilde{y}_i(\varkappa_i, -T) &= y_{i0}(\varkappa_i + z^- + \lambda_i T), \\ \tilde{y}_i(\varkappa_i, 0) &= y_{iT}(\varkappa_i + z^-), \end{aligned} \tag{2.14}$$

where $\tau_i \in [-T, 0]$ and $\varkappa_i \in [0, z^+ - z^-]$. If y_{i0} and y_{iT} are prespecified, a family of integral constraints is defined, namely,

$$c_i(z_i) = y_{iT}(\varkappa_i + z^-) - y_{i0}(\varkappa_i + z^- + \lambda_{iT}) = \int_{-T}^{0} \tilde{w}_i(\varkappa_i, \tau_i)\, d\tau_i. \tag{2.15}$$

This set of constraints is readily expressed in terms of the original variables (z, t) by defining the family of characteristic lines

$$\Gamma_i(z_0, t) = z_0 - \lambda_i t, \qquad t \in [0, T] \qquad z_0 \in [z^- + \lambda_i T, z^+ + \lambda_i T]. \tag{2.16}$$

Then (2.15) may be rewritten as a line integral along Γ_i:

$$y_{iT}(z_0 - \lambda_i T) = y_{i0}(z_0) + \int_0^T w_i(z_0 - \lambda_i\tau, \tau)\, d\tau. \tag{2.17}$$

These are the control constraints ordained by the specification of a fixed terminal state.

Provided $\mathbf{x}_0 \in C^1(R)$, $\mathbf{x}_T \in C^1(Z_T)$ and $\mathbf{u} \in C^1(D)$, a unique solution $\mathbf{x} \in C^1(D)$ will exist. These continuity conditions on the initial data and controls may be relaxed if one considers existence and uniqueness (in the appropriate sense) of weak solutions. The controls derived below are discontinuous on the boundaries of $\mathscr{D}_i \doteq \{(\varkappa_i, \tau_i) | \varkappa_i \in [0, z^+ - z^-], \tau_i \in [-T, 0]\}$, but may be arbitrarily closely approximated by controls $\mathbf{u} \in C^1(D)$, and should properly be interpreted as the limit of a sequence of such approximate controls, provided C^1 solutions are sought. The discontinuity of the optimal controls, it should be remarked, is such that the terminal solution $\mathbf{x}_T \in C^1(Z_T)$.

3. Control to a terminal linear variety. Define a set of admissible controls,

$$U = \{\mathbf{u} \in L_2^n(D) | \mathbf{x}(z, T) \in S\}, \tag{3.1}$$

where $\mathbf{x}(z, t)$ is the solution of the initial value problem (2.1)–(2.2), and

$$S = \left\{ \mathbf{x}_T \in L_2^n(Z_T) \,\middle|\, \int_{Z_T} \mathbf{L}\mathbf{x}_T(z)\, dz = \mathbf{f} \right\} \tag{3.2}$$

is a terminal linear variety specified by the parameters of $\mathbf{L} \in R^{m \times n}$ and $\mathbf{f} \in R^m$. The admissible control $\mathbf{u}^0 \in U$ is sought such that the control energy functional

$$J(\mathbf{u}) = \int_D \mathbf{u}'(z, t)\mathbf{R}\mathbf{u}(z, t)\, dz\, dt \tag{3.3}$$

takes its minimum value for $\mathbf{u} = \mathbf{u}^0$. In (3.3), $\mathbf{R}$ is assumed to be a positive definite $n \times n$ matrix; the domain D may be taken as the entire strip $R \times [0, T]$, although (2.17) shows that each control is constrained by (3.2) only on a subset of this domain (since Z_T is only a subset of R). Only the union of such subsets ($i = 1, 2, \cdots, n$) need actually be included in D, since $\mathbf{u}^0 = 0$ minimizes (3.3) absolutely in regions where $\mathbf{u}$ is unconstrained.

This control problem may be converted into a constrained minimum norm problem in $L_2^n(D)$ by expressing the admissible controls (3.1) explicitly and by converting (3.3) into a norm on this space. The original problem is then solved by applying the projection theorem [2] to the minimum norm problem.

An explicit expression for the control constraint imposed by (3.2) follows by integrating (2.17) over z_0:

$$\begin{aligned} \int_{z^- + \lambda_i T}^{z^+ + \lambda_i T} y_{iT}(z_0 - \lambda_i T)\, dz_0 &= \int_{z^- + \lambda_i T}^{z^+ + \lambda_i T} y_{i0}(z_0)\, dz_0 \\ &\quad + \int_{z^- + \lambda_i T}^{z^+ + \lambda_i T} \int_0^T w_i(z_0 - \lambda_i \tau, \tau)\, d\tau\, dz_0 . \end{aligned} \tag{3.4}$$

Defining D_i as the image of $\mathscr{D}_i$ in the original coordinate system (i.e., the parallelogram bounded by the lines $t = 0$, $t = T$, $\Gamma_i(z^-, t)$ and $\Gamma_i(z^+, t)$) and by introducing a characteristic function of D_i,

$$d_i(z, t) = \begin{cases} 1 & \text{if} \quad (z, t) \in D_i, \\ 0 & \text{if} \quad (z, t) \notin D_i, \end{cases} \tag{3.5}$$

equation (3.4) may be rewritten as

$$\int_{Z_T} y_{iT}(z)\, dz = y_{i0} + \int_D d_i(z, t) w_i(z, t)\, dz\, dt, \qquad i = 1, 2, \cdots, n, \tag{3.6}$$

where $y_{i0} = \int_{z^- + \lambda_i T}^{z^+ + \lambda_i T} y_{i0}(z)\, dz$. Augmenting equations (3.6) a more compact form is obtained:

$$\int_{Z_T} \mathbf{y}_T(z) = \mathbf{y}_0 + \int_D \mathbf{D}(z, t)\mathbf{w}(z, t)\, dz\, dt, \tag{3.7}$$

where $\mathbf{D}(z, t) = \operatorname{diag}[d_1(z, t), \cdots, d_n(z, t)]$ and $\mathbf{y}_0 \doteq \operatorname{col}[y_{10}, \cdots, y_{n0}]$. By left-multiplying (3.7) by $\mathbf{L}\mathbf{V}^{-1}$ and recalling (2.5)–(2.6), a set of (m) linear integral constraints on the (transformed) control $\mathbf{w}$ results:

$$\int_D \mathbf{L}\mathbf{V}^{-1}\mathbf{D}(z,t)\mathbf{w}(z,t)\,dz\,dt = \mathbf{L}\mathbf{V}^{-1}\mathbf{y}_0 - \mathbf{f}. \tag{3.8}$$

The control energy (3.3) takes the form of a norm on $L_2^n(D)$ if $\mathbf{R}$ is factored as $\mathbf{R} = \tilde{\mathbf{R}}'\tilde{\mathbf{R}}$ ($\tilde{\mathbf{R}}$ invertible), and another transformed control function

$$\tilde{\mathbf{w}} = \tilde{\mathbf{R}}\mathbf{u} = \tilde{\mathbf{R}}\mathbf{V}^{-1}\mathbf{w} \tag{3.9}$$

is introduced; then the energy becomes:

$$J(\tilde{\mathbf{w}}) = \int_D \tilde{\mathbf{w}}'\tilde{\mathbf{w}}\,dz\,dt. \tag{3.10}$$

With the natural inner product on $L_2^n(D)$, $J(\tilde{\mathbf{w}}) = \|\tilde{\mathbf{w}}\|^2$ and

$$U = \{\mathbf{u} \in L_2^n(D) | \mathbf{u} = \tilde{\mathbf{R}}^{-1}\tilde{\mathbf{w}};\ \langle \tilde{\mathbf{r}}_i, \tilde{\mathbf{w}}\rangle = \tilde{f}_i;\ i = 1, 2, \cdots, m\}, \tag{3.11}$$

where

$$\tilde{\mathbf{r}}_i(z,t) = [\tilde{\mathbf{R}}^{-1\prime}\mathbf{V}'\mathbf{D}(z,t)\mathbf{V}^{-1\prime}\mathbf{L}']_i \quad (i\text{th column}),$$

$$\tilde{f}_i = [\mathbf{L}\mathbf{V}^{-1}\mathbf{y}_0 - \mathbf{f}]_i.$$

Applying the projection theorem to the minimization of $\|\tilde{\mathbf{w}}\|^2$ subject to the constraints $\langle \tilde{\mathbf{r}}_i, \tilde{\mathbf{w}}\rangle = \tilde{f}_i$; $i = 1, 2, \cdots, m$, we obtain the following theorem.

THEOREM 3.1. *The control* $\mathbf{u} = \mathbf{u}^0$ *satisfying* (3.1) *and minimizing* (3.3) *exists if and only if the constraints* (3.11) *are consistent,* (*if so*) *is unique and is given by*

$$\mathbf{u}^0(z,t) = \tilde{\mathbf{R}}^{-1}\tilde{\mathbf{w}}^0(z,t) = \tilde{\mathbf{R}}^{-1}\left(\sum_{i=1}^{m} \alpha_i\tilde{\mathbf{r}}_i\right), \tag{3.12}$$

where $\{\alpha_i\}$ *are solutions of the normal equations*

$$\begin{bmatrix} \langle \tilde{\mathbf{r}}_1, \tilde{\mathbf{r}}_1\rangle \cdots \langle \tilde{\mathbf{r}}_m, \tilde{\mathbf{r}}_1\rangle \\ \vdots \\ \langle \tilde{\mathbf{r}}_1, \tilde{\mathbf{r}}_m\rangle \cdots \langle \tilde{\mathbf{r}}_m, \tilde{\mathbf{r}}_m\rangle \end{bmatrix} \begin{bmatrix} \alpha_1 \\ \vdots \\ \alpha_m \end{bmatrix} = \begin{bmatrix} \tilde{f}_1 \\ \vdots \\ \tilde{f}_m \end{bmatrix}. \tag{3.13}$$

In the general case, the ith optimal control function $u_i^0(z,t)$ is given by a linear combination of the characteristic functions $d_j(z,t)$, $j = 1, 2, \cdots, n$, defined by (3.5).

Example. For $\mathbf{L} = \mathbf{I}$ and $m = n$,

$$\langle \tilde{\mathbf{r}}_j, \tilde{\mathbf{r}}_i\rangle = \int_D \tilde{\mathbf{r}}_j'\tilde{\mathbf{r}}_i\,dz\,dt = \int_D [\mathbf{V}^{-1}\mathbf{D}(z,t)\mathbf{V}\mathbf{R}^{-1}\mathbf{V}'\mathbf{D}(z,t)\mathbf{V}'^{-1}]_{ji}\,dz\,dt$$

so

$$\boldsymbol{\alpha} = \left[\int_D (\mathbf{V}^{-1}\mathbf{D}(z,t)\mathbf{V}\mathbf{R}^{-1}\mathbf{V}'\mathbf{D}(z,t)\mathbf{V}'^{-1})\,dz\,dt\right]^{-1}\tilde{\mathbf{f}}.$$

4. Control to a terminal state (special cost functional). Define a set of admissible controls,

$$U = \{\mathbf{u} \in PC^n(D) | \mathbf{x}(z,T) \in S\}, \tag{4.1}$$

where $\mathbf{x}(z, t)$ is the solution of the initial value problem (2.1)–(2.2) with $\mathbf{x}_0 \in PC^n(R)$, and $PC^n(X)$ denotes the space of n-vector-valued piecewise continuous functions on X. The terminal state target set is

$$S = \{\mathbf{x}_T \in PC^n(Z_T) | \mathbf{x}_T = \mathbf{x}_T(z) \text{ a.e.}\}, \tag{4.2}$$

where $\mathbf{x}_T(z)$ is given in $PC^n(Z_T)$. An admissible (optimal) control $\mathbf{u}^0 \in U$ is sought such that the control energy functional (3.3) takes its minimum value for $\mathbf{u} = \mathbf{u}^0$. In (3.3), $\mathbf{R}$ is now assumed to have the special form

$$\mathbf{R} = \mathbf{V}'\mathbf{I}\mathbf{V}, \tag{4.3}$$

where $\mathbf{V}$ is a solution of the eigenvalue problem (2.4) and $\mathbf{I}$ is a positive diagonal matrix which may be taken as the identity matrix without loss of generality (the solution is independent of the weightings on the diagonal). In this case, the optimal control is given by the following.

THEOREM 4.1. *The control* $\mathbf{u} = \mathbf{u}^0$ *satisfying* (4.1) *and minimizing* (3.3) *exists, is unique* (*in* $PC^n(D)$) *and is given by*

$$\mathbf{u}^0(z, t) = \mathbf{V}^{-1}\mathbf{w}^0(z, t), \tag{4.4}$$

where the elements of $\mathbf{w}^0(z, t)$ *are given by*

$$w_i^0(z, t) = \begin{cases} 0 & \text{if } (z, t) \notin D_i, \\ (\lambda_i T)^{-1}[\mathbf{V}(\mathbf{x}_T(z + \lambda_i(T - t)) - \mathbf{x}_0(z + \lambda_i T))]_i & \text{if } (z, t) \in D_i \end{cases} \tag{4.5}$$

and D_i *is the parallelogram defined prior to* (3.5), *for* $i = 1, 2, \cdots, n$.

Proof. The control constraints (on $w_i(z, t)$) are given by (2.17). Using the factorization of (3.9)–(3.10) above, conclude that $\tilde{\mathbf{R}} = \mathbf{V}$ and $\tilde{\mathbf{w}} = \mathbf{w}$, and the cost may be written

$$J(\mathbf{w}) = \int_D \mathbf{w}'\mathbf{w}\, dz\, dt = \sum_{i=1}^{n} J_i(w_i), \tag{4.6}$$

where

$$J_i(w_i) = \int_D w_i^2(z, t)\, dz\, dt. \tag{4.7}$$

Now the constraints (2.17) are also decoupled, and hence $w_i^0(z, t)$ may be chosen to minimize only $J_i(w_i)$; thus the problem has been reduced to (n) single-variable optimization problems. Furthermore, the constraints (2.17) apply only in the regions D_i, and it is possible to take

$$w_i^0(z, t) \equiv 0, \qquad (z, t) \in (D - D_i), \tag{4.8}$$

giving the first part of (4.5); (4.7) is now equivalent to

$$J_i(w_i) = \int_{D_i} w_i^2(z, t)\, dz\, dt. \tag{4.9}$$

Alternatively, the characteristic functions (3.5) could be used as above. In the characteristic coordinates (2.11)–(2.12), the image ($\mathscr{D}_i$) of D_i is a rectangle; transforming to these coordinates and partitioning the $\mathscr{z}_i$-axis into N segments of length $\Delta\varkappa = (z^+ - z^-)/N$, the constraint equation (2.17) may be written

$$(4.10)\qquad \int_{-T}^{0} w_i(\varkappa_i, \tau_i)\, d\tau_i\varkappa = [\mathbf{V}(\mathbf{x}_T(z^- + \varkappa_i) - \mathbf{x}_0(z^- + \varkappa_i + \lambda_i T))]_i,$$

where the values $\varkappa_i = j\Delta\varkappa, j = 0, 1, \cdots, N$, are considered in particular, and (4.9) is approximated by

$$(4.11)\qquad J_{iN}(w_i) = (1/N)(z^+ - z^-) \sum_{j=0}^{N} \int_{-T}^{0} w_i^2(j\Delta\varkappa, \tau_i)\, d\tau_i.$$

Piecewise continuity of initial and final data guarantees piecewise continuity of the right-hand side of (4.10), and hence

$$(4.12)\qquad \lim_{N\to\infty} J_{iN}(w_i) = J_i(w_i).$$

But $J_{iN}(w_i)$ is minimized by independently minimizing

$$(4.13)\qquad J_{ij}(w_i) = \int_{-T}^{0} w_i^2(j\Delta\varkappa, \tau_i)\, d\tau_i, \qquad j = 0, 1, \cdots, N,$$

subject to

$$(4.14)\qquad \int_{-T}^{0} w_i(j\Delta\varkappa, \tau_i)\, d\tau_i = [\mathbf{V}(x_T(j\Delta\varkappa + z^-) - \mathbf{x}_0(j\Delta\varkappa + z^- + \lambda_i T))]_i.$$

In the limit as $N \to \infty$, the minimum value of (4.9) is seen to be achieved by using the control generated by the family of solutions to the optimization problem (4.13)–(4.14) with $j\Delta\varkappa$ becoming a continuous variable in the limit. This problem is readily re-expressed in the original coordinates (z, t) and solved in the inner product space $L_2([0, T])$ with a z as a parameter, viz., minimize $\|w_i(z + \lambda_i(T - t), t)\|^2$ subject to the constraint

$$(4.15)\qquad \langle w_i(z + \lambda_i(T - t), t), 1\rangle = [\mathbf{V}(\mathbf{x}_T(z) - \mathbf{x}_0(z + \lambda_i T))]_i$$

for each $z \in Z_T$. The solution of this simple problem yields the second part of (4.5). The resulting control function is evidently piecewise continuous on D, the discontinuities occurring on the characteristic lines emanating from (a) the endpoints of the terminal domain, Z_T, (b) points of discontinuity in the initial data, and (c) points of discontinuity in the terminal data.

A special case of interest occurs when Z_T consists of the neighborhood of a single point, say z_T, in the terminal domain. In this event, nonzero controls are obtained on the neighborhoods of the (n) characteristic lines intersecting at z_T. Due to the special form of the cost functional (4.3), such "pointwise" controls may be linearly superimposed to generate the optimal controls for a discrete set of terminal points, or for an entire line segment. For more general cost functionals, however, this convenient additivity property fails to hold because the energy functional will not decouple along the (n) families of characteristics; this case is treated in the following section.

As a physical interpretation of pointwise control, consider the case where the equations (2.1) represent the dynamics of a long flexible beam and it is desired to control the terminal state of the beam at a certain point. The above solution indicates that control may be achieved by delivering specially contrived impulsive forces and moments to the beam at a finite number of points. Such impulses then travel at the wave speeds of the system and meet with the desired result at the terminal point.

5. Control to a terminal state (general cost functional). The control problem is as stated in § 4, with the assumption (4.3) replaced by the less restrictive condition that $\mathbf{R}$ be symmetric positive definite. Using the transformations (2.8)–(2.9), the control set U is still described by (2.17), as a set of linear integral constraints on each transformed control function w_i over the characteristic domain D_i. In this case, however, the energy functional (3.3) cannot be written as a sum of independent functionals on these domains; the controls $w_i, i = 1, 2, \cdots, n$, must now be jointly determined.

The following proposition outlines an iterative algorithm utilizing the characteristic domains for solving the general problem. Alternative algorithms could be derived by (i) discretizing the original problem and solving the finite-dimensional approximation to the distributed optimization problem (finite element method), or (ii) by discretizing the necessary conditions for optimality derived for the continuous problem. The proposed algorithm differs from (i) and (ii) in that the special role of the characteristic domains D_i is preserved (optimal controls will in general be discontinuous on the boundaries of D_i, $i = 1, 2, \cdots, n$), and convergence may be more rapid. It is shown that the proposed algorithm yields reduction in control energy at each iteration, although this result is short of a convergence proof.

PROPOSITION 5.1. The following algorithm is proposed for the solution of (2.1)–(2.2), (4.1)–(4.2), and (3.3) where $\mathbf{R}$ is symmetric positive definite:

(i) Select an initial (transformed) control

$$\mathbf{w}^0 = (w_1^0, w_2^0, \cdots, w_n^0)'.$$

(ii) Find w_1^1 to minimize

$$J_1^0(w) = J(w, w_2^0, \cdots, w_n^0). \tag{5.1}$$

(iii) Similarly, find w_j^1 to minimize (for $j = 2, 3, \cdots, n$)

$$J_j^0(w) = J(w_1^1, \cdots, w_{j-1}^1, w, w_{j+1}^0, \cdots, w_n^0). \tag{5.2}$$

(iv) Repeat (ii), (iii) for $J_j^1(w)$, $J_j^2(w)$, etc. until convergence is obtained.

THEOREM 5.1. *Regarding the algorithm given in Proposition* 5.1,

(a) *The initial control* $\mathbf{w}^0$ *may be chosen as a solution of Theorem* 4.1.

(b) *The solution of step* (ii) (*or* (iii)) *exists and is unique.*

(c) *If* $w_2^0, \cdots, w_n^0$ *are optimal, then the control* $\mathbf{w} = (w_1^1, w_2^0, \cdots, w_n^0)$ *is optimal, where* w_1^1 *is the solution of step* (iii).

(d)

$$J^{k+1}(\mathbf{w}^{k+1}) \leqq J^k(\mathbf{w}^k). \tag{5.3}$$

Proof. (a) This point is included mainly as a remark that the rate of convergence in general depends on the initial control estimate $\mathbf{w}^0$. By choosing the scale of the individual eigenvectors of $\mathbf{A}$ (i.e., the columns of $\mathbf{V}$), it is possible to find a solution of (2.4) which minimizes $\|\mathbf{R} - \mathbf{V}'\mathbf{V}\|^2$; using this value of $\mathbf{V}$ in

Theorem 4.1, a reasonable initial estimate, $\mathbf{w}^0$ (given by (4.5)), of the optimal control is obtained.

(b) The solution of the optimization problem of (ii) is derived below; the solution of (iii) follows by rearranging indices. Recalling (3.9), (3.10), the control energy may be rewritten as

$$J(\tilde{\mathbf{w}}) = \int_D \left(\sum_{i=1}^n \tilde{w}_i^2 \right) dz\,dt. \tag{5.4}$$

From (3.9),

$$\tilde{w}_i = \sum_{j=1}^n \alpha_{ij} w_j, \qquad \alpha_{ij} \doteq (\tilde{\mathbf{R}}\mathbf{V}^{-1})_{ij}. \tag{5.5}$$

Referring to (5.1),

$$\begin{aligned} J_1^0(w) &= \int_D \sum_{i=1}^n \left(\alpha_{i1} w + \sum_{j=2}^n a_{ij} w_j^0 \right)^2 dz\,dt \\ &= \int_D \left\{ \left(\sum_{i=1}^n \alpha_{i1} \right)^2 w^2 + 2 \left(\sum_{i=1}^n \alpha_{i1} \sum_{j=2}^n \alpha_{ij} w_j^0 \right) w \right. \\ &\qquad \left. + \sum_{i=1}^n \left(\sum_{j=2}^n \alpha_{ij} w_j^0 \right)^2 \right\} dz\,dt. \end{aligned} \tag{5.6}$$

Define for $i = 1, 2, \cdots, n$,

$$\beta_i = \alpha_{i1}, \tag{5.7}$$

$$\hat{w}_i^0 = \sum_{j=2}^n \alpha_{ij} w_j^0, \tag{5.8}$$

$$\hat{w}^0 = \left(\sum_{i=1}^n \beta_i^2 \right)^{-1} \left(\sum_{i=1}^n \beta_i \hat{w}_i^0 \right). \tag{5.9}$$

Using (5.7)–(5.9) in (5.6) and completing the square, obtain

$$J_1^0(w) = \left(\sum_{i=1}^n \beta_i^2 \right) \int_D (w + \hat{w}^0)^2 \,dz\,dt + \hat{J}_1^0, \tag{5.10}$$

where $\hat{J}_1^0$ is a constant depending on the given functions $w_2^0, \cdots, w_n^0$:

$$\hat{J}_1^0 = \int_D \left[\sum_{i=1}^n \hat{w}_i^{0\,2} - \left(\sum_{i=1}^n \beta_i^2 \right) \hat{w}^{0\,2} \right] dz\,dt. \tag{5.11}$$

The scale factor $(\sum_{i=1}^n \beta_i^2)$ and the additive constant $\hat{J}_1^0$ may be ignored in the minimization of (5.10). Furthermore, w is constrained by (2.17) for $i = 1$ only on D_1 and hence outside this region its optimal value is given explicitly as

$$w_1^1(z, t) = -\hat{w}^0(z, t), \qquad (z, t) \notin D_1. \tag{5.12}$$

The integration of (5.10) may then be taken over D_1; transforming to characteristic coordinates, an argument precisely analogous to (4.10)–(4.15) demonstrates that the control energy decouples along the characteristics (piecewise continuity of the initial and terminal data being used at this point), and hence an equivalent optimization problem (on $L_2(0, T)$) is:

$$(5.13)\qquad \min_{w\in W} \|w(z + \lambda_1(T - t), t) + \hat{w}^0(z + \lambda_1(T - t), t)\|^2,$$

where

$$(5.14)\qquad W = \{w \in PC(D) | \langle w(z + \lambda_1(T - t), t), 1\rangle = c_1(z)\}$$

and $c_1(z) \doteq [\mathbf{V}(\mathbf{x}_T(z) - \mathbf{x}_0(z + \lambda_1 T))]_1$ for all $z \in [z^- + \lambda_1 T, z^+ + \lambda_1 T]$. The solution of (5.13)–(5.14) is

$$(5.15)\qquad w_1^1(z, t) = w_1^*(z, t) - \hat{w}^0(z, t), \qquad (z, t) \in D_1,$$

where $w_1^*(z, t)$ is the unique element of minimum norm in the linear subspace

$$W^* = \{w^* \in PC(D) | \langle w^*(z + \lambda_1(T - t), t), 1\rangle$$
$$= c_1(z) + \langle \hat{w}^0(z + \lambda_1(T - t), t), 1\rangle\}.$$

The solution of this problem is an application of the projection theorem [6, pp. 46–77]; in summary

$$(5.16)\qquad w_1^1(z, t) = \begin{cases} -\hat{w}^0(z, t) & \text{if } (z, t) \in D - D_1, \\ -\hat{w}^0(z, t) + (\lambda_1 T)^{-1}\left(c_1(z + \lambda_1 T)\right. \\ \qquad \left. + \int_0^T \hat{w}^0(z + \lambda_1(T - t), t)dt\right) & \text{if } (z, t) \in D_1, \end{cases}$$

where c_1 and $\hat{w}^0$ are given by (5.14) and (5.9).

(c) This follows directly from (b). The same result applies at any step of the algorithm, viz., if all but one of the control functions of $\mathbf{w}^k$ is optimal, optimality is achieved at the next step. Thus in special cases, convergence to the optimal control may be achieved in a finite number of iterations.

(d) This result also follows directly from (b) and from the construction of the algorithm. Unfortunately, (5.3) does not preclude a "limit cycle" behavior of the algorithm, e.g., when equality is achieved but $\mathbf{w}^{k+1} \neq \mathbf{w}^k$.

An important distinction between the results of Theorems 4.1 and 5.1 is that in the former case, the optimal controls are zero in regions where they are unconstrained by the terminal state objective, whereas in the latter case the optimal controls are nontrivial throughout the entire domain D, although discontinuities still occur on the boundaries of the characteristic domains D_i.

6. Generalizations and conclusions. The above examples, although not devoid of interest in practical applications of control, are indeed "concocted" special cases in which the method of characteristics can be used to advantage in deriving explicit expressions for optimal controls. Certain (well-posed) initial boundary value problems may be incorporated in the above framework, but most other generalizations are not so successful. Decoupling fails in general when the coefficients of $\mathbf{A}$ are functions of z or t. Controllability fails when $\mathbf{u}$ in (2.1) is replaced by $\mathbf{Bu}$ where $\mathbf{B}$ has rank less than n, unless the terminal set is restricted. The discussion of the preceding sections indicates that optimal controls for most minimum-energy state transfer problems cannot be derived explicitly by solving ordinary differential equations arising from the method of characteristics.

Nevertheless, the explicit solutions derived above do share certain common properties with optimal controls for more general problems involving first order linear hyperbolic systems. The most striking feature of these controls is that they are discontinuous along characteristics emanating (backward in time) from the boundary of the terminal surface (endpoints of the line segment Z_T, in this case). It is in fact reasonable to conjecture that this same property holds for the cases described in the preceding paragraph. Such discontinuities could lead to numerical difficulties in computer algorithms devised for the solution of these more general problems. In addition to this insight, the limitations of decoupling and use of the method of characteristics have been explored in detail, revealing a conflict which cannot always be resolved between decoupling of the equations and decoupling of the control energy functional.

REFERENCES

[1] P. R. GARABEDIAN, *Partial Differential Equations*, John Wiley, New York, 1967.

[2] I. B. RHODES, *Dynamic Systems, Control and Optimization*, to be published.

[3] J. L. LIONS, *Optimal Control of Systems Governed by Partial Differential Equations*, Springer-Verlag, New York, 1971.

[4] C. BARDOS, *Problèmes aux limites pour les équations aux derivées partielles du premier ordre à coefficients réels; théorèmes d'approximation; application à l'équation de transport*, Doctoral thesis, University of Paris, June, 1969.

[5] D. L. LUKES AND D. L. RUSSELL, *The quadratic criterion for distributed systems*, this Journal, 7 (1969), pp. 101–121.

[6] D. G. LUENBERGER, *Optimization by Vector Space Methods*, John Wiley, New York, 1969.

SIAM J. CONTROL
Vol. 11, No. 1, February 1973

NECESSARY CONDITIONS FOR OPTIMIZATION PROBLEMS WITH HYPERBOLIC PARTIAL DIFFERENTIAL EQUATIONS*

M. B. SURYANARAYANA†

Abstract. In this paper we consider a Mayer-type optimization problem with side conditions described by nonlinear hyperbolic partial differential equations and Darboux-type initial data in a rectangle G. The state variables are assumed to belong to a Sobolev space in G while the controls are assumed to be only measurable in G with values in a fixed closed subset of a Euclidean space. Using the results of an earlier paper, we first state existence and uniqueness theorems for the solutions of the original Darboux problem describing the state equations, as well as existence theorems for the solutions of the conjugate problem (yielding the multipliers). We derive then an increment formula for the cost functional and corresponding estimates. Finally we prove, under rather general hypotheses, a necessary condition of the Pontryagin-type for the minimum of the cost functional. Illustrative examples are provided.

1. Introduction.

In the present paper we consider a system of nonlinear hyperbolic partial differential equations (state equations) of the form

$$\begin{aligned} &\partial z^i/\partial x\partial y = f_i(x, y, z, z_x, z_y, v), \quad (x, y) \in G, \\ &i = 1, \cdots, n, \quad z(x, y) = (z^1, \cdots, z^n), \quad v(x, y) = (v^1, \cdots, v^m), \\ &G = [a \leqq x \leqq a + h, b \leqq y \leqq b + k], \end{aligned} \tag{1.1}$$

with Darboux-type boundary conditions

$$\begin{aligned} z(x, b) &= \varphi(x), \quad a \leqq x \leqq a + h, \\ z(a, y) &= \psi(y), \quad b \leqq y \leqq b + k, \end{aligned} \tag{1.2}$$

and with constraints

$$v(x, y) \in U. \tag{1.3}$$

We are concerned with the minimum of a functional of the form

$$I[z, v] = \sum_{i=1}^{n} A_i z^i(a + h, b + k). \tag{1.4}$$

Here $\varphi(x) = (\varphi^1, \cdots, \varphi^n)$, $a \leqq x \leqq a + h$, and $\psi(y) = (\psi^1, \cdots, \psi^n)$, $b \leqq y \leqq b + k$, are given absolutely continuous functions (AC) in the respective intervals, with $\varphi(a) = \psi(b)$. The control space U above is a given fixed set of the u-space E_m. The constants A_i, $i = 1, \cdots, n$, are given.

The minimum of the functional $I[z, v]$ is sought in suitable classes Ω of pairs $z(x, y) = (z^1, \cdots, z^n)$, $v(x, y) = (v^1, \cdots, v^m)$, $(x, y) \in G$, satisfying (1.1), (1.2), (1.3), the functions z^i belonging to a Sobolev space $W_p^1(G)$ on G, $1 \leqq p \leqq +\infty$, and continuous on G, and the functions v^j being measurable on G. In the present paper we give a Pontryagin-type necessary condition for the minimum.

* Received by the editors May 18, 1971.

† Department of Mathematics, Eastern Michigan University, Ypsilanti, Michigan 48197. This research was supported in part by the United States Air Force Office of Scientific Research under Project AFOSR-69-1662 at The University of Michigan.

First, in § 3 we obtain an existence and uniqueness statement for the solution $z(x, y) = (z^1, \cdots, z^n)$, $(x, y) \in G$, $z \in (W_p^1(G))^n$, of the Darboux problem (1.1)–(1.2) (the original problem) for a given p, $1 \leqq p \leqq +\infty$, for given φ, ψ, and for a given measurable function $v(x, y) = (v^1, \cdots, v^m)$, $(x, y) \in G$. We derive this existence and uniqueness statement from our previous paper [8] on multidimensional integral equations of the Volterra type.

The optimization problem (1.1)–(1.4) can be written in the form proposed by Cesari [2] with state equations of the Dieudonné–Rashevsky type, the Hamiltonian function H then containing $2n$ multipliers λ_i, μ_i, $i = 1, \cdots, n$. As shown in [2], these $2n$ multipliers are expected to satisfy a suitable system of linear partial differential equations and corresponding boundary conditions (the conjugate problem).

In § 4 we formulate the conjugate problem pertinent to the optimization problem (1.1)–(1.4), and for the first time we prove, in the present situation, an existence theorem for the solutions λ_i, μ_i of the conjugate problem. In other words, we prove, under hypotheses, that there are multipliers λ_i, μ_i, $i = 1, \cdots, n$, in $L_\infty(G)$, satisfying in a suitable sense the partial differential equations and boundary conditions pertaining to the conjugate problem of problem (1.1)–(1.4). As in § 3, again we derive the existence statement from our previous paper [8] on multidimensional integral equations of the Volterra type.

In § 5 we give a new proof of the increment formula of [2], under a set of hypotheses different from those in [2]. In § 6 we derive as in [2] the Pontryagin-type necessary condition for the optimization problem (1.1)–(1.4) with the existence of suitable multipliers actually proved.

In § 7 we make a number of remarks on the results obtained, particularly in relation to the previous papers by Cesari [2] and A. I. Egorov [3]. In particular, we show that the present necessary condition yields—under strong smoothness hypotheses—the necessary condition previously proved by A. I. Egorov [3]. On the other hand we show (§ 7, Example 3) that these smoothness hypotheses under which Egorov's condition has been proved are not known a priori, while our necessary condition holds.

2. Notations. If $(X, \|\cdot\|)$ denotes any normed linear space, then X^n, $n \geqq 1$, denotes the Cartesian product of X with itself, n times; for $x = (x^1, \cdots, x^n) \in X^n$, we define $\|x\| = \sum_{i=1}^n \|x^i\|$. If $x \in E^n$, n-dimensional Euclidean space, then we take $\|x\| \equiv |x| = \sum_{i=1}^n |x^i|$. We shall denote by Γ, a rather arbitrary family of measurable control functions. Precisely, let Γ be any set of measurable functions; $v: G \to U$, $v = (v^1, \cdots, v^m)$, with the following property:

(*) For every function $v \in \Gamma$, any point $u \in U$, and any closed subset $S \subset G$, the control function v_ε, defined by $v_\varepsilon = v$ in $G - S$, and $v_\varepsilon = u$ in S, belongs to Γ. Thus, every constant function $v: G \to \{u\}$, $u \in U$, belongs to Γ.

For functions $\varphi \in L_p(G)$, $1 \leqq p \leqq +\infty$, we denote by $\|\varphi\|$ or $\|\varphi\|_p$ the usual L_p-norm; in particular, $\|\varphi\|_\infty = \operatorname{ess\,sup} |\varphi|$. For functions in a Sobolov space $[W_p^1(G)]^n$ in G, say, $z(x, y) = (z^1, \cdots, z^n)$, we shall denote by $z_x = (z_x^1, \cdots, z_x^n)$ and $z_y = (z_y^1, \cdots, z_y^n)$, the usual generalized first order partial derivatives of z, and we take $\|z\|_W = \|z\|_{W_p^1(G)} = \|z\|_p + \|z_x\|_p + \|z_y\|_p$.

3. The original problem. We shall need the following hypotheses:

(H_1). The functions $\varphi(x) = (\varphi^1, \cdots, \varphi^n)$ and $\psi(y) = (\psi^1, \cdots, \psi^n)$ are defined and absolutely continuous on $[a, a+h]$ and $[b, b+k]$, respectively. The derivatives φ_x and ψ_y which exist almost everywhere belong to $L_p([a, a+h])$ and $L_p([b, b+k])$, respectively, for some p, $1 \leqq p \leqq +\infty$. Furthermore, $\varphi(a) = \psi(b)$.

(H_2). The function $f = f(x, y, z_1, z_2, z_3, u) = (f_1, \cdots, f_n)$ is defined on $G \times E^{3n} \times U$, and each f_i is continuous in u and measurable in (x, y) for fixed $(z_1, z_2, z_3) \in E^{3n}$.

(H_3). For each $v \in \Gamma$ the function $s_0(x, y) = f(x, y, 0, 0, 0, v(x, y))$ belongs to $L_p(G)$ with p as in (H_1).

(H_4). The functions f_i are differentiable as functions of (z_1, z_2, z_3) and the derivatives $\partial f_i/\partial z_1^j$, $\partial f_i/\partial z_2^j$, $\partial f_i/\partial z_3^j$, $i, j = 1, \cdots, n$, are continuous in (z_1, z_2, z_3) for fixed (x, y, u). Clearly they are measurable in (x, y).

(H_5). There are functions $K_{jr}(x, y, u)$, $j = 1, 2, 3$, $r = 1, \cdots, n$, such that for all $(x, y, u) \in G \times U$ and $(z_1, z_2, z_3) \in E^{3n}$ we have

$$|\partial f_i/\partial z_j^r(x, y, z_1, z_2, z_3, u)| \leqq K_{jr}(x, y, u) \tag{3.1}$$

and such that $K_{jr}(x, y, v(x, y)) \in L_\infty(G)$ for $v \in \Gamma$, $j = 1, 2, 3$, $i, r = 1, \cdots, n$.

We state below an existence theorem (Theorem 3.1) for the solution z of the Darboux problem (1.1)–(1.2) and a theorem (Theorem 3.2) concerning their behavior. We refer to [8] (or [7]) for proofs of these and other statements. Theorem 3.1 provides norm estimates on the solution z as an element of $W_p^1(G)$, along with pointwise estimates on z, z_x, and z_y. Theorem 3.2 shows the dependence of the solution on the data.

THEOREM 3.1. *Let $v \in \Gamma$ be given. If* (H_1)–(H_5) *hold, then there exists a unique $z \in (W_p^1(G))^n$ (with p as in* (H_1)*), continuous on G, satisfying* (1.2), *for which the generalized partial derivatives z_x, z_y, z_{xy} exist and satisfy* (1.1) *a.e. in G. Furthermore, there are constants B_1 and B_2 depending only on h, k, p and on $K = \{\|K_{ij}(x, y, v(x, y))\|_\infty; i = 1, 2, 3; j = 1, \cdots, n\}$ such that*

$$\|z\|_W \leqq B_1[k^{1/p}(\|\varphi_x\|_p + \tfrac{1}{2}\|\varphi\|_p) + h^{1/p}(\|\psi_y\|_p + \tfrac{1}{2}\|\psi\|_p) + (h + k)\|s_0(x, y)\|_p], \tag{3.2}$$

$$\begin{aligned} |z(x, y)| \leqq \frac{1}{2}\Big[& \|\varphi\|_c + \|\psi\|_c + BB_2(h + k) \\ & + e^{K(h+k)}\Big(B + \iint_G s_0(\alpha, \beta)\, d\alpha\, d\beta\Big)\Big], \end{aligned} \tag{3.3}$$

$$|z_x(x, y)| \leqq \theta_1(x) + BB_2, \quad |z_y(x, y)| \leqq \theta_2(y) + BB_2, \tag{3.4}$$

where

$$B = \|\varphi_x\|_p h^{1/q} + \|\psi_y\|_p k^{1/q} + Khk(\|\varphi\|_c + \|\psi\|_c) + \iint_G s_0(\alpha, \beta)$$

and

$$\theta_1(x) = e^{Kk}\Big[|\varphi_x(x)| + Kk|\varphi(x)| + \int_b^{b+k} s_0(x, \beta)\, d\beta\Big],$$

$$\theta_2(y) = e^{Kh}\Big[|\psi_y(y)| + Kh|\psi(y)| + \int_a^{a+h} s_0(\alpha, y)\, d\alpha\Big].$$

The existence of the solution and the norm estimate (3.2) follow from [8, Appendix, Theorem 5, A.2] while pointwise estimates are a consequence of the absolute continuity (in the sense of Tonelli) of the solution z, and a repeated application of Gronwall's lemma (see [8, Appendix, (A.10)] or [7]).

THEOREM 3.2. *For* $i = 1, 2$, *let* z_i *denote the solution of* (1.1)–(1.2) *corresponding to the data* (φ_i, ψ_i) *satisfying* (H$_1$) *and control function* v_i *in* Γ. *Let* $z = z_1 - z_2$, $\varphi = \varphi_1 - \varphi_2$, $\psi = \psi_1 - \psi_2$ *and* $s(x, y) = |f(x, y, z_1, z_{1x}, z_{1y}, v_1) - f(x, y, z_1, z_{1x}, z_{1y}, v_2)|$. *With this notation, the above inequalities* (3.2), (3.3), (3.4) *again hold with* s_0 *replaced by* s *and no further changes.* (See [8, Appendix, (A.10)] or [7].)

Furthermore, if $\varphi_1 = \varphi_2$, $\psi_1 = \psi_2$ *and* $v_1 = v_2$ *outside a square* $S = [\bar{x} - \delta, \bar{x} + \delta] \times [\bar{y} - \delta, \bar{y} + \delta] \subset G$, *then the pointwise estimates become*

$$
\begin{aligned}
|z(x, y)| &\equiv |z_1(x, y) - z_2(x, y)| \leqq B_1 \iint_S s(\alpha, \beta)\, d\alpha\, d\beta, \\
|z_x(x, y)| &\leqq e^{Kk} \int_{\bar{y}-\delta}^{\bar{y}+\delta} s(x, \beta)\, d\beta + B_2 \iint_S s(\alpha, \beta)\, d\alpha\, d\beta, \\
|z_y(x, y)| &\leqq e^{Kb} \int_{\bar{x}-\delta}^{\bar{x}+\delta} s(\alpha, y)\, d\alpha + B_2 \iint_S s(\alpha, \beta)\, d\alpha\, d\beta,
\end{aligned} \tag{3.5}
$$

where B_1 *and* B_2 *depend only on* h, k, *and on* $K = \max\{\|K_{ij}(x, y, v_r(x, y))\|_\infty, i = 1, 2, 3; j = 1, \cdots, n; r = 1, 2\}$.

We shall need in the sequel these particular pointwise estimates.

Remark 1. In view of the uniqueness of the solution z of the Darboux problem (1.1)–(1.2) for any given element $v \in \Gamma$, we shall denote the functional $I[z, u]$ of (1.4) simply by $I[v]$, or $I : \Gamma \to E_1$.

Remark 2. By introducing the notation $z_1 = z$; $z_2 = z_x$; $z_3 = z_y$, the Darboux problem (1.1)–(1.2) can be written in the equivalent Dieudonné–Rashevsky form:

$$
\begin{aligned}
&z_{1x} = z_2, \qquad z_{3x} = f(x, y, z_1, z_2, z_3, v), \qquad z_{1y} = z_3, \\
&z_{2y} = f(x, y, z_1, z_2, z_3, v)
\end{aligned} \tag{3.6}
$$

with boundary conditions,

$$
\begin{aligned}
&z_1(x, b) = \varphi(x), \qquad z_2(x, b) = \varphi_x(x), \qquad z_1(a, y) = \psi(y), \\
&z_3(a, y) = \psi_y(y).
\end{aligned} \tag{3.7}
$$

It is to be noted that even though the system (3.6) seems overdetermined (four equations in three unknowns), it is not actually so, since the second and the fourth equations are equivalent.

4. The conjugate problem. Cesari [2] has proved Pontryagin-type necessary conditions for problems of optimization with state equations in the Dieudonné–Rashevsky form $z_{ix} = f_i(x, y, z, v)$, $z_{iy} = g_i(x, y, z, v)$, $z = (z_1, \cdots, z_n)$, $i = 1, 2, \cdots, n$; and taking as Hamiltonian the expression $H = \lambda_1 f_1 + \cdots + \lambda_n f_n + \mu_1 g_1 + \cdots + \mu_n g_n$. By assuming the z_i, λ_i, μ_i to be in suitable Sobolev spaces, Cesari [2] showed that the multipliers λ_i, μ_i should satisfy the "conjugate problem," i.e., partial differential equations of the form $\lambda_{ix} + \mu_{iy} = -\partial H/\partial z_i$, $i = 1, 2, \cdots, n$, along with boundary conditions, which are complementary to those for $z_1, \cdots, z_n$

and in relation to the cost functional under consideration. In view of Remark 2 of § 3, we use here the same Hamiltonian with the remark that since z_{2x} and z_{3y} do not appear in (3.6) we take $\lambda_2 = \mu_3 = 0$, and the Hamiltonian reduces to

$$H = \lambda_1 z_2 + \lambda_3 f + \mu_1 z_3 + \mu_2 f, \tag{4.1}$$

where $\lambda_i = (\lambda_i^1, \cdots, \lambda_i^n)$, $\mu_i = (\mu_i^1, \cdots, \mu_i^n)$ and the products are inner products in E_n. By taking the cost functional I in (1.4) in the equivalent form (Cesari [2])

$$I = \frac{1}{2}\int_a^{a+h} A z_2(x, b+k)\,dx + \frac{1}{2}\int_b^{b+k} A z_3(a+h, y)\,dy, \tag{4.2}$$

where $A = (A_1, \cdots, A_n)$, the conjugate problem becomes

$$\begin{aligned} \lambda_{1x}^i + \mu_{1y}^i &= -\sum_{j=1}^n (\lambda_3^j + \mu_2^j)\frac{\partial f_j}{\partial z_1^i}, \\ \mu_{2y}^i &= -\lambda_1^i - \sum_{j=1}^n (\lambda_3^j + \mu_2^j)\frac{\partial f_j}{\partial z_2^i}, \\ \lambda_{3x}^i &= -\mu_1^i - \sum_{j=1}^n (\lambda_3^j + \mu_2^j)\frac{\partial f_j}{\partial z_3^i}, \qquad i = 1, 2, \cdots, n; \end{aligned} \tag{4.3}$$

$$\begin{aligned} \lambda_1(a+h, y) &= \mu_1(x, b+k) = 0, \\ \mu_2(x, b+k) &= \lambda_3(a+h, y) = A/2. \end{aligned} \tag{4.4}$$

In the present paper, we show first that the conjugate problem (4.3), (4.4) is equivalent to a system of two-dimensional Volterra-type linear integral equations of the type we have studied in a previous paper [8]. The results obtained there will enable us to prove in this paper the existence of multipliers λ_i, μ_i as solutions of the conjugate problem in a suitable class of functions, in $L_\infty(G)$ (not a Sobolev space).

In order to obtain the equivalent system of integral equations, we treat λ_{1x} as arbitrary and formally integrate both sides of (4.3) as follows (in conjunction with boundary conditions in (4.4)):

$$\begin{aligned} \lambda_1^i(x, y) &= \int_{a+h}^x \lambda_{1x}^i(\alpha, y)\,d\alpha, \\ \mu_1^i(x, y) &= -\int_{b+k}^y \lambda_{1x}^i(x, \beta) - \int_{b+k}^y \left(w \cdot \frac{\partial f}{\partial z_1^i}\right)(x, \beta)\,d\beta, \\ \lambda_3^i(x, y) &= \tfrac{1}{2}A_i + \int_{a+h}^x \int_{b+k}^y \lambda_{1x}^i(\alpha, \beta)\,d\alpha\,d\beta \\ &\quad + \int_{a+h}^x \int_{b+k}^y \left(w \cdot \frac{\partial f}{\partial z_1^i}\right)(\alpha, \beta)\,d\alpha\,d\beta - \int_{a+h}^x \left(w \cdot \frac{\partial f}{\partial z_3^i}\right)(\alpha, y)\,d\alpha, \\ \mu_2^i(x, y) &= \tfrac{1}{2}A_i - \int_{b+k}^y \int_{a+h}^x \lambda_{1x}^i(\alpha, \beta)\,d\alpha\,d\beta - \int_{b+k}^y \left(w \cdot \frac{\partial f}{\partial z_2^i}\right)(x, \beta)\,d\beta, \end{aligned} \tag{4.5}$$

where w stands for $\lambda_3 + \mu_2$. It is clear that w satisfies the integral equation $w = Tw$,

where

$$(4.6)\qquad (Tw)^i(x,y) = A_i + \int_x^{a+h}\int_y^{b+k} w\cdot\frac{\partial f}{\partial z_1^i} + \int_y^{b+k} w\cdot\frac{\partial f}{\partial z_2^i} + \int_x^{a+h} w\cdot\frac{\partial f}{\partial z_3^i}.$$

THEOREM 4.1. *If* (H_2), (H_4), *and* (H_5) *hold, if* $v \in \Gamma$ *and* z *is the corresponding solution of the Darboux problem* (1.1), *then there exist infinitely many sets of solutions* $\lambda_1, \lambda_3, \mu_1, \mu_2$ *in* $[L_\infty(G)]^n$, *with* $\lambda_1(x,y), \lambda_3(x,y), (x,y) \in G$, *AC with respect to* x *for almost all* y, $\mu_1(x,y), \mu_2(x,y), (x,y) \in G$, *AC with respect to* y *for almost all* x, λ_1, λ_3, μ_1, μ_2 *satisfying the boundary conditions* (4.4), *and having generalized partial derivatives* $\lambda_{1x}, \lambda_{3x}, \mu_{1y}, \mu_{2y}$.

Proof. As a consequence of [8, § 5, Thm. 3], it is seen that there is a unique $w \in [L_\infty(G)]^n$ with $w = Tw$, where T is defined by (4.6). The conclusion of the theorem follows now by defining the functions $\lambda_1, \mu_1, \lambda_3, \mu_2$ as in (4.5), in terms of the unique solution w of (4.6), and an arbitrarily chosen $[L_\infty(G)]^n$-function λ_{1x}.

Remarks. (a) Since $w = \lambda_3 + \mu_2$ is uniquely determined as the fixed point of T, for different choices of λ_{1x}, we still get the same $\lambda_3 + \mu_2$.

(b) The solutions of (4.5) need not belong to a Sobolev class since, for example, $\partial f/\partial z_3^i$ and hence λ_3^i as given in (4.5) need not possess derivatives with respect to y. (See § 7, Example 3.)

(c) If f does not depend on (z_1, z_2, z_3), then by choosing $\lambda_{1x} = 0$, it is seen that a possible set of multipliers is given by the constant functions $\lambda_1 = 0 = \mu_1$, $\lambda_3 = \mu_2 = A/2$. If $\partial f/\partial z_r^i$, $r = 1, 2, 3$, are continuous in (x, y), as is the case when they depend on z only, then the multipliers can be chosen to be continuous. Finally, if f is linear in (z_1, z_2, z_3) with coefficients analytic in (x, y), then the multipliers can be chosen to be analytic in (x, y) (see [8]).

5. The increment formula and an error estimate. Let v_0 and v_ε be any two elements of Γ, the set of control functions. Let z and z_ε be the solutions of (1.1)–(1.2) corresponding to v_0 and v_ε, respectively. Let $(\lambda, \mu) = (\lambda_1, \lambda_3, \mu_1, \mu_2)$ and $(\lambda_\varepsilon, \mu_\varepsilon) = (\lambda_{1\varepsilon}, \lambda_{3\varepsilon}, \mu_{1\varepsilon}, \mu_{2\varepsilon})$ be solutions of (4.3)–(4.4) corresponding to (v_0, z) and $(v_\varepsilon, z_\varepsilon)$ respectively, $(\lambda_1, \lambda_3, \mu_1, \mu_2) \in [L_\infty(G)]^{4n}$ as in Theorem 4.1.

In the sequel, when there is no confusion, the symbol $H(u)$ stands for the expression

$$(5.1)\qquad \begin{aligned} H(u) &= H(x, y, z(x,y), z_x(x,y), z_y(x,y), u, \lambda(x,y), \mu(x,y)) \\ &= H(x, y, z_1(x,y), z_2(x,y), z_3(x,y), u, \lambda(x,y), \mu(x,y)), \end{aligned}$$

where z, λ, μ are related to v_0 and u denotes a point of U. Also, for the sake of simplicity, we shall denote by $\dot{z}$ the expression

$$\dot{z} = (z(x,y), z_x(x,y), z_y(x,y)) = (z_1(x,y), z_2(x,y), z_3(x,y)).$$

In any case, we have $z_1 = z, z_2 = z_x, z_3 = z_y$.

In order to obtain a necessary condition of the Pontryagin type we express the increment $I[v_\varepsilon] - I[v_0]$ in terms of the integral of $H(v_\varepsilon(x,y))$ over G.

To this end, let us observe that, by simple calculations involving integration by parts of the expression

$$\int_a^{a+h}\int_b^{b+k} [\lambda_1(z_{1\varepsilon x} - z_{1x}) + \mu_1(z_{1\varepsilon y} - z_{1y}) + \lambda_3(z_{3\varepsilon x} - z_{3x}) + \mu_2(z_{2\varepsilon y} - z_{2y})]\,dx\,dy$$

and the boundary conditions (1.2) and (4.5), we obtain

$$I[v_\varepsilon] - I[v_0] = \eta + \int\int_G [H(v_\varepsilon(x, y)) - H(v_0(x, y))]\,dx\,dy,$$

where

$$\eta = \sum_{j=1}^{3} \int\int_G \left(\frac{\partial H}{\partial z_j}(x, y, \dot{z}_\theta, v_\varepsilon, \lambda, \mu) - \frac{\partial H}{\partial z_j}(x, y, \dot{z}, v_0, \lambda, \mu)\right)(z_{j\varepsilon} - z_j)\,dx\,dy$$

and

$$z_{j\theta}(x, y) = z_j(x, y) + \theta(x, y)[z_{j\varepsilon}(x, y) - z_j(x, y)], \qquad 0 \leqq \theta(x, y) \leqq 1.$$

For details we refer to Cesari [2], or the author [7].

Error estimate. It is clear that if f_i (in the state equations (1.1)) are linear, i.e., of the form $Az + Bz_x + Cz_y + D(x, y, u)$ where A, B, C are matrix-valued functions on G, then η reduces to zero. For the nonlinear case, we shall now obtain an estimate on η, and for this purpose, we need the following hypotheses:

(H_6). There exists a function $M(x, y, u)$, $(x, y, u) \in G \times U$, with $M(x, y, v(x, y)) \in L_4(G)$ for any $v \in \Gamma$, such that, for $(x, y, z_1, z_2, z_3, u) \in G \times E^{3n} \times U$ and $1 \leqq p < +\infty$, we have $|f(x, y, z_1, z_2, z_3, u)| \leqq M(x, y, u) + B_3[|z_1| + |z_2| + |z_3|]^{p/4}$ for some constant $B_3 \geqq 0$. For $p = +\infty$, we require $|f| \leqq M(x, y, u) + \phi(|z_1| + |z_2| + |z_3|)$ for some function $\phi(\xi) \geqq 0$, $0 \leqq \xi < +\infty$ with $\phi(\xi) \leqq K\xi$ for some K.

(H_7). There exist functions $K'_{ij}(x, y, u)$, $i = 1, 2, 3$; $j = 1, \cdots, n$, such that $K'_{ij}(x, y, v(x, y)) \in L_\infty(G)$ for any $v \in \Gamma$, and such that for $(x, y, u) \in G \times U$ and $z, \bar{z} \in E^{3n}$, $i = 1, 2, 3$; $j = 1, \cdots, n$, we have

$$(5.2) \qquad \left|\frac{\partial f}{\partial z_i^j}(x, y, z_1, z_2, z_3, u) - \frac{\partial f}{\partial z_i^j}(x, y, \bar{z}_1, \bar{z}_2, \bar{z}_3, u)\right| \leqq K'_{ij}(x, y, u) \sum_{s=1}^{3} |z_s - \bar{z}_s|.$$

Remark. As before, let $s(x, y)$ denote $|f(x, y, \dot{z}(x, y), v_1(x, y)) - f(x, y, \dot{z}(x, y), v_2(x, y))|$, where $\dot{z} = (z, z_x, z_y)$ and z is the solution of (1.1), (1.2) corresponding to v_1 and $v_1, v_2 \in \Gamma$. Then it is seen from (H_3), (H_4) and (H_5) that $s \in L_p(G)$; (p as in (H_1)). Indeed,

$$|f(x, y, \dot{z}(x, y), v_1(x, y))| \leqq K|\dot{z}(x, y)| + |f(x, y, 0, v_1(x, y))|,$$

where $K = \max\{\|K_{jr}(x, y, v_1(x, y))\|_\infty : j = 1, 2, 3;\ r = 1, \cdots, n\}$; (see ($H_5$)). Since $|\dot{z}| \in L_p$ and $|f(x, y, 0, v_1(x, y))| \in L_p$ (by (H_3)), it follows that $s \in L_p(G)$.

The assumption (H_6) is made only to guarantee that in addition, $s \in L_4(G)$. The same conclusion can be made under the following hypothesis:

(H'_6): There is a function $M(x, y)$ defined on G such that (i) $M(x, y)v(x, y) \in L_4(G)$ for every $v \in \Gamma$, and (ii) for $(x, y, z_1, z_2, z_3) \in G \times E^{3n}$ and $u_1, u_2 \in U$, we have

$$|f(x, y, z_1, z_2, z_3, u_1) - f(x, y, z_1, z_2, z_3, u_2)| \leqq M(x, y)|u_1 - u_2|.$$

Indeed,

$$s(x, y) = |f(x, y, \dot{z}(x, y), v_1(x, y)) - f(x, y, \dot{z}(x, y), v_2(x, y))|$$
$$\leqq M(x, y)|v_1(x, y) - v_2(x, y)| \leqq M|v_1| + M|v_2|$$

and $s \in L_4(G)$.

Further, if (H_6') and (H_7) hold with $M(x, y) \in L_\infty(G)$ and $\Gamma \subset [L_\infty(G)]^m$, then statement (3.5) in Theorem 3.2 can be replaced by

$$(5.3) \qquad |z(x, y)| + |z_x(x, y)| + |z_y(x, y)| \leqq B_4 \delta \|v_1 - v_2\|_\infty, \quad (x, y) \in G,$$

where the constant B_4 depends only on $\|M\|_\infty$, h, k and on all K_{ij}, K'_{ij}.

We shall denote below by u an arbitrary fixed point $u \in U$. Let v_0 be an element of Γ, let $z(x, y)$, $(x, y) \in G$, be the corresponding solution of the Darboux problem (1.1)–(1.2), and let $\dot{z}$ denote the $3n$-vector function $\dot{z}(x, y) = (z, z_x, z_y) = (z_1, z_2, z_3)$ as above. Let $(\bar{x}, \bar{y})$ be an interior point of G. Let $\delta_0 > 0$ be the minimum distance of $(\bar{x}, \bar{y})$ from the boundary of G. Let K/n denote the maximum of the $12\,n$ numbers $\|K_{ij}(x, y, u)\|_\infty$, $\|K'_{ij}(x, y, u)\|_\infty$, $\|K_{ij}(x, y, v_0(x, y))\|_\infty$, $\|K'_{ij}(x, y, v_0(x, y))\|_\infty$, $i = 1, 2, 3$; $j = 1, \cdots, n$; where the K_{ij} are as in (H_5) and the K'_{ij} as in (H_7).

By (H_6) the function

$$s(x, y; u) = |f(x, y, \dot{z}(x, y), u) - f(x, y, \dot{z}(x, y), v_0(x, y))|$$

belongs to $L_4(G)$. Thus, given $\xi > 0$ there is a $\delta > 0$ such that

$$\int\int_C |s(x, y; u)|^2 \, dxdy \leqq \xi^2 \quad \text{and} \quad \int\int_C |s(x, y; u)|^4 \, dxdy \leqq \xi^2$$

for every measurable set $C \subset G$ of measure $\leqq 4\delta^2$. We may well assume $0 \leqq \delta \leqq \delta_0$.

Let S_δ denote the square $[\bar{x} - \delta \leqq x \leqq \bar{x} + \delta, \bar{y} - \delta \leqq y \leqq \bar{y} + \delta]$ and let C be any closed subset of S_δ. Let v_ε be the function defined by $v_\varepsilon = v_0$ in $G - C$, and $v_\varepsilon = u$ in C. Then the function

$$s(x, y) = |f(x, y, \dot{z}(x, y), v_\varepsilon(x, y)) - f(x, y, \dot{z}(x, y), v_0(x, y))|$$

is zero outside C and equals $s(x, y; u)$ in C. Noting that $v_\varepsilon \in \Gamma$, we denote by z_ε the solution of problem (1.1)–(1.2) relative to v_ε, and as usual we write $\dot{z}_\varepsilon = (z_\varepsilon, z_{\varepsilon x}, z_{\varepsilon y}) = (z_{1\varepsilon}, z_{2\varepsilon}, z_{3\varepsilon})$. Inequalities (3.5) yield in this case

$$(5.4) \qquad \begin{aligned} |z_{j\varepsilon}(x, y) - z_j(x, y)| \leqq B\bigg[\int\int_{S_\delta} s(\alpha, \beta)\, d\alpha\, d\beta &+ \int_{\bar{y}-\delta}^{\bar{y}+\delta} s(x, \beta)\, d\beta \\ &+ \int_{\bar{x}-\delta}^{\bar{x}+\delta} s(\alpha, y)\, d\alpha \bigg] \end{aligned}$$

for all $(x, y) \in G, j = 1, 2, 3$, independently of the particular closed set $C \subset S_\delta$, and where B depends only on h, k and K. This inequality and the fact that $s \in L_4(G)$ under (H_6)(or (H_6')) can now be used to estimate η.

The integrand in the expression for η can be written as

$$\left[\sum_{i=1}^{n}\left|\lambda_3^i+\mu_2^i\right|\left|\frac{\partial f_i}{\partial z_j}(x,y,\dot z_\theta,v_\varepsilon)-\frac{\partial f_i}{\partial z_j}(x,y,\dot z,v_\varepsilon)\right|\right.$$
$$\left.+\sum_{i=1}^{n}\left|\lambda_3^i+\mu_2^i\right|\left|\frac{\partial f_i}{\partial z_j}(x,y,\dot z,v_\varepsilon)-\frac{\partial f_i}{\partial z_j}(x,y,\dot z,v_0)\right|\right]\cdot(z_{j\varepsilon}-z_j)$$

which yields, using (H_7),

$$|\eta|\leqq\|\lambda_3+\mu_2\|_\infty\cdot K\cdot\int\int_G\sum_{i=1}^{3}\sum_{j=1}^{3}|z_{i\theta}-z_i||z_{j\varepsilon}-z_j|(x,y)\,dx\,dy$$
$$(5.5)\qquad +\|\lambda_3+\mu_2\|_\infty\int\int_G\sum_{j=1}^{3}\left|z_{j\varepsilon}-z_j\right|\left|\frac{\partial f}{\partial z_j}(x,y,\dot z,v_\varepsilon)\right.$$
$$\left.-\frac{\partial f}{\partial z_j}(x,y,\dot z,v_0)\right|dx\,dy.$$

But, since the integrand in the last term in (5.5) is zero outside S_δ, we have $|\eta|\leqq\|\lambda_3+\mu_2\|_\infty\cdot K\cdot(\eta_1+2\eta_2)$, where

$$\eta_1=\int\int_G\sum_{i,j=1}^{3}|z_{i\theta}-z_i||z_{j\varepsilon}-z_j|(x,y)\,dx\,dy,$$
$$\eta_2=\int\int_{S_\delta}\sum_{j=1}^{3}|z_{j\varepsilon}-z_j|(x,y)\,dx\,dy.$$

To obtain an estimate on η, we observe that $|z_{i\theta}-z_i|\leqq|z_{i\varepsilon}-z_i|$ and then, by (5.4), we get

$$|\eta_1|\leqq\int\int_G\left(\sum_{i=1}^{3}|z_{i\varepsilon}-z_i|(x,y)\right)^2dx\,dy$$
$$\leqq 36B^2\int\int_G\left[\int\int_{S_\delta}s(\alpha,\beta)\,d\alpha\,d\beta+\int_{\bar y-\delta}^{\bar y+\delta}s(x,\beta)\,d\beta+\int_{\bar x-\delta}^{\bar x+\delta}s(\alpha,y)\,d\alpha\right]^2dx\,dy.$$

Using Hölder's inequality and the fact that $s\in L_4(G)$ under (H_6), it is seen that

$$(5.6)\qquad |\eta_1|\leqq M_2\delta^2\left[\int\int_{S_\delta}s^2(\alpha,\beta)\,d\alpha\,d\beta+\left(\int\int_{S_\delta}s^4(\alpha,\beta)\,d\alpha\,d\beta\right)^{1/2}\right]$$

for some positive constant M_2. Similarly, using (5.4) and Hölders' inequality, we get

$$|\eta_2|\leqq\int\int_{S_\delta}3B\left[\int\int_{S_\delta}s(\alpha,\beta)\,d\alpha\,d\beta+\int_{\bar y-\delta}^{\bar y+\delta}s(x,\beta)\,d\beta+\int_{\bar x-\delta}^{\bar x+\delta}s(\alpha,y)\,d\alpha\right]dx\,dy$$
$$(5.7)\qquad\leqq 3B(4\delta^2+2\delta+2\delta)\int\int_{S_\delta}s(\alpha,\beta)\,d\alpha\,d\beta$$
$$\leqq 6B(4\delta+4)\delta^2\left(\int\int_{S_\delta}s^2(\alpha,\beta)\,d\alpha\,d\beta\right)^{1/2}.$$

Using (5.6) and (5.7), the inequality (5.5) can now be written as

$$|\eta| \leqq M\delta^2\left[\int\int_{S_\delta} s^2(\alpha,\beta)\,d\alpha\,d\beta + \left(\int\int_{S_\delta} s^2(\alpha,\beta)\,d\alpha\,d\beta\right)^{1/2} + \left(\int\int_{S_\delta} s^4(\alpha,\beta)\,d\alpha\,d\beta\right)^{1/2}\right], \tag{5.8}$$

where M is a constant depending only on K, B, and $\|\lambda_3 + \mu_2\|_\infty$.

Given $\varepsilon > 0$, let us now choose a positive number ζ with $0 < \zeta^2 \leqq \zeta < \varepsilon/6M$. Let $\delta > 0$ be chosen as before with $\int\int_{S_\delta} s^2 \leqq \zeta^2$ and $\int\int_{S_\delta} s^4 \leqq \zeta^2$. Then $|\eta| \leqq M\delta^2(\zeta^2 + \zeta + \zeta) \leqq M\delta^2 \cdot 3 \cdot (\varepsilon/6M) = \varepsilon\delta^2/2$. In conclusion, if $v_0 \in \Gamma$, $u \in U$, and $\varepsilon > 0$ are given, then there exists a $\delta > 0$ such that for a function $v_\varepsilon = v_0$ outside S_δ and $v_\varepsilon = u$ in a closed subset of S_δ,

$$I(v_\varepsilon) - I(v_0) = \eta + \int\int_G [H(v_\varepsilon(x, y)) - H(v_0(x, y))]\,dx\,dy \tag{5.9}$$

with $|\eta| \leqq \delta^2\,\varepsilon/2$.

6. A necessary condition for optimality. In this section, we shall state and prove a necessary condition for optimality, analogous to the one-dimensional Pontryagin's necessary condition. We need the concept of "minimum condition" for the class of problems under consideration, and this is made precise in the following definition.

DEFINITION 6.1. Let $v_0 \in \Gamma$. Then v_0 is said to satisfy the *minimum condition* if there is a set $B \subset G$ with meas B = meas G such that for $(x, y) \in B$, we have $H(v_0(x, y)) \leqq H(u)$ for all $u \in U$.

We recall that $H(u)$ stands for $H(x, y, z(x, y), z_x(x, y), z_y(x, y), u, \lambda(x, y), \mu(x, y))$, where z and $(\lambda, \mu) = (\lambda_1, \lambda_3, \mu_1, \mu_2)$ satisfy (1.1), (1.2) and (4.3), (4.4) respectively. The following hypothesis is needed in the proof of the necessary condition:

(H_8). The functions $f_i = f_i(x, y, z_1, z_2, z_3, u)$, $i = 1, \cdots, n$, are continuous on $G \times E^{3n} \times U$.

Remark. In the proof of the necessary condition, the inequality (3.1) of (H_5) is needed only for $(z_1, z_2, z_3) = (z(x, y), z_x(x, y), z_y(x, y))$, where z is an optimal trajectory. Further, the hypothesis (H_6) can be replaced by (H_6').

THEOREM 6.1 (Pontryagain-type necessary condition). *Let $v_0 \in \Gamma$ be optimal for I; i.e., $I(v_0) \leqq I(v)$ for all $v \in \Gamma$. Let conditions* (H_1)–(H_8) *hold. Then there exists a unique function $z \in [W_p^1(G)]^n$ satisfying the Darboux problem* (1.1)–(1.2) *and ∞-many sets of multipliers $(\lambda_1, \lambda_3, \mu_1, \mu_2) \in (L_\infty(G))^{4n}$ satisfying* (4.4)–(4.5) *with v replaced by v_0. With this z and any of these sets of multipliers, the optimal control v_0 necessarily satisfies the minimum condition.*

Proof. The existence of z and of $\lambda_1, \lambda_3, \mu_1, \mu_2$ under hypotheses (H_1)–(H_5) has been shown in § 3 and § 4. Before proving the necessary condition, let us note that throughout this proof, $z_1 = z$, $z_2 = z_x$, $z_3 = z_y$, $(\lambda, \mu) = (\lambda_1, \lambda_3, \mu_1, \mu_2)$ have the same meaning and they correspond to v_0. Further, as in § 5, $H(u) = H(x, y, u) = H(x, y, z(x, y), z_x(x, y), z_y(x, y), u, \lambda(x, y), \mu(x, y))$.

For each natural number n, let C_n be a closed subset of G such that (i) $\operatorname{meas}(C_n) > (1 - n^{-1}) \operatorname{meas} G$, and (ii) on C_n the functions v_0, $\dot{z} = (z_1, z_2, z_3)$,

$(\lambda, \mu) = (\lambda_1, \lambda_3, \mu_1, \mu_2)$ are all continuous. Let C'_n be the set of all points of density of C_n so that meas $C'_n =$ meas C_n and the functions v_0, z, λ, μ are continuous on C'_n with respect to itself. Now, for any $u \in U$, let $R(x, y; u) = H(x, y, v_0(x, y)) - H(x, y, u)$. Then, this function is continuous on C'_n for each n. Let $B =$ (interior of G) $\cap$ $(\bigcup_{n=1}^{\infty} C'_n)$. Then meas $B \geqq$ meas $C'_n > (1 - n^{-1})$ meas G for all n and hence meas $B \geqq$ meas G. Further, since $B \subset G$, it follows that meas $B =$ meas G. We shall prove that B is the required set, i.e., for $(x, y) \in B$, we have $H(x, y, v_0(x, y)) \leqq H(x, y, u)$ for all $u \in U$. Let (x_0, y_0) be an arbitrary point of B. Then there exists an N such that $(x_0, y_0) \in C'_N$. Now, let us choose $\delta_1, \delta_2, \delta_3, \delta_4$ as follows:

(i) Since $(x_0, y_0) \in C'_N$, it is a point of density for C'_N and hence there is a $\delta_1 > 0$ such that $0 < \delta < \delta_1$ implies

$$\text{meas}\,(C'_N \cap S_\delta(x_0, y_0)) > \tfrac{1}{2}\,\text{meas}\,(S_\delta(x_0, y_0)),$$

where, as before, $S_\delta(x_0, y_0)$ is a square of side length 2δ with center at (x_0, y_0).

(ii) Let us suppose that the minimum condition does not hold at (x_0, y_0). Then, there is a $u \in U$ with $\varepsilon = R(x_0, y_0; u) > 0$. Using the continuity of $R(x, y; u)$ we obtain a $\delta_2 > 0$ such that $|R(x, y; u) - R(x_0, y_0; u)| < \varepsilon/2$ whenever $(x, y) \in C_N$ and $|(x, y) - (x_0, y_0)| < 2\delta_2$. Thus, $R(x, y; u) > \varepsilon/2$ for all $(x, y) \in C_N \cap S_{\delta_2}(x_0, y_0)$.

(iii) The function $s(x, y; u) = |f(x, y, \dot{z}(x, y), u) - f(x, y, \dot{z}(x, y), v_0(x, y))|$, with u as in (ii), belongs to $L_4(G)$ (by (H_6)); and hence there exists a $\delta_3 > 0$ such that for $0 < \delta < \delta_3$, we have

$$\iint_{S_\delta} s^2(u, \alpha, \beta)\, d\alpha\, d\beta < \zeta^2$$

and

$$\iint_{S_\delta} s^4(u, \alpha, \beta)\, d\alpha\, d\beta < \zeta^2,$$

where ζ is any number with $0 < \zeta^2 \leqq \zeta \leqq \varepsilon/6M$ (ε as in (ii) and M as in the equality (5.8)).

(iv) Since (x_0, y_0) is in the interior of G, there is a $\delta_4 > 0$ such that $S_\delta = S_\delta(x_0, y_0) \subset G$ for $0 < \delta < \delta_4$.

Let $\sigma > 0$ be such that $\sigma < \min(\delta_1, \delta_2, \delta_3, \delta_4)$ and let v_ε be a function defined by $v_\varepsilon(x, y) = u$ if $(x, y) \in C_N \cap S_\sigma$ and $= v_0(x, y)$ otherwise. Clearly, v_ε is an element of Γ. Also, $R(x, y; v_\varepsilon(x, y))$ is zero outside $C_N \cap S_\sigma$ and is $> \varepsilon/2$ for all $(x, y) \in C_N \cap S_\sigma$. Thus

$$\begin{aligned} I[v_\varepsilon] - I[v_0] &= \eta + \iint_G [H(v_\varepsilon(x, y)) - H(v_0(x, y))]\, dx\, dy \\ &= \eta - \iint_{C_N \cap S_\sigma} R(x, y; v_\varepsilon(x, y))\, dx\, dy \\ &< \eta - \tfrac{1}{2}\varepsilon\,\text{meas}\,(C_N \cap S_\sigma) < \eta - 4^{-1}\,\varepsilon\,\text{meas}\,S_\sigma \\ &= \eta - \varepsilon\sigma^2, \end{aligned}$$

where $|\eta| < \varepsilon\sigma^2/2$ from § 5. Thus, $I[v_\varepsilon] - I[v_0] < -\varepsilon\sigma^2/2 < 0$. This is contrary to the assumption that v_0 is optimal. The contradiction arose because of (ii).

It follows that for any $(x, y) \in B$, we have $H(x, y, v_0(x, y)) \leqq H(x, y, u)$ for all $u \in U$. This concludes the proof of the theorem.

7. Discussion and examples. In this section we shall discuss the Pontryagin-type necessary condition given in Theorem 6.1 in relation to the results of Cesari [2] and A. I. Egorov [3]. We shall first show that our results yield those of A. I. Egorov under conditions of smoothness. We shall also give examples where our necessary condition applies. In particular, Example 3 of (D) below will show that our results are actually more general than those of A. I. Egorov.

(A) *The linear case.* If the state equations are linear, i.e., of the form $z_{xy} = Az + Bz_x + Cz_y + D(x, y, u)$, where A, B, C are matrix-valued functions on G, then we have seen that the increment formula reduces to

$$I(v_\varepsilon) - I(v_0) = \iint_G [H(v_\varepsilon(x, y)) - H(v_0(x, y))]\, dx\, dy.$$

Now, if a control $v_0 \in \Gamma$ satisfies the minimum condition, then in particular $H(v_0(x, y)) \leqq H(v_\varepsilon(x, y))$ a.e. in G and hence $I(v_0) \leqq I(v_\varepsilon)$ for all v_ε in Γ; i.e., v_0 is optimal for I. Thus, the necessary condition is also sufficient in the linear case. For the existence of solutions for the Goursat problem (1.1), (1.2) (as well as the conjugate problem (4.3), (4.4)) in this case, we may require that the matrix-valued functions $A(x, y)$, $B(x, y)$, $C(x, y)$ be in $L_\infty(G)$ and that $D(x, y, u)$ be continuous in u. Further, we shall require $D(x, y, v(x, y))$ to be in $[L_p(G)]^n$ for $v \in \Gamma$.

(B) *Various types of cost functional.*

(B1) It is clear that the cost functional (1.4) or $I[z, v] = \sum_{i=1}^n A_i z^i(a + h, b + k)$ can be written in the Lagrange form

$$J[z, v] = \iint_G f_0(x, y, z(x, y), z_x(x, y), z_y(x, y), v(x, y))\, dx\, dy$$

with $f_0 = \sum_{i=1}^n A_i f_i$ and the f_i as in (1.1).

However, the Lagrange problem of the minimum of $J[z, v]$ with f_0 not necessarily equal to $\sum A_i f_i$, and z, v satisfying (1.1)–(1.3), can always be formulated as the Mayer problem (1.1)–(1.4) by suitable transformations. This is done, as usual, by introducing a new variable z^0 with $z^0_{xy} = f_0(x, y, z, z_x, z_y, v)$, $z^0(a, y) = 0$, $z^0(x, b) = 0$. Then, the functional J can be written in the form $J = z^0(a + h, b + k)$ (cf. Cesari [2] or the author [7]).

(B2) The general Mayer problem with cost functional $I'[z, v] = \phi(z(a + h, b + k))$ with z, v satisfying (1.1)–(1.3) and an arbitrary $\phi(\zeta)$, $\zeta \in E_n$ (twice continuously differentiable), can be reduced to problem (1.1)–(1.4). As above, this is usually done by introducing a new variable z^0 satisfying

$$z^0_{xy} = \sum_{i,j=1}^n \frac{\partial^2 \varphi}{\partial z^i \partial z^j} z^i_x z^i_y + \sum_{i=1}^n \frac{\partial \varphi}{\partial z^i} f_i(x, y, z, z_x, z_y, v)$$

and

$$z^0(a, y) = \phi(\psi^1(y), \cdots, \psi^n(y)), \qquad z^0(x, b) = \phi(\varphi^1(x), \cdots, \varphi^n(x)),$$

where $\varphi(x) = (\varphi^1, \cdots, \varphi^n)$, $\psi(y) = (\psi^1, \cdots, \psi^n)$ are the initial data as in (1.2). (See A. I. Egorov [3] and the author [7].)

(B3) The problem of minimizing

$$J[z, v] = \int_a^{a+h} F(x, z(x, b + k), z_x(x, b + k))\, dx$$

can be reduced to that of $z^0(a + h, b + k)$, where z^0 is defined by

$$z^0_{xy} = \sum_{i=1}^n \frac{\partial F}{\partial z^i} z^i_y + \frac{\partial F}{\partial z^i_x} f_i(x, y, z, z_x, z_y, v)$$

and

$$z^0(a, y) = 0, \qquad z^0(x, b) = \int_a^x F(\alpha, \varphi(\alpha), \varphi'(\alpha))\, d\alpha$$

(see A. I. Egorov [3]).

(B4) Let J denote any linear combination of the functionals mentioned above in (B1), (B2), (B3). It is clear that the functional J can be reduced to the form (1.4) by suitable addition of an auxiliary variable z^0.

(C) *Comparison with A. I. Egorov's results.* The optimization problem (1.1)–(1.4) was studied by A. I. Egorov [3] where he proposed a necessary condition in terms of the Hamiltonian

$$H(x, y, z, z_x, z_y, v, \theta) = \sum_{i=1}^n \theta^i f_i(x, y, z, z_x, z_y, v)$$

and multipliers $\theta(x, y) = (\theta^1, \cdots, \theta^n)$ satisfying the Goursat-type problem

$$\theta^i_{xy} = \frac{\partial H}{\partial z^i} - \frac{\partial}{\partial x}\frac{\partial H}{\partial z^i_x} - \frac{\partial}{\partial y}\frac{\partial H}{\partial z^i_y}, \qquad (x, y) \in G, \quad i = 1, \cdots, n, \tag{7.1}$$

with boundary conditions

$$\theta^i_x = -(\partial H/\partial z^i_y) \quad \text{for } y = b + k, \qquad \theta^i_y = -(\partial H/\partial z^i_x) \quad \text{for } x = a + h, \tag{7.2}$$

$$\theta^i(a + h, b + k) = A_i, \qquad i = 1, \cdots, n. \tag{7.3}$$

In (7.1) the derivatives are evaluated at (z, z_x, z_y, v, θ) and the numbers A_i in (7.3) are those in (1.4).

In [3] the control variables v_i are assumed to be piecewise continuous. In the present paper the controls v_i are assumed to be only measurable, and we proved, under our general assumption (H_1)–(H_5), that the functions z^i belong to a Sobolev class $W^1_p(G)$ and are continuous in G. In any case, the derivatives $(\partial/\partial x)(\partial H/\partial z^i_x)$, $(\partial/\partial y)(\partial H/\partial z^i_y)$ which appear in (7.1) need not in general exist, as Example 3 below in (D) will show.

On the other hand, under suitable regularity conditions, equations (7.1)–(7.3) can be derived from the conjugate problem (4.3)–(4.4), by defining $\theta = (\theta^1, \cdots, \theta^n)$ in terms of the multipliers $\lambda_1, \lambda_3, \mu_1, \mu_2$.

We need the following assumptions:

(**) For a given optimal pair (z, v) and corresponding multipliers $\lambda_1, \lambda_3, \mu_1, \mu_2$, let us assume that the following partial derivatives exist as generalized derivatives:

$$\frac{\partial}{\partial x}\sum_{j=1}^n (\lambda^j_3 + \mu^j_2)\frac{\partial f_j}{\partial z^i_x}, \qquad \frac{\partial}{\partial y}\sum_{j=1}^n (\lambda^j_3 + \mu^j_2)\frac{\partial f_j}{\partial z^i_y}, \qquad i = 1, \cdots, n.$$

From (4.5) and (**) it follows that λ^i_{3xy} and μ^i_{2xy}, $i = 1, \cdots, n$, exist as generalized derivatives, and in view of (4.3) we have

$$\lambda^i_{3xy} = -\mu^i_{1y} - \frac{\partial}{\partial y}\sum_{j=1}^{n}(\lambda^j_3 + \mu^j_2)\frac{\partial f}{\partial z^i_y},$$

$$\mu^i_{2xy} = -\lambda^i_{1x} - \frac{\partial}{\partial x}\sum_{j=1}^{n}(\lambda^j_3 + \mu^j_2)\frac{\partial f_i}{\partial z^i_x},$$

a.e. in G, $i = 1, \cdots, n$. Thus, if $\theta = \lambda_3 + \mu_2$, that is, $\theta^i = \lambda^i_3 + \mu^i_2$, $i = 1, \cdots, n$, then θ_{xy} exists a.e. in G as a generalized derivative and, in view of (4.3), we get (7.1) with $H = \sum_i \theta^i f_i$. Again, from (4.5) we get

$$(7.4) \qquad \theta^i_x = \lambda^i_{3x} = -\mu^i_1 - (\partial H/\partial z^i_y) = -(\partial H/\partial z^i_y) \quad \text{for } y = b + k,$$

and analogously,

$$(7.5) \qquad \theta^i_y = -(\partial H/\partial z^i_x) \quad \text{for } x = a + h, \qquad i = 1, \cdots, n.$$

Finally,

$$\theta(a + h, b + k) = (\lambda_3 + \mu_2)(a + h, b + k) = A.$$

Thus, under assumption (**), the sums $\theta^i = \lambda^i_3 + \mu^i_2$, $i = 1, \cdots, n$, act as multipliers θ^i described by (7.1)–(7.2). It is of interest to note that $\theta = \lambda_3 + \mu_2$ is obtained as the fixed point of the contraction operator T (see Remark (a) in § 4), and thus θ is unique, in harmony with the uniqueness of Egorov's solution θ of (7.1)–(7.2).

(D) *Examples.*

Example 1 [3, p. 560]. Let $G = \{(x, y)|0 \leqq x \leqq 1,\ 0 \leqq y \leqq 1\}$ and consider the problem of minimizing

$$S = \int_0^1\int_0^1 (1 - 2y)z(x, y)\,dx\,dy$$

with side condition

$$z_{xy} = -2z - 2z_x - 2z_y - v,$$

and boundary conditions

$$z(x, 0) = z(0, y) = 0;$$

here z is a scalar and v is a control variable with values $[0, 1]$. To obtain the multipliers (and use Theorem 6.1), we first introduce a new set of variables z_1, $i = 1, \cdots, 6$, defined by $z_1 = z$, $z_2 = z_x$, $z_3 = z_y$, $z_4 = \int_0^x\int_0^y (1 - 2\beta)z_1(\alpha, \beta)\,d\alpha\,d\beta$, $z_5 = z_{4x}$, $z_6 = z_{4y}$. Then S attains a minimum together with $z_4(1, 1)$ and the side conditions are now

$$z_{2y} = z_{3x} = z_{1xy} = -2z_1 - 2z_2 - z_3 - v, \qquad z_{5y} = z_{6x} = z_{4xy} = (1 - 2y)z_1.$$

The corresponding conjugate problem is

$$\begin{aligned}
&\lambda_{1x} + \mu_{1y} = -\partial H/\partial z_i = 2(\lambda_3 + \mu_2) - (1 - 2y)(\lambda_6 + \mu_5),\\
&\mu_{2y} = -\partial H/\partial z_2 = -\lambda_1 + 2(\lambda_3 + \mu_2),\\
&\lambda_{3x} = -\partial H/\partial z_3 = -\mu_1 + (\lambda_3 + \mu_2),\\
&\lambda_{4x} + \mu_{4y} = -\partial H/\partial z_4 = 0,\\
&\mu_{5y} = -\partial H/\partial z_5 = -\lambda_4, \qquad \lambda_{6x} = -\partial H/\partial z_6 = -\mu_4;
\end{aligned}$$

with boundary conditions

$$\lambda_i(1, y) = 0 = \mu_j(x, 1) \quad \text{for } i \neq 6 \text{ and } j \neq 5,$$

$$\lambda_6(1, y) = \mu_5(x, 1) = 1/2.$$

Here $H = \lambda_1 z_2 + \mu_1 z_3 + (\lambda_3 + \mu_2)(-2z_1 - 2z_2 - z_3 - v) + \lambda_4 z_5 + \mu_4 z_6 + (\lambda_6 + \mu_5)(1 - 2y)z_1$. It is seen from the boundary conditions that one can take $\lambda_4 = \mu_4 = 0, \mu_5 = \lambda_6 = 1/2$ on G.

Further, if $\theta = \mu_2 + \lambda_3$ and $\zeta = \theta_x - \theta$, then by formal differentiation of the above equations, we get $\theta_{xy} = \theta_y + 2(\theta_x - \theta) + (1 - 2y)$ or $\zeta_y = 2\zeta + 1 - 2y$, $\zeta(x, 1) = 0$. Solving for ζ as a function of y, we get $\zeta(x, y) = y - e^{2(y-1)}$. Thus, $\theta_x - \theta = y - e^{2(y-1)}$. Solving for θ as a function of x, we get $\theta(x, y) = (e^{x-1} - 1) \cdot (y - e^{2(y-1)})$. It is clear that for $(x, y) \in G$, $e^{x-1} \leqq 1$ and hence $\theta(x, y) \geqq 0$ or $\leqq 0$ according as $y \leqq y_0$ or $y \geqq y_0$, where $y_0 = e^{2y_0-2}$. But then, H as a function of v is a minimum for $v_0(x, y)$, where v_0 is a function on G defined by: $v_0(x, y) = 1$ for $y \leqq y_0$; $= -1$ for $y \geqq y_0$.

Now, to obtain the value of the functional, we first solve the following for z: $z_{xy} = -2z - 2z_x - z_y - v_0$ with $z(x, 0) = z(0, y) = 0$. It is seen that z is given by

$$z(x, y) = \begin{cases} 2^{-1}(1 - e^{-x})(e^{-2y} - 1) & \text{for } y \leqq y_0,\\ 2^{-1}(1 - e^{-x})(1 + e^{-2y} - 2e^{2y_0 - 2y}) & \text{for } y \geqq y_0, \end{cases}$$

and the functional takes the value

$$S = e^{-1}(y_0^2 + \tfrac{1}{2}e^{-2} - y_0), \quad \text{where } y_0 = e^{2y_0-2}.$$

This is the optimum value obtained in [3] also.

Example 2 [3, p. 561]. Find the minimum of the functional

$$S = \int_0^1 z(x, 1)\,dx - \int_0^1 z(1, y)\,dy,$$

where the side conditions on z are given by

(7.6) $\quad z_{xy} = v; \quad |v| \leqq 1; \quad 0 \leqq x \leqq 1; \quad 0 \leqq y \leqq 1; \quad z(x, 0) = z(0, y) = 0.$

If z_0 is a variable satisfying the relations

(7.7) $\quad z_{0xy} = z_y - z_x \quad \text{and} \quad z_0(0, y) = z_0(x, 0) = 0,$

then the above optimization problem reduces to the problem of minimizing $z_0(1, 1)$ with (7.6), (7.7) as side conditions. The conjugate problem is described in terms of the multipliers $\lambda_1, \mu_1, \lambda_3, \mu_2, \lambda_4, \mu_4, \lambda_6, \mu_5$; $\lambda_{1x} + \mu_{1y} = 0$; $\mu_{2y} =$

$-\lambda_1 + (\lambda_6 + \mu_5); \lambda_{3x} = -\mu_1 - (\lambda_6 + \mu_5); \lambda_{4x} + \mu_{4y} = 0; \mu_{5y} = -\lambda_4; \lambda_{6x} = -\mu_4$; with boundary conditions $\lambda_i(1, y) = 0$; $\mu_j(x, 1) = 0$ for $i \neq 6$, and $j \neq 5$; $\lambda_6(1, y) = 1/2$; $\mu_5(x, 1) = 1/2$.

In order to obtain a set of solutions, we may introduce the auxiliary equations $\lambda_1 = 0$ and $\lambda_4 = 0$ on G. Then, we obtain $\mu_1 = 0, \mu_4 = 0, \mu_5 = 1/2$ and $\lambda_6 = 1/2$ on G. Also, $\mu_2 = (y - 1)$ and $\lambda_3 = -(x - 1)$. Thus, the Hamiltonian reduces to $H = (y - x)v + (z_y - z_x)$. It follows that as a function of v alone H is a minimum at $v_0(x, y)$, where v_0 is defined on G as follows:

$$v_0(x, y) = \begin{cases} -1 & \text{for } 0 \leqq x < y \leqq 1, \\ 1 & \text{for } 0 \leqq y < x \leqq 1. \end{cases}$$

Substituting in (7.6), and integrating we obtain

$$z(x, y) = \begin{cases} -xy + \phi(x) & \text{for } 0 \leqq x < y \leqq 1, \\ xy + \psi(y) & \text{for } 0 \leqq y < x \leqq 1, \end{cases} \tag{7.8}$$

where ϕ and ψ are absolutely continuous functions defined on $[0, 1]$ with $\phi(0) = \psi(0) = 0$. Now ϕ and ψ are to be chosen so that the two expressions of (7.8) coincide for $x = y$. Thus $\phi(y) - y^2 = y^2 + \psi(y)$, i.e., $\psi(y) = \phi(y) - 2y^2$. Hence $z(x, y) = -xy + \phi(x)$ for $0 \leqq x < y \leqq 1$; $= xy - 2y^2 + \phi(y)$ for $0 \leqq y < x \leqq 1$; where ϕ is some arbitrary absolutely continuous function defined on $[0, 1]$ with $\phi(0) = 0$. The corresponding value of the functional is

$$\begin{aligned} S &= \int_0^1 z(x, 1)\, dx - \int_0^1 z(1, y)\, dy \\ &= \int_0^1 [\phi(x) - x]\, dx - \int_0^1 [\phi(y) + y - 2y^2]\, dy = -1/3. \end{aligned}$$

This again is in harmony with the optimum obtained in [3].

Example 3 [7, p. 207]. Let G be the rectangle $[0, 1] \times [0, 1]$. Let us consider the problem of minimizing the functional $S = z(1, 1)$ with side conditions and constraints

$$z_{xy} = (1 + z_x)v, \quad z(x, 0) = 0, \quad z(0, y) = 0, \quad -1 \leqq v \leqq 1. \tag{7.9}$$

Let us first observe that for any $v(x, y)$ in $L_1(G)$, the solution of (7.9) is given by

$$z(x, y) = \int_0^x \left[-1 + \exp\left(\int_0^y v(\alpha, \beta)\, d\beta \right) \right] d\alpha. \tag{7.10}$$

Now, since $v(\alpha, \beta) \geqq -1$ for all (α, β),

$$\int_0^1 v(\alpha, \beta)\, d\beta \geqq -1$$

and hence

$$\int_0^1 \exp\left(\int_0^1 v(\alpha, \beta)\, d\beta \right) d\alpha \geqq 1/e.$$

Thus, $S = z(1, 1) \geqq -1 + (1/e)$ for any admissible pair (z, v) satisfying (7.9). It follows that the function $v_0(x, y)$ defined by $v_0(x, y) = -1$ for almost all $(x, y) \in G$ is optimal for S, and vice versa.

In order to verify that v_0 satisfies the minimum condition, we formulate the conjugate problem $\lambda_{1x} + \mu_{1y} = -\partial H/\partial z_1 = 0$, $\mu_{2y} = -\partial H/\partial z_2 = -\lambda_1 - v(\lambda_3 + \mu_2)$, $\lambda_{3x} = -\partial H/\partial z_3 = -\mu_1$; with boundary conditions $\lambda_1(1, y) = \mu_1(x, 1) = 0$, $\lambda_3(1, y) = \mu_2(x, 1) = 1/2$. Here the Hamiltonian H is given by $H = \lambda_1 z_2 + \mu_1 z_3 + (\lambda_3 + \mu_2)f$, where $z_1 = z$, $z_2 = z_x$, $z_3 = z_y$, $f = (1 + z_x)v$. The multipliers corresponding to v_0 are obtained as solutions of the above system of equations with v replaced by $v_0(x, y)$. Clearly, we may choose $\lambda_1 = \mu_1 = 0$ and $\lambda_3 = 1/2$ on G. But then $\mu_{2y} = -v_0(\mu_2 + 2^{-1})$, $\mu_2(x, 1) = 1/2$. A solution of this equation is given by

$$\mu_2(x, y) = \exp\left(\int_y^1 v_0(x, \beta)\, d\beta\right) - \frac{1}{2}.$$

Substituting in the Hamiltonian, we get

$$H(x, y, z, u, \lambda, \mu) = u(1 + z_x) \exp\left(\int_y^1 v_0(x, \beta)\, d\beta\right),$$

where z corresponds to v_0. Now, if $v_0(x, y) = -1$ on G, then

$$H = u(1 + z_x) \exp(y - 1) = u\left(\exp \int_0^y v_0(x, \beta)\, d\beta\right) \exp(y - 1) = \frac{u}{e}.$$

Clearly, $H(u) \geqq H(v_0(x, y))$ for $(x, y) \in G$ and $u \in U = [-1, 1]$.

Remarks. It is to be observed that in the above example, the optimal solution $v_0(x, y) = -1$ happens to be smooth, and as is mentioned in § 7(c), Egorov's condition also holds. However, Example 3 can be easily modified into another one for which Egorov's necessary condition cannot be applied. Indeed, if $w(x)$, $0 \leqq x \leqq 1$, is a fixed, continuous, positive nowhere differentiable function (such a function exists), we consider instead of (7.9) the equation $z_{xy} = (1 + z_x)v \cdot w$ with the same boundary conditions and constraints as above. Then (7.10) is replaced by

$$z(x, y) = \int_0^x \left[-1 + \exp\left(w(\alpha) \int_0^y v(\alpha \cdot \beta)\, d\beta\right)\right] d\alpha$$

and the optimal control is still $v_0(x, y) = -1$ a.e. in G. Here, Egorov's Hamiltonian H (see [3]) is given by $\theta(1 + z_x)vw$ and the second order derivative $(H_{z_x})_x$ required in (7.1) does not exist.

Also of interest, in the above example, is the fact that the multiplier $\mu_2(x, y)$ need not have a partial derivative with respect to x. Thus, in general, the multipliers need not have partial derivatives with respect to both the variables; as such they may not belong to a Sobolev class.

Acknowledgment. The author is greatly indebted to Professor Lamberto Cesari for his valuable guidance and constant encouragement during the writing of this paper.

REFERENCES

[1] A. G. Butkovskii, A. I. Egorov and K. A. Lurie, *Optimal control of distributed parameter systems* (a survey of Soviet publications), this Journal, 6 (1968), pp. 437–476.

[2] L. Cesari, *Optimization with partial differential equations in Dieudonné–Rashevsky form and conjugate problems*, Arch. Rational Mech. Anal, 33 (1969), pp. 339–357.

[3] A. I. Egorov, *Optimal control of processes in certain parameter systems*, Automat. Remote Control, 25 (1964), pp. 557–566.

[4] ———, *Necessary conditions for optimality for systems with distributed parameters*, Mat. Sb., 69 (1966), pp. 371–421.

[5] C. B. Morrey, *Multiple Integrals in the Calculus of Variations*, Springer, Berlin, 1966.

[6] S. L. Sobolev, *Applications of Functional Analysis in Mathematical Physics*, Amer. Math. Soc. Transl., vol. 7, Providence, R.I., 1963.

[7] M. B. Suryanarayana, *Optimization problems with hyperbolic partial differential equations*, Doctoral thesis, The University of Michigan, Ann Arbor, 1969.

[8] ———, *On multidimensional integral equations of Volterra type*, Pacific J. Math., 41 (1972), pp. 809–828.

SIAM J. CONTROL
Vol. 11, No. 1, February 1973

A NEW APPROACH TO THE THEORY OF CANONICAL DECOMPOSITION OF LINEAR DYNAMICAL SYSTEMS*

P. D'ALESSANDRO,† A. ISIDORI† AND A. RUBERTI‡

Abstract. In this paper, a new approach to the structure theory of linear dynamical systems is developed.

Based on the structural properties of influenceability and invisibility and without introducing any limiting hypotheses, the existence of a canonical decomposition of the state space into subspaces of constant dimensions is proved.

Furthermore, the existence of the canonical form of the equations, and also the uniqueness of this form within an equivalence class are proved, the latter also being determined. The general structure theory obtained in this way comprises all the hitherto known results and in particular yields a constant decomposition for time-invariant systems.

1. Introduction. The first contributions to the structure theory of linear dynamical systems were due to R. E. Kalman [2], [3]. In these works, starting from the properties of controllability and observability at a given time t, he decomposed the state space into the direct sum of four linear subspaces and gave the corresponding canonical form of the state equations for that time t. Further contributions, regarding both the refinement of the statements of the decomposition theorems and the proofs of these theorems, were due to Kalman himself and to L. Weiss [6]. In particular, considering also the properties of reachability and constructibility (at a given time t), many possibilities of decomposition were indicated [6]; nevertheless, formal and complete proofs were not always given. Still basing himself on the properties of controllability and observability, at a given time t, L. Weiss returns to the problem in [5], but formulates some restrictive hypotheses that are substantially equivalent to assuming the dimensions of the subspaces into which the state space is decomposed to be invariant in time. For this case he supplies both the proof of the decomposition theorem and the algorithms for the determination of the canonical forms; the results of this work comprise the case of the analytical systems and therefore also that of the time-invariant ones.

D. C. Youla, in [7], approaches the aforementioned problem in a different manner. He starts from the requirement of constructing all the realizations of a given realizable weighting pattern, and subsequently shows how the realization associated with any factorization of a given realizable weighting pattern can be divided into four subsystems in parallel, of which one corresponds to the minimal realization. Analyzing the properties of controllability and observability of these subsystems, he then interprets the division under consideration as a version of R. E. Kalman's canonical decomposition. However, one must here point out that the structural properties considered in the latter are different from those that are implicitly present in D. C. Youla's decomposition (cf. § 5).

* Received by the editors July 13, 1971, and in revised form November 29, 1971. This work was supported in part by the National Council of Research of Italy.

† Istituto di Automatica, University of Rome, Rome, Italy.

‡ Istituto di Automatica, University of Rome, Rome, Italy and Fondazione U. Bordoni, Rome.

Two ever-present requirements in the literature about structure theory are the following:

(a) the need for finding a decomposition of the state space into subspaces of constant dimension;

(b) developing a theory that comprises the known results for the time-invariant case and, more particularly, supplies a constant decomposition for this case (see, for example, the comments in [5] about the results developed in [7]).

In the present paper, introducing new structural properties, we develop a theory that satisfies the abovementioned requirements. Moreover, we also solve the problem of analyzing all representations of the state equations in canonical form.

It seems useful here to stress that these new properties, whose introduction is in itself justified by the fact that it permits the development of a theory of decomposition that is valid for all linear dynamical systems, are also significant by virtue of the fact that they characterize in a unique manner the minimal realizations of a given weighting pattern (cf. § 5).

2. Structural properties of influenceability and invisibility.

Consider the linear system described by the equations

$$\begin{aligned} \dot{x}(t) &= A(t)x(t) + B(t)u(t), \\ y(t) &= C(t)x(t), \end{aligned} \tag{1}$$

where $x(t) \in R^n$, $u(t) \in R^p$, $y(t) \in R^q$, and $A(\cdot)$, $B(\cdot)$, $C(\cdot)$ are matrices of continuous functions of t; furthermore, let $X(\cdot)$ be a fundamental matrix of solutions for the homogeneous equation associated with (1), and let $\Phi(t, \tau) = X(t)X^{-1}(\tau)$ be the corresponding state transition matrix.

At a given time t, the usually considered properties of the state, with reference to the input, are those of controllability and reachability. For the purposes of solving the problem of canonical decomposition it seems useful to introduce the following property.

DEFINITION 1. A state x of the system (1) is *influenceable*[1] *at time* t if it is possible to express it as the sum of states, each of which is at least controllable or reachable at the same time t.

Remark. If the system is such that controllability and reachability imply each other at each time t, then influenceability coincides with these two properties. In particular, this happens in some classes of systems, including the time-invariant ones, the analytical ones, etc. It is therefore clear that a decomposition theory based on the property of influenceability will yield the already known ones in the abovementioned cases.

Let us now consider the Gramian matrix

$$P(t; \eta, \zeta) = \int_{\eta}^{\zeta} \Phi(t, \tau)B(\tau)B^T(\tau)\Phi^T(t, \tau)\, d\tau. \tag{2}$$

[1] The need for streamlining the presentation has led the authors, somewhat unwillingly, to introduce two new terms for the structural properties used in the treatment. Indeed, the word "invisible" used in Definition 2 has already been used by other authors as a synonym for "unobservable."

When $\eta = t$ and $\zeta > t$, this coincides with the controllability matrix at time t; when $\eta = t$ and $\zeta < t$, it coincides with the reachability matrix at time t.

As is known, the range of the controllability (reachability) matrices, i.e., $\mathscr{R}[P(t;t,\zeta)]$, is a nondecreasing function with ζ increasing (decreasing) for each fixed t [5]. Analogous properties hold for the range of $P(t;\eta,\zeta)$. One can therefore state the following lemma.

LEMMA 1. *If the interval* $[\eta',\zeta']$ *contains the interval* $[\eta,\zeta]$, *one has*

$$\mathscr{R}[P(t;\eta',\zeta')] \supseteq \mathscr{R}[P(t;\eta,\zeta)] \quad \textit{for all } t. \tag{3}$$

Proof. The proof can be done following the lines of a proof given in [5] about the controllability matrix. $P(t;\eta,\zeta)$ is a Gramian matrix and has the property that, for each x, $0 \leqq x^T P(t;\eta,\zeta)x \leqq x^T P(t;\eta',\zeta')x$ if $[\eta',\zeta']$ contains $[\eta,\zeta]$. Therefore $x \in \mathscr{N}[P(t;\eta',\zeta')] \Rightarrow x \in \mathscr{N}[P(t;\eta,\zeta)]$. This implies $\mathscr{N}[P(t;\eta',\zeta')] \subseteq \mathscr{N}[P(t;\eta,\zeta)]$, and therefore, by orthogonal complementation, also implies equation (3).

Using this lemma, one can state the following condition of influenceability.

THEOREM 1. *A state* x *is influenceable at time* t *if and only if there exist two values* t_1 *and* t_2 *such that* $x \in \mathscr{R}[P(t;t_1,t_2)]$.

Proof. One can limit oneself to considering the case $t_1 < t_2$. As regards sufficiency, one can first of all note that

$$x \in \mathscr{R}[P(t;t_1,t_2)] \Rightarrow \exists x_1 \in \mathscr{R}[P(t;t_1,t)], \exists x_2 \in \mathscr{R}[P(t;t,t_2)], \tag{4}$$

where $x_1 + x_2 = x$.

If $t_1 < t < t_2$, x_1 is reachable at time t and x_2 is controllable at time t; x is therefore influenceable at time t. If $t \leqq t_1$ $(t \geqq t_2)$, x is controllable (reachable) at time t and therefore influenceable at time t.

As regards necessity, if x is controllable (reachable) at t, there exists a $t_2 > t$ $(t_2 < t)$ such that $x \in \mathscr{R}[P(t;t,t_2)]$; the condition is therefore satisfied with $t_1 = t$. If x is the sum of two states, one of which (x_c) is controllable and the other (x_r) is reachable, i.e., if $x = x_c + x_r$, there exists a $t_2 > t$ and a $t_1 < t$ such that

$$x_c \in \mathscr{R}[P(t;t,t_2)] \quad \text{and} \quad x_r \in \mathscr{R}[P(t;t_1,t)]. \tag{5}$$

Since, by virtue of Lemma 1,

$$\mathscr{R}[P(t;t_1,t_2)] \supseteq \mathscr{R}[P(t;t,t_2)] \cup \mathscr{R}[P(t;t_1,t)], \tag{6}$$

it follows that $x \in \mathscr{R}[P(t;t_1,t_2)]$.

For the space $\mathscr{P}(t)$ of all influenceable states at time t one can state the following theorem.

THEOREM 2. *The space* $\mathscr{P}(t)$ *of all influenceable states at each time* t *can be expressed in the form*

$$\mathscr{P}(t) = \mathscr{R}[P(t;\hat{\eta},\hat{\zeta})], \tag{7}$$

where $\hat{\eta}$, $\hat{\zeta}$ *is any pair of values of* η, ζ *for which, at an arbitrary fixed time* t_0, $P(t_0;\eta,\zeta)$ *has maximum rank. Furthermore,*

$$\dim \mathscr{P}(t) = \text{const.} \triangleq n_p. \tag{8}$$

Proof. It follows from Theorem 1 that

$$\mathscr{P}(t) = \bigcup_{\eta<\zeta} \mathscr{R}[P(t;\eta,\zeta)]. \tag{9}$$

By virtue of Lemma 1, on the other hand, if $\hat{\eta}$, $\hat{\zeta}$ is a pair of values of η, ζ for which $P(t_0;\eta,\zeta)$ has maximum rank, one can write

$$\bigcup_{\eta<\zeta} \mathscr{R}[P(t_0;\eta,\zeta)] = \mathscr{R}[P(t_0;\hat{\eta},\hat{\zeta})]. \tag{10}$$

When $t \neq t_0$, the relationship

$$P(t;\eta,\zeta) = \Phi(t,t_0)P(t_0;\eta,\zeta)\Phi^T(t,t_0) \tag{11}$$

applies, and it can be seen from this that (10) holds for each t_0 with the same $\hat{\eta}$, $\hat{\zeta}$ and (7) is thus proved. Since $\Phi(t,t_0)$ is nonsingular, (8) follows from (11) evaluated for $\eta = \hat{\eta}$ and $\zeta = \hat{\zeta}$.

Proceeding now to consider the structural properties of the state at a given time t with reference to the output, it seems useful to introduce the concept of invisibility. In this connection, referring to the properties of unobservability and unconstructibility [4], one can give the following definition.

DEFINITION 2. A state x of the system (1) is *invisible at time t* if it is unobservable and unconstructible at time t.

A remark analogous to the one made in connection with Definition 1 can be repeated for this case.

The Gramian matrix to be considered in relation with Definition 2 is

$$Q(t;\eta,\zeta) = \int_{\eta}^{\zeta} \Phi^T(\tau,t)C^T(\tau)C(\tau)\Phi(\tau,t)\,d\tau. \tag{12}$$

By means of considerations analogous to those made in connection with (2), writing $\mathscr{N}[Q]$ for the null space of the matrix Q, one can state the following lemma and Theorems 3 and 4.

LEMMA 2. *If the interval $[\eta',\zeta']$ contains the interval $[\eta,\zeta]$, one has*

$$\mathscr{N}[Q(t;\eta',\zeta')] \subseteq \mathscr{N}[Q(t;\eta,\zeta)] \quad \textit{for all } t. \tag{13}$$

THEOREM 3. *A state x is invisible at time t if and only if*

$$x \in \mathscr{N}[Q(t;\eta,\zeta)] \quad \textit{for all } \eta,\zeta.$$

Proof. One can limit oneself to considering the case $\eta < \zeta$. As regards sufficiency, it is derived immediately from

$$x \in \mathscr{N}[Q(t;\eta,\zeta)]\ \forall\eta, \forall\zeta \geqq \eta \ \Rightarrow \ \begin{cases} x \in \mathscr{N}[Q(t;t,\zeta)] & \forall\zeta \geqq t, \\ x \in \mathscr{N}[Q(t;\eta,t)] & \forall\eta \leqq t, \end{cases} \tag{14}$$

in which the conditions on the right-hand side imply the unobservability and unconstructibility of x at time t. As regards necessity, it is immediately noted that the implication in (14) also applies in the other direction, since

$$Q(t;\eta,\zeta) = Q(t;t,\zeta) + Q(t;\eta,t). \tag{15}$$

THEOREM 4. *The space $\mathscr{Q}(t)$ of all the invisible states at each time t can be expressed in the form*

$$\mathscr{Q}(t) = \mathscr{N}[Q'(t;\bar{\eta},\bar{\zeta})], \tag{16}$$

where $\bar{\eta}$, $\bar{\zeta}$ is any pair of values of η, ζ for which, at an arbitrarily fixed time t_0, $Q(t_0;\eta,\zeta)$ has maximum rank. Furthermore,

$$\dim \mathscr{Q}(t) = \text{const.} \triangleq n_q. \tag{17}$$

Remark. The concept of invisible state has a simple physical counterpart; if the state of a system is invisible at t_0 and the input is zero over $(-\infty, +\infty)$, then the corresponding output is zero over $(-\infty, +\infty)$. From another point of view, if the state x_0 of a system is invisible at t_0, then $\Phi(t, t_0)x_0$ is unobservable (unconstructible) at $t \geqq t_0$ $(t \leqq t_0)$. A similar interpretation can be given to the concept of influenceable state; if the state x_0 of a system is influenceable at t_0, then $\Phi(t, t_0)x_0$ is controllable (reachable) at t sufficiently small (large).

3. Decomposition of the state space. In the structure theory of linear dynamical systems it is usual to consider not only the decomposition of the state space of system (1) into four subspaces whose elements do or do not have two suitable structural properties (such as controllability and unobservability, for example), but also the transformation of the equations (1) into the special form that corresponds to this decomposition. In some approaches [2] the starting point is the decomposition of the state space, while in others [5] it is the transformation of the equations. In this latter case it seems natural for the transformation to be performed by means of an admissible change of coordinates, i.e., one that leaves the transformed equations continuous. It is clear, however, that there will be a corresponding requirement when one starts from the decomposition of the state space.

In fact, if one wants to obtain continuous differential equations when the union of bases in the individual subspaces is assumed as a basis in the state space, it is necessary for the decompositions at different instants of time to be related to each other "in an admissible manner." In other words, one has to guarantee suitable conditions of consistency, and the authors deem it desirable to formalize these in a definition.

DEFINITION 3. A decomposition of the state space $\mathscr{X}$ of the system (1), i.e.,

$$\mathscr{X} = \mathscr{A}(\cdot) \oplus \mathscr{B}(\cdot) \oplus \mathscr{C}(\cdot) \oplus \mathscr{D}(\cdot), \tag{18}$$

will be said to be *canonical with respect to the properties p_1 and p_2* if

(i) at every time t,
each state $x_a \in \mathscr{A}(t)$ has properties p_1 and p_2;
each state $x_b \in \mathscr{B}(t)$ has property p_1 and not property p_2;
each state $x_c \in \mathscr{C}(t)$ has property p_2 and not property p_1;
each state $x_d \in \mathscr{D}(t)$ has neither property p_1 nor property p_2;

(ii) the direct sum of any two of the considered subspaces maintains the property that is common to them (first condition of consistency)[2];

[2] This condition is implicitly present in the procedures that are usually proposed for performing a decomposition in accordance with condition (i). However, the authors wanted to state it explicitly in order to stress the fact that it constitutes a natural choice that is intended to reduce the degree of arbitrariness contained in (18).

(iii) there exists a nonsingular $n \times n$ matrix $T(\cdot,\cdot)$ of C^1-functions of two variables such that, writing $\mathscr{S}(\cdot)$ for any one of the four subspaces $\mathscr{A}(\cdot)$, $\mathscr{B}(\cdot)$, $\mathscr{C}(\cdot)$, $\mathscr{D}(\cdot)$, one has $\mathscr{S}(t') = \{y = T(t', t)x : x \in \mathscr{S}(t)\}$ for all t and t' (second condition of consistency).

One can immediately verify that condition (iii) implies that the dimensions of the four subspaces considered are constant with respect to t. It is known that this is not generally the case when the properties to which reference has been made are those of controllability and unobservability at time t and in general, therefore, it is not possible to obtain a canonical decomposition in the sense of Definition 3 on the basis of these properties. On the other hand, if one makes reference to the properties of influenceability and invisibility, it will be shown that one can obtain a canonical decomposition.

In this connection, having defined the subspaces $\mathscr{A}(t)$, $\mathscr{B}(t)$, $\mathscr{C}(t)$, $\mathscr{D}(t)$ of the state space $\mathscr{X}$ of the system (1) at each time t by means of the relationships (introduced by R. E. Kalman in [2])

$$\begin{aligned} \mathscr{A}(t) &= \mathscr{P}(t) \cap \mathscr{Q}(t), \\ \mathscr{P}(t) &= \mathscr{A}(t) \oplus \mathscr{B}(t), \\ \mathscr{Q}(t) &= \mathscr{A}(t) \oplus \mathscr{C}(t), \\ \mathscr{X} &= \mathscr{A}(t) \oplus \mathscr{B}(t) \oplus \mathscr{C}(t) \oplus \mathscr{D}(t), \end{aligned} \tag{19}$$

one can prove the following theorem.

THEOREM 5. *It is always possible, among all the infinite decompositions that satisfy equations* (19), *to obtain a decomposition of the state space of system* (1) *that is canonical with respect to the properties of influenceability and invisibility.*

Proof. Equations (19), by virtue of their construction, define decompositions of the state space of system (1) that respect condition (i) of Definition 3, referring to the properties of influenceability and invisibility. Furthermore, one can immediately note that these decompositions also respect condition (ii). As regards condition (iii), let us begin by noting on the basis of (7) and (11) that the relationship

$$\mathscr{P}(t') = \{y = \Phi(t', t)x : x \in \mathscr{P}(t)\} \tag{20}$$

holds for the subspace $\mathscr{P}(\cdot)$ of all the influenceable states.

Analogously, the relationship

$$\mathscr{Q}(t') = \{y = \Phi(t', t)x : x \in \mathscr{Q}(t)\} \tag{21}$$

holds for the subspace $\mathscr{Q}(\cdot)$ of all invisible states.

Consequently, the space $\mathscr{A}(\cdot)$ defined by the first equation of (19) also satisfies a relationship of the same type, i.e.,

$$\mathscr{A}(t') = \{y = \Phi(t', t)x : x \in \mathscr{A}(t)\} \tag{22}$$

and it therefore respects condition (iii) because $\Phi(t', t)$ is a matrix of C^1-functions in t and t'. For the other subspaces it is now possible to impose respect of condition (iii) by construction; in fact, if $\mathscr{B}(t)$, $\mathscr{C}(t)$, $\mathscr{D}(t)$ are arbitrarily fixed subspaces that satisfy (19) at time t, the corresponding subspaces $\mathscr{B}(t')$, $\mathscr{C}(t')$, $\mathscr{D}(t')$ obtained from these by means of a relationship of the type (22) will still satisfy (19) at time t'.

Remark. It should be pointed out that, with the exception of $\mathscr{A}(\cdot)$, the subspaces that satisfy equations (19) are not uniquely identified; furthermore, they do not generally satisfy the second condition of consistency. This condition can be satisfied by means of the construction procedure implicitly indicated in the proof of Theorem 5. It must be stressed, however, that this procedure is not the only one that can be adopted. In the case of time-invariant systems, for example, the matrix $T(\cdot,\cdot)$ that appears in the second condition of consistency can be a constant matrix.

The calculation of the dimensions of the four subspaces that have been introduced can be carried out by starting directly from the Gramian matrices defined by (2) and (12). In this connection, bearing in mind the definitions of n_p and n_q given by (8) and (17), one can state the following theorem.

THEOREM 6.[3] *The subspaces defined by equations* (19) *have constant dimensions equal to*

$$\begin{aligned} n_a &= n_p - n_0, \\ n_b &= n_0, \\ n_c &= n_q + n_0 - n_p, \\ n_d &= n - n_0 - n_q, \end{aligned} \tag{23}$$

where

$$n_0 = \operatorname{rank}\,[Q(t;\bar{\eta},\bar{\zeta})P(t;\hat{\eta},\hat{\zeta})]. \tag{24}$$

Proof. $\mathscr{A}(\cdot)$ has constant dimension n_a because it has the property (22). Since the relationship

$$\operatorname{rank} QP = \operatorname{rank} P - \dim \mathscr{R}[P] \cap \mathscr{N}[Q] \tag{25}$$

holds for any two matrices P and Q, bearing in mind equations (19), one obtains

$$\dim \mathscr{B}(t) = \operatorname{rank}\,[Q(t;\bar{\eta},\bar{\zeta})P(t;\hat{\eta},\hat{\zeta})] = n_0. \tag{26}$$

Equations (23) follow from this equation and, once again, from (19). The constancy of n_b, n_c, n_d follows from the constancy of n_a, n_p, n_q and from equations (23).

4. Canonical form of the equations. In connection with the decompositions of the state space, as already mentioned, it is usual to consider the possibility of writing equations (1) in the special form that is known as "canonical." Two theorems will now be stated with regard to this special form, one concerning existence and the other uniqueness, both within the framework of the theory developed up to this point.

THEOREM 7. *Assuming as a basis of the state space the union of suitable bases in the four subspaces* $\mathscr{A}(\cdot)$, $\mathscr{B}(\cdot)$, $\mathscr{C}(\cdot)$, $\mathscr{D}(\cdot)$ *of any canonical decomposition with respect to the properties of influenceability and invisibility, the equations* (1) *take*

[3] It is stressed that this theorem is stated quite generally for all decompositions implicitly defined by (19) and not just for the canonical ones.

the form

$$(27)\qquad \begin{pmatrix}\dot{x}_a(t)\\ \dot{x}_b(t)\\ \dot{x}_c(t)\\ \dot{x}_d(t)\end{pmatrix}=\begin{pmatrix}A_{aa}(t) & A_{ab}(t) & A_{ac}(t) & A_{ad}(t)\\ 0 & A_{bb}(t) & 0 & A_{bd}(t)\\ 0 & 0 & A_{cc}(t) & A_{cd}(t)\\ 0 & 0 & 0 & A_{dd}(t)\end{pmatrix}\begin{pmatrix}x_a(t)\\ x_b(t)\\ x_c(t)\\ x_d(t)\end{pmatrix}+\begin{pmatrix}B_a(t)\\ B_b(t)\\ 0\\ 0\end{pmatrix}u(t),$$

$$y(t)=(0\quad C_b(t)\quad 0\quad C_d(t))(x_a^T(t)\quad x_b^T(t)\quad x_c^T(t)\quad x_d^T(t))^T,$$

where $[x_a^T(t)\ \ 0\ \ 0\ \ 0]^T, [0\ \ x_b^T(t)\ \ 0\ \ 0]^T, [0\ \ 0\ \ x_c^T(t)\ \ 0]^T$, *and* $[0\ \ 0\ \ 0\ \ x_d^T(t)]^T$ *are coordinates of vectors belonging respectively to* $\mathscr{A}(t)$, $\tilde{\mathscr{B}}(t)$, $\tilde{\mathscr{C}}(t)$, $\tilde{\mathscr{D}}(t)$ *and the coefficient matrices are continuous* (*canonical forms*).

Remark. With reference to equations (27), the term "canonical form" will henceforth be used both for the equations themselves and for the triplet of coefficient matrices.

Proof. Let $\{z_i(t_0)\}$, $i = 1,\cdots,n$, be a basis in the state space obtained as a union of bases in the subspaces $\mathscr{A}(t_0)$, $\tilde{\mathscr{B}}(t_0)$, $\tilde{\mathscr{C}}(t_0)$, $\tilde{\mathscr{D}}(t_0)$, and let $\{T(t,t_0)z_i(t_0)\}$, $i = 1,\cdots,n$, be a basis in the state space at time t, $T(t,t_0)$ being the matrix considered in condition (iii) of Definition 3. This basis, by construction, is still a union of bases in $\mathscr{A}(t), \tilde{\mathscr{B}}(t), \tilde{\mathscr{C}}(t), \tilde{\mathscr{D}}(t)$. Consequently, the coordinate transformation $[x_a^T(t)\quad x_b^T(t)\quad x_c^T(t)\quad x_d^T(t)]^T = T(t)x(t)$ is defined at each time t, $T(t)$ being a nonsingular $n \times n$ matrix of functions of class C^1.

The vectors $[x_a^T(t)\ \ 0\ \ 0\ \ 0]^T$, $[0\ \ x_b^T(t)\ \ 0\ \ 0]^T$, $[0\ \ 0\ \ x_c^T(t)\ \ 0]^T$ and $[0\ \ 0\ \ 0\ \ x_d^T(t)]^T$ are coordinates of vectors belonging respectively to $\mathscr{A}(t)$, $\tilde{\mathscr{B}}(t), \tilde{\mathscr{C}}(t)$, $\tilde{\mathscr{D}}(t)$. Since $T(t)$ is a matrix of C^1-functions, the coefficient matrices of the representation of equations (1) in the new basis are continuous; it remains to show that they take the form (27).

As a result of the choice of the new basis in the state space, bearing in mind (7), (16) and (19), one obtains

$$(28)\qquad T(t)P(t;\bar{\eta},\bar{\zeta})T^T(t)=\begin{pmatrix}\tilde{P}_{aa} & \tilde{P}_{ab} & 0 & 0\\ \tilde{P}_{ba} & \tilde{P}_{bb} & 0 & 0\\ 0 & 0 & 0 & 0\\ 0 & 0 & 0 & 0\end{pmatrix}$$

and

$$(29)\qquad [T^{-1}(t)]^TQ(t;\bar{\eta},\bar{\zeta})T^{-1}(t)=\begin{pmatrix}0 & 0 & 0 & 0\\ 0 & \tilde{Q}_{bb} & 0 & \tilde{Q}_{bd}\\ 0 & 0 & 0 & 0\\ 0 & \tilde{Q}_{db} & 0 & \tilde{Q}_{dd}\end{pmatrix}.$$

The matrices on the main diagonals of the right-hand sides of (28) and (29) are respectively $n_a \times n_a$, $n_b \times n_b$, $n_c \times n_c$, and $n_d \times n_d$. Moreover, by virtue of Theorems 2 and 4, equations (28) and (29) also hold if one considers any pair of values η, ζ in place of $\hat{\eta}$, $\hat{\zeta}$ (in place of $\bar{\eta}$, $\bar{\zeta}$), such that $[\eta, \zeta] \supseteq [\hat{\eta}, \hat{\zeta}]$ ($[\eta, \zeta] \supseteq [\bar{\eta}, \bar{\zeta}]$).

Starting from (28) and (29) expanded in this way, it is quite easy to deduce that the coefficient matrices of the representation of equations (1) in the new basis take the form (27). For this purpose one can follow, for example, the procedure adopted in [5].

Proceeding now to deal with the problem of the uniqueness of the canonical form (27), it should first of all be pointed out that the uniqueness is to be understood with reference to a suitable equivalence relation. The starting point here lies in the fact that for the system defined by equations (1) one usually considers as equivalent representations all the equations corresponding to the triplets of coefficient matrices that are "algebraically equivalent" to the given one $[A(\cdot), B(\cdot), C(\cdot)]$ [3], [7]. Algebraic equivalence between triplets of matrices is performed by means of transformations belonging to a group; the latter may conveniently be formalized by [1]

$$\mathscr{T}_a = \{\theta : \theta[A(t), B(t), C(t)] = [T(t)A(t)T^{-1}(t) + \dot{T}(t)T^{-1}(t), T(t)B(t), C(t)T^{-1}(t)]\}, \tag{30}$$

where $T(\cdot)$ is a nonsingular $n \times n$ matrix of C^1-functions.

Let $\mathscr{R}$ be the set of all triplets of matrices that are algebraically equivalent to the triplet $[A(\cdot), B(\cdot), C(\cdot)]$ of equations (1), and let $\mathscr{R}_c$ be the subset of $\mathscr{R}$ made up of the triplets of matrices that take the canonical form (27). One can state the following theorem.

THEOREM 8. *The set $\mathscr{R}_c$ of all triplets of matrices that assume the canonical form* (27) *is an equivalence class with respect to the subgroup $\mathscr{T}_c$, of all algebraic equivalence transformations that satisfy the condition*

$$T(t) = \begin{pmatrix} T_{aa}(t) & T_{ab}(t) & T_{ac}(t) & T_{ad}(t) \\ 0 & T_{bb}(t) & 0 & T_{bd}(t) \\ 0 & 0 & T_{cc}(t) & T_{cd}(t) \\ 0 & 0 & 0 & T_{dd}(t) \end{pmatrix}, \tag{31}$$

where the matrices on the main diagonal are respectively $n_a \times n_a$, $n_b \times n_b$, $n_c \times n_c$, $n_d \times n_d$ (in other words, the canonical form of the (1) *is unique modulo the equivalence defined by $\mathscr{T}_c$).*

Proof. Given a triplet of matrices in the canonical form (27), one can immediately observe that the triplet of matrices obtained from it by means of a transformation of the type defined by (30) with condition (31), is still canonical.

It remains to demonstrate that the transformation from a triplet of matrices in canonical form to any other algebraically equivalent triplet, still in canonical form, satisfies (31). Let these triplets be $[A(\cdot), B(\cdot), C(\cdot)]$ and $[\bar{A}(\cdot), \bar{B}(\cdot), \bar{C}(\cdot)]$; the corresponding Gramian matrices $P(t; \hat{\eta}, \hat{\zeta})$ and $\bar{P}(t; \hat{\eta}, \hat{\zeta})$ have the form (28), and, since the two triplets are algebraically equivalent, are related by

$$\bar{P}(t; \hat{\eta}, \hat{\zeta}) = T(t)P(t; \hat{\eta}, \hat{\zeta})T^T(t). \tag{32}$$

Partitioning $T(t)$ in the form

$$(33) \qquad T(t) = \begin{pmatrix} [T(t)]_{11} & [T(t)]_{12} \\ [T(t)]_{21} & [T(t)]_{22} \end{pmatrix},$$

where $[T(t)]_{11}$ is $(n_a + n_b) \times (n_a + n_b)$, and performing the transformations indicated in (32), one obtains conditions that can only be satisfied by putting

$$[T(t)]_{21} = 0 \quad \text{for all } t.$$

With this result, and others that can be deduced by applying analogous arguments to the Gramian matrices $Q(t; \bar{\eta}, \zeta)$ and $\bar{Q}(t; \bar{\eta}, \zeta)$, the structure of (31) is justified.

Remark. As regards the case of time-invariant systems, it should be pointed out that the equivalence class considered in Theorem 8 comprises both the constant and nonconstant canonical forms; by virtue of the very definition of an equivalence class, one can pass from the constant to the nonconstant forms, and vice versa, by means of a transformation of the type (30). It is also easy to note that the constant canonical forms, in turn, constitute an equivalence class with respect to a constant transformation with a structure analogous to (31).

5. Conclusions. Starting from the structural properties of influenceability and invisibility, a general theory of canonical decomposition has been developed. This theory supplies subspaces of constant dimensions in all cases, i.e., it does not call for any limiting hypotheses. Furthermore, the theory yields all the previously obtained results. This is due to the fact that, in all cases in which it had hitherto proved possible to effect a decomposition into subspaces with constant dimensions, or at least in those known to the authors, the structural properties assumed for the purposes of these decompositions were such as to be implied by thė two properties, influenceability and invisibility, on which the authors based the theory presented in this paper. In the framework of this theory the authors have proved the existence of the canonical decomposition, the existence of the canonical form of the equations, and also the uniqueness of this form within an equivalence class, the latter having likewise been determined.

As regards the type of treatment adopted, it should be pointed out that account was taken of the conceptually important distinction between the problems of existence of the canonical forms and the algorithmic problems related to the construction of these forms. In tackling the problems of existence, in fact, one does not have to worry about algorithmic problems, and this makes it possible to give compact proofs as in the case examined in this paper. On the other hand, once the existence of the decompositions is proved, one can choose the most suitable algorithm in each case; in the time-invariant case, for example, since the existence of a constant canonical form is known from the general theory, it can be sought with one of the many available algorithms. It should also be stressed that it is not difficult to show that the theorem of existence (Theorem 7) can be proved with the help of Dolezal's theorem, just as is done by L. Weiss in the case of the problem he treats in [5]. In this manner, therefore, it is quite clear that none of the advantages outlined by Weiss would be lost.

It seems interesting to make a few remarks regarding the correlation between the problem of decomposition treated in this paper and the important problem of the minimal realization of a given weighting pattern. It is obvious that the weighting pattern of the system depends only on the "b" part of the canonical form of the equations. On the other hand, starting from D. C. Youla's fundamental work [7], it is easy to prove the more interesting result that the realizations of a given weighting pattern are minimal if and only if all the elements of the state space are influenceable and visible at every instant of time (i.e., if and only if they are "completely influenceable" and "completely visible").

REFERENCES

[1] P. D'Alessandro, A. Isidori and A. Ruberti, *Equivalence transformations in linear dynamical systems*, Internat. J. System Sci., 2 (1971), pp. 177–188.

[2] R. E. Kalman, *Canonical structure of linear dynamical systems*, Proc. Nat. Acad. Sci. U.S.A., 48 (1962), pp. 596–600.

[3] ———, *Mathematical description of linear dynamical systems*, this Journal, 1 (1963), pp. 152–192.

[4] R. E. Kalman, P. L. Falb and A. M. Arbib, *Topics in Mathematical System Theory*, McGraw-Hill, New York, 1969.

[5] L. Weiss, *On the structure theory of linear differential systems*, this Journal, 6 (1968), pp. 659–680.

[6] L. Weiss and R. E. Kalman, *Contribution to linear system theory*, Internat. J. Engrg. Sci., 3 (1965), pp. 141–171.

[7] D. C. Youla, *The synthesis of linear dynamical systems from prescribed weighting pattern*, SIAM J. Appl. Math., 14 (1966), pp. 527–549.

SIAM J. Control
Vol. 11, No. 1, February 1973

SUR UN PROCEDE D'OPTIMISATION UTILISANT SIMULTANEMENT LES METHODES DE PENALISATION ET DES VARIATIONS LOCALES. I*

PIERRE LORIDAN†

Abstract. We consider a strictly convex, continuous functional J defined on a real reflexive Banach space V, and we wish to minimize J on the closed, convex set K defined by

$$K = \{v \in V | G(v) \leqq 0\},$$

where G denotes a convex, continuous functional on V.

To arrive at the solution $u \in K$ of this problem, we suppose that there exists a sequence of finite-dimensional spaces V_h whose union is dense in V, and we consider the functional

$$J_h(v_h) = \begin{cases} J(v_h) & \text{if} \quad G(v_h) \leqq 0, \\ J(v_h) + \dfrac{G^2(v_h)}{r(h)} & \text{if} \quad G(v_h) > 0, \end{cases}$$

where $r(h)$ is an increasing function of h such that $\lim_{h \to 0} r(h) = 0$. Then J_h is obtained by simultaneously using penalty and discretization techniques.

By applying to J_h the method of local variations described in [3] with a step $\rho = \rho(h)$, we obtain a "stationary" point u_h^ρ, and we show that $u_h^\rho \to u$ strongly as $h \to 0$, when J is strongly convex. When the strong convexity does not hold, we apply the procedure to the "regularized" functional J_h, and we then demonstrate the weak convergence of the method.

What is of interest in this procedure is that one gives a direct approximation of u, using a single parameter h.

Introduction. L'objet de ce travail est l'étude d'un procédé d'approximation utilisant simultanément plusieurs méthodes d'optimisation de telle sorte que l'approximation soit obtenue par la variation d'un seul paramètre.

Cette étude étant assez longue nous avons jugé utile de diviser notre travail en deux parties. Seule la première partie fait l'objet de cet article, suivant le plan:

1. Hypothèses.
2. Le problème (P).
3. Description du procédé d'approximation.
4. Résultats préliminaires.
5. Convergence du procédé.
6. Applications.
7. Cas des fonctionnelles régularisées.

La deuxième partie fera l'objet d'une publication ultérieure. Elle comportera notamment l'extension des résultats précédents au cas de plusieurs contraintes ainsi qu'une étude plus particulière dans le cadre des fonctions de n variables où certaines estimations d'erreur seront fournies.

1. Hypothèses. 1. On considère une fonctionnelle J définie sur V espace de Banach réflexif, vérifiant:

(J1) J admet une dérivée-Fréchet notée $J'(u)$ pour tout $u \in V$:

$$J(v) - J(u) = (J'(u), v - u) + \omega_1(u, v - u)$$

* Received by the editors May 12, 1971, and in revised form March 15, 1972.

† Laboratoire de Bio-mathematiques, La Salpêtrière—L.E.N.A., 47 Boulevard de l'Hôpital, 75—Paris (13 ème), France.

avec $\omega_1(u, v - u) \leqq k_1(u)\|v - u\|^\alpha$, $\alpha > 1$, k_1 transformant les suites bornées en bornés. $(\cdot, \cdot)$ désigne le produit scalaire mettant en dualité V et V' dual topologique de V. $\|\cdot\|$ désigne la norme sur l'espace de Banach V.

(J2) J est foretement convexe, c'est à dire: il existe une constante $C > 0$ telle que:

$$J(v) - J(u) \geqq (J'(u), v - u) + C\|v - u\|^2, \quad \text{pour tout } v \in V, \quad \text{pour tout } u \in V.$$

(J3) $J(v) \geqq 0$ pour tout $v \in V$.

(J4) L'optimum global de J sur V est atteint en dehors du convexe K défini ci-après.

2. D'autre part, soit G une fonctionnelle définie sur V, vérifiant:

(G1) G est continue et convexe.

(G2) $\lim_{\|v\| \to +\infty} G(v) = +\infty$.

(G3) G admet une dérivée-Fréchet $G'(u)$ pour tout $u \in V$ avec $G(v) - G(u) = (G'(u), v - u) + \omega_2(u, v - u)$ et $\omega_2(u, v - u) \leqq k_2(u)\|v - u\|^\beta$, $\beta > 1$, k_2 transformant les suites bornées en bornés.

(G4) Sur tout borné B, il existe une constante $M_B > 0$, telle que $|G(v) - G(u)| \leqq M_B\|v - u\|$ pour tout $(u, v) \in B \times B$.

3. Le convexe K est défini par:

$$K = \{v \in V | G(v) \leqq 0\}.$$

On suppose vérifiée l'hypothèse:

(K1) Il existe au moins un élément $v_0 \in K$ vérifiant $G(v_0) < 0$.

Remarque 1.1. D'après les propriétés de G, K est un convexe formé et borné.

Remarque 1.2. L'hypothèse (J3) n'est pas une restriction car on peut toujours s'y ramener en ajoutant à J une constante convenable.

L'hypothèse (J4) est raisonnable sinon le problème (P) ci-après se ramènerait à un problème sans contrainte.

2. Le problème (P). Il s'agit de trouver $u \in K$ tel que $J(u) = \inf_{v \in K} J(v)$.

Remarque 2.1. D'après (J1) et (J2), J est continue et strictement convexe donc, en particulier semi-continue inférieurement pour la topologie faible de V. Le convexe K fermé et borné étant faiblement compact, J atteint donc son minimum sur K en un point u, unique en raison de la stricte convexité de J.

Le problème (P) admet donc une solution et une seule. De plus, grâce à (J4), on peut montrer que $G(u) = 0$.

3. Description du procédé d'approximation. On suppose qu'il existe une suite de sous-espaces $\{V_h\}$ de dimension finie $N(h)$ contenus dans V et dont la réunion est dense dans V (avec h paramètre positif destiné à tendre vers zéro).

Si $\{e_1, e_2, \cdots, e_{N(h)}\}$ désigne une base de V_h, tout vecteur $v_h \in V_h$ s'écrira:

$$v_h = \sum_{i=1}^{N(h)} \alpha_i e_i, \qquad \alpha_i \in R.$$

On pose $\theta(h) = \sup(\|e_1\|, \|e_2\|, \cdots, \|e_{N(h)}\|)$. Soit d'autre part $\{V_h^\rho\}$, $\rho > 0$, une suite d'ensembles dont la réunion est dense dans V_h. Plus précisément, V_h^ρ désigne l'ensemble des vecteurs de V_h s'écrivant $v_h^\rho = \sum_{i=1}^{N(h)} m_i \rho e_i$ avec $m_i \in Z$.

Dans la suite, ρ sera pris fonction de h et on posera $\rho_n = \rho(h_n)$ pour $h = h_n$, $n \in N$; $\lim_{h \to 0} \rho(h) = 0$.

On considère alors le problème (P_h) associé à la première discrétisation:

(P_h) trouver $u_h \in V_h$ tel que $J_h(u_h) = \inf_{v_h \in V_h} J_h(v_h)$ avec:

$$J_h(v_h) = \begin{cases} J(v_h) & \text{si} \quad G(v_h) \leqq 0, \\ J(v_h) + \dfrac{G^2(v_h)}{r(h)} & \text{si} \quad G(v_h) > 0, \end{cases}$$

avec $r(h)$ fonction croissante de h et $\lim_{h \to 0} r(h) = 0$. Dans la suite on posera $r_n = r(h_n)$ pour $h = h_n$, $n \in N$.

On sait que (P_h) "approche" (P) au sens donné dans [16].

En considérant alors le deuxième réseau de discrétisation, on applique la méthode des variations locales décrite par Y. Cherruault dans [7], pour approcher (P_h), mais, de plus, pour calculer le point de stationnarité $u_{r_n}^{\rho_n}$ correspondant à $h = h_n$, on part du point de stationnarité $u_{r_{n-1}}^{\rho_{n-1}}$ précédemment obtenu pour $h = h_{n-1}$ ($h_n < h_{n-1}$; $h_n \to 0$ quand $n \to +\infty$). On se propose alors de montrer que les points $u_{r_n}^{\rho_n}$, $n \in N$, "approchent" la solution u du problème (P) en un sens à préciser.

Comme ρ et r sont pris fonction de h, le procédé qui vient d'être d'écrit constitue bien une *approximation de* (P) *en une seule étape.*

Remarque 3.1. D'après ce qui précède pour calculer le point de stationnarité correspondant à $h = h_n$, on part du point de stationnarité $u_{r_{n-1}}^{\rho_{n-1}}$ précédemment obtenu pour $h = h_{n-1}$. En pratique, il faudra d'abord interpoler à partir du point $u_{r_{n-1}}^{\rho_{n-1}}$ pour obtenir un point à $N(h_n)$ coordonnées. Ceci est en particulier justifié si l'on choisit une suite $\{h_n\}$ de terme général $h_n = h_0/2^n$, $n \in N$, $h_0 > 0$ fixé, car on peut alors affirmer que le réseau de discrétisation de pas h_{n-1} est inclus dans le réseau de pas h_n, autrement dit $V_{h_{n-1}} \subset V_{h_n}$.

Remarque 3.2. La démonstration de la convergence du procédé exige la connaissance d'un certain nombre de résultats préliminaires que nous présentons sous forme de lemmes et de corollaires.

Dans ce qui suit nous poserons:

$$H(v) = \begin{cases} 0 & \text{si} \quad G(v) \leqq 0, \\ G(v) & \text{si} \quad G(v) > 0. \end{cases}$$

4. Résultats préliminaires. Avec les notations précédentes, posons, pour simplifier l'écriture, $u_{r_n}^{\rho_n} = u_n$.

LEMME 4.1. *La suite* $r_n J_{r_n}(u_n)$ *est décroissante.*

En effet, soit deux valeurs successives de $r: r_k$ et r_{k-1} avec $r_k < r_{k-1}$ et soit $H(v) = \sup(0, G(v))$. D'après le choix du processus on a:

$$J_{r_k}(u_k) \leqq J_{r_k}(u_{k-1}),$$

d'où:

$$r_k J_{r_k}(u_k) \leqq r_k J_{r_k}(u_{k-1}). \tag{4.1}$$

Mais

$$(4.2)\qquad r_k J_{r_k}(u_{k-1}) = r_k J(u_{k-1}) + H^2(u_{k-1}).$$

D'après l'hypothèse (J3), on a:

$$r_k J(u_{k-1}) < r_{k-1} J(u_{k-1})$$

et la relation (4.2) devient

$$(4.3)\qquad r_k J_{n_k}(u_{k-1}) < r_{k-1} J(u_{k-1}) + H^2(u_{k-1}).$$

Comme le second membre de (4.3) n'est autre que $r_{k-1} J_{r_{k-1}}(u_{k-1})$, la comparaison de (4.1) et de (4.3) donne le résultat annoncé $r_k J_{r_k}(u_k) < r_{k-1} J_{r_{k-1}}(u_{k-1})$.

COROLLAIRE 4.1. *Quand $n \to \infty$ (i.e., $r_n \to 0$) la suite $r_n J_{r_n}(u_n)$ converge et est en particulier bornée supérieurement.*

Ce résultat est évident car la suite est décroissante d'après le lemme 4.1 et est bornée inférieurement par 0 (conséquence de l'hypothèse (J3)).

COROLLAIRE 4.2. *La suite $H^2(u_n)$ est bornée.*

En effet, d'après (J3) on a $r_n J_{r_n}(u_n) > H^2(u_n)$ et cette inégalité montre que $H^2(u_n)$ est bornée d'après le corollaire 4.1.

COROLLAIRE 4.3. *L'ensemble des éléments $u_n = u_{r_n}^{\rho_n}$, $n \in N$, est borné.*

Preuve. En effet, si $u_n \in K$ la propriété est évidente puisque K est borné. Si $u_n \notin K$, alors d'après le corollaire 4.2 et l'hypothèse (G2), on a $\|u_n\| \leqq C$ (car dans le cas contraire on aurait $G^2(u_n) \to +\infty$ ce qui est contraire à la propriété $H^2(u_n)$ est bornée, H coïncidant avec G si $u_n \notin K$).

En particulier, comme V est réflexif, on pourra extraire de la suite u_n, une sous-suite faiblement convergente.

LEMME 4.2. *Si une suite $\{u_n\}$ converge faiblement vers s, si $G(u_n)$ est borné supérieurement par $M > 0$, pour tout n, alors sur tout borné B, il existe une constante C_B telle que*

$$H^2(v) - H^2(u_n) < 2H(u_n)(G'(u_n), v - u_n) + C_B\|v - u_n\|^{\beta_0}$$

avec $\beta_0 = \inf(\beta, 2)$, *pour tout $v \in B$ (β correspondant à l'hypothèse* (G3)).

Nous distinguerons plusieurs cas pour la démonstration et utiliserons l'hypothèse (G4).

1° *cas.* $(v, u_n) \in \complement K \times \complement K$ ou bien $(v, u_n) \in K \times \complement K$ on a alors soit $H^2(v) - H^2(u_n) = G^2(v) - G^2(u_n)$ pour la première éventualité, soit $H^2(v) - H^2(u_n) \leqq G^2(v) - G^2(u_n)$ dans la deuxième éventualité (puisqu'alors $H(v) = 0$ et $H(u_n) = G(u_n)$).

Dans ce 1er cas, nous avons donc

$$H^2(v) - H^2(u_n) \leqq G^2(v) - G^2(u_n).$$

Pour majorer $G^2(v) - G^2(u_n)$ nous utiliserons l'hypothèse (G3):

$$\begin{aligned}G^2(v) - G^2(u_n) &= [G(v) + G(u_n)][G(v) - G(u_n)]\\ &= [2G(u_n) + (G'(u_n), v - u_n) + \omega_2(u_n, v - u_n)][(G'(u_n), v - u_n) + \omega_2(u_n, v - u_n)]\\ &= 2G(u_n)(G'(u_n), v - u_n) + 2G(u_n)\omega_2(u_n, v - u_n) + [(G'(u_n), v - u_n)\\ &\qquad + \omega_2(u_n, v - u_n)];\end{aligned}$$

d'où puisque $G(u_n) = H(u_n)$ dans les deux éventualités:

$$(4.4)\qquad \begin{aligned} G^2(v) - G^2(u_n) = 2H(u_n)(G'(u_n), v - u_n) + 2G(u_n)\omega_2(u_n, v - u_n) \\ + [G(v) - G(u_n)]^2. \end{aligned}$$

On a supposé $G(u_n) < M$; de plus G étant convexe, $\omega_2(u, v - u)$ est positif. D'où: $2G(u_n)\omega_2(u_n, v - u_n) < 2Mw_2(u_n, v - u_n)$. D'autre part, en utilisant l'hypothèse (G4) on a:

$$(G(v) - G(u_n))^2 \leqq M_B^2\|v - u_n\|^2.$$

Ces deux majorations jointes à l'égalité (4.4) permettent d'en déduire une majoration de $G^2(v) - G^2(u_n)$ et donc de $H^2(v) - H^2(u_n)$, en notant que $k_2(u_n) \leqq C$:

$$H^2(v) - H^2(u_n) \leqq 2H(u_n)(G'(u_n), v - u_n) + 2MC\|v - u_n\|^\beta + M_B^2\|v - u_n\|^2,$$

inégalité que nous laissons provisoirement sous cette forme.

2° *cas.* $(v, u_n) \in \complement K \times K$, c'est à dire $H(v) = G(v) > 0$, $G(u_n) \leqq 0$, $H(u_n) = 0$. Alors

$$\begin{aligned} H^2(v) - H^2(u_n) &= (H(v) + H(u_n))(H(v) - H(u_n)) \\ &< (H(v) + H(u_n))(G(v) - G(u_n)), \end{aligned}$$

et d'autre part $H(v) + H(u_n) < G(v) - G(u_n)$ d'où

$$H^2(v) - H^2(u_n) < (G(v) - G(u_n))^2$$

soit, en utilisant l'hypothèse (G4),

$$(4.5)\qquad H^2(v) - H^2(u_n) < M_B^2\|v - u_n\|^2.$$

D'autre part, puisque $H(u_n) = 0$ on ne change pas l'inégalité (4.5) en écrivant:

$$H^2(v) - H^2(u_n) < 2H(u_n)(G'(u_n), v - u_n) + M_B^2\|v - u_n\|^2$$

et, d'autre part, comme toutes les constantes qui apparaissent sont positives, on peut aussi majorer le 2^d membre en y ajoutant $2MC\|v - u_n\|^\beta$ ce qui conduit à une écriture identique à celle du 1° cas.

3° *cas.* $(v, u_n) \in K \times K$. Alors $H^2(v) - H^2(u_n) = 0$ et l'on peut évidemment majorer $H^2(v) - H^2(u_n)$ par une expression identique à celle obtenue dans le 1° et le 2° cas. En résumé pour tout (v, u_n) vérifiant les conditions du lemme 4.2 on peut écrire:

$$H^2(v) - H^2(u_n) \leqq 2H(u_n)(G'(u_n), v - u_n) + 2MC\|v - u_n\|^\beta + M_B^2\|v - u_n\|^2.$$

En posant $\beta_0 = \inf(\beta, 2)$ on peut écrire:

$$\begin{aligned} 2MC\|v - u_n\|^\beta + M_B^2\|v - u_n\|^2 \leqq \|v - u_n\|^{\beta_0}[2MC(A + C_1)^{\beta - \beta_0} \\ + M_B^2(A + C_1)^{2 - \beta_0}], \end{aligned}$$

d'où, en notant C_B le coefficient qui apparaît dans la dernière inégalité, on a bien:

$$(4.6)\qquad H^2(v) - H^2(u_n) \leqq 2H(u_n)(G'(u_n), v - u_n) + C_B\|v - u_n\|^{\beta_0}$$

avec $\beta_0 = \inf(\beta, 2)$.

En pratique, dans les exemples rencontrés, on a $\beta_0 = \beta = 2$.

5. Convergence du procédé. Sur V_h on définit deux normes: d'une part $\|v_h\|_h = \|v_h\|$ et, d'autre part, en notant α_i la $i$ème coordonnée de v_h dans V_h (par rapport à une base $\{e_i\}$, $i = 1, 2, \cdots, N(h)$):

$$|v_h|_h = \max_{1 \leqq i \leqq N(h)} |\alpha_i|,$$

V_h étant de dimension finie, ces deux normes sont équivalentes et en particulier, il existe une constante $S_1(h)$ telle que

$$S_1(h)|v_h|_h \leqq \|v_h\|.$$

A ce sujet nous renvoyons à la remarque figurant à la fin de ce travail.

HYPOTHÈSES 5.1. Nous supposons que toutes les hypothèses du paragraphe 1 sont vérifiées et que de plus:

$$\text{(a)} \qquad \lim_{h\to 0} \frac{N(h)[\theta(h)]^{\alpha}[\rho(h)]^{\alpha-1}}{S_1(h)} = 0, \qquad \alpha > 1,$$

$$\text{(b)} \qquad \lim_{h\to 0} \frac{N(h)[\theta(h)]^{\beta_0}[\rho(h)]^{\beta_0-1}}{S_1(h)r(h)} = 0, \qquad \beta_0 > 1,$$

(c) $\theta(h)\rho(h)$ est borné quand $h \to 0$ (en pratique cette condition sera souvent réalisée par le choix: $\lim_{h\to 0} \theta(h)\rho(h) = 0$). $\theta(h)$ a la signification donnée au paragraphe 3.

Comme nous serons amenés à majorer $H^2(u_n + \rho e_i) - H^2(u_n)$ et $H^2(u_n - \rho e_i) - H^2(u_n)$ dans les démonstrations qui suivent (pour une sous-suite u_n extraite de la suite $u_n = u_{r_n}^{\rho_n}$ définie au paragraphe 3), nous allons d'abord montrer que l'on peut appliquer les résultats du lemme 4.2.

LEMME 5.1. *$\|u_n + \rho e_i\|$ et $\|u_n - \rho e_i\|$ sont bornées pour tous les e_i vecteurs de base V_h.*

En effet:

$$\|u_n + \rho e_i\| \leqq \|u_n\| + \|\rho e_i\| \leqq \|u_n\| + \rho(h)\theta(h).$$

Or: d'après la condition (c), $\rho(h)\theta(h) < M$ et comme $\|u_n\|$ est bornée, on en déduit que $\|u_n + \rho e_i\|$ est bornée. Même remarque pour $\|u_n - \rho e_i\|$.

Remarque 5.1. On sait que $G(u_n)$ est borné d'après le corollaire 4.2. D'après le lemme 5.1, $u_n + \rho e_i$ et $u_n - \rho e_i$ sont dans un borné: on pourra donc appliquer le lemme 4.2 en faisant successivement $v = u_n + \rho e_i$ et $v = u_n - \rho e_i$ chaque fois que la sous-suite u_n est faiblement convergente.

THÉORÈME 5.1. *La solution u_h du problème* (P$_h$) *défini au paragraphe* 3 *converge fortement vers u et $J_h(u_h) \to J(u)$ quand $h \to 0$.*

Ce théorème est une simple adaptation d'un résultat que nous avons donné dans [16]. Nous en donnons brièvement la démonstration pour la commodité du lecteur.

Soit $\bar{u}_h$ le point réalisant le minimum de J sur le convexe $K_h = K \cap V_h$. En utilisant la continuité de J et de G et le fait que l'intérieur de K n'est pas vide (en vertu de l'hypothèse (K1)) on montre facilement avec l'hypothèse de densité de $\bigcup_{h>0} V_h$ dans V que: $J(\bar{u}_h) \to J(u)$ et que $\bar{u}_h \to u$ fortement quand $h \to 0$ (la convergence forte venant de la forte convexité de J).

En considérant ensuite la fonctionnelle J_h (paragraphe 3) on a d'une part $J(\bar{u}_h) = J_h(\bar{u}_h)$ et d'autre part d'après la définition de u_h: $J_h(u_h) \leqq J_h(\bar{u}_h)$, ce qui donne encore $J(u_h) \leqq J(\bar{u}_h)$. Comme $J(\bar{u}_h)$ est bornée, il en est de même de $J(u_h)$ et puisque $\lim_{\|v\| \to +\infty} J(v) = +\infty$ (en raison de la forte convexité de J), on déduit par un raisonnement classique que $\|u_h\|$ est bornée et donc que l'on peut extraire une sous-suite $u_{h'}$ convergeant faiblement vers un élément s quand $h' \to 0$. On montre ensuite que $s \in K$ (cf. [16]) et l'on a: $J(u) \leqq J(s) \leqq \lim_{h' \to 0} \inf J(u_{h'}) \leqq \lim_{h' \to 0} J(\bar{u}_{h'}) = J(u)$. Il en résulte que $J(u_{h'}) \to J(u)$ et que $s = u$. Reprenant alors le même raisonnement pour n'importe quelle sous-suite faiblement convergente, on montre que la suite u_h elle-même converge vers u (fortement, en raison de la forte convexité de J).

Enfin pour montrer que $J_h(u_h) \to J(u)$ quand $h \to 0$, on utilise l'inégalité:

$$J(u_h) \leqq J_h(u_h) \leqq J(\bar{u}_h),$$

ce qui achève la démonstration.

Remarque 5.2. Nous utiliserons ce théorème dans la démonstration suivante. D'autre part, lorsque la forte convexité de J n'est pas vérifiée, mais en supposant $\lim_{\|v\| \to +\infty} J(v) = +\infty$, on a la convergence faible de u_h vers u: ce résultat interviendra dans l'étude du cas des fonctionnelles régularisées (paragraphe 7).

THÉORÈME 5.2. *Si toutes les hypothèses* 5.1 *sont vérifiées, alors la suite des points de stationnarité* $u_{r_n}^{\rho_n}$ *converge fortement vers* u *solution du problème* (P) *quand* $n \to +\infty$.

Démonstration. Posons $u_n = u_{r_n}^{\rho_n}$ pour simplifier l'écriture. En un point de stationnarité u_n, on a par définition:

$$\begin{aligned} 0 &\leqq J_{r_n}(u_n + \rho e_i) - J_{r_n}(u_n), \\ 0 &\leqq J_{r_n}(u_n - \rho e_i) - J_{r_n}(u_n). \end{aligned} \qquad i = 1, 2, \cdots, N(h),$$

Notons que $J_r'(v) = J'(v) + (2/r)H(v)G'(v)$. Alors, en utilisant la remarque 5.1 (autrement dit le lemme 4.2 et l'hypothèse (J1)), on a:

$$0 \leqq J_{r_n}(u_n + \rho e_i) - J_{r_n}(u_n) \leqq \rho(J_{r_n}'(u_n), e_i) = \omega_1(u_n, \rho e_i) + \frac{C_B\|\rho e_i\|^{\beta_0}}{r_n},$$

$$0 \leqq J_{r_n}(u_n - \rho e_i) - J_{r_n}(u_n) \leqq -\rho(J_{r_n}'(u_n), e_i) + \omega_1(u_n, -\rho e_i) + \frac{C_B\|\rho e_i\|^{\beta_0}}{r_n}.$$

On en déduit:

$$-\frac{\omega_1(u_n, \rho e_i)}{\rho} - \frac{C_B\|\rho e_i\|^{\beta_0}}{\rho r_n} \leqq (J_{r_n}'(u_n), e_i) \leqq \frac{\omega_1(u_n, \rho e_i)}{\rho} + \frac{C_B\|\rho e_i\|^{\beta_0}}{\rho r_n}$$

soit, en tenant compte de l'hypothèse (J1) et de l'écriture de $J_r'(v)$:

$$(5.1) \qquad (J'(u_n), e_i) + \frac{2}{r_n}H(u_n)(G'(u_n), e_i) \leqq C_1\rho^{\alpha-1}\|e_i\|^{\alpha} + C_B\frac{\rho^{\beta_0-1}}{r_n}\|e_i\|^{\beta_0}, \qquad i = 1, 2, \cdots, N(h),$$

où l'on a posé $\rho = \rho_n = \rho(h_n)$.

Soit u_h la solution du problème (P_h) défini au paragraphe 3. Pour $h = h_n$ nous noterons encore u_h la solution pour simplifier l'écriture. Cet élément est caractérisé par:

$$(J'(u_h), v_h) + \frac{2H(u_h)}{r(h)}(G'(u_h), v_h) = 0 \quad \text{pour tout} \quad v_h \in V_h.$$

Si nous retranchons cette expression du 1er membre de (5.1) avec $v_h = e_i$, nous obtenons:

$$\begin{aligned}(J'(u_n) - J'(u_h), e_i) + \frac{2}{r_n}[H(u_n)(G'(u_n), e_i) - H(u_h)(G'(u_h), e_i)]\\ \leqq C_1\rho^{\alpha-1}\|e_i\|^{\alpha} + C_B\frac{\rho^{\beta_0 - 1}}{r_n}\|e_i\|^{\beta_0}, \qquad i = 1, 2, \cdots, N(h),\end{aligned} \tag{5.2}$$

que nous pouvons encore noter de manière plus condensée:

$$(J_r'(u_n) - J_r'(u_h), e_i) \leqq C_1\rho^{\alpha-1}\|e_i\|^{\alpha} + C_B\frac{\rho^{\beta_0 - 1}}{r_n}\|e_i\|^{\beta_0}, \tag{5.3}$$

et en utilisant $\theta(h)$:

$$(J_r'(u_n) - J_r'(u_h), e_i) \leqq C_1\rho^{\alpha-1}[\theta(h)]^{\alpha} + C_B\frac{\rho^{\beta_0 - 1}}{r_n}[\theta(h)]^{\beta_0}. \tag{5.4}$$

Considérons maintenant l'élément $u_n - u_h$; cet élément appartenant à V_h on peut trouver des nombres réels λ_i, $i = 1, 2, \cdots, N(h)$, tels que:

$$u_n - u_h = \sum_{i=1}^{N(h)} \lambda_i e_i. \tag{5.5}$$

La fonctionnelle J_r étant fortement convexe, on remarque que:

$$2C\|u_n - u_h\|^2 \leqq (J_r'(u_n) - J_r'(u_h), u_n - u_h) \tag{5.6}$$

et en utilisant (5.5):

$$2C\|u_n - u_h\| \leqq (J_r'(u_n) - J_r'(u_h), \sum_{1}^{N(h)} \lambda_i e_i). \tag{5.7}$$

On peut majorer le second membre de (5.7) grâce à la relation (5.4), ce qui donne:

$$\begin{aligned}(J_r'(u_n) - J_r'(u_h), \sum_{1}^{N(h)} \lambda_i e_i) \leqq \sum_{1}^{N(h)} \lambda_i(J_r'(u_n) - J_r'(u_h), e_i)\\ \leqq N(h)\sup|\lambda_i|\left[C_1\rho^{\alpha-1}(\theta(h))^{\alpha} + C_B\frac{\rho^{\beta_0 - 1}}{r}(\theta(h))^{\beta_0}\right].\end{aligned} \tag{5.8}$$

Mais, par définition,

$$\sup_i|\lambda_i| = |u_n - u_h|_h \leqq \frac{\|u_n - u_h\|}{S_1(h)}.$$

En reportant dans (5.8), puis revenant à (5.7) on obtient après division par

$\|u_n - u_h\|$:

$$2C\|u_n - u_h\| \leqq \frac{N(h)}{S_1(h)}\left[C_1\rho^{\alpha-1}(\theta(h))^\alpha + C_B\frac{\rho^{\beta_0-1}}{r}(\theta(h))^{\beta_0}\right]. \tag{5.9}$$

En utilisant alors les conditions (a) et (b) figurant dans les hypothèses 5.1, on remarque que le second membre de (5.9) tend vers zéro et donc

$$\|u_n - u_h\| \to 0 \quad \text{quand} \quad h_n \to 0.$$

D'autre part, d'après le théorème 5.1 on a $\|u_h - u\| \to 0$ quand $h \to 0$. Alors puisque $\|u_n - u\| \leqq \|u_n - u_h\| + \|u_h - u\|$ on en déduit la convergence forte de u_n vers u quand $n \to +\infty$ (c'est à dire, quand $h_n \to 0$) ce qui achève la démonstration.

COROLLAIRE 5.1. *$J(u_n) \to J(u)$ et $H(u_n) \to 0$ quand $n \to +\infty$.*

C'est une conséquence de la continuité de J et de H.

Remarque 5.3. Ce qui précède montre l'importance des fonctionnelles fortement convexes pour la démonstration de la convergence du procédé.

Lorsque la forte convexité n'est pas réalisée, on modifie le procédé en régularisant la fonctionnelle pénalisée, le terme de régularisation dépendant lui-même du paramètre de discrétisation: avant d'en faire l'étude, nous présentons des applications des résultats précédents.

6. Applications. Parmi les applications du procédé à la résolution de problèmes d'optimisation convexe nous pouvons signaler plus particulièrement l'approximation de la solution de certaines inéquations variationnelles.

Plus précisément, soit V un espace de Hilbert réel, de norme notée $\|\cdot\|$, soit V' son dual topologique, le produit scalaire mettant V et V' en dualité étant noté $(\cdot,\cdot)$.

Soit $a(u, v)$ une forme bilinéaire sur V, continue, symétrique et coercive, la coercivité se traduisant par: "Il existe une constante $C > 0$ telle que $a(v, v) \geqq C\|v\|^2$ pour tout $v \in V$." Soit K un convexe fermé borné défini par une fonctionnelle G vérifiant les hypothèses du paragraphe 1 (l'exemple le plus simple est fourni par $K = \{v \in V | \, \|v\|^2 \leqq M, M > 0\}$).

Le problème est le suivant: f étant donné dans V', résoudre l'inéquation variationnelle $a(u, v - u) \geqq (f, v - u)$ pour tout $v \in K$.

Il est bien connu que ce problème équivaut à la minimisation de la fonctionnelle $J(v) = a(v, v) - 2(f, v)$ sur K.

On vérifie facilement que J est dérivable et fortement convexe. De plus, suivant la remarque 1.2, on peut toujours se ramener au cas $J(v) \geqq 0$. Enfin, si le problème est bien posé, il est sous-entendu que l'hypothèse (J4) est vérifiée: en effet, dans le cas contraire, l'inéquation variationnelle se ramènerait à l'équation $a(u, v) = (f, v)$ pour tout $v \in V$ et le problème serait sans contraintes.

Les conditions de convergence du procédé sont donc remplies (avec d'ailleurs $\alpha = 2$).

7. Cas des fonctionnelles régularisées.

HYPOTHÈSES 7.1. Nous supposons vérifiées toutes les hypothèses du paragraphe 1 *sauf* l'hypothèse (J2) de forte convexité que nous remplaçons par les conditions:

(J5) $\lim_{\|v\|\to\infty} J(v) = +\infty$,

(J6) J est strictement convexe.

De plus on suppose que l'on peut approcher V comme au paragraphe 3, et que V est un Hilbert.

Description du procédé. Au lieu d'appliquer la méthode des variations locales à la fonctionnelle J_h définie au paragraphe 3, on va l'appliquer à la fonctionnelle régularisée:

$$J_{h,\varepsilon}(v_h) = J_h(v_h) + \varepsilon(h)\|v_h\|^2,$$

où $\varepsilon(h)$ est une fonction positive croissante de h vérifiant $\lim_{h\to 0} \varepsilon(h) = 0$.

On vérifie facilement que $J_{h,\varepsilon}(v_h)$ est dérivable et fortement convexe sur V_h pour tout h donné. De plus, cette fonctionnelle admet un minimum unique sur V_h et nous écrirons:

$$J_{h,\varepsilon}(u_\varepsilon) = \inf_{v_h \in Vh} J_{h,\varepsilon}(v_h),$$

où, pour simplifier l'écriture, nous avons posé $\varepsilon = \varepsilon(h)$. Nous noterons u_ε^ρ le point de stationnarité de la fonctionnelle $J_{h;\varepsilon}$ obtenu par application de la méthode des variations locales pour $\rho = \rho(h)$ et nous nous proposons de montrer la convergence faible de u_ε^ρ vers u quand $h \to 0$.

Cette démonstration utilisera le résultat intermédiaire suivant.

THÉORÈME 7.1. *Quand* $h \to 0$:

(a) $J_h(u_\varepsilon) \to J(u)$,

(b) $J_{h,\varepsilon}(u_\varepsilon) \to J(u)$,

(c) $u_\varepsilon \to u$ *faiblement.*

Démonstration. Rappelons d'abord que la solution u_h du problème (P_h):

$$J_h(u_h) = \inf_{v_h \in Vh} J_h(v_h)$$

vérifie:

$$J_h(u_h) \to J(u) \quad \text{quand} \quad h \to 0,$$

$$u_h \to u \quad \text{faiblement}$$

(cf. la remarque 5.2 faisant suite au théorème 5.1). Alors, par définition, on a:

$$J_h(u_h) \leqq J_h(u_\varepsilon) \leqq J_{h,\varepsilon}(u_\varepsilon) \leqq J_{h,\varepsilon}(u_h),$$

la dernière inégalité provenant du fait que u_ε réalise le minimum de $J_{h,\varepsilon}(v_h)$. Comme $u_h \to u$ faiblement, la suite u_h est faiblement bornée, donc fortement bornée et, d'après le choix de $\varepsilon(h)$, il en résulte que $\varepsilon(h)\|u_h\|^2 \to 0$ quand $h \to 0$.

D'autre part, $J_h(u_h) \to J(u)$ quand $h \to 0$ et, puisque $J_{h,\varepsilon}(u_h) = J_h(u_h) + \varepsilon(h)\|u_h\|^2$, il en résulte que: $J_{h,\varepsilon}(u_h) \to J(u)$ quand $h \to 0$.

Ayant vu précédemment que: $J_h(u_h) \leqq J_h(u_\varepsilon) \leqq J_{h,\varepsilon}(u_\varepsilon) \leqq J_{h,\varepsilon}(u_h)$ on en déduit facilement les résultats (a) et (b) du théorème.

De l'inégalité $J_{h,\varepsilon}(u_\varepsilon) \leqq J_{h,\varepsilon}(u_h)$ on tire:

$$\varepsilon(h)\|u_\varepsilon\|^2 \leqq J_h(u_h) + \varepsilon(h)\|u_h\|^2 - J_h(u_\varepsilon)$$

et puisque $J_h(u_h) - J_h(u_\varepsilon) \leqq 0$, on en déduit:

$$\varepsilon(h)\|u_\varepsilon\|^2 \leqq \varepsilon(h)\|u_h\|^2,$$

d'où $\|u_\varepsilon\|^2 \leqq \|u_h\|^2$.

Il en résulte que $\|u_\varepsilon\|$ est aussi bornée et puisque V est réflexif, on peut extraire de la suite u_ε une sous-suite $u_{\varepsilon'}$ convergeant faiblement vers s quand $h' \to 0$ (on a posé $\varepsilon' = \varepsilon(h')$).

Montrons d'abord que $s \in K$: en effet dans le cas contraire, il existerait $\lambda > 0$ tel que $H^2(u_{\varepsilon'}) \geqq \lambda$ quand $h' \to 0$ et en conséquence on aurait $H^2(u_{\varepsilon'})/r(h') \to +\infty$ quand $h' \to 0$ et $J_{h'}(u_{\varepsilon'})$ tendrait vers $+\infty$ ce qui contradirait l'assertion (b).

Par conséquent on doit avoir $s \in K$. Montrons maintenant que $s = u$: J convexe et continue est aussi semi-continue inférieurement pour la topologie faible de V donc:

$$\liminf_{h'\to 0} J(u_{\varepsilon'}) \geqq J(s) \geqq J(u).$$

Mais, d'autre part $\lim_{h'\to 0} \sup J(u_{\varepsilon'}) \leqq \lim_{h'\to 0} J_{h'}(u_{\varepsilon'}) = J(u)$. Par conséquent, $\lim_{h'\to 0} J(u_{\varepsilon'}) = J(u)$ et $J(s) = J(u)$. Comme u est unique, l'égalité $J(s) = J(u)$ entraîne $s = u$ puisque l'on a déjà montré que $s \in K$.

On peut alors reprendre le même raisonnement pour n'importe quelle sous-suite faiblement convergente extraite de la suite u_ε; il en résulte, suivant un raisonnement classique que la suite u_ε elle-même converge faiblement vers u quand $h \to 0$, ce qui achève la démonstration.

COROLLAIRE 7.1. *Quand* $h \to 0$,

(a) $J(u_\varepsilon) \to J(u)$,
(b) $H^2(u_\varepsilon)/r(h) \to 0$,
(c) $H(u_\varepsilon) \to 0$.

La première partie (a) vient de la démonstration du théorème précédent. Le résultat (b) est une conséquence de (a) et de la convergence de $J_h(u_\varepsilon)$ vers $J(u)$. Enfin, le résultat (c) vient de (b) et du fait que $r(h) \to 0$. Démontrons maintenant la convergence du procédé utilisant la méthode des variations locales. Dans ce qui suit, rappelons que u_ε^ρ désigne un point de stationnarité et que les notations sont celles rencontrées dans les paragraphes précédents.

THÉORÈME 7.2. *Si en plus des hypothèses* 7.1, *on suppose* $\theta(h)\rho(h)$ *borné et*:

$$\text{(a)} \qquad \lim_{h\to 0} \frac{N(h)\rho(h)[\theta(h)]^2}{S_1(h)} = 0,$$

$$\text{(b)} \qquad \lim_{h\to 0} \frac{N(h)[\rho(h)]^{\alpha-1}[\theta(h)]^\alpha}{S_1(h)\varepsilon(h)} = 0, \qquad \alpha > 1,$$

$$\text{(c)} \qquad \lim_{h\to 0} \frac{N(h)[\rho(h)]^{\beta_0-1}[\theta(h)]^{\beta_0}}{S_1(h)\varepsilon(h)r(h)} = 0, \quad \beta_0 = \min(\beta, 2), \qquad \beta > 1;$$

alors la suite $u_\varepsilon^\rho \to u$ *faiblement quand* $h \to 0$.

Remarque 7.1. Les conditions précédentes se simplifient lorsque $\alpha = \beta = 2$:

Il suffit alors de réaliser la condition (c) qui s'écrit:

$$\lim_{h\to 0} \frac{N(h)\rho(h)[\theta(h)]^2}{S_1(h)\varepsilon(h)r(h)} = 0;$$

en effet, comme $r(h) \to 0$, on en déduit que

$$\frac{N(h)\rho(h)[\theta(h)]^2}{S_1(h)\varepsilon(h)}$$

doit tendre vers zéro ("plus vite" que $r(h)$), ce qui réalise la condition (b) où $\alpha = 2$. De même la condition (b) étant réalisée avec $\varepsilon(h) \to 0$ on en déduit la réalisation de la condition (a).

Nous reviendrons donc qu'habituellement la condition (c) est la plus importante.

Démonstration du théorème 7.2. Nous suivrons les calculs faits dans la démonstration du théorème 5.2. En désignant par $\langle\, ,\rangle$ le produit scalaire sur l'espace de Hilbert V, on peut écrire:

$$\|u_\varepsilon\|^2 - \|u_\varepsilon^\rho\|^2 = 2\langle u_\varepsilon^\rho, u_\varepsilon - u_\varepsilon^\rho\rangle + \|u_\varepsilon - u_\varepsilon^\rho\|^2$$

et par conséquent:

$$(J'_{h,\varepsilon}(u_\varepsilon^\rho), u_\varepsilon^\rho - u_\varepsilon) = (J'_h(u_\varepsilon^\rho), u_\varepsilon^\rho - u_\varepsilon) + 2\varepsilon(h)\langle u_\varepsilon^\rho, u_\varepsilon^\rho - u_\varepsilon\rangle.$$

De plus le point u_ε minimisant $J_{h,\varepsilon}(v_h)$ sur V_h vérifie

$$(J'_{h,\varepsilon}(u_\varepsilon), u_\varepsilon^\rho - u_\varepsilon) = 0$$

et

$$(J'_{h,\varepsilon}(u_\varepsilon^\rho), u_\varepsilon^\rho - u_\varepsilon) = (J'_{h,\varepsilon}(u_\varepsilon^\rho) - J'_{h,\varepsilon}(u_\varepsilon), u_\varepsilon^\rho - u_\varepsilon).$$

On peut minorer le second membre de cette dernière égalité par $2\varepsilon(h)\|u_\varepsilon^\rho - u_\varepsilon\|^2$ car la fonctionnelle $J_{h,\varepsilon}(v_h)$ est fortement convexe en raison de la présence de terme de régularisation. D'après ce qui précède on a donc:

$$2\varepsilon(h)\|u_\varepsilon^\rho - u_\varepsilon\|^2 \leqq (J'_{h,\varepsilon}(u_\varepsilon^\rho), u_\varepsilon^\rho - u_\varepsilon).$$

On montre ensuite, en suivant exactement le type de calculs développés dans la démonstration du théorème 5.2 que:

$$\begin{aligned}&(J'_{h,\varepsilon}(u_\varepsilon^\rho), u_\varepsilon^\rho - u_\varepsilon)\\ &\qquad \leqq \frac{N(h)}{S_1(h)}\left(C_1\rho^{\alpha-1}(\theta(h))^\alpha + C_B\frac{\rho^{\beta_0-1}}{r}(\theta(h))^{\beta_0} + \varepsilon(h)\rho(h)(\theta(h))^2\right)\|u_\varepsilon^\rho - u_\varepsilon\|.\end{aligned}$$

Puis en comparant les deux inégalités on obtient finalement après division par $2\varepsilon(h)\|u_\varepsilon^\rho - u_\varepsilon\|$:

$$\|u_\varepsilon^\rho - u_\varepsilon\| \leqq \frac{N(h)}{2\varepsilon(h)S_1(h)}\left(C_1\rho^{\alpha-1}(\theta(h))^\alpha + C_B\frac{\rho^{\beta_0-1}}{r(h)}(\theta(h))^{\beta_0}\right) + \frac{N(h)\rho(h)(\theta(h))^2}{S_1(h)}$$

et en utilisant les conditions (a), (b), (c) du théorème, on a:

$$\|u_\varepsilon^\rho - u_\varepsilon\| \to 0 \quad \text{quand} \quad h \to 0.$$

Enfin, pour achever la démonstration, on utilise le théorème 7.1 assurant la convergence faible de u_ε vers u.

Comme $\|u_\varepsilon^\rho - u_\varepsilon\| \to 0$ on en déduit que $u_\varepsilon^\rho \to u$ faiblement quand $h \to 0$.

Remarque 7.2. Toutes les démonstrations précédentes demeurent valables si, à l'espace V, on associe une suite d'espaces V_h de dimension finie $N(h)$ ainsi qu'une suite d'opérateurs linéaires continus injectifs p_h tels que $\bigcup_{h>0} p_h V_h$ soit dense dans V (voir [1]).

Outre la norme $|v_h|_h$ déjà définie sur V_h, on peut prendre $\|v_h\|_h = \|p_h v_h\|$. Sur V_h ces deux normes sont équivalentes et il existe deux constantes $S_1(h)$ et $S_2(h)$ telles que:

$$S_1(h)|v_h|_h \leqq \|v_h\|_h \leqq S_2(h)|v_h|_h.$$

Dans les conditions de convergence (a), (b), (c) des hypothèses 5.1, il suffit alors de remplacer $\theta(h)$ par $S_2(h)$ et la convergence de $u_n = r_{r_n}^{\rho_n}$ vers u est alors à prendre au sens: $p_{h_n} u_n \to u$ quand $n \to +\infty$.

Remarque 7.3. Tout ce qui précède complète l'étude que nous avons faite dans [17].

Remarque 7.4. Au paragraphe 4, nous avons utilisé la propriété $\lim_{\|v\| \to +\infty} G(v) = +\infty$ pour démontrer que la suite des points de stationnarité est bornée. Lorsque cette condition n'est pas vérifiée, il y a lieu de modifier la fonctionnelle pénalisée: cette étude sera présentée dans la deuxième partie.

Pour d'autres résultats nous renvoyons à [18].

REFERENCES

[1] J. P. Aubin, *Approximation des espaces de distributions et des opérateurs différentiels*, Bulletin de la Société Mathématique de France, Mémoire n° 12, 1967.

[2] A. Auslender, *Méthodes numériques pour la résolution des problèmes d'optimisation avec contraintes*, Thèse, Université de Grenoble, 1969.

[3] A. V. Balakrishnan, *On a new computing technique in optimal control*, this Journal, 6 (1968), pp. 149–173.

[4] A. Bensoussan and P. Kenneth, *Sur l'analogie entre les méthodes de régularisation et de pénalisation*, Revue Française d'Informatique et de Recherche Opérationnelle, Série Rouge, 13 (1968), pp. 13–26.

[5] N. Bourbaki, *Espaces vectoriels topologiques*, Hermann, Paris, 1953.

[6] J. Céa, *Optimisation, théorie et algorithmes*, Dunod, Paris, 1971.

[7] Y. Cherruault, *Une méthode directe de minimisation et applications*, Revue d'Informatique et de Recherche Opérationnelle, 10 (1968), pp. 31–52.

[8] J. W. Daniel, *On the approximate minimization of functionals*, Math. Comp., 23 (1969), pp. 573–581.

[9] ———, *On the convergence of a numerical method for optimal control problems*, J. Optimization Theory and Applications, 4 (1969), pp. 330–342.

[10] ———, *The Approximate Minimization of Functionals*, Prentice-Hall, Englewood Cliffs, New Jersey, 1971.

[11] S. de Julio, *Numerical solution of dynamical optimization problems*, this Journal, 8 (1970), pp. 135–147.

[12] A. Fiacco and G. McCormick, *Non-Linear Programming: Sequential Unconstrained Minimization Techniques*, John Wiley, New York, 1968.

[13] P. Haarhoff and J. Buys, *A new method for the optimization of a non-linear function subject to non-linear constraints*, Comput. J., 13 (1970), pp. 178–184.

[14] P. Kenneth, M. Sibony and J. P. Yvon, *La méthode de pénalisation et ses applications aux problèmes de controle optimal*, Cahiers de l'I.R.I.A., n⁰ 2, Rocquencourt, France, 1970.

[15] J. L. LIONS, *Quelques méthodes de résolution des problèmes aux limites non linéaires*, Gauthier-Villars, Paris, 1969.

[16] P. LORIDAN, *Approximation de problème d'optimisation avec contraintes*, Rapport, Département de Mathématiques, Université de Lille, Lille, 1967.

[17] ———, *Sur la minimisation de fonctionnelles convexes par pénalisation*, Revue d'Informatique et de Recherche Opérationnelle, R-1 (1971), pp. 117–133.

[18] P. LORIDAN, Thèse, Université de Paris VI, to appear.

[19] G. MEYER AND E. POLAK, *Abstract models for the synthesis of optimization algorithms*, this Journal, 9 (1971), pp. 547–560.

[20] E. POLAK, *Computational Methods in Optimization: A Unified Approach*, Academic Press, New York, 1970.

[21] A. TIKHONOV, *Methods for the regularization of optimal control problems*, Soviet Math. Dokl., 6 (1965), pp. 761–763.

[22] M. M. VAINBERG, *Variational Methods for the Study of Nonlinear Operators*, Holden-Day, San Francisco, 1964.

[23] D. WILDE, *Méthodes de recherche d'un optimum*, Dunod, Paris, 1966.

[24] J. P. YVON, *Application de la pénalisation à la résolution d'un problème de controle optimal*, Cahiers de l'I.R.I.A., n^0 2, Rocquencourt, France, 1970.

SIAM J. Control
Vol. 11, No. 1, February 1973

SUR UN PROCEDE D'OPTIMISATION UTILISANT SIMULTANEMENT LES METHODES DE PENALISATION ET DES VARIATIONS LOCALES. II*

PIERRE LORIDAN†

Abstract. This paper completes the study made in a preceding article [18]. Here, we consider optimization problems with several constraints, i.e., we seek to minimize a functional J on a closed, convex set K in a real, reflexive Banach space, K being defined in terms of q functionals G_i, $i = 1, \cdots, q$:

$$K = \{v \in V | G_i(v) \leqq 0, \quad i = 1, 2, \cdots, q\}.$$

To arrive at the solution u, we consider

$$J_{r(h)}(v_h) = J(v_h) + \frac{1}{r(h)} \sum_{i=1}^{q} H_i^2(v_h),$$

where

$$H_i(v_h) = \begin{cases} 0 & \text{if} \quad G_i(v_h) \geqq 0, \\ G_i(v_h) & \text{if} \quad G_i(v_h) > 0, \end{cases} \qquad i = 1, \cdots, q,$$

and $v_h \in V_h$, a finite-dimensional space satisfying $\bigcup_{h>0} V_h$ is dense in V. By applying the method of local variations to $J_{r(h)}(v_h)$, we demonstrate the strong convergence of the procedure as in [18]. Finally, in the case of functions of N variables, we give error estimates and show that the convergence can be of the order of r.

Introduction. Dans un précédent article (cf. [18]) nous avons étudié un procédé d'approximation de problèmes d'optimisation convexe avec une seule contrainte. Nous montrons, dans cette deuxième partie, qu'il est possible d'étendre les résultats au cas de plusieurs contraintes. Puis nous nous plaçons dans le cadre des fonctions n variables où il est facile de présenter des estimations d'erreurs concernant l'utilisation du procédé.

Le plan d'étude de cette seconde partie est le suivant:

1. Hypothèses.
2. Le problème (P).
3. Description du procédé d'approximation.
4. Résultats préliminaires.
5. Convergence du procédé.
6. Cas des fonctionnelles régularisées.
7. Cas des fonctions de N variables.
8. Estimations d'erreur.

1. Hypothèses. Soit K un convexe fermé dans un espace de Banach réflexif V dont la norme est notée $\| \cdot \|$.

1. On considère une fonctionnelle J définie sur V et vérifiant les hypothèses (déjà données dans [18] et que nous rappelons ici pour la commodité du lecteur):

(J1) J admet une dérivée-Fréchet notée $J'(u)$ pour tout $u \in V$:

$$J(v) - J(u) = (J'(u), v - u) + \omega_1(u, v - u)$$

* Received by the editors May 12, 1971, and in revised form March 15, 1972.

† Laboratoire de Bio-mathematiques, La Salpêtrière—L.E.N.A., 47 Boulevard de l'Hôpital, 75—Paris (13 ème), France.

avec $\omega_1(u, v - u) \leqq k_1(u)\|v - u\|^\alpha$, $\alpha > 1$, k_1 transformant les suites bornées en bornés. $(\cdot, \cdot)$ désigne le produit scalaire mettant en dualité V et V' dual topologique de V. $\|\cdot\|$ désigne la norme sur l'espace de Banach V.

(J2) J est fortement convexe, c'est à dire: il existe une constante $C > 0$ telle que:

$$J(v) - J(u) \geqq (J'(u), v - u) + C\|v - u\|^2, \quad \text{pour tout} \quad v \in V, \quad \text{pour tout} \quad u \in V.$$

(J3) $J(v) \geqq 0$ pour tout $v \in V$.

(J4) L'optimum global de J sur V est atteint en dehors du convexe K.

2. Soient q fonctionnelles G_i, $i = 1, 2, \cdots, q$, définies sur V vérifiant:

(G1) G_i est continue, convexe $(i = 1, 2, \cdots, q)$.

(G2) Il existe au moins un indice i pour lequel $\lim_{\|v\| \to +\infty} G_i(v) = +\infty$.

(G3) G_i admet une dérivée de Fréchet $G_i'(u)$ pour tout $u \in V$ avec $G_i(v) - G_i(u) = (G_i'(u), v - u) + \omega_i(u, v - u)$ et $\omega_i(u, v - u) \leqq k_i(u)\|v - u\|^{\beta_i}$, $\beta_i > 1$, $k_i(u)$ transformant les suites bornées en bornés, $i = 1, 2, \cdots, q$.

(G4) Sur tout borné B il existe une constante $M_{B,i}$, $i = 1, 2, \cdots, q$, telle que $|G_i(v) - G_i(u)| \leqq M_{B,i}\|v - u\|$, $i = 1, 2, \cdots, q$.

3. Ces fonctionnelles définissent le convexe K de la manière suivante:

$$K = \{v \in V | G_i(v) \leqq 0, i = 1, 2, \cdots, q\};$$

nous supposerons $K \neq \emptyset$.

2. Le problème (P). Il s'agit de trouver $u \in K$ tel que $J(u) \leqq J(v)$ pour tout $v \in K$. Comme le convexe K est fermé et borné en raison des hypothèses (G1) et (G2), l'hypothèse (J1) assure l'existence de u (continuité de J) tandis que l'hypothèse (J2) assure l'unicité.

De plus, en utilisant l'hypothèse (J4) on montre que u appartient à la frontière du convexe K.

On se propose d'approcher la solution u du problème (P) par un procédé dont la description faite au paragraphe suivant est une extension au cas de plusieurs contraintes de ce que nous avons déjà présenté dans [18].

3. Description du procédé d'approximation. On suppose qu'il existe une suite de sous-espaces $\{V_h\}$ de dimension finie $N(h)$, contenus dans V et dont la réunion est dense dans V (avec h paramètre positif destiné à tendre vers zéro). Si $\{e_1, e_2, \cdots, e_{N(h)}\}$ désigne une base de V_h, tout vecteur $v_h \in V_h$ s'écrira:

$$v_h = \sum_{i=1}^{N(h)} \alpha_i e_i, \qquad \alpha_i \in R.$$

On pose $\theta(h) = \sup(\|e_1\|, \|e_2\|, \cdots, \|e_{N(h)}\|)$. Soit d'autre part $\{V_h^\rho\}$, $\rho > 0$, une suite d'ensembles dont la réunion est dense dans V_h. Plus précisément, V_h^ρ désigne l'ensemble des vecteurs de V_h s'écrivant $v_h^\rho = \sum_{i=1}^{N(h)} m_i \rho\, e_i$, avec $m_i \in Z$. Dans la suite ρ sera pris fonction de h et on posera $\rho_n = \rho(h_n)$, $\rho_{n+1} < \rho_n$ si $h_{n+1} < h_n$; $\lim_{h \to 0} \rho(h) = 0$.

On considère le problème (P_h) associé à la première discrétisation:

(P_h) trouver $u_h \in V_h$ tel que $J_{r(h)}(u_h) = \inf_{v_h \in V_h} J_{r(h)}(v_h)$, où la fonctionnelle

$J_{r(h)}$ est définie par:

$$J_{r(h)}(v_h) = J(v_h) + \frac{1}{r(h)} \sum_{i=1}^{q} H_i^2(v_h)$$

avec:

$$H_i(v_h) = \begin{cases} 0 & \text{si} \quad G_i(v_h) \leqq 0, \\ G_i(v_h) & \text{si} \quad G_i(v_h) > 0, \end{cases} \qquad i = 1, 2, \cdots, q,$$

où $r(h)$ est une fonction croissante de h telle que $\lim_{h \to 0} r(h) = 0$. Dans la suite on posera $r_n = r(h_n)$, $n \in N$.

On sait que (P_h) "approche" (P) au sens suivant (cf. [16]):

$$\left.\begin{array}{ll} u_h \to u & \text{faiblement} \\ J(u_h) \to J(u) & \end{array}\right\} \quad \text{quand} \quad h \to 0.$$

De plus $u_h \to u$ fortement en raison de la forte convexité de J.

En considérant alors le deuxième réseau de discrétisation relatif au paramètre $\rho = \rho(h)$, on applique la méthode des variations locales décrite par Y. Cherruault dans [7] pour approcher la solution u_h de (P_h). De plus, pour calculer le point de stationnarité u_n correspondant à $h = h_n$, on part du point de stationnarité u_{n-1} précédemment obtenu pour $h = h_{n-1}$. En pratique il faudra interpoler à partir du point u_{n-1} pour obtenir un point à $N(h_n)$ coordonnées. Ceci est en particulier justifié si l'on choisit une suite $\{h_n\}$ de terme général $h_n = h_0/2^n$, $n \in N$, car on peut alors affirmer que le réseau de discrétisation de pas h_{n-1} est inclus dans le réseau de pas h_n. Comme dans [18], on se propose de montrer que les points de stationnarité u_n, $n \in N$, "approchent" la solution u quand $n \to +\infty$, le procédé décrit réalisant ainsi une approximation de (P) en une seule étape.

4. Résultats préliminaires.

LEMME 4.1. *La suite $r_n J_{r_n}(u_n)$ est décroissante.*

(Démonstration analogue à celle du lemme 4.1 de [18], en posant $H^2(v) = \sum_{i=1}^{q} H_i^2(v)$.)

LEMME 4.2. *La suite $H^2(u_n)$ est bornée.*

En effet, d'après l'hypothèse (J3), on a $r_n J_{r_n}(u_n) \geqq H^2(u_n)$ et le résultat se déduit alors du lemme 4.1.

COROLLAIRE 4.1. *L'ensemble des u_n est borné.*

En effet si $u_n \in K$ la propriété est évidente puisque K est borné. Si $u_n \notin K$, alors le résultat est une conséquence du lemme 4.2 car $\lim_{\|v\| \to +\infty, v \notin K} H^2(v) = +\infty$ (d'après (G2)).

Remarque 4.1. En particulier, comme V est un Banach réflexif, on pourra extraire de la suite u_n une sous-suite faiblement convergente que nous noterons encore u_n.

LEMME 4.3. *Si la suite $\{u_n\}$ converge faiblement, alors sur tout borné B_i, il existe une constante C_{B_i} telle que:*

$$H_i^2(v) - H_i^2(u_n) < 2H_i(u_n)(G_i'(u_n), v - u_n) + C_{B_i}\|v - u_n\|^{\alpha_i}$$

pour tout $v \in B_i$, $\alpha_i = \inf(\beta_i, 2)$, β_i correspondant à l'hypothèse (G3), $i = 1, 2, \cdots, q$ (adaptation du lemme 4.2 présenté dans [18]).

5. Convergence du procédé. Sur V_h on définit deux normes: d'une part $\|v_h\|_h = \|v_h\|$ et, d'autre part, en notant γ_i la $i$ème coordonnée de v_h dans V_h (par rapport à une base $\{e_i\}$, $i = 1, 2, \cdots, N(h)$):

$$|v_h|_h = \max_{i \leqq i \leqq N(h)} |\gamma_i|.$$

V_h étant de dimension finie, ces deux normes sont équivalentes et en particulier, il existe une constante $S_1(h)$ telle que

$$S_1(h)|v_h|_h \leqq \|v_h\|.$$

HYPOTHÈSES 5.1. Nous supposons vérifiées les hypothèses du paragraphe 1 et de plus:

$$\text{(a)} \qquad \lim_{h \to 0} \frac{N(h)[\theta(h)]^{\alpha}[\rho(h)]^{\alpha-1}}{S_1(h)} = 0, \qquad \alpha > 1,$$

$$\text{(b)} \qquad \lim_{h \to 0} \frac{N(h)[\theta(h)]^{\alpha_i}[\rho(h)]^{\alpha_i-1}}{S_1(h)r(h)} = 0, \qquad \alpha_i > 1, \quad i = 1, 2, \cdots, q.$$

(c) $\theta(h)\rho(h)$ est borné quand $h \to 0$.

Rappelons que $\theta(h) = \sup(\|e_1\|, \|e_2\|, \cdots, \|e_{N(h)}\|)$.

THÉORÈME 5.1. *Avec les hypothèses précédentes la suite des points de stationnarité u_n converge fortement vers u quand $n \to +\infty$.*

La démonstration est analogue à celle développée au paragraphe 5 de [18]. En suivant le même type de calculs on montre d'une part que l'on a:

$$(J'_{r_n}(u_n), u_n - u_h) \leqq \frac{N(h)}{S_1(h)}\|u_n - u_h\| \left(C_1 \rho^{\alpha-1}(\theta(h))^{\alpha} + \frac{1}{r_n} \sum_{i=1}^{q} C_{B,i} \rho^{\alpha_i - 1}(\theta(h))^{\alpha_i} \right),$$

où u_h est la solution du problème (P_h).

Utilisant d'autre part la caractérisation de u_h et la forte convexité de $J_{r(h)}(v_h)$ on montre que:

$$2C\|u_n - u_h\|^2 \leqq (J'_{r_n}(u_n), u_n - u_h).$$

En comparant les deux inégalités précédentes on obtient après division par $\|u_n - u_h\|$:

$$2C\|u_n - u_h\| \leqq \frac{N(h)}{S_1(h)} \left(C_1 \rho^{\alpha-1}(\theta(h))^{\alpha} + \frac{1}{r_n} \sum_{i=1}^{q} C_{B,i} \rho^{\alpha_i - 1}(\theta(h))^{\alpha_i} \right).$$

En utilisant alors les conditions (a) et (b) des hypothèses 5.1, on en déduit que: $\|u_n - u_h\| \to 0$ quand $n \to +\infty$ (c'est à dire, quand $h = h_n \to 0$). Ayant par ailleurs rappelé au paragraphe 3 que $u_h \to u$ fortement on en déduit que $\|u_n - u\| \to 0$ quand $n \to +\infty$.

Remarque 5.1. Grâce au théorème précédent, on est désormais en mesure de démontrer la convergence du procédé dans le cas d'une seule contrainte $G(v)$ lorsque la condition $\lim_{\|v\| \to +\infty} G(v) = +\infty$ n'est pas réalisée.

Il suffit en effet de faire l'hypothèse supplémentaire: sur tout borné B, il existe une constante $M_B > 0$ telle que

$$|J(v) - J(u)| \leqq M_B\|v - u\| \quad \text{pour tout} \quad (u, v) \in B \times B$$

et d'appliquer le procédé décrit au paragraphe 3 à la fonctionnelle

$$J_r(v) = J(v) + \frac{H^2(v)}{r} + \frac{S^2(v)}{r}$$

avec:

$$H(v) = \begin{cases} 0 & \text{si} \quad G(v) \leqq 0, \\ G(v) & \text{si} \quad G(v) > 0, \end{cases}$$

$$S(v) = \begin{cases} 0 & \text{si} \quad J(v) - J(v_0) \leqq 0, \\ J(v) - J(v_0) & \text{si} \quad J(v) > J(v_0), \end{cases}$$

où v_0 est un élément fixé arbitrairement dans $K = \{v \in V | G(v) \leqq 0\}$. En posant $G_1(v) = G(v)$, $G_2(v) = J(v) - J(v_0)$ on voit que le problème ainsi défini est un problème à deux contraintes G_i, $i = 1, 2$, les G_i vérifiant alors toutes les hypothèses (G1) à (G4) (paragraphe 1).

Il en résulte que l'on peut se passer de l'hypothèse $\lim_{\|v\| \to +\infty} G(v) = +\infty$ dans l'étude que nous avons présentée dans [18].

Remarque 5.2. Dans le cas où J n'est pas fortement convexe on peut régulariser le problème d'optimisation comme dans [18]: nous en rappelons brièvement le principe au paragraphe suivant ainsi que les conditions de convergence.

6. Cas des fonctionnelles régularisées.

HYPOTHÈSES 6.1. Nous supposons vérifiées toutes les hypothèses du paragraphe 1 sauf l'hypothèse de forte convexité que nous remplaçons par les conditions:

(J5) $\lim_{\|v\| \to +\infty} J(v) = +\infty$,

(J6) J est strictement convexe.

De plus, on suppose que l'on peut approcher V comme au paragraphe 3 et on se place dans le cas où V est un Hilbert.

Description du procédé. Au lieu d'appliquer la méthode des variations locales à la fonctionnelle $J_{r(h)}(v_h)$ définie au paragraphe 3, on va l'appliquer à la fonctionnelle régularisée:

$$J_{h,\varepsilon}(v_h) = J_{r(h)}(v_h) + \varepsilon(h)\|v_h\|^2,$$

où $\varepsilon(h)$ est une fonction positive, croissante de h vérifiant $\lim_{h \to 0} \varepsilon(h) = 0$. Comme dans [18], on sera amené à utiliser l'élément $u_\varepsilon \in V_h$ vérifiant

$$J_{h,\varepsilon}(u_\varepsilon) = \inf_{v_h \in V_h} J_{h,\varepsilon}(v_h).$$

On a un théorème analogue au théorème 7.1 de [18]:

THÉORÈME 6.1. *Quand* $h \to 0$:

(a) $J_h(u_\varepsilon) \to J(u)$,

(b) $J_{h,\varepsilon}(u_\varepsilon) \to J(u)$,

(c) $u_\varepsilon \to u$ *faiblement.*

Nous renvoyons à [18] pour une démonstration précise.

THÉORÈME 6.2. *Si en plus des hypothèses* 6.1, *on suppose* $\theta(h)\rho(h)$ *borné et*:

(a) $$\lim_{h\to 0}\frac{N(h)\rho(h)[\theta(h)]^2}{S_1(h)} = 0,$$

(b) $$\lim_{h\to 0}\frac{N(h)[\rho(h)]^{\alpha-1}[\theta(h)]^{\alpha}}{S_1(h)\varepsilon(h)} = 0, \qquad \alpha > 1,$$

(c) $$\lim_{h\to 0}\frac{N(h)[\rho(h)]^{\alpha_i-1}[\theta(h)]^{\alpha_i}}{S_1(h)\varepsilon(h)r(h)} = 0,$$

$$\alpha_i = \min(\beta_i, 2), \quad \beta_i > 1, \quad i = 1, 2, \cdots, q,$$

alors la suite $u_\varepsilon^\rho \to u$ *faiblement quand* $h \to 0$ (*où* u_ε^ρ *désigne le point de stationnarité de* $J_{h,\varepsilon}(v_h)$ *obtenu par application de la méthode des variations locales pour* $\rho = \rho(h)$).

La démonstration est analogue à celle développée dans [18]. De plus, comme dans [18], on peut simplifier les conditions précédentes dans le cas $\alpha = \beta_i = 2$, $i = 1, 2, \cdots, q$: il suffit alors de poser

$$\lim_{h\to 0}\frac{N(h)\rho(h)[\theta(h)]^2}{S_1(h)\varepsilon(h)r(h)} = 0.$$

Remarque 6.1. Ce qui précède est encore valable si à l'espace V on associe une suite d'espaces V_h de dimension finie $N(h)$ ainsi qu'une suite d'opérateurs linéaires continus injectifs p_h tels que $\bigcup_{h>0} p_h V_h$ soit dense dans V (voir [1]).

Outre la norme $|v_h|_h$ déjà définie sur V_h, on peut prendre $\|v_h\|_h = \|p_h v_h\|$. Sur V_h ces deux normes sont équivalentes et il existe deux constantes $S_1(h)$ et $S_2(h)$ telles que:

$$S_1(h)|v_h|_h \leqq \|v_h\|_h \leqq S_2(h)|v_h|_h.$$

Dans les conditions de convergence qui précèdent il suffit alors de remplacer $\theta(h)$ par $S_2(h)$ et les résultats de convergence sont à prendre au sens $p_h u_n \to u$ quand $n \to \infty$.

7. Cas des fonctions de N variables.

HYPOTHÈSES 7.1.

(J1) Soit J une fonction de N variables: $J: R^N \to R$, continue, strictement convexe.

(J2) L'optimum global de J est en dehors de K (convexe défini plus loin).

(J3) $J(v) \geqq 0$ pour tout $v \in R^N$.

(J4) J est continuement différentiable avec: $J(v) - J(u) = J'(u)\cdot(v - u) + \omega_1(u, v - u)$, $J'(u)\cdot(v - u)$ désignant le produit scalaire des vecteurs $v - u$ et $J'(u)$ (R^N étant identifié à son dual). De plus, on suppose qu'il existe $\alpha > 1$ tel que $\omega_1(u, v - u)/\|v - u\|^\alpha \to 0$ quand $\|v - u\| \to 0$.

Soit G une fonction de N variables sur R^N, vérifiant:

(G1) $G: R^N \to R$, continue, convexe.

(G2) $\lim_{\|v\|\to\infty} G(v) = +\infty$.

(G3) G est continuement différentiable (donc aussi G^2) et l'on suppose que l'on peut écrire: $G^2(v) - G^2(u) = 2G(u)G'(u)\cdot(v - u) + \omega_2(u, v - u)$ avec l'existence d'un nombre $\beta > 1$ tel que $\omega_2(u, v - u)/\|v - u\|^\beta \to 0$ quand $\|v - u\| \to 0$.

Remarquons qu'une telle écriture peut s'obtenir à partir d'une hypothèse analogue faite sur G: il suffit de reprendre les calculs faits dans le cadre des fonctionnelles, au paragraphe 4 (cf. [18]).

Le convexe K est défini par:

(K1) $K = \{v \in R^N | G(v) \leqq 0\}$ et on suppose qu'il existe au moins un élément v_0 vérifiant $G(v_0) < 0$. On cherche toujours à approcher la solution u du problème (P): trouver $u \in K$ tel que $J(u) \leqq J(v)$ pour tout $v \in K$.

Approximation de (P). L'approximation décrite ici est plus simple que dans le cas des fonctionnelles puisqu'il n'y a qu'une seule discrétisation, provenant de l'utilisation de la méthode des variations locales:

On considère la fonction:

$$J_r(v) = J(v) + \frac{1}{r}[\sup(0, G(v))]^2, \qquad r > 0,$$

et on cherche une approximation de l'optimum de J_r par la méthode des variations locales en prenant $\rho = \rho(r)$, $\rho(r) \to 0$ quand $r \to 0$. De plus, comme dans le cas des fonctionnelles, pour minimiser $J_{r_k}(v)$ on part du point de stationnarité précédemment obtenu, soit $u_{K-1} = u_{r_{k-1}}^{\rho_{k-1}}$. Les lemmes et corollaires du paragraphe 4 se transposent aisément et on peut donc montrer que la suite $\{u_k\}$ des points de stationnarité est bornée, c'est à dire que l'on peut extraire une sous-suite encore notée u_k qui converge vers s dans R^N (nous supposerons par exemple R^N muni de la norme euclidienne). Posons $\rho_n = \rho(r_n)$ et $u_n = u_{r_n}^{\rho_n}$. On a alors le théorème suivant.

THÉORÈME DE CONVERGENCE. *Si les hypothèses* 7.1 *sont vérifiées et si*:

$$\lim_{n\to\infty} \frac{r_n}{\rho_n^{\beta-1}} \to \lambda \neq 0, \qquad \lambda \in R^*,$$

alors

$$u_n \to u \quad \textit{solution de } (\mathrm{P}).$$

Preuve. Soit $\{e_i\}$, $i = 1, 2, \cdots, N$, une base de R^N. En un point de stationnarité u_n, des calculs analogues à ceux faits pour les fonctionnelles, donnent:

$$(7.1) \qquad |J'_{r_n}(u_n)\cdot e_i| \leqq \frac{|\omega_1(u_n, \rho_n e_i)|}{\rho_n} + \frac{|\omega_2(u_n, \rho\, e_i)|}{\rho_n r_n}$$

et pour $v \in R^N$, on peut trouver des $v_i \in R$, $|v_i| \leqq C$ tels que $v = \sum_1^N v_i e_i$. Alors en utilisant l'inégalité (7.1):

$$(7.2) \qquad \begin{aligned} |J'_{r_n}(u_n)\cdot v| &< \sum_1^N v_i |J'_{r_n}(u_n)\cdot e_i| \leqq \cdots \\ &\leqq CN \sup \frac{|\omega_1(u_n, \rho_n e_i)|}{\rho_n} + \frac{|\omega_2(u_n, \rho_n e_i)|}{\rho_n r_n}. \end{aligned}$$

Alors d'après (J4), $|\omega_1(u_n, \rho_n e_i)|/\rho_n \to 0$ pour tout i; de même, pour tout i, on aura $|\omega_2(u_n, \rho_n e_i)|/\rho_n r_n \to 0$ car $\omega_2/\rho_r = \omega_2/\rho^\beta \cdot \rho^{\beta-1}/r$; or $\omega_2/\rho^\beta \to 0$ d'après (G3) et $\rho^{\beta-1}/r \to 1/\lambda \in R$ d'après l'hypothèse faite dans l'énoncé du théorème. Il en résulte que le dernier membre de l'inégalité (7.2) tend vers zéro et donc $|J'_{r_n}(u_n)\cdot v| \to 0$.

Nous allons maintenant expliciter cette propriété, pour vérifier les conditions d'optimalité sur K.

Montrons d'abord que $H(u_n)/r_n$ admet une limite finie, (avec $H(v) = \sup(0, G(v))$). Montrons que *l'approximation obtenue* est externe, c'est à dire que $u_n \in \complement K$ au moins à partir d'un certain rang: en effet, dans le cas contraire, si u_n appartenait à K pour tout n assez grand on aurait $H(u_n) = 0$ et on aurait:

$$J'_r(u_n) = J'(u_n)$$

et d'après ce qui précède $|J'(u_n) \cdot v|$ tendrait vers zéro pour tout v quand $n \to +\infty$. A la limite on aurait donc $J'(s) \cdot v = 0$ pour tout v et s serait l'optimum global de J. Comme d'autre part K est fermé on devrait aussi avoir $s \in K$, ce qui contredirait l'hypothèse (J4) donc on a bien $u_n \in \complement K$ quand $n \to +\infty$.

D'où: $H(u_n) = G(u_n) > 0$ et, comme G est continu, $G(u_n) \to G(s) \geqq 0$. D'autre part, l'hypothèse (K1) permet d'affirmer que l'optimum global de G est à l'intérieur de K en raison de la convexité de G. Comme s ne vérifie pas $G(s) < 0$, s ne peut être l'optimum de G et donc $G'(s) \neq \theta$ (θ désigne le vecteur nul dans R^N). En conséquence, il existe au moins un élément $v \in R^N$, tel que $G'(s) \cdot v \neq 0$. Considérons alors l'inégalité:

$$\left\| J'(u_n) \cdot v \right| - \frac{2}{r_n} H(u_n) \left| G'(u_n) \cdot v \right\| < |J'_{r_n}(u_n) \cdot v|. \tag{7.3}$$

On a déjà vu que le second membre $\to 0$.

D'autre part, d'après les hypothèses (J4) et (G3), la continuité des gradients donne:

$$J'(u_n) \cdot v \to J'(s) \cdot v,$$

$$G'(u_n) \cdot v \to G'(s) \cdot v \neq 0$$

d'après ce qui précède. Grâce à (7.3), on en déduit alors que $(2/r_n)H(u_n) \to \lambda_0 > 0$, λ_0 fini. Cette limite est d'ailleurs indépendante de v puisque u_n ne dépend pas de v. Donc, on peut affirmer que pour tout $v \in R^N$, il existe $\lambda_0 > 0$ tel que:

$$J'(s) \cdot v + \lambda_0 G'(s) \cdot v = 0 \quad \text{pour tout} \quad v.$$

Comme $(2/r_n)H(u_n) \to \lambda_0$ et que $r_n \to 0$, on peut affirmer que $G(u_n) \to 0$ (puisque $H(u_n) = G(u_n)$). Comme $G(u_n) \to G(s)$ il en résulte que $G(s) = 0$. Donc s vérifie les conditions suffisantes d'optimalité sur K et par conséquent $s = u$. On en déduit alors que la suite u_n elle-même converge vers u ce qui achève la démonstration du théorème.

Corollaire 7.1.

(a) $J(u_n) \to J(u)$,

(b) $G^2(u_n)/r_n \to 0$,

(c) $J_{r_n}(u_n) \to J(u)$.

La première partie (a) résulte de la continuité de J. Pour le résultat (b), on remarque que $G^2(u_n)/r_n = G(u_n) \cdot G(u_n)/r_n$ or $G(u_n) \to 0$ et $G(u_n)/r_n \to \lambda_0/2 \in R^+$, d'où le résultat. Enfin le résultat (c) est une conséquence des deux précédents et de la définition de J_r.

Nous nous proposons maintenant de donner des estimations de l'erreur commise en remplaçant le calcul de u par u_n. Nous allons nous placer dans le cas où G est fortement convexe. Démontrons au préalable le lemme suivant.

LEMME 7.1. *Si G est fortement convexe sur $\complement K$, c'est à dire si*

$$G(v) - G(u) \geqq G'(u)\cdot(v-u) + \alpha\|v-u\|^2, \qquad \alpha > 0,$$

alors:

$$G^2(v) - G^2(u) \geqq 2G(u)G'(u)\cdot(v-u) + 2G(u)\alpha\|v-u\|^2$$

$$\textit{pour tout} \quad (u,v)\in\complement K \times \complement K.$$

Preuve. $(u,v)\in\complement K\times\complement K \Rightarrow G(v)+G(u) > 0$ et

$$\begin{aligned} G^2(v) - G^2(u) &\geqq [G(v)+G(u)][G'(u)\cdot(v-u)+\alpha\|v-u\|^2] \\ &\geqq [2G(u)+G'(u)\cdot(v-u)+\alpha\|v-u\|^2][G'(u)\cdot(v-u)+\alpha\|v-u\|^2] \\ &\geqq 2G(u)G'(u)\cdot(v-u)+2G(u)\alpha\|v-u\|^2+[G'(u)\cdot(v-u)+\alpha\|v-u\|^2]^2 \end{aligned}$$

et comme le dernier terme est positif on en déduit

$$G^2(v) - G^2(u) \geqq 2G(u)G'(u)\cdot(v-u) + 2G(u)\alpha\|v-u\|^2. \tag{7.4}$$

8. Estimations d'erreur. Pour effectuer de telles estimations nous passons par l'intermédiaire de u_r, optimum global de J_r, c'est à dire:

$$J_r(u_r) = \inf_{v\in R^N} J_r(v).$$

Nous avons donné dans [17] des propriétés de u_r et celles-ci s'adaptent aisément au cas des fonctions de N variables. En particulier, on sait que $u_r \in \complement K$ ainsi que u_n. Appliquons, l'inégalité (7.4) pour de tels éléments, en faisant successivement $v = u_n$, $u = u_r$, puis $v = u_r$, $u = u_n$:

$$G^2(u_n) - G^2(u_r) \geqq 2G(u_r)G'(u_r)\cdot(u_n-u_r) + 2\alpha G(u_r)\|u_n-u_r\|^2,$$

$$G^2(u_r) - G^2(u_n) \geqq 2G(u_n)G'(u_n)\cdot(u_r-u_n) + 2\alpha G(u_n)\|u_n-u_r\|^2.$$

En faisant la somme on obtient:

$$0 \geqq [2G(u_n)G'(u_n) - 2G(u_n)G'(u_n)]\cdot(u_n-u_r) + 2\alpha\|u_n-u_r\|^2[G(u_r)+G(u_n)], \tag{8.1}$$

d'où:

$$[G(u_n)G'(u_n) - G(u_r)G'(u_r)]\cdot(u_n-u_r) \geqq \alpha\|u_n-u_r\|^2[G(u_r)+G(u_n)]. \tag{8.2}$$

D'autre part, J étant convexe strictement, J' est strictement monotone, c'est à dire:

$$[J'(u_n) - J'(u_r)]\cdot(u_n-u_r) > 0. \tag{8.3}$$

Multiplions l'inégalité (8.2) par $2/r$ et ajoutons à (8.3): nous faisons alors apparaître le gradient de J_r:

$$J_r'(v) = J'(v) + \frac{2}{r}G(v)G'(v)$$

et, compte-tenu du fait que $J_r'(u_r) = 0$, les inégalités (8.2) et (8.3) ainsi combinées, conduisent à:

$$J'_r(u_n)\cdot(u_n - u_r) \geqq \frac{2\alpha}{r}\|u_n - u_r\|^2[G(u_r) + G(u_n)].$$

Or, on sait que $J'_r(u_n)\cdot v \to 0$ pour tout $v \in R^N$ donc $\|J'_r(u_n)\| \to 0$ et en majorant le 1er membre de l'inégalité précédente par $\|J'_r(u_n)\| \cdot \|u_n - u_r\|$, on obtient

$$\|J'_r(u_n)\| \geqq \frac{2\alpha}{r}\|u_n - u_r\|\,[G(u_n) + G(u_r)]. \tag{8.4}$$

Comme

$$\lim_{r\to 0}\frac{G(u_r)}{r} = \lim_{r_n\to 0}\frac{G(u_n)}{r_n} = \frac{\lambda_0}{2} > 0,$$

cette inégalité jointe à la propriété $\|J'_r(u_n)\| \to 0$ montre que $\|u_n - u_r\| \to 0$ et comme $u_r \to u$ (cf. [17]) on en déduit une autre démonstration de convergence du procédé.

Mais de plus, on peut fournir, grâce à cette inégalité, une estimation de l'erreur $\|u_n - u\|$: l'inégalité (8.4) donne en effet:

$$\|u_n - u_r\| \leqq \frac{r}{2\alpha}\frac{\|J'_r(u_n)\|}{G(u_n) + G(u_r)} \leqq \frac{r}{2\alpha}\frac{\|J'_r(u_n)\|}{G(u_n)}. \tag{8.5}$$

D'autre part, on a montré dans [17] que:

$$\|u_r - u\| < \sqrt{\frac{G(u_r)}{\alpha}}. \tag{8.6}$$

Comme $\|u_n - u\| \leqq \|u_n - u_r\| + \|u_r - u\|$, on en déduit, d'après (8.5) et (8.6):

$$\|u_n - u\| \leqq \frac{r_n}{2\alpha}\frac{\|J'_r(u_n)\|}{G(u_n)} + \sqrt{\frac{G(u_r)}{\alpha}}. \tag{8.7}$$

Remarque importante. Le deuxième membre de (8.7) dépendant de u_r qui est inconnu, cette estimation peut paraître illusoire à première vue, c'est à dire inutilisable en pratique. En fait, il n'en est rien car dans [17], nous avons montré qu'il est possible de déterminer r pour que $G(u_r)/\alpha$ soit $< 10^{-p}$, $p \in N$, ce qui montre l'intérêt particulier de la formule (8.7). En pratique, il suffira de faire décroître r jusqu'à la valeur désirée, on en déduira la valeur de ρ correspondante ainsi que le point de stationnarité u_n.

On peut, si on le désire, fournir une estimation d'erreur sur J dans le cas où on suppose:

$$|J(v) - J(u)| \leqq |J'(u)\cdot(v - u)| + C\|v - u\|^\alpha, \qquad \alpha > 1.$$

Cette inégalité appliquée à $v = u$ et $u = u_n$ donne:

$$|J(u) - J(u_n)| \leqq \|J'(u_n\| \cdot \|u - u_n\| + C\|u - u_n\|^\alpha$$

et on se sert alors de l'évaluation (8.7).

Estimation d'erreur dans le cas où G^2 est fortement convexe sur $\complement K$. On suppose donc qu'il existe $\alpha > 0$ tel que:

$$[G(v)G'(v) - G(u)G'(u)] \cdot (v - u) \geqq 2\alpha\|v - u\|^2$$

pour tout $(u, v) \in \complement K \times \complement K$. Alors en reprenant ce qui précéde, on obtient: $\|J'_r(u_n)\| \geqq (2\alpha/r)\|u_n - u_r\|$ et comme $\|J'_r(u_n)\| \to 0$ on en déduira que $\|u_n - u_r\| = o(r)$ ce qui donne une indication sur la rapidité de convergence car d'autre part, on peut montrer que l'on a: $G^2(u_r) \geqq \alpha\|u_r - u\|^2$ (car $G(u) = 0$); et comme $G^2(u_r)/r^2 \to (\lambda_0/2)^2$, $\|u_r - u\| = O(r)$, d'où $\|u_n - u\| = O(r)$.

On trouvera d'autres résultats dans [19].

REFERENCES

[1] J. P. Aubin, *Approximation des espaces de distributions et des opérateurs différentiels*, Bulletin de la Société Mathématique de France, Mémoire n^0 12, 1967.

[2] A. Auslender, *Méthodes numériques pour la résolution des problèmes d'optimisation avec contraintes*, Thèse, Université de Grenoble, 1969.

[3] A. V. Balakrishnan, *On a new computing technique in optimal control*, this Journal, 6 (1968), pp. 149–173.

[4] A. Bensoussan and P. Kenneth, *Sur l'analogie entre les méthodes de régularisation et se pénalisation*, Revue Française d'Informatique et de Recherche Opérationnelle, Série Rouge, 13 (1968), pp. 13–26.

[5] N. Bourbaki, *Espaces vectoriels topologiques*, Hermann, Paris, 1953.

[6] J. Céa, *Optimisation, théorie et algorithmes*, Dunod, Paris, 1971.

[7] Y. Cherruault, *Une méthode directe de minimisation et applications*, Revue d'Informatique et de Recherche Opérationnelle, 10 (1968), pp. 31–52.

[8] J. W. Daniel, *On the approximate minimization of functionals*, Math. Comp., 23 (1969), pp. 573–581.

[9] ———, *On the convergence of a numerical method for optimal control problems*, J. Optimization Theory and Applications, 4 (1969), pp. 330–342.

[10] ———, *The Approximate Minimization of Functionals*, Prentice-Hall, Englewood Cliffs, N.J., 1971.

[11] S. de Julio, *Numerical solution of dynamical optimization problems*, this Journal, 8 (1970), pp. 135–147.

[12] A. Fiacco and G. McCormick, *Non-Linear Programming: Sequential Unconstrained Minimization Techniques*, John Wiley, New York, 1968.

[13] P. Haarhoff and J. Buys, *A new method for the optimization of a non-linear function subject to non-linear constraints*, Comput. J., 13 (1970), pp. 178–184.

[14] P. Kenneth, M. Sibony and J. P. Yvon, *La méthode de pénalisation et ses applications aux problèmes de controle optimal*, Cahiers de l'I.R.I.A., n^0 2, Rocquencourt, France, 1970.

[15] J. L. Lions, *Quelques méthodes de résolution des problèmes aux limites non linéaires*, Gauthier-Villars, Paris, 1969.

[16] P. Loridan, *Approximation de problème d'optimisation avec contraintes*, Rapport, Département de Mathématiques, Université de Lille, 1968.

[17] ———, *Sur la minimisation de fonctionnelles convexes par pénalisation*, Revue d'Informatique et de Recherche Opérationnelle, R-1 (1971), pp. 117–133.

[18] ———, *Sur un procédé d'optimisation utilisant simultanément les méthodes de pénalisation et des variations locales I*, this Journal, 11 (1973), pp. 159–172.

[19] ———, Thèse, Université de Paris VI, to appear.

[20] G. Meyer and E. Polak, *Abstract models for the synthesis of optimization algorithms*, this Journal, 9 (1971), pp. 547–560.

[21] E. Polak, *Computational Methods in Optimization: A Unified Approach*, Academic Press, New York, 1970.

[22] A. TIKHONOV, *Methods for the regularization of optimal control problems*, Soviet Math. Dokl., 6 (1965), pp. 761–763.

[23] D. WILDE, *Méthodes de recherche d'un optimum*, Dunod, Paris, 1966.

[24] J. P. YVON, *Application de la pénalisation à la résolution d'un problème de controle optimal*, Cahiers de l'I.R.I.A., n^0 2, Rocquencourt, France, 1970.

SIAM J. CONTROL
Vol. 11, No. 1, February 1973

FURTHER COMMENTS ON THE PAPER, "OPTIMAL CONTROL OF PROCESSES DESCRIBED BY INTEGRAL EQUATIONS. I" BY V. R. VINOKUROV*

V. R. VINOKUROV†

The remarks in [1] on the sufficiency of the maximum principle were brought forth by an inaccurate translation of my paper. In the original article [2], the following definition is given on page 23:

We shall say that $x(t)$, $u(t)$ satisfy the maximum principle if there exist a constant vector $\mu = (\mu_1, \mu_2, \cdots, \mu_l)$ and a piecewise-smooth scalar function $\lambda(t)$ such that, for an optimal trajectory $x(t)$ and a control $u(t)$ for which the functionals (1.4) vanish, relations (2.1)–(2.6) are satisfied, and, for almost all $t \in [0, T]$,

$$H(x, z, u, t) = \min_{v \in \omega(x)} H(x, z, v, t), \tag{2.7}$$

$$d\lambda(t)/dt \leqq 0. \tag{2.8}$$

In the translation of my article [3], on page 327, line 1, the words "optimal trajectory $x(t)$ and a control $u(t)$" were mistranslated as "optimal trajectory $x(t)$ and optimal control $u(t)$." Thus the word "optimal" in the translation modifies both "trajectory" and "control," whereas, in the original, it modified only trajectory. For this reason, the sufficiency result makes no sense.

Actually, an optimal trajectory can be realized by a nonoptimal control. The sufficiency part of the maximum principle is stated as follows: If the maximum principle is satisfied, and if the trajectory is optimal, then the control is optimal too (see [3, Theorem 2.1, p. 327]). Therefore, it is sufficient to prove that the inequality $I_0(x, u_0) \leqq I_0(x, u)$ holds, and it is not necessary to show that $I_0(x_0, u_0) \leqq I_0(x, u)$.

Relation (2.2) implies that the Jacobian (2.13) does not vanish.

There are two other errors in translation (as well as the typographical mistake which was corrected in [4]) in [3]. Namely, on p. 327, line 17, "in" should be replaced by "outside of," and on p. 336, line 10 from the bottom, "optimal control" should be replaced by "control."

REFERENCES

[1] L. W. NEUSTADT AND J. WARGA, *Comments on the paper "Optimal Control of Processes Described by Integral Equations. I" by V. R. Vinokurov*, this Journal, 8 (1970), p. 572.

[2] V. R. VINOKUROV, *Optimal control of processes described by integral equations. I*, Izv. Vyssh. Uchebn. Zaved. Matematika, 1967, no. 7, pp. 21–33 (in Russian).

[3] ———, *Optimal control of processes described by integral equations. I*, this Journal, 7 (1969), pp. 324–336 (English translation from the Russian).

[4] *Erratum: Optimal control of processes described by integral equations. I, by V. R. Vinokurov*, this Journal, 8 (1970), p. 440.

Note. Both in our role as Editors of this journal and as authors of [1], we wish to apologize to the author for the misunderstanding we had of his paper because of a misinterpretation by the translators of the original Russian article. L. W. NEUSTADT and J. WARGA.

* Received by the editors October 19, 1971.

† Orsk, U.S.S.R.

SIAM J. CONTROL
Vol. 11, No. 1, February 1973

ERRATUM: GLOBAL CONTROLLABILITY OF NONLINEAR SYSTEMS*

D. L. LUKES†

Theorem 2.1 is true and the method of proof is çorrectly based upon showing the existence of a solution to equations (2.5)–(2.7). The shortcoming of the alleged fixed point $x^{\infty}(t), u^{\infty}(t), y^{\infty}$ obtained on p. 116 by extraction of a uniformly convergent subsequence is that it will in general not satisfy (2.5)–(2.7) since it arises as a limit as $k_i \to \infty$ and not as $k \to \infty$.

The corrected fixed-point argument proceeds as follows: Substitute (2.7) into (2.5)–(2.6) so that the relevant equations can be dealt with in the form

$$(u, x)(t) = N(u, x)(t)$$

in which N is a nonlinear operator on the Banach space E of continuous maps $t \to (u, x)(t)$ from $[0, T]$ into R^{r+n} with the sup-norm. A preliminary translation of the origin in E allows N to be taken of the form

$$N(u, x)(t) = [C(t)\tilde{N}(u, x)(T), \tilde{N}(u, x)(t) + D(t)\tilde{N}(u, x)(T)]$$

in which the matrices $C(t), D(t)$ are continuously differentiable in t,

$$\tilde{N}(u, x)(t) = \int_0^t e^{A(t-\omega)}\tilde{h}(u(\omega), x(\omega), \omega)\, d\omega$$

and $\tilde{h}$ is the induced translation of h (inheriting the continuity and boundedness properties of h). The following facts can now be verified:

(i) $\|N(u, x)\|$ is bounded on E.

(ii) $N: E \to E$ is continuous.

(iii) For all c, sufficiently large, N maps E into the subset of E,

$$E_c = [(u, x) \in E: |(u, x)(t + \Delta) - (u, x)(t)| \leqq c|\Delta|;\ \text{all } t, \Delta].$$

(iv) E_c is a compact and convex subset of the Banach space E.

The proof of (i) follows from the continuity of $C(t), D(t)$ on $[0, T]$ and the boundedness of $|\tilde{h}|$ on R^{r+n+1}. Statement (ii) is based upon the continuity and hence uniform continuity of h on compact subsets of R^{r+n+1}. The invariance of E_c under N for c sufficiently large can be shown by writing $C(t)$ and $D(t)$ as integrals of their derivatives in the formula for $N(u, x)$ and then once again applying the boundedness of $|\tilde{h}|$. Showing E_c closed and convex is a simple exercise. Since E_c is a bounded subset of E, once its closed character has been established, then by noting that it is an equi-uniformly-continuous family the Arzela–Ascoli theorem can be invoked to conclude its compactness.

A direct consequence of (i)–(iv) is that for all sufficiently large c the restriction of N to E_c provides a continuous map of a compact and convex subset of a Banach space into itself. Hence the Schauder theorem applies, establishing the existence of the required fixed point. This in turn determines a solution to (2.5)–(2.7) and concludes the proof.

* This Journal, 10 (1972), pp. 112–126. Received by the editors March 29, 1972.

† Department of Applied Mathematics and Computer Science, University of Virginia, Charlottesville, Virginia 22901.

SIAM J. CONTROL
Vol. 11, No. 2, May 1973

NECESSARY AND SUFFICIENT CONDITIONS FOR OPTIMAL CONTROL OF SEMI-MARKOV JUMP PROCESSES*

LAWRENCE D. STONE†

Abstract. A class of controlled semi-Markov jump processes is defined in this paper. Conditions are found which guarantee that satisfaction of the dynamic programming equations for stochastic control is necessary and sufficient for the minimization of the expected discounted cost of a controlled semi-Markov process over a random time. Terminal costs are included, and the controls are allowed to depend on both the present state of the process and the length of time the process has been in that state.

1. Introduction. The problem under consideration in this paper is that of finding necessary and sufficient conditions for a control to minimize the expected discounted cost of a controlled semi-Markov process over a random time. The setting of our results is more general than is usually considered for control problems involving semi-Markov processes in that we allow the control to depend on the present state of the process and the length of time the process has been in that state. Moreover, we allow terminal costs and do not restrict ourselves to finite state spaces.

Our approach is to consider a semi-Markov process as a two-dimensional Markov process and to find an equation (i.e., (4.9)) involving the infinitesimal generator of this two-dimensional process. Under specified conditions, we show that satisfaction of this equation is a necessary and sufficient condition for a control to be optimal (i.e., to minimize cost). Equation (4.9) is a form of the dynamic programming conditions for stochastic control problems. In order to obtain our result, we define the notion of a controlled semi-Markov process in § 2 and compute its infinitesimal generator in § 3. Section 4 contains the results on necessity and sufficiency. These results are also specialized to the case of undiscounted cost and to Markov jump process.

The principle of dynamic programming has been applied to the optimal control of discrete time Markov processes by many authors, e.g., Howard [9], Derman [4] and [5], Blackwell [2], Astrom [1] and Ross [15]. In the case of continuous time Markov processes, most of the work on optimal stochastic control has dealt with diffusion processes. A review of the literature on optimal control of diffusion processes is given in [7] where there is a discussion of the dynamic programming conditions for optimal control of diffusion processes. These conditions involve the infinitesimal generator of the diffusion and are the analogue, for stochastic control, of Bellman's principle of optimality. Although there are results showing that these equations are necessary or sufficient for the optimal control of certain classes of diffusion processes, very few results of this type exist for continuous time Markov processes which are not diffusions or for non-Markovian processes. Some exceptions to this are to be found in [14] and in a remark on page 527 of [12]. Theorem 4.5 of this paper extends the results on the necessity and sufficiency of the dynamic programming conditions for optimal control to semi-Markov jump processes.

* Received by the editors October 19, 1971, and in revised form May 18, 1972.

† Daniel H. Wagner, Associates, Paoli, Pennsylvania 19301. This work was supported by the Office of Naval Research under Contract N00014-70-C-0232.

Optimal control of semi-Markov processes has been studied in [3], [10], [11], [13] and [16]. However, these references deal with a more restricted class of processes and a more restricted class of controls than considered here. In particular, all of these references except [16] restrict their attention to processes which have only a finite number of states, and all except [3] restrict the class of controls to those which depend only on the state of the process and not on the length of time the process has remained in that state. In addition the control problem which we consider involves terminal conditions while the references mentioned in this paragraph deal with stationary problems, for the most part.

2. Controlled semi-Markov jump processes. Let $\{X_t:0 \leqq t < \infty\}$ be a semi-Markov jump process (see [17] for definition) defined on the probability space $(\Omega, \mathscr{A}, P)$. In this paper we make the slight generalization that we allow the state space Σ of $\{X_t\}$ to be an arbitrary set with a topology $\mathscr{T}_1$. Let $\mathscr{S}_1$ be the Borel σ-field formed from $\mathscr{T}_1$. We assume that $\mathscr{S}_1$ contains all singleton subsets of Σ. (The need for $\mathscr{S}_1$ to contain all singletons will be made clear later.) Let R^+ be the nonnegative real numbers and $\mathscr{T}_2$ the usual topology on R^+. Let $\mathscr{S}_2$ be the set of Borel sets of R^+. Let Y_0 be a nonnegative random variable defined on $(\Omega, \mathscr{A}, P)$ and for $t \geqq 0$, define

$$v_t = \inf\{s > t : X_{t+s} \neq X_t\},$$

$$Y_t = \begin{cases} Y_0 + t & \text{for } t < v_0, \\ t - \sup[s:0 \leqq s \leqq t \text{ and } X_s \neq X_t] & \text{for } t \geqq v_0. \end{cases}$$

For convenience of notation, we denote v_0 by v and let $Z_t = (X_t, Y_t)$ for $t \geqq 0$. In [17], v is required to be finite with probability one. In this paper we do not make that requirement.

We use $P_{x,s}$ and $E_{x,s}$ to denote, respectively, probability and expectation conditioned on $(X_0, Y_0) = (x, s)$. Define

$$\hat{a}(x, s) = P_{x,0}[v \leqq s] \qquad \text{for } x \in \Sigma \text{ and } s \in R^+,$$

$$\hat{k}(x, s, \Gamma) = P_{x,0}[X_v \in \Gamma | v = s] \quad \text{for } (x, s) \in \Sigma \times R^+ \text{ and } \Gamma \in \mathscr{S}_1.$$

Then $\hat{a}$ and $\hat{k}$ give the jump time and after jump distributions of the semi-Markov process. Moreover,

$$P_{x,0}[X_v \in \Gamma] = \int_0^\infty \hat{k}(x, s, \Gamma)\hat{a}(x, ds).$$

One may think of the functions $\hat{a}$ and $\hat{k}$ as specifying the semi-Markov process. This is the point of view we adopt for control problems.

We suppose that for each $x \in \Sigma$ there is a family of measures on $\mathscr{S}_1$,

$$\{k(x, b, \cdot):b \in B\},$$

where B is some index set. We assume that B is endowed with a topology and that $\mathscr{S}_3$ is the corresponding Borel σ-field. A control u is an ordered pair of functions (γ, β) with

$$\gamma:\Sigma \times R^+ \to R^+ \quad \text{and} \quad \beta:\Sigma \times R^+ \to B$$

such that:

(i) γ and β are Borel measurable (i.e., $\mathscr{S}_1 \times \mathscr{S}_2$ measurable where $\mathscr{S}_1 \times \mathscr{S}_2$ denotes the product σ-field generated by $\mathscr{S}_1$ and $\mathscr{S}_2$),

(ii) for $x \in \Sigma$, there exists $t(x) > 0$ such that $\int_0^{t(x)} \gamma(x, s)\, ds < \infty$,

(iii) $k(\cdot, \beta, \Gamma)$ is Borel measurable for each $\Gamma \in \mathscr{S}_1$.

Let U be the class of controls which satisfy (i)–(iii). A control $u = (\gamma, \beta) \in U$ determines a semi-Markov process $\{X_t^u\}$ in the following manner. Let

$$a_\gamma(x, t) = 1 - \exp\left(-\int_0^t \gamma(x, s)\, ds\right) \quad \text{for } (x, t) \in \Sigma \times R^+.$$

Suppose that $(X_0^u, Y_0^u) = (x, s)$. Then at time 0, the process starts in state x and remains there a random time ξ_1, such that

$$P_{x,0}\{\xi_1 \leqq t\} = \frac{a_\gamma(x, s + t) - a_\gamma(x, s)}{(1 - a_\gamma(x, s))}.$$

By (ii) we may define the sample paths of $\{X_t^u\}$ to be right continuous. That is at time $\xi_1 > 0$, the process transitions to the state $X_{\xi_1}^u$, where

$$P_{x,0}\{X_{\xi_1}^u \in \Gamma\} = k(x, \beta(x, \xi_1), \Gamma).$$

The process stays in state $X_{\xi_1}^u$ for a random time $\xi_2 > 0$ such that $P_{x,0}\{\xi_2 \leqq t\} = a_\gamma(X_{\xi_1}^u, t)$ and at time $\xi_1 + \xi_2$ transitions to $X_{\xi_1+\xi_2}^u$, where

$$P_{x,0}\{X_{\xi_1+\xi_2}^u \in \Gamma\} = k(X_{\xi_1}^u, \beta(X_{\xi_1}^u, \xi_2), \Gamma).$$

The process $\{X_t^u\}$ continues in this way transitioning to a new state after remaining a random time in the present state.

For $\omega \in \Omega$ and $u \in U$, let $\zeta(\omega)$ be the first cluster point of the jump times of the sample path $\{X_t^u(\omega): t \geqq 0\}$. If there is no cluster point, we let $\zeta(\omega) = \infty$. We call ζ the terminal time of the process $\{X_t^u\}$. The process $\{Z_t^u\} = \{X_t^u, Y_t^u\}$ with the terminal time ζ is a two-dimensional Markov process with stationary Borel measurable transition probabilities.

THEOREM 2.1. *If $u \in U$, then $\{X_t^u\}$ is a semi-Markov jump process, and $\{Z_t^u\}$ is a strongly measurable strong Markov process.*

Proof. In order to satisfy the definition of a semi-Markov process, we must show that $\{Z_t^u\}$ is a strong Markov process with right-continuous sample paths. Having shown this, it follows automatically that $\{Z_t^u\}$ is strongly measurable (see [6, vol. I, p. 98]). By our definition of $\{X_t^u\}$, the sample paths of $\{Z_t^u\}$ are right continuous in the $\mathscr{T}_1 \times \mathscr{T}_2$ topology regardless of the topology $\mathscr{T}_1$ on Σ. In particular, one may choose $\mathscr{T}_1 = \mathscr{D}$, the discrete topology on Σ.

Since $\mathscr{S}_1$ contains all the singleton subsets of Σ, it follows that $\{Z_t^u\}$ is a right-continuous Markov process on the topological measurable space $(\Sigma \times R^+, \mathscr{D} \times \mathscr{T}_2, \mathscr{S}_1 \times \mathscr{S}_2)$ (see [6, vol. II, p. 222] for the definition of the topological measurable space).

For $t \geqq 0$, let

$$\rho_t^u(x, s, \Gamma, r) = P_{x,s}[X_t^u \in \Gamma; Y_t^u \leqq r] \quad \text{for } \Gamma \in \mathscr{S}_1 \quad \text{and} \quad r \geqq 0.$$

Let F be the set of bounded measurable (with respect to $\mathscr{S}_1 \times \mathscr{S}_2$) functions $f:\Sigma \times R^+ \to R$, where R is the real numbers. For $f \in F$, define

$$\|f\| = \sup_{(x,s)\in\Sigma\times R^+} |f(x,s)|,$$

$$\mathbf{T}_t^u f(x,s) = \int_{\Sigma\times R^+} f(y,r)\rho_t^u(x,s,dy,dr).$$

In order to use Theorem 3.10 of [6] to prove our theorem, we show that for each $t \in R^+$, T_t^u maps functions $f \in F$ which are continuous in the $\mathscr{D} \times \mathscr{T}_2$ topology into continuous functions. To prove this, it is sufficient to show that for each $(x,s) \in \Sigma \times R^+$ and $t > 0$,

$$\lim_{h\to 0} \mathbf{T}_t^u f(x,s) - \mathbf{T}_t^u f(x,s+h) = 0$$

whenever $f(x,\cdot)$ is continuous in the $\mathscr{T}_2$ topology. Let $t > 0$ and fix (x,s). Consider the case in which $(X_0, Y_0) = (x,s)$. Then with probability

$$[1 - a_y(x,s+h)]/[1 - a_y(x,s)],$$

$(X_h^u, Y_h^u) = (x, s+h)$, where $h > 0$. Thus for $h > 0$,

$$\rho_{t+h}^u(x,s,\Gamma,r) = \frac{1 - a_y(x,s+h)}{1 - a_y(x,s)}\rho_t^u(x,s+h,\Gamma,r)$$
$$+ \frac{a_y(x,s+h) - a_y(x,s)}{1 - a_y(x,s)}G_h(\Gamma,r),$$

where $G_h(\Gamma,r)$ gives the probability that $X_t \in \Gamma$ and $Y_t \leqq r$ given that a jump occurred in the interval $[0,h]$. Solving the above equation for $\rho_t^u(x,s+h,\Gamma,r)$, it follows that

$$\mathbf{T}_t^u f(x,s+h) = \frac{1 - a_y(x,s)}{1 - a_y(x,s+h)}\mathbf{T}_{t+h}^u f(x,s)$$
$$- \frac{a_y(x,s+h) - a_y(x,s)}{1 - a_y(x,s+h)}\int_{\Sigma\times R^+} f(y,r)G_h(dy,dr)$$
$$= \mathbf{T}_{t+h}^u f(x,s) + w(h),$$

where $w(h) \to 0$ as $h \to 0+$. Hence, for $h > 0$,

$$\text{(2.1)}\quad \begin{aligned} |\mathbf{T}_t^u f(x,s) - \mathbf{T}_t^u f(x,s+h)| &\leqq |\mathbf{T}_t^u f(x,s) - \mathbf{T}_{t+h}^u f(x,s)| + o(1) \\ &\leqq |E_{x,s}[f(X_t^u, Y_t^u) - f(X_{t+h}^u, Y_{t+h}^u)]| + o(1). \end{aligned}$$

By the right continuity of $\{(X_t^u, Y_t^u)\}$ and the continuity of $f(x,\cdot)$ for $x \in \Sigma$, it follows that $\lim_{h\to 0+} f(X_{t+h}^u, Y_{t+h}^u) = f(X_t^u, Y_t^u)$ a.s. Thus, the dominated convergence theorem and (2.1) yield that $\lim_{h\to 0+} |\mathbf{T}_t^u f(x,s) - \mathbf{T}_t^u f(x,s+h)| = 0$.

To show that $\lim_{h\to 0+} |\mathbf{T}_t^u f(x,s) - \mathbf{T}_t^u f(x,s-h)| = 0$ for $s > 0$, one observes that

$$\rho_t^u(x,s-h,\Gamma,r) = \frac{1 - a_y(x,s)}{1 - a_y(x,s-h)}\rho_{t-h}^u(x,s,\Gamma,r)$$
$$+ \frac{a_y(x,s) - a_y(x,s-h)}{1 - a_y(x,s-h)}G_h(\Gamma,r),$$

and that

$$|\mathbf{T}_t^u f(x,s) - \mathbf{T}_t^u f(x, s-h)| \leqq |E_{x,s}[f(X_t^u, Y_t^u) - f(X_{t-h}^u, Y_{t-h}^u)]| + o(1).$$

Note that

$$\lim_{h \to 0+} (X_{t-h}(\omega), Y_{t-h}(\omega)) \to (X_t(\omega), Y_t(\omega)) \tag{2.2}$$

unless there is a jump at time t. Since the probability of having a jump at time t is 0, (2.2) holds for a.e. $\omega \in \Omega$. For each ω for which (2.2) holds, we have $X_{t-s}^u(\omega) = X_t^u(\omega)$ for $t - Y_t^u(\omega) \leqq s \leqq t$. Since $f(x, \cdot)$ is continuous for $x \in \Sigma$, it follows that $f(X_{t-h}^u, Y_{t-h}^u) \to f(X_t^u, Y_t^u)$ a.s. As before, we use the dominated convergence theorem to conclude that $\lim_{h\to 0+} |\mathbf{T}_t^u f(x,s) - \mathbf{T}_t^u f(x, s-h)| = 0$. It follows that $T_t^u f$ is continuous in the $\mathscr{D} \times \mathscr{T}_2$ topology.

By what we have just shown, $\{Z_t^u\}$ is a Feller process on the topological measure space $(\Sigma \times R^+, \mathscr{D} \times \mathscr{T}_2, \mathscr{S}_1 \times \mathscr{S}_2)$, and by Theorem 3.10 of [6], it is strong Markov. This proves the theorem.

We call $\{X_t^u\}$ a *controlled semi-Markov jump process.*

We suppose there is a cost rate $c(x, s, u(x, s))$ associated with being in state x for time s when using control $u(x, s)$. When convenient, we shall use the variable z to represent a point $(x, s) \in \Sigma \times R^+$. Thus, $c(z, u(z))$ will often be used in place of $c(x, s, u(x, s))$. In addition we suppose there is a function $C: \Sigma \times R^+ \to R$ such that $C(z)$ gives the cost of stopping the process in state z. Let τ be a stopping time of $\{X_t^u\}$. Fix $\lambda > 0$ and define for $u \in U$,

$$\varphi_u(z) = E_z\left[\int_0^\tau e^{-\lambda t} c(Z_t^u, u(Z_t^u))\, dt + e^{-\lambda\tau} C(Z_\tau^u)\right], \qquad z \in \Sigma \times R^+. \tag{2.3}$$

For the case of undiscounted cost, we define for $u \in U$,

$$\hat{\varphi}_u(z) = E_z\left[\int_0^\tau c(Z_t^u, u(Z_t^u))\, dt + C(Z_\tau^u)\right], \qquad z \in \Sigma \times R^+. \tag{2.4}$$

In order that the above integrals be well-defined, we assume that c and C are Borel measurable. To allow for the possibility that $\tau(\omega) = \zeta(\omega)$ for some ω we adjoin a state z_0 to $\Sigma \times R^+$ and define $Z_\zeta = z_0$. The function C is assumed to be defined on this extended state space.

The dependence of τ on u will be suppressed. Let U' be a specified set of controls. We are interested in controls $u^* \in U'$ such that

$$\varphi_{u^*}(x, s) = \min_{u \in U'} \varphi_u(x, s), \qquad (x, s) \in \Sigma \times R^+. \tag{2.5}$$

Such a control is called *optimal in* U'. In order to find necessary and sufficient conditions for a function φ_{u^*} to satisfy (2.5), we discuss the infinitesimal operator of a semi-Markov jump process.

3. Infinitesimal operator of a semi-Markov jump process. If $\{X_t^u\}$ is a controlled semi-Markov jump process, then the process $\{(X_t^u, Y_t^u)\}$ is a two-dimensional, right-continuous Markov process with stationary transition probabilities. Thus, we may define $\mathbf{A}^u$, the weak infinitesimal operator for this process, in the manner given in [6]. Let $\tilde{F}$ be the set of all $f \in F$ such that

$$\lim_{t \to 0+} \mathbf{T}_t^u f(x,s) = f(x,s) \quad \text{for } (x,s) \in \Sigma \times R^+.$$

The weak infinitesimal operator is defined by

$$\mathbf{A}^u f = \lim_{t \to 0+} (\mathbf{T}_t^u f - f)/t,$$

whenever the limit of the right side exists in the weak sense and is a member of $\tilde{F}$. We say that $\lim_{n \to \infty} f_n = f$ in the weak sense if:

(a) $\lim_{n \to \infty} f_n(x,s) = f(x,s)$ for each $(x,s) \in \Sigma \times R^+$, and

(b) $\|f_n\|$, $n = 1, 2, \cdots$, are bounded.

Let $f^+(x,\cdot)$ denote the right-hand derivative of $f(x,\cdot)$. If $\gamma(x,\cdot)$ is right continuous at s, then

$$\gamma(x,s) = \frac{a_\gamma^+(x,s)}{1 - a_\gamma(x,s)}.$$

In this case $\gamma(x,s)$ is the jump rate of the process $\{X_t^u\}$ given that it has been in state x for a time s. Recall that $k(x, \beta(x,s), \Gamma)$ gives the probability, when using control $\beta(x,s)$, that the transition from x will be into the set Γ given that the transition takes place after being in x for a time s. Thus if $\gamma(x,\cdot)$ is right continuous, it follows from [8, § 411] that

$$a_\gamma(x,t) = \int_0^t a_\gamma^+(x,s)\, ds$$

and

$$P_{x,0}\{X_\nu^u \in \Gamma \text{ and } \nu \leqq t\} = \int_0^t k(x, \beta(x,s), \Gamma)\gamma(x,s) \exp\left(-\int_0^s \gamma(x,r)\, dr\right) ds.$$

We now compute $\mathbf{A}^u f$ for a particular class of functions f and controlled semi-Markov jump processes.

LEMMA 3.1. *Let $u = (\gamma, \beta) \in U$ be such that for $x \in \Sigma$:*

(i) *$\gamma(x,\cdot)$ is right continuous and $\gamma \leqq M \in R^+$,*

(ii) *$\beta(x,\cdot)$ is piecewise constant, right continuous, and has only a finite number of discontinuities.*

Suppose $f \in F$ is such that f^+ exists and is bounded, and $f^+(x,\cdot)$ is right continuous for all $x \in \Sigma$. Then

$$\mathbf{A}^u f(x,s) = f^+(x,s) + \gamma(x,s)\left[\int_\Sigma f(y,0)k(x,\beta(x,s),dy) - f(x,s)\right] \tag{3.1}$$

for $(x,s) \in \Sigma \times R^+$.

Proof. Since $\gamma(x,s) = a_\gamma^+(x,s)/[1 - a(x,s)]$, it follows that $\gamma \leqq M$ implies $a_\gamma(x,h) \leqq Mh$ and

$$\frac{a_\gamma^+(x,s+r)}{1 - a_\gamma(x,s)} \leqq \frac{a_\gamma^+(x,s+r)}{1 - a_\gamma(x,s+r)} \leqq M. \tag{3.2}$$

Thus,

$P_{x,s}$ [more than one jump in $[0, h]$]

$$= \int_0^h \int_\Sigma \frac{a_\gamma^+(x, s + r)}{1 - a_\gamma(x, s)} a_\gamma(y, h - r) k(x, \beta(x, s + r), dy)\, dr$$

$$\leqq \int_0^h M^2(h - r)\, dr \leqq M^2 h^2,$$

and the probability of having more than one jump in $[0, h]$ is $o(h)$ as $h \to 0$. For r small enough, $\beta(x, s + r) = \beta(x, s)$. Thus for small $h > 0$,

$\mathbf{T}_h^u f(x, s) - f(x, s)$

$$= \frac{1 - a_\gamma(x, s + h)}{1 - a_\gamma(x, s)} f(x, s + h) + \int_0^h \int_\Sigma f(y, h - r)(1 - a_\gamma(y, h - r))$$

$$\cdot \frac{a_\gamma^+(x, s + r)}{1 - a_\gamma(x, s)} k(x, \beta(x, s), dy)\, dr - f(x, s) + o(h)$$

$$= \frac{(1 - a_\gamma(x, s + h)) f(x, s + h) - (1 - a_\gamma(x, s)) f(x, s)}{1 - a_\gamma(x, s)}$$

$$+ \int_\Sigma \int_0^h f(y, h - r)(1 - a_\gamma(y, h - r)) \frac{a_\gamma^+(x, s + r)}{1 - a_\gamma(x, s)}\, dr\, k(x, \beta(x, s), dy) + o(h).$$

By the dominated convergence theorem,

$$\lim_{h \to 0+} \frac{1}{h}(\mathbf{T}_h^u f(x, s) - f(x, s))$$

$$= f^+(x, s) - \frac{a_\gamma^+(x, s)}{1 - a_\gamma(x, s)} f(x, s) + \int_\Sigma \frac{a_\gamma^+(x, s)}{1 - a_\gamma(x, s)} f(y, 0) k(x, \beta(x, s), dy)$$

$$= f^+(x, s) + \gamma(x, s)\left[\int_\Sigma f(y, 0) k(x, \beta(x, s), dy) - f(x, s)\right].$$

The boundedness of f and γ guarantee that $\|\mathbf{T}_h^u f - f\|/h$ is uniformly bounded as $h \to 0$. Thus the above limit holds in the weak sense. Observe that any function $g \in F$ such that $g(x, \cdot)$ is right continuous for all $x \in \Sigma$ is contained in $\tilde{F}$. Thus the right-hand side of (3.1) is in $\tilde{F}$, and the lemma is proved.

4. Necessary and sufficient conditions for optimal control of semi-Markov jump processes. In this section we prove the main result of the paper, Theorem 4.5, which gives conditions under which (4.9) is necessary and sufficient for optimality of a control u^* within a specified class of controls. Throughout this section, we let Δ be a subset of $\Sigma \times R^+$ such that $S(x) = \{s : (x, s) \in \Delta\}$ is a closed right half-line for $x \in \Sigma$. Moreover, we define

$$\pi(\Delta) = \{x : (x, s) \in \Delta \text{ for some } s \in R^+\},$$

$$\tau = \min\{\zeta, \inf[t : t \geqq 0 \text{ and } Z_t^u \in \Delta]\}.$$

We suppress the dependence of τ on u. Note that τ is a Markov or stopping time of $\{Z_t^u\}$.

THEOREM 4.1. *Let $U' \subset U$. Suppose that there exists $f \in F$ such that f is in the domain of $\mathbf{A}^u$ for $u \in U'$ and*

$$\lambda f(z) = \inf_{u \in U'} \{\mathbf{A}^u f(z) + c(z, u(z))\} \quad \textit{for } z \in \Sigma \times R^+ - \Delta, \tag{4.1}$$

$$f(z) = C(z) \qquad \textit{for } z \in \Delta. \tag{4.2}$$

Then

$$f(z) \leqq \inf_{u \in U'} \varphi_u(z) \quad \textit{for } z \in \Sigma \times R^+. \tag{4.3}$$

Proof. For f in the domain of $\mathbf{A}^u$, Theorem 1.7 of [6] yields that

$$f = \mathbf{R}_\lambda^u(\lambda f - \mathbf{A}^u f),$$

where

$$\mathbf{R}_\lambda^u f(z) = \int_0^\infty e^{-\lambda t}\mathbf{T}_t f(z)\,dt \quad \text{for } z \in \Sigma \times R^+.$$

Thus, by Theorem 5.1 of [6],

$$E_z[e^{-\lambda\tau} f(Z_\tau^u)] - f(z) = E_z\left[\int_0^\tau e^{-\lambda t}[\mathbf{A}^u f(Z_t^u) - \lambda f(Z_t^u)]\,dt\right]. \tag{4.4}$$

For $z \in \Delta$, (4.3) follows trivially from the fact that $\varphi^u(z) = C(z)$ for $z \in \Delta$ and $u \in U$. Let $u \in U'$. By (4.1),

$$c(z, u(z)) + A^u f(z) - \lambda f(z) \geqq 0 \quad \text{for } z \in \Sigma \times R^+ - \Delta. \tag{4.5}$$

Integrating the left-hand side of (4.5), we obtain

$$E_z\left[\int_0^\tau e^{-\lambda r} c(Z_r^u, u(Z_r^u))\,dr + \int_0^\tau e^{-\lambda r}[\mathbf{A}^u f(Z_r^u) - \lambda f(Z_r^u)]\,dr\right] \geqq 0$$

$$\text{for } z \in \Sigma \times R^+ - \Delta.$$

By (4.4) and (4.2), we obtain

$$E_z\left[\int_0^\tau e^{-\lambda r} c(Z_r^u, u(Z_r^u))\,dr + e^{-\lambda\tau} C(Z_\tau^u)\right] \geqq f(z) \quad \text{for } z \in \Sigma \times R^+ - \Delta.$$

Since u is an arbitrary member of U', we have $\varphi_u(z) \geqq f(z)$ for $u \in U'$ and $z \in \Sigma \times R^+ - \Delta$. The theorem follows.

For the undiscounted case we prove the following corollary.

COROLLARY 4.2. *Let $U' \subset U$. Suppose there exists $f \in F$ such that f is in the domain of $\mathbf{A}^u$ for $u \in U'$ and*

$$0 = \inf_{u \in U'} \{\mathbf{A}^u f(z) + c(z, u(z))\} \quad \textit{for } z \in \Sigma \times R^+ - \Delta,$$

$$f(z) = C(z) \qquad \textit{for } z \in \Delta.$$

If $E_z[\tau] < \infty$ *for* $z \in \Sigma \times R^+$, *then*

$$f(z) \leqq \inf_{u \in U'} \hat{\varphi}_u(z) \quad \textit{for } z \in \Sigma \times R^+.$$

Proof. The corollary follows in the same manner as Theorem 4.1 by the use of the corollary to Theorem 5.1 of [6].

In Theorem 4.3 below, we find a necessary condition for optimality which we shall use in the proof of our main theorem. The operator **Q** used in the proof of Theorem 4.3 is a generalization to continuous time processes of the operator T used by Blackwell in [2] for discrete time processes.

Fix h such that $0 < h \leqq \infty$ and define the following stopping times:

$$\sigma_1(x, s) = \min\{h, \delta_1, \tau\},$$

where δ_1 gives the waiting time for the first jump of the process $\{X_t\}$ given $(X_0, Y_0) = (x, s)$. Although σ_1 and δ_1 depend on the control u, we shall not indicate this dependence. In a similar fashion, we define

$$\sigma_n(Z_{\sigma_{n-1}}) = \min\{h, \delta_n, \tau\} \quad \text{for } n \geqq 2,$$

where δ_n is the waiting time for the first jump of the process after σ_{n-1}. For convenience of notation, we shall write σ_n for $\sigma_n(Z_{\sigma_{n-1}})$ and let

$$\xi_0 = 0, \qquad \xi_n = \sigma_1 + \cdots + \sigma_n \qquad \text{for } n \geqq 1.$$

Define

$$W(x, s, u) = E_{x,s} \int_0^{\sigma_1} e^{-\lambda t} c(x, s + t, u(x, s + t))\, dt \quad \text{for } u \in U, (x, s) \in \Sigma \times R^+.$$

In addition, we let

$$H_u(x, s, \Lambda, r) = P_{x,s}[Z^u_{\sigma_1} \in \Lambda \text{ and } \sigma_1 \leqq r],$$

where $(x, s) \in \Sigma \times R^+$, Λ is a measurable subset of $\Sigma \times R^+$, $r \geqq 0$, and $u \in U$. Let $U' \subset U$ and let u^* be optimal in U'. Define

$$\psi(x, s, U') = \inf_{u \in U'} \left\{ W(x, s, u) + \int e^{-\lambda r} \varphi_{u^*}(z) H_u(x, s, dz, dr) \right\}$$
$$\text{for } (x, s) \in \Sigma \times R^+,$$

where an integral without an indicated range is understood to run over $\Sigma \times R^+ \times R^+$.

We say that $U' \subset U$ is closed under one point exchanges if $u_1, u_2 \in U'$ imply

$$u(y, r) \equiv \begin{cases} u_1(y, r) & \text{for } y = x \text{ and } r \geqq s, \\ u_2(y, r), & \text{otherwise}, \end{cases}$$

is a member of U', where $(x, s) \in \Sigma \times R^+$ is arbitrary. Since singleton subsets of Σ are measurable, U is closed under one point exchanges. Similarly, the following sets are closed under one point exchanges:

$$U_1 = \{u : u = (\gamma, \beta) \in U, \gamma(x, \cdot) \text{ is right continuous for } x \in \Sigma \text{ and } \gamma \leqq M\},$$

$$U_2 = \{u : u \in U \text{ and } c(x, \cdot, u(x, \cdot)) \text{ is right continuous for } x \in \Sigma\},$$

$$U_1 \cap U_2,$$

where M is a real number and $c(x, \cdot, u(x, \cdot))$ is understood to be a function defined on R^+.

THEOREM 4.3. *Suppose $U' \subset U$ is closed under one point exchanges. Let C and c be bounded, and let $\tau(\omega) < \zeta(\omega)$ for a.e. $\omega \in \Omega$ such that $\zeta(\omega) < \infty$. If u^* is optimal in U', then*

$$\varphi_{u^*}(z) = \psi(z, U') \quad \textit{for } z \in \Sigma \times R^+. \tag{4.6}$$

Proof. We observe that for $(x, s) \in \Sigma \times R^+$,

$$\varphi_{u^*}(x, s) = W(x, s, u^*) + \int e^{-\lambda r} \varphi_{u^*}(z) H_{u^*}(x, s, dz, dr),$$

and thus $\varphi_{u^*}(x, s) \geqq \psi(x, s, U')$.

Suppose that for some $(x', s') \in \Sigma \times R^+$, $\varphi_{u^*}(x', s') > \psi(x', s', U')$. Then let $v \in U'$ be such that

$$\varphi_{u^*}(x', s') > W(x', s', v) + \int e^{-\lambda r} \varphi_{u^*}(z) H_v(x', s', dz, dr). \tag{4.7}$$

Define

$$\hat{u}(y, r) = \begin{cases} v(y, r) & \text{for } y = x' \text{ and } r \geqq s', \\ u^*(y, r), & \text{otherwise.} \end{cases}$$

Since U' is closed under one point exchanges, $\hat{u} \in U'$. For $f \in F$, define the operator $\mathbf{Q} : F \to F$ as follows:

$$\mathbf{Q}f(z) = W(z, \hat{u}) + \int e^{-\lambda r} f(z') H_{\hat{u}}(z, dz', dr) \quad \text{for } z \in \Sigma \times R^+.$$

Let $\mathbf{Q}^1 = \mathbf{Q}$, and define $\mathbf{Q}^{n+1} f = \mathbf{Q}(\mathbf{Q}^n f)$ for $n \geqq 1$. Then we have

$$\begin{aligned} \mathbf{Q}^2 \varphi_{u^*}(x, s) &= W(x, s, \hat{u}) + \int e^{-\lambda r} \mathbf{Q} \varphi_{u^*}(z) H_{\hat{u}}(x, s, dz, dr) \\ &= W(x, s, \hat{u}) + E_{x,s}[e^{-\lambda \xi_1} W(Z^{\hat{u}}_{\xi_1}, \hat{u})] + E_{x,s}[e^{-\lambda \xi_2} \varphi_{u^*}(Z^{\hat{u}}_{\xi_2})] \end{aligned}$$

and

$$\mathbf{Q}^n \varphi_{u^*}(x, s) = E_{x,s}\left[\sum_{j=1}^{n} \exp(-\lambda \xi_{j-1}) W(Z^{\hat{u}}_{\xi_{j-1}}, \hat{u})\right] + E_{x,s}[\exp(-\lambda \xi_n) \varphi_{u^*}(Z^{\hat{u}}_{\xi_n})]. \tag{4.8}$$

Note that $\mathbf{Q}^n \varphi_{u^*}$ gives the cost of using $\hat{u}$ up to time ξ_n plus the terminal cost given by the second term on the right of (4.8). Observe that as $n \to \infty$, the first term on the right-hand side of (4.8) approaches

$$E_{x,s} \int_0^\tau e^{-\lambda t} c(Z^{\hat{u}}_t, \hat{u}(Z^{\hat{u}}_t))\, dt.$$

If $\tau(\omega) < \infty$, then there exists $N(\omega)$ such that for $n \geqq N(\omega)$, $\xi_n(\omega) = \tau(\omega)$, by virtue of the assumption that $\tau(\omega) < \zeta(\omega)$ whenever $\zeta(\omega) < \infty$. This combined with the boundedness of C and c gives

$$\lim_{n\to\infty} e^{-\lambda\xi_n(\omega)}\varphi_{u^*}(Z^{\hat{u}}_{\xi_n}(\omega)) = \begin{cases} e^{-\lambda\tau(\omega)}C(Z^{\hat{u}}_{\tau}(\omega)) & \text{for a.e. } \omega \in \Omega \text{ such that } \tau(\omega) < \infty, \\ 0 & \text{for } \tau(\omega) = \infty. \end{cases}$$

It now follows that

$$\lim_{n\to\infty} \mathbf{Q}^n\varphi_{u^*}(x, s) = \varphi_{\hat{u}}(x, s) \quad \text{for } (x, s) \in \Sigma \times R^+.$$

From (4.7) it follows that

$$\varphi_{u^*}(x', s') > \mathbf{Q}\varphi_{u^*}(x', s').$$

Since $f \leqq g$ implies $\mathbf{Q}f \leqq \mathbf{Q}g$, we have

$$\varphi_{u^*}(x', s') > \mathbf{Q}\varphi_{u^*}(x', s') \geqq \lim_{n\to\infty} \mathbf{Q}^n\varphi_{u^*}(x', s') = \varphi_{\hat{u}}(x', s')$$

which contradicts the optimality of u^* in U'. Hence $\varphi_{u^*}(x, s) \leqq \psi(x, s, U')$ for $(x, s) \in \Sigma \times R^+$, and the theorem is proved.

Let $\hat{\psi}$ be the function obtained by setting $\lambda = 0$ in the definition ψ. Then one may prove the following corollary which gives a necessary condition for an optimal control in the undiscounted case.

COROLLARY 4.4. *Let the conditions of Theorem* 4.3 *be satisfied. In addition, assume that* $\tau(\omega) < \infty$ *for a.e.* $\omega \in \Omega$. *If* $u^* \in U'$ *and*

$$\hat{\varphi}_{u^*}(z) = \inf_{u\in U'} \hat{\varphi}(z) \quad \textit{for } z \in \Sigma \times R^+,$$

then

$$\hat{\varphi}_{u^*}(z) = \hat{\psi}(z, U') \quad \textit{for } z \in \Sigma \times R^+.$$

Proof. Since $\tau(\omega) < \infty$ for a.e. $\omega \in \Omega$, one may use the same method of proof as given for Theorem 4.3 with $\lambda = 0$. This proves the corollary.

For $u = (\gamma, \beta) \in U$, define the operator $\mathbf{V}^u$ on functions $f \in F$ such that $f^+(x, \cdot)$ exists for $x \in \Sigma$ as follows:

$$\mathbf{V}^u f(x, s) = f^+(x, s) + \gamma(x, s)\left[\int_\Sigma f(y, 0)k(x, \beta(x, s), dy) - f(x, s)\right]$$

for $(x, s) \in \Sigma \times R^+$. Note that $\mathbf{A}^u$ given in Lemma 3.1 is a restriction of $\mathbf{V}^u$.

The following theorem gives the main result of this paper. This theorem shows that under the stated conditions, the dynamic programming conditions for control of a stochastic process are necessary and sufficient for the optimality of u^*. An analogue of this theorem for undiscounted cost is given in the corollary below.

THEOREM 4.5. *Let* $U' \subset U$ *be the set of controls* $u = (\gamma, \beta)$ *such that* $\gamma \leqq M \in R^+$ *and for each* $x \in \Sigma$,

(i) $\gamma(x, \cdot)$ *is right continuous,*

(ii) $\beta(x, \cdot)$ *is piecewise constant, right continuous, and has only a finite number of discontinuities.*

Suppose that:

(iii) *C and c are bounded*,

(iv) *$c(x, \cdot, u(x, \cdot))$ is right continuous for $u \in U'$ and $x \in \Sigma$*,

(v) *for $x \in \pi(\Delta)$, $C^+(x, \cdot)$ exists, is bounded, and is right continuous.*

Then u^* is optimal in U' if and only if $u^* \in U'$ and

$$\lambda\varphi_{u^*}(z) = \min_{u \in U'} \{c(z, u(z)) + \mathrm{A}^u\varphi_{u^*}(z)\} \quad \textit{for } z \in \Sigma \times R^+ - \Delta. \tag{4.9}$$

Proof. Suppose u^* is optimal in U' and that $(x, s) \in \Sigma \times R^+ - \Delta$. Since $\{s : (x, s) \in \Delta\}$ is closed, we may choose h so small that

$$\sigma_1 = \min\{h, \delta_1\}.$$

Observe that U' is closed under one point exchanges and that since $\gamma \leqq M$, $\zeta(\omega) = \infty$ for a.e. $\omega \in \Omega$. One may check that the remaining hypotheses of Theorem 4.3 are satisfied. Let $u = (\gamma, \beta) \in U'$. By Theorem 4.3, we have for $(x, s) \in \Sigma \times R^+ - \Delta$,

$$\begin{aligned}\varphi_{u^*}(x, s) &\leqq W(x, s, u) + \int e^{-\lambda r}\varphi_{u^*}(z)H_u(x, s, dz, dr)\\ &= \frac{1 - a_\gamma(x, s + h)}{1 - a_\gamma(x, s)}\left[\int_0^h e^{-\lambda r}c(x, s + r, u(x, s + r))\,dr + e^{-\lambda h}\varphi_{u^*}(x, s + h)\right]\\ &\quad + \int_0^h \frac{a_\gamma^+(x, s + r)}{1 - a_\gamma(x, s)}\left[\int_0^r e^{-\lambda w}c(x, s + w, u(x, s + w))\,dw\right.\\ &\qquad\left. + e^{-\lambda r}\int_\Sigma \varphi_{u^*}(y, 0)k(x, \beta(x, s + r), dy)\right] dr.\end{aligned}$$

Hence, for positive h small enough,

$$\begin{aligned}&\frac{\varphi_{u^*}(x, s) - e^{-\lambda h}\varphi_{u^*}(x, s + h)}{h}\\ &\qquad \leqq \frac{1 - a_\gamma(x, s + h)}{1 - a_\gamma(x, s)}\frac{1}{h}\int_0^h e^{-\lambda r}c(x, s + r, u(x, s + r))\,dr\\ &\qquad\quad + \frac{1}{h}\int_0^h \frac{a_\gamma^+(x, s + r)}{1 - a_\gamma(x, s)}\left[\int_0^r e^{-\lambda w}c(x, s + w, u(x, s + w))\,dw\right.\\ &\qquad\quad\left. + e^{-\lambda r}\int_\Sigma \varphi_{u^*}(y, 0)k(x, \beta(x, s + r), dy) - e^{-\lambda h}\varphi_{u^*}(x, s + h)\right] dr.\end{aligned} \tag{4.10}$$

Since c and C are bounded, so is φ_{u^*}. This combined with the boundedness of γ gives that the right-hand side of (4.10) is bounded by a positive number, say K. Then for h small enough, $\varphi_{u^*}(x, s) - e^{-\lambda h}\varphi_{u^*}(x, s + h) \leqq hK$, and

$$\lim_{h \to 0+} e^{-\lambda h}\varphi_{u^*}(x, s + h) = \varphi_{u^*}(x, s).$$

Note that assumption (ii) of the theorem guarantees that for h small enough,

$\beta(x, s + r, \cdot) = \beta(x, s, \cdot)$ for $0 \leqq r \leqq h$ so that

$$\lim_{r \downarrow 0} \int_{\Sigma} \varphi_{u^*}(y, 0)k(x, \beta(x, s + r), dy) = \int_{\Sigma} \varphi_{u^*}(y, 0)k(x, \beta(x, s), dy).$$

If $\varphi_{u^*}^{+}$ exists, then by taking the limit as $h \to 0+$ in (4.10), we obtain

$$-\varphi_{u^*}^{+}(x, s) + \lambda\varphi_{u^*}(x, s)$$

$$\leqq c(x, s, u(x, s)) + \gamma(x, s)\left[\int_{\Sigma} \varphi_{u^*}(y, 0)k(x, \beta(x, s), dy) - \varphi_{u^*}(x, s)\right]. \tag{4.11}$$

Since the u chosen above is an arbitrary member of U', we have

$$\lambda\varphi_{u^*}(x, s) \leqq \inf_{u \in U'} \{c(x, s, u(x, s)) + \mathbf{V}^u\varphi_{u^*}(x, s)\} \quad \text{for } (x, s) \in \Sigma \times R^+ - \Delta \tag{4.12}$$

provided $\varphi_{u^*}^{+}$ exists.

We have already shown that $\varphi_{u^*}(x, \cdot)$ is right continuous at (x, s) for $(x, s) \in \Sigma \times R^+ - \Delta$. Thus, from (4.11) (taking $u = u^*$ and observing that equality holds), we obtain that $\varphi_{u^*}^{+}(x, s)$ exists and is bounded for $(x, s) \in \Sigma \times R^+ - \Delta$ and furthermore that $\varphi_{u^*}^{+}(x, \cdot)$ is right continuous at s. Since $\varphi_{u^*}(x, s) = C(x, s)$ for $(x, s) \in \Delta$, we have that φ_{u^*} satisfies the conditions on f in Lemma 3.1. Thus, for $u \in U'$, Lemma 3.1 yields that φ_{u^*} is in the domain of $\mathbf{A}^u$ and that $\mathbf{A}^u\varphi_{u^*} = \mathbf{V}^u\varphi_{u^*}$. Making this substitution in (4.12) we find that (4.9) is necessary for u^* to be optimal in U'.

Suppose $u^* \in U'$ and that φ_{u^*} satisfies (4.9). Clearly, $\varphi_{u^*}(z) = C(z)$ for $z \in \Delta$. By the preceding paragraph φ_{u^*} is in the domain of $\mathbf{A}^u$ for $u \in U'$. Thus Theorem 4.1 yields that u^* is optimal in U'. This proves the theorem.

COROLLARY 4.6. *Let U' be a class of controls u which satisfy the conditions* (i)–(v) *of Theorem* 4.5 *and for which*

$$E_z[\tau] < \infty \quad \text{for } z \in \Sigma \times R^+. \tag{4.13}$$

Suppose $u^* \in U'$. Then

$$\hat{\varphi}_{u^*}(z) = \inf_{u \in U'} \hat{\varphi}_u(z) \quad \text{for } z \in \Sigma \times R^+$$

if and only if

$$0 = \min_{u \in U'} \{c(z, u(z)) + \mathbf{A}^u\hat{\varphi}_{u^*}(z)\} \quad \text{for } z \in \Sigma \times R^+ - \Delta. \tag{4.14}$$

Proof. Since (4.13) holds, we have $\tau(\omega) < \infty$ for a.e. $\omega \in \Omega$. Thus we may follow the proof of Theorem 4.5 with $\lambda = 0$ and apply Corollary 4.4 to show that (4.14) is necessary for the optimality of u^*. The boundedness of C and c combined with (4.13) imply that $\hat{\varphi}_{u^*}$ is bounded. An argument essentially identical to that given in the proof of Theorem 4.1 shows that $\hat{\varphi}_{u^*}$ is in the domain of $\mathbf{A}^u$ for $u \in U'$. The sufficiency of (4.14) then follows from Corollary 4.2.

Remark 4.7. Let $\tilde{U}$ be the class of controls $u = (\gamma, \beta)$ such that

$$\gamma : \Sigma \to R^+, \qquad \beta : \Sigma \to B,$$

where γ and β are Borel measurable, and $k(\cdot, \beta, \Gamma)$ is Borel measurable for each $\Gamma \in \mathscr{S}_1$.

In this case the jump time distributions for $\{X_t^u\}$ become negative exponential and the after jump distributions, $k(x, \beta(x), \cdot)$, do not depend on the length of time the process $\{X_t^u\}$ spends in state x. Thus the process is Markovian.

Suppose, in addition, that $\Delta = \Sigma' \times [0, \infty]$ for some $\Sigma' \subset \Sigma$ and that for $u \in \tilde{U}$,

$$\varphi_u(x) = E_x\left[\int_0^\tau e^{-\lambda t}\tilde{c}(X_t, u(X_t))\,dt + e^{-\lambda\tau}\tilde{C}(X_\tau)\right] \quad \text{for } x \in \Sigma,$$

$$\hat{\varphi}_u(x) = E_x\left[\int_0^\tau \tilde{c}(X_t, u(X_t))\,dt + \tilde{C}(X_\tau)\right] \qquad \text{for } x \in \Sigma,$$

where $\tilde{c}$ and $\tilde{C}$ are bounded Borel functions. For $u \in \tilde{U}$, let $\tilde{\mathbf{A}}^u$ be the weak infinitesimal operator of $\{X_t^u\}$. Then

$$\tilde{\mathbf{A}}^u f(x) = \gamma(x)\left[\int_\Sigma f(y)k(x, \beta(x), dy) - f(x)\right]$$

for any bounded Borel measurable function $f: \Sigma \to R$. A straightforward modification of the proof of Theorem 4.5 shows that if γ is bounded, u^* is optimal in $\tilde{U}$ if and only if

$$\lambda\varphi_{u^*}(x) = \inf_{u \in \tilde{U}} \{\tilde{c}(x, u(x)) + \tilde{\mathbf{A}}^u \varphi_{u^*}(x)\} \quad \text{for } x \in \Sigma. \tag{4.15}$$

For the above situation, Kushner [12, p. 527] has observed that (4.15) is sufficient for optimality.

Similarly, for the undiscounted case, one may show that $\gamma < M \in R^+$ and $E[\tau] < \infty$ implies that u^* is optimal in $\tilde{U}$ if and only if

$$0 = \inf_{u \in \tilde{U}} \{\tilde{c}(x, u(x)) + \tilde{\mathbf{A}}^u \hat{\varphi}_{u^*}(x)\} \quad \text{for } x \in \Sigma.$$

REFERENCES

[1] K. J. ASTROM, *Optimal control of Markov processes with incomplete state information*, J. Math. Anal. Appl., 10 (1965), pp. 174–205.

[2] D. BLACKWELL, *Discounted dynamic programming*, Ann. Math. Statist., 36 (1965), pp. 226–235.

[3] S. S. CHITOPEKAR, *Continuous time Markovian sequential control processes*, this Journal, 7 (1969), pp. 367–389.

[4] C. DERMAN, *Denumerable state Markovian decision processes—average cost criterion*, Ann. Math. Statist., 37 (1966), pp. 1545–1553.

[5] ———, *Markovian sequential control processes—denumerable state space*, J. Math. Anal. Appl., 10 (1965), pp. 295–302.

[6] E. B. DYNKIN, *Markov Processes*, Academic Press, New York, 1965.

[7] W. H. FLEMING, *Optimal continuous-parameter stochastic control*, SIAM Rev., 11 (1969), pp. 470–509.

[8] E. W. HOBSON, *The Theory of Functions of a Real Variable and the Theory of Fourier's Series*, vol. I, Dover, New York, 1957.

[9] R. A. HOWARD, *Dynamic Programming and Markov Processes*, John Wiley, New York, 1960.

[10] ———, *Dynamic Probabilistic Systems*, John Wiley, New York, 1971.

[11] W. S. JEWELL, *Markov-renewal programming. I and II*, Operations Res., 11 (1963), pp. 938–971. 938–971.

[12] H. J. KUSHNER, *Optimal discounted stochastic control for diffusion processes*, this Journal, 5 (1967), pp. 520–531.

[13] S. OSAKI AND H. MINE, *Linear programming algorithms for semi-Markovian decision processes*, J. Math. Anal. Appl., 22 (1968), pp. 356–381.

[14] R. RISHEL, *Necessary and sufficient dynamic programming conditions for continuous time stochastic optimal control*, this Journal, 8 (1970), pp. 559–571.

[15] S. M. ROSS, *Nondiscounted denumerable Markovian decision models*, Ann. Math. Statist., 39 (1968), pp. 412–423.

[16] ———, *Applied Probability Models with Optimization Applications*, Holden-Day, San Francisco, 1970.

[17] L. D. STONE, *Distribution of the supremum functional for continuous state space semi-Markov processes*, Ann. Math. Statist., 40 (1969), pp. 844–853.

SIAM J. Control
Vol. 11, No. 2, May 1973

STABILITY CONSIDERATIONS FOR A VOLTERRA INTEGRAL EQUATION WITH DISCONTINUOUS NONLINEARITY*

HAJIME MAEDA†

Abstract. The asymptotic behavior of the solutions of a Volterra integral equation with discontinuous nonlinearity is investigated. The global existence of solutions is proved and some sufficient conditions for local and/or global stability are obtained by means of the Popov-type method. The sufficient conditions obtained here are the same ones obtained for finite-dimensional systems. Also the linearization problem for relay feedback systems is discussed.

1. Introduction. The equation governing a feedback system containing a relay element gives an ordinary differential equation with discontinuous right-hand side and/or a Volterra integral equation with discontinuous operator. When considering the stability problem of such a feedback system, one generally encounters the difficulty of whether or not a solution exists. For this reason, little attention has been focused on the stability problems of relay feedback systems. However, the well-known Lyapunov method and the Popov-like method, both of which are based on the qualitative estimates of the solutions, could be used to investigate the asymptotic behavior of solutions, supposing the solutions of the equations do exist and possess a "good property." By assuming the existence of solutions, some stability criteria for relay feedback systems have been obtained [1]–[3] which are concerned with the global stability problems.

Concerning the local stability problems of the feedback systems containing an ideal relay, which is characterized by the sign function, Tsypkin's conjecture is famous [4]. It states that if the linearized system of the original relay feedback system is asymptotically stable for large gains including infinity, the original relay system is locally asymptotically stable. The intuitive reasoning for his conjecture is due to the fact that the saturating nonlinearity, which is linear near the origin, approaches the ideal relay characteristic as the linear gain increases without bound. D. V. Anosov proved, by using the fact that the solutions of the linear differential equations decay exponentially or equivalently that there exists a quadratic Lyapunov function for the linearized system, that the conjecture is affirmative for finite-dimensional systems [5]. It would seem likely that the conjecture is also true for Volterra integral equations, provided the solutions of the equations exist. The linearization problem of Volterra integral equations with differentiable nonlinearities has been investigated in [6]. But the problem for discontinuous nonlinearities has not been treated up to now, partially because the solutions of a Volterra integral equation, though it is linear, need not decay exponentially.

In this paper we investigate the asymptotic behavior of the solutions of a Volterra integral equation with discontinuous nonlinearity having discontinuity at the origin, and we give several sufficient conditions for stability.

In § 3 the main results are given. We shall consider a Volterra integral equation with discontinuous nonlinearity and give a sufficient condition for local stability

* Received by the editors September 23, 1971, and in revised form March 20, 1972.

† Department of Mechanical Engineering, Osaka University, Toyonaka, Osaka, Japan.

under very general assumptions on the integral kernel. The method is the well-known Popov-type technique.

In § 4 we investigate the stability criterion obtained in § 3 and derive the well-known Tsypkin's conditions for stability under somewhat restrictive assumptions on the integral kernel.

In § 5 we give a sufficient condition for global stability which is an extension of the results of [1]–[3].

In the Appendix we give a detailed proof for the global existence of the solutions to an integral equation with discontinuous nonlinearity considered here.

2. Mathematical preliminaries and assumptions. In the sequel we consider a Volterra integral equation of the form

$$x(t) = z(t) - \int_0^t h(t - s)\phi(x(s))\,ds, \qquad t > 0, \tag{1}$$

where $h(t)$ and $z(t)$ are measurable functions defined on $(0, \infty)$, and $\phi(x)$ is a locally integrable function on $(-\infty, \infty)$ and has discontinuity of the first kind at $x = 0$.

Following [7], we define the concept of the solutions of (1) as follows.

Definition. For (1), a function $x(t)$ defined on $(0, \infty)$ is said to be a *solution* if

(i) $x(t)$ is continuous on $(0, \infty)$,

(ii) $x(t)$ satisfies

$$x(t) = z(t) - \int_0^t h(t - s)u(s)\,ds, \qquad t > 0, \tag{2}$$

for some function $u(t)$ defined almost everywhere in $(0, \infty)$ and satisfying

$$m_x\{\phi(x(t))\} \leqq u(t) \leqq M_x\{\phi(x(t))\} \tag{3}$$

for almost all $0 < t < \infty$, where

$$m_x\{\phi(x)\} = \lim_{\delta \to 0} \operatorname*{ess\,inf}_{|z-x| \leqq \delta} \phi(z),$$

$$M_x\{\phi(x)\} = \lim_{\delta \to 0} \operatorname*{ess\,sup}_{|z-x| \leqq \delta} \phi(z).$$

This concept of solutions of the integral equation with discontinuous nonlinearity is similar to Filippov's [8] which is defined for ordinary differential equations with discontinuous right-hand side. It should be noted that the concept defined above coincides with the usual one, provided the nonlinearity is continuous.

A sufficient condition guaranteeing the global existence (but not uniqueness) of a solution of (1) is given by

(a) $z(t)$ is a continuous function on $(0, \infty)$ and

(b) $\int_0^t |h(s)|\,ds < \infty$ for each $t \in (0, \infty)$.

The existence of a solution to a perturbed integral equation such as (1) is not considered in [7]; therefore, in the Appendix, we give a detailed proof for the existence of a solution, which is done in an analogous way to that for ordinary differential equations [8].

In the sequel we shall need the following notation. Let L_p denote the space of pth power integrable functions on $(0, \infty)$ with norm

$$\|x\|_p = \left(\int_0^\infty |x(t)|^p \, dt\right)^{1/p}, \qquad p = 1, 2, \cdots,$$

and let L_∞ denote the space of essentially bounded functions defined on $(0, \infty)$ with norm

$$\|x\|_\infty = \operatorname*{ess\,sup}_{t>0} |x(t)|.$$

We define

$$G(x) = \operatorname*{ess\,sup}_{|z|<x} |\phi(z)|, \qquad G(0) = \max\{|\phi(0+)|, |\phi(0-)|\}. \tag{4}$$

Observe that $G(x)$ is nondecreasing for $x \geqq 0$.

Concerning (1) we assume:

A1(a). $G(x) < \infty$ for each $0 \leqq x \leqq \rho$, $0 < \rho < \infty$.

A1(b). $x\phi(x) \geqq \gamma|x|$, $\gamma > 0$, a.e. in $|x| < \rho$.

A2. $h(t)$ is an absolutely continuous function on $(0, \infty)$ such that (a) $h \in L_1$, (b) $\dot{h} \in L_1$.

A3. $z(t)$ is an absolutely continuous function on $(0, \infty)$.

Here $\dot{h}(t)$ is the time derivative of $h(t)$, which exists for almost all $t > 0$ since $h(t)$ is assumed to be absolutely continuous. Note that A2(b) implies $h \in L_\infty$. In view of A2(a) and A3, a solution of (1) exists and is absolutely continuous on $(0, \infty)$.

Consider the linear integral equation of the form

$$\tilde{z}(t) = z(t) - k\int_0^t h(t-s)\tilde{z}(s)\, ds, \qquad t > 0. \tag{5}$$

It is well known that the unique absolutely continuous solution of (5) has the form

$$\tilde{z}(t) = z(t) - k\int_0^t g(t-s)z(s)\, ds, \qquad t > 0, \tag{6}$$

where $g(t)$ is the resolvent kernel determined by

$$g(t) = h(t) - k\int_0^t h(t-s)g(s)\, ds, \qquad t > 0. \tag{7}$$

We assume that:

A4. For some $k \geqq 0$ the solution $g(t)$ determined by (7) exists for all $t > 0$ and $g \in L_1$.

Note that if assumption A2 is satisfied, a necessary and sufficient condition for $g(t)$ to be in L_1 is that $\inf_{\operatorname{Re} s \geqq 0} |1/k + H(s)| > 0$ (see [6], [9]).

3. A general sufficient condition for local stability. In this section we investigate the asymptotic behavior of solutions of (1) with "small" $z(t)$ and give a sufficient condition under which the solution must be "small" as long as the perturbation term is "small" enough. The main result of the present paper is the following theorem.

Theorem 1. *Let assumptions* A1–A3 *be satisfied. If the assumption* A4 *and*

$$\text{A5.} \qquad \operatorname{Re}\left[(1 + j\omega q)\frac{H(j\omega)}{1 + kH(j\omega)}\right] \geqq 0, \qquad -\infty < \omega < \infty,$$

are fulfilled for some $k \geqq 0$ and some $0 < q < \infty$, where $H(j\omega)$ is the Fourier transform of $h(t)$, then for any $0 < \varepsilon < \varepsilon_1 = \min(\rho, \gamma/k)$ there exist $\delta_1 = \delta_1(\varepsilon) > 0$ ($\delta_1(\varepsilon) \to 0, \varepsilon \to 0$) and $\delta = \delta(\varepsilon) > 0$ ($\delta(\varepsilon) \to 0, \varepsilon \to 0$) such that $\|z + q\dot{z}\|_1 \leqq \delta_1$ and $|z(0+)| \leqq \delta$ imply $|x(t)| < \varepsilon$ for $t > 0$, $\|x\|_1 \leqq \varepsilon$ ($\|x\|_2 \leqq \varepsilon$, if $k > 0$) and $x(t) \to 0$ as $t \to \infty$.

Before proving the theorem we need the following lemma.

Lemma 1. *Let $\|\tilde{z} + q\dot{\tilde{z}}\|_1 \leqq c_1$ and $q(\int_0^{z(0+)} \phi(\sigma)\,d\sigma - (k/2)|z(0+)|^2) \leqq c_2$. If assumptions* A1–A5 *are satisfied, then the solution of* (1) *exists and satisfies the following relation as long as $|x(t)| \leqq \rho$, $0 < t \leqq T$:*

$$\tag{8} \begin{aligned} &\int_0^T (\gamma|x(t)| - k|x(t)|^2)\,dt + q\left(m\gamma|x(T)| - \frac{k}{2}|x(T)|^2\right) \\ &\qquad \leqq c_1\left(G(\sup_{0<t\leqq T}|x(t)|) + k\sup_{0<t\leqq T}|x(t)|\right) + c_2, \end{aligned}$$

for $0 < m \leqq 1$.

Proof of Lemma 1. As a notational convenience let $\langle a, b\rangle_T = \int_0^T a(t)b(t)\,dt$, $(a * b)(t) = \int_0^t a(t - s)b(s)\,ds$ and $a_T(t)$ be a truncated function of $a(t)$, i.e., $a_T(t) = a(t)$ for $0 < t \leqq T$, $a_T(t) = 0$ for $t > T$. It should be noted that, by means of the resolvent kernel $g(t)$ determined by (7) and $\tilde{z}(t)$ by (6), the solution $x(t)$ of (1) can be represented as follows:

$$\tag{9} x(t) = \tilde{z}(t) - g * (u - kx)(t), \qquad t > 0.$$

Hence, it satisfies

$$\tag{10} \dot{x}(t) = \dot{\tilde{z}}(t) - g(0+)(u(t) - kx(t)) - \dot{g} * (u - kx)(t)$$

for almost all $t > 0$. Now observe that the assumptions A2(b) and A4 imply $\|\dot{g}\|_1 \leqq \|\dot{h}\|_1 + k\{|h(0+)| + \|\dot{h}\|_1\}\|g\|_1 < \infty$ and so $\dot{g}(t)$ is Fourier transformable and it is given by $\mathscr{F}[\dot{g}] = j\omega\hat{g}(j\omega) - g(0+)$ where $\hat{g}(j\omega) = \mathscr{F}[g] = (H(j\omega)/1 + kH(j\omega))$.

By virtue of equations (9), (10), the Parseval formula, assumption A5 and the well-known inequality $|\langle a, b\rangle| \leqq \|a\|_\infty\|b\|_1$, we estimate

$$\tag{11} \begin{aligned} &\langle x, u - kx\rangle_T + q\langle \dot{x}, u - kx\rangle_T \\ &\quad = \langle \tilde{z} + q\dot{\tilde{z}}, u - kx\rangle_T - \langle (g + q(g(0+) + \dot{g})) * (u - kx), u - kx\rangle_T \\ &\quad = \langle \tilde{z} + q\dot{\tilde{z}}, u - kx\rangle_T - \frac{\operatorname{Re}}{2\pi}\int_{-\infty}^{\infty} (1 + j\omega q)\hat{g}(j\omega)|\hat{u}_T(j\omega) - k\hat{x}_T(j\omega)|^2\,d\omega \\ &\quad \leqq \|\tilde{z} + q\dot{\tilde{z}}\|_1\left\{G\left(\sup_{0<t\leqq T}|x(t)|\right) + |k|\sup_{0<t\leqq T}|x(t)|\right\} \\ &\quad \leqq c_1\left\{G\left(\sup_{0<t\leqq T}|x(t)|\right) + k\sup_{0<t\leqq T}|x(t)|\right\}, \end{aligned}$$

where $\hat{u}_T(j\omega)$ and $\hat{x}_T(j\omega)$ are the Fourier transforms of $u_T(t)$ and $x_T(t)$ respectively.

On the other hand, by using the assumption A1 and the fact $x(0+) = z(0+)$, we estimate

$$\langle x, u - kx\rangle_T + q\langle \dot{x}, u - kx\rangle_T \geqq \int_0^T (\gamma|x| - k|x|^2)\,dt$$

$$(12) \qquad + q\left(\int_0^{x(T)} \phi\,d\sigma - \frac{k}{2}|x(T)|^2\right) - q\left(\int_0^{z(0+)} \phi\,d\sigma - \frac{k}{2}|z(0+)|^2\right)$$

$$\geqq \int_0^T (\gamma|x| - k|x|^2)\,dt + q\left(\gamma|x(T)| - \frac{k}{2}|x(T)|^2\right) - c_2.$$

Combining (11) and (12), we obtain, in view of $m \leqq 1$, the desired form (8). Thus Lemma 1 is established.

Proof of Theorem 1. We first prove Theorem 1 for $k > 0$. Let $0 < \varepsilon < \varepsilon_1 = \min(\rho, \gamma/k)$. Choose numbers $0 < m < 1$, $c_1 = c_1(\varepsilon) > 0$ and $c_2 = c_2(\varepsilon) > 0$ so as to satisfy

$$(13\text{a}) \qquad \gamma m/k \leqq \varepsilon,$$

$$(13\text{b}) \qquad q\left(\gamma m|x| - \frac{k}{2}|x|^2\right) \geqq c_1[G(|x|) + k|x|] + c_2 \quad \text{for } |x| = \frac{\gamma m}{k},$$

$$(13\text{c}) \qquad \frac{c_1[G(\gamma m/k) + k(\gamma m/k)] + c_2}{(1-m)\gamma} \leqq \varepsilon.$$

It is clear that $c_1 = c_1(\varepsilon) \to 0$ and $c_2 = c_2(\varepsilon) \to 0$ as $\varepsilon \to 0$ since $G(0) \geqq \gamma > 0$. For each $c_2(\varepsilon)$ we can find a $0 < \delta_0 = \delta_0(\varepsilon) < \gamma m/k$ such that

$$(14) \qquad 0 \leqq q\left(\int_0^x \phi(\sigma)\,d\sigma - \frac{k}{2}|x|^2\right) \leqq c_2 \quad \text{for any } |x| \leqq \delta_0.$$

Since $|z(0+)| \leqq \delta_0$ implies $|x(0+)| \leqq \delta_0 < \gamma m/k$, and the solution $x(t)$ is continuous, it follows that $|x(t)| < \gamma m/k$ for sufficiently small $t > 0$. Now suppose there exists a $t^* > 0$ such that $\sup_{0<t\leqq t^*}|x(t)| = |x(t^*)| = \gamma m/k$. Then from (13b) and Lemma 1, i.e., (8), we get

$$0 \geqq c_1[G(|x(t^*)|) + k|x(t^*)|] + c_2 - q\left(\gamma m|x(t^*)| - \frac{k}{2}|x(t^*)|^2\right)$$

$$\geqq \int_0^{t^*} (\gamma|x| - k|x|^2)\,dt \geqq 0,$$

which contradicts the fact that $x(t)$ is a continuous function with $|x(0+)| < \gamma m/k$ and $|x(t^*)| = \gamma m/k$. Hence there is no such t^*, i.e., $|x(t)| < \gamma m/k \leqq \varepsilon$ for all $t > 0$. In view of the fact that $|x(t)| < \gamma m/k$ for all $t > 0$, we get from (8) and (13c), $\int_0^t \gamma|x(s)|\,ds \leqq k\int_0^t |x(s)|^2\,ds + (1-m)\gamma\varepsilon \leqq \gamma m\int_0^t |x(s)|\,ds + (1-m)\gamma\varepsilon$, which implies $\|x\|_1 \leqq \varepsilon$ and $\|x\|_2^2 \leqq (\gamma m/k)\|x\|_1 \leqq \varepsilon^2$ by (13a). We have thus proved that for any $0 < \varepsilon < \varepsilon_1$ there exist $c_1 > 0$ $(c_1(\varepsilon) \to 0, \varepsilon \to 0)$ and $\delta_0 > 0$ $(\delta_0(\varepsilon) \to 0, \varepsilon \to 0)$ such that $\|\tilde{z} + q\dot{\tilde{z}}\|_1 \leqq c_1$ and $|z(0+)| \leqq \delta_0$ imply $\|x\|_p \leqq \varepsilon$, $p = 1, 2, \infty$.

In order to complete the proof for $k > 0$ we have to show the existence of $\delta_1 > 0$ and $\delta > 0$ stated in Theorem 1. Let $\delta_1 > 0$ be a constant satisfying $\delta_1 < c_1/(1 + k\|g\|_1)$. Then there exists $\delta' > 0$ such that $\|\tilde{z} + q\dot{\tilde{z}}\|_1 \leqq c_1$ for $|z(0+)| \leqq \delta'$, which is easily shown by the relation $\|\tilde{z} + q\dot{\tilde{z}}\|_1 \leqq (1 + k\|g\|_1)\|z + q\dot{z}\|_1 + k|z(0+)|\|g\|_1$. Let $\delta = \min(\delta_0, \delta')$. Then the constants δ_1 and δ determined as above are the desired constants. $\delta_1(\varepsilon) \to 0$ and $\delta(\varepsilon) \to 0$ as $\varepsilon \to 0$ is clear. Note that $z + q\dot{z} \in L_1, q > 0$, implies that $z, \dot{z} \in L_1$, hence $\lim_{t\to\infty} z(t) = 0$. Therefore, $x - z = -h * u \in L_1$ since $x \in L_1$ and $z \in L_1$, and $(h * u)(t)$ is uniformly continuous since $h \in L_1$ and $|u(t)| \leqq G(\gamma m/k) < \infty$(see [10]). Therefore, $(h * u)(t) \to 0$ as $t \to \infty$, which implies $x(t) = z(t) - (h * u)(t) \to 0$ as $t \to \infty$ (see [10]). Thus the theorem has been proved for $k > 0$.

We shall next prove the theorem for $k = 0$. Lemma 1 holds for $k = 0$, i.e., letting $k = 0$ and $m = 1$ in (8) leads to

$$\int_0^T \gamma|x(t)|\,dt + q\gamma|x(T)| \leqq c_1 G\left(\sup_{0<t\leqq T}|x(t)|\right) + c_2, \tag{15}$$

as long as $|x(t)| \leqq \rho, 0 < t \leqq T$, where $c_1 \geqq \|z + q\dot{z}\|_1, c_2 \geqq \int_0^{z(0+)}\phi(\sigma)\,d\sigma$. Given $0 < \varepsilon \leqq \rho$, choose $c_1 = c_1(\varepsilon)$ and $c_2 = c_2(\varepsilon)$ so as to satisfy

$$q\gamma|x| \geqq c_1 G(|x|) + c_2, \quad \text{for } |x| = \varepsilon, \tag{16a}$$

$$(c_1 G(\varepsilon) + c_2)/\gamma \leqq \varepsilon. \tag{16b}$$

It is easily shown that $c_1(\varepsilon) \to 0$ and $c_2(\varepsilon) \to 0$ as $\varepsilon \to 0$. For each $c_2 = c_2(\varepsilon)$ determined above we can find a $0 < \delta = \delta(\varepsilon) < \varepsilon$ such that

$$q\int_0^x \phi(\sigma)\,d\sigma \leqq c_2 \quad \text{for any } |x| \leqq \delta. \tag{17}$$

Note that $|x(t)| < \varepsilon$ for sufficiently small $t > 0$ if $|z(0+)| < \delta$. Now suppose there exists a $t^* > 0$ such that $\sup_{0<t\leqq t^*}|x(t)| = |x(t^*)| = \varepsilon$. Then from (16a) and (15) we get $0 \geqq c_1[G(|x(t^*)|)] + c_2 - q\gamma|x(t^*)| \geqq \int_0^{t^*}|x(t)|\,dt \geqq 0$, which is a contradiction. Hence $|x(t)| < \varepsilon$ for all $t > 0$. From (15) and (16b) we obtain $\|x\|_1 \leqq (c_1G(\varepsilon) + c_2)/\gamma \leqq \varepsilon$. $c_1 = \delta_1(\varepsilon)$ and $\delta = \delta(\varepsilon)$ are desired constants. The property $x(t) \to 0$, $t \to \infty$ can be shown in the way stated above. Thus the theorem is completely proved.

We remark that the result of Theorem 1 is local (not necessarily global) in the sense that if the perturbation term $z(t)$ is "small" enough, more precisely $\|z\|_1$, $\|\dot{z}\|_1$ and $|z(0+)|$ are small enough, then the solution is "small", i.e., $\|x\|_\infty$ and $\|x\|_1(\|x\|_2)$ are small. In order to extend the results to assure the global stability, which means that for any $z(t)$ satisfying $z \in L_1$ and $\dot{z} \in L_1$ the solution is bounded and integrable, i.e., $x \in L_\infty$ and $x \in L_1$, we need the assumption $k = 0$. This is clear from the proof of Theorem 1. However, the proof of the theorem suggests that we can estimate a region of asymptotic stability, i.e., we can estimate quantities $\alpha_1 \geqq \|z\|_1, \alpha_2 \geqq \|\dot{z}\|_1$ and $\delta \geqq |z(0+)|$ by which the asymptotic stability of the null solution is guaranteed.

Before considering the global asymptotic stability, we shall investigate the local stability conditions obtained here in the following section.

4. Tsypkin's conditions for local asymptotic stability. In this section we shall show that Tsypkin's conditions for stability obtained for finite-dimensional systems are also applicable for Volterra integral equations. In order to investigate the stability conditions of Theorem 1, we need the following additional condition:

A6. If $h(0+) = 0$, then $\dot{h}(t)$ and $\ddot{h}(t)$ are absolutely continuous functions defined on $(0, \infty)$ such that $\ddot{h}(t), \dddot{h}(t) \in L_1$.

THEOREM 2. *Let assumptions* A1–A3 *and* A6 *be satisfied. Let $H(s)$ be the Laplace transform of $h(t)$. If the following conditions hold, then the conclusion of Theorem* 1 *is valid*:

A7. $$\inf_{\operatorname{Re} s \geqq 0} |H(s)| > 0,$$

A8. $$\begin{aligned} &h(0+) > 0 \quad \textit{if} \quad h(0+) \neq 0, \\ &\dot{h}(0+) > 0 \quad \textit{and} \quad \ddot{h}(0+) < 0 \quad \text{if} \quad h(0+) = 0. \end{aligned}$$

This result shows that Tsypkin's conditions for stability are applicable not only for finite-dimensional systems but also for distributed parameter systems.

Before we prove the theorem the following remarks are needed: Since $H(s)$ considered here need not be holomorphic in the annulus $|s| > R$ for sufficiently large R, the Laurent expansion used in [4] is, in general, not applicable. However, the properties $h^{(i)} \in L_1, i = 0, 1, \cdots, n$, and $|h^{(i)}(0+)| < \infty, i = 0, 1, \cdots, n-1$, admit the following expansion:

$$H(s) = \sum_{k=1}^{n} h^{(k-1)}(0+)s^{-k} + H_1(s), \qquad \operatorname{Re} s \geqq 0, \quad s \neq 0, \tag{18}$$

$$|s^n H_1(s)| = o(|s|^{-1}), \qquad \operatorname{Re} s \geqq 0, \quad s \to \infty,$$

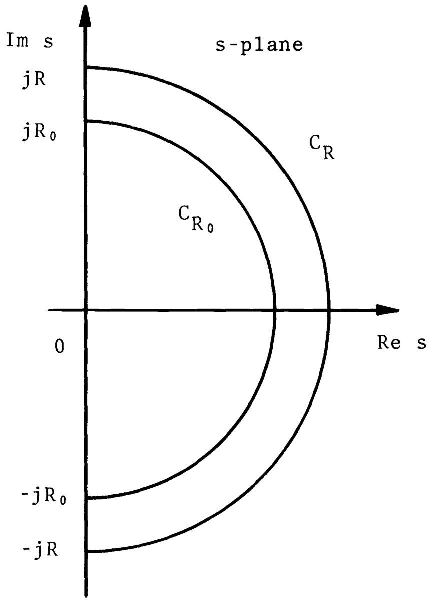
which can be shown by the formula for the Laplace transform of the nth derivative, that is by setting $H_1(s) = s^{-n}\mathscr{L}[h^{(n)}(t)]$ in the formula.

Proof of Theorem 2. Let a_i and ω_i be nonnegative numbers such that $\operatorname{Re}[H(j\omega_i)] = -a_i$ and $\operatorname{Im}[H(j\omega_i)] = 0$. If such an a_i does not exist, define $a_i = +\infty$. In the formula (18), setting $s = j\omega$ and $n = 1, 2$, we get $\operatorname{Im}[H(j\omega)] < 0$ for large ω since $h(0+) > 0$ or $\dot{h}(0+) > 0$. Therefore, $\inf_i a_i = a$, $0 \leqq a \leqq \infty$ exists but A7 implies $a > 0$. Note that for each $\mu \in (0, a)$ there exists $R_0 > 0$ such that

$$|H(s)| < \mu \quad \text{for all } s \in \{s; |\arg s| \leqq \pi/2, |s| \geqq R_0\}.$$

Let C_R denote the contour of Fig. 1 consisting of a semicircle $|s| = R$, $|\arg s| \leqq \pi/2$, joined by the line segment $(-jR, jR)$ of the imaginary axis. Let $R \geqq R_0$. Then the image $H(C_R)$ in the H-plane does not pass through the point $(-\mu, j0)$ and also does not encircle the point, since the image $H(C_R)$ does not encircle and pass through the origin in the H-plane (by A7 and the principle of the argument [11]). Therefore, for the function defined by $f(s) = \mu + H(s)$, the increment of the argument of $f(s)$ is equal to zero, namely $\Delta_{C_R} \arg f(s) = 0$ for $R \geqq R_0$. Since the function $f(s)$ is meromorphic in $\operatorname{Re} s \geqq 0$ and has neither zeros nor poles on C_R it follows by the principle of argument, that

$$2\pi Z_f = \Delta_{C_R} \arg f(s) = 0 \quad \text{for } R \geqq R_0,$$



where Z_f is the numbers of zeros of $f(s)$ in the interior of C_R. Thus we get

$$\inf_{\operatorname{Re} s \geqq 0} |\mu + H(s)| > 0, \qquad 0 < \mu < a.$$

This and the well-known Paley–Wiener theorem [9] imply that the assumption A4 is satisfied for $k' > 1/a$.

Next let $X(\omega) = \operatorname{Re} H(j\omega)$ and $Y(\omega) = \operatorname{Im} H(j\omega)$, $-\infty < \omega < \infty$. It is clear from (18) and assumptions A6, A8 that

$$X(\omega) - \omega q Y(\omega) = qh(0+) + o(|\omega|^{-1}), \qquad \omega \to \infty,$$

if $h(0+) \neq 0$, or

$$X(\omega) - \omega q Y(\omega) = \{-\dot{h}(0+) + q|\ddot{h}(0+)| + o(|\omega|^{-1})\}/\omega^2, \qquad \omega \to \infty,$$

if $h(0+) = 0$. Therefore let $q > 0\,(h(0+) \neq 0)$ or let $q > \dot{h}(0+)/|\ddot{h}(0+)|\,(h(0+) = 0)$. Then there exists an ω_0 such that $X(\omega) - \omega q Y(\omega) \geqq 0$ for $\omega_0 \leqq |\omega| < \infty$. Let $k'' \geqq \sup_\omega |X - \omega q Y|/\inf_\omega |H(j\omega)|^2$ (the right-hand side is finite by virtue of A2 and A7). Then it follows, by direct calculation, that $\operatorname{Re}[(1 + j\omega q)H(j\omega)/(1 + k''H(j\omega))] \geqq 0$, for all $-\infty < \omega < \infty$. It is clear that the assumptions A4, A5 of Theorem 1 are satisfied for $k > \max\{k', k''\}$. Thus the proof of Theorem 2 is established.

Note that the sufficient conditions A7, A8 for stability are the same ones suggested by Tsypkin, which were proved by D. V. Anosov for finite-dimensional systems.

Now we prove that the stability conditions A7, A8 can be deduced from the stability conditions for linear systems.

THEOREM 3. *Let the assumptions* A1–A3 *and* A6 *be satisfied. If the following conditions hold, then the conclusion of Theorem* 1 *is valid*:

$$\text{A9.} \qquad h(0+) \neq 0 \quad \textit{or} \quad \dot{h}(0+)\ddot{h}(0+) \neq 0 \quad (h(0+) = 0),$$

$$\text{A10.} \qquad \inf_{\operatorname{Re} s \geqq 0} |\mu + H(s)| > 0, \qquad 0 \leqq \mu \leqq \mu_0, \quad \mu_0 > 0.$$

Proof of Theorem 3. We only show that the assumption A8 can be deduced from A9, A10. First note that $H(\sigma) > 0$ for real $\sigma \geqq 0$. This can be shown as follows. From A10, the function $\mu + H(s)$ has the constant sign for real $s = \sigma \geqq 0$ and $0 \leqq \mu \leqq \mu_0$. Let $\mu > 0$ and $\sigma \to \infty$. It follows that $\mu + H(\sigma) \to \mu > 0, \sigma \to \infty$. Thus $\mu + H(\sigma) > 0$ for $\sigma \geqq 0$, where μ is an arbitrary number belonging to $[0, \mu_0]$. Therefore the statement given above is obtained. From A9 and the fact that $\lim_{\sigma \to \infty} \sigma H(\sigma) = h(0+)$, it follows that $h(0+) > 0$ if $h(0+) \neq 0$. Similarly, when $h(0+) = 0$, $\lim_{\sigma \to \infty} \sigma^2 H(\sigma) = \dot{h}(0+)$ implies $\dot{h}(0+) > 0$. Finally, let $n = 3$ and $s = \sigma + j\omega$, $\sigma \geqq 0$, in (18). Then we get $\operatorname{Re} H(\sigma + j\omega) = [-\dot{h}(0+) + o(|\omega|^{-1})]/\omega^2 < 0, |\omega| \to \infty$, and

$$\operatorname{Im} H(\sigma + j\omega) = \frac{\omega^3(\ddot{h}(0+) - 2\sigma\dot{h}(0+)) - (3\sigma^2\ddot{h}(0+) + 2\sigma^3\dot{h}(0+) + o(|\omega|^{-1}))\omega}{(\sigma^2 + \omega^2)^3},$$

$$|\omega| \to \infty.$$

Suppose $\ddot{h}(0+) > 0$. It is clear by the above equation that there exists ω_0 such that $\operatorname{Im} H(\sigma + j\omega_0) = 0$, and $\omega_0 \to \infty$ as $\sigma \to \ddot{h}(0+)/2\dot{h}(0+) > 0$. Therefore $\operatorname{Re} H(\sigma + j\omega_0) \to 0$ as $\sigma \to \ddot{h}(0+)/2\dot{h}(0+)$. This implies that for sufficiently small $\mu > 0$, $\inf_{\operatorname{Re} s \geqq 0} |\mu + H(s)| = 0$, which contradicts A10. Thus the proof of Theorem 3 is established.

We remark that the stability conditions A7, A8 are equivalent to A10 if the assumption A9 is satisfied. The fact that A7, A8 implies A10 is clear by the proof of Theorem 2.

We also remark that under the constraint $h \in L_1$, the assumption A10 with $0 < \mu \leqq \mu_0$ is a necessary and sufficient condition for $g(t) = \mathscr{L}^{-1}[H(s)/(1 + kH(s))]$ to be absolutely integrable over $(0, \infty)$ for each $1/\mu_0 \leqq k < \infty$, or equivalently a necessary and sufficient condition for $x = z - kh * x$ to be L_2 input-output stable, i.e., $\|x\|_2 \leqq K\|z\|_2$ (see [12]). Thus we may conclude that under the constraints A2, A6, A9, if the linearized system of the relay feedback system is L_2 input-output stable for sufficiently large gain including infinity (for L_2-stability for infinite gain it is understood that $\inf_{\operatorname{Re} s \geqq 0} |H(s)| > 0$), then the original relay feedback system is stable in the sense of Theorem 1.

5. A sufficient condition for global stability. Thus far we have been concerned with the asymptotic behavior of the solutions of the discontinuous integral equation with "small" perturbations and we have derived sufficient conditions for stability. In this section we are concerned with global stability. As stated in § 3, in order to investigate the asymptotic behavior of the solution in the global sense, using the approach proposed here, we need $k = 0$.

We assume that:

A1(a)′ $\qquad x\phi(x) \geqq \gamma|x|, \qquad \gamma > 0 \text{ a.a.} \quad -\infty < x < \infty,$

A1(b)′ $\qquad G(x)/x \to 0 \quad \text{as } |x| \to \infty.$

Note that A1(a)′ implies that the graph of the characteristic $\phi(\cdot)$ must lie in the first and third quadrants. The key result is as follows

THEOREM 4. *Let assumptions* A1(a)′, A1(b)′ *and* A2, A3 *be satisfied. If*

A11. $\qquad \operatorname{Re}[(1 + j\omega q)H(j\omega)] \geqq 0, \qquad -\infty < \omega < \infty,$

is fulfilled for some $0 < q < \infty$, *then for any* $z(t)$ *satisfying* $\|z + q\dot{z}\|_1 < \infty$, $x \in L_\infty$, $x \in L_1$ *and* $x(t) \to 0$ *as* $t \to \infty$.

Proof of Theorem 4. Note that A1(b)′ implies that for any $\varepsilon > 0$ there exists a nonnegative number $b = b(\varepsilon)$ such that

$$G(|x|) \leqq \varepsilon|x| + b, \qquad -\infty < x < \infty. \tag{19}$$

Note also that the solution of (1) satisfies (15), i.e.,

$$\int_0^T \gamma|x(t)|\,dt + q\gamma|x(T)| \leqq \|z + q\dot{z}\|_1 G\left(\sup_{0<t\leqq T}|x(t)|\right) + \int_0^{x(0+)} \phi\,d\sigma. \tag{20}$$

Let $0 < \varepsilon < q\gamma/\|z + q\dot{z}\|_1$. Combining (19) and (20), and letting $\sup_{0<t\leqq T}|x(t)| = |x(t^*)|$, we get by direct calculation,

$$|x(t^*)| \leqq \frac{b + \int_0^{x(0+)}\phi\,d\sigma}{q\gamma - \|z + q\dot{z}\|_1\varepsilon} = M_1 < \infty,$$

which implies $\sup_{t>0}|x(t)| \leqq M_1$. From (20) it follows clearly that

$$\int_0^T |x(t)|\,dt \leqq \frac{\|z + q\dot{z}\|_1 G(M_1) + \int_0^{x(0+)}\phi d\sigma}{\gamma} = M_2 < \infty$$

which implies $\|x\|_1 \leqq M_2$. The last statement of the theorem can be proved in the same way as in Theorem 1.

The results obtained here are global, i.e., the solution $x(t)$ is bounded, integrable over $(0, \infty)$ and tends to zero as long as the perturbation $z(t)$ satisfies $z \in L_1$, $\dot{z} \in L_1$.

It should be noted that the advantage of Theorem 4 is that we can apply the well-known, Popov-type, stability criterion of the form A11, obtained in [1]–[3], for the wide class of feedback systems with discontinuous nonlinearities. Also, it could be shown that if we consider a more restrictive stability criterion $\operatorname{Re}[(1 + j\omega q)H(j\omega)] \geqq \delta > 0$, we can delete A1(b)′.

6. Conclusion. In this paper we have treated the local and global stability problems of a Volterra integral equation with discontinuous nonlinearity and we have shown that under mild conditions, Tsypkin's conjecture, which was proved for finite-dimensional systems, is also affirmative for an integral equation. By the results obtained here, it is reasonable to apply Tsypkin's conditions for stability for a feedback system consisting of an ideal relay and distributed parameter system.

Appendix. In this Appendix we assume, concerning (1):

(I) $\phi(x)$ is a locally integrable function on $-\infty < x < \infty$.
(II) $z(t)$ is a continuous function on $(0, \infty)$.
(III) $h(t)$ is a measurable function satisfying

$$\int_0^t |h(s)|\,ds < \infty \quad \text{for each } t > 0.$$

The key result of the Appendix is the following theorem.

THEOREM A.1. *Suppose the assumptions* (I), (II) *and* (III) *are satisfied. Then at least one solution of* (1) *exists in the sense of the definition on* $(0, \infty)$.

The theorem is an immediate consequence of the following two lemmas.

LEMMA A.1 (Local existence of the solution). *Under the assumptions* (I), (II) *and* (III) *at least one solution of* (1) *exists on* $(0, t_0)$, $0 < t_0 < \infty$.

Proof of Lemma A.1. Let $x_k(t)$ be a unique continuous solution of the equation

$$\text{(A.1)} \qquad x_k(t) = z(t) - \int_0^t h(t - s)\phi_k(x_k(s))\,ds, \qquad t > 0,$$

where $\phi_k(x) = (1/(2r_k))\int_{x-r_k}^{x+r_k} \phi(\sigma)\,d\sigma$, $\{r_k\}$ is a sequence of positive numbers such that $r_k \to 0$ as $k \to \infty$. Such a solution $x_k(t)$ exists for all $t > 0$ by virtue of (II), (III) and the fact that $\phi_k(\cdot)$ is continuous on $(-\infty, \infty)$. Note that since for any $m > 0$ there exists a number k_0 such that $0 < r_k < m$ for all $k > k_0$, it follows that $|\phi_k(\sigma)| \leqq M$ for $|\sigma| \leqq 2m, k > k_0$ where $M = \operatorname{ess\,sup}_{|\sigma| \leqq 3m} |\phi(\sigma)|$. Let $m = \sup_{0 < t \leqq t_1} |z(t)|$, and $P(t) = \int_0^t |h(s)|\,ds$ and let t_0 be a positive number determined by the relation $P(t_0)M \leqq m$. Since $P(0) = 0$ and $P(t)$ is continuous, monotone nondecreasing, the existence of such a t_0 is clear. Then by (A.1), we obtain $|x_k(t)| \leqq m + P(t)M \leqq 2m, 0 < t \leqq t_0, k > k_0$, which implies $|\phi_k(x_k(t))| \leqq M, 0 < t \leqq t_0, k > k_0$. That is, the sequence $\{x_k(t)\}$ is uniformly bounded on $(0, t_0]$ and equicontinuous since

$$|x_k(t + \delta) - x_k(t)| \leqq |z(t + \delta) - z(t)| + M\left\{\int_0^{t_0} |h(s + \delta) - h(s)|\,ds + \int_0^\delta |h(s)|\,ds\right\}$$
$$= \alpha(\delta) \to 0 \quad \text{as } \delta \to 0.$$

Applying the lemma of Ascoli–Arzela, the sequence $\{x_k\}$ contains a subsequence $\{x_{k_i}\}$ converging uniformly to a continuous limit function $x(t)$ on $(0, t_0]$. Let $u_i(t) = \phi_{k_i}(x_{k_i}(t)), k_i > k_0$. The sequence $\{u_i(t)\}$ is uniformly bounded on $(0, t_0]$, i.e., $|u_i(t)| \leqq M, 0 < t \leqq t_0, k_i > k_0$, which guarantees the existence of a weakly convergent subsequence [13]. Let us denote it again by $\{u_i(t)\}$ and the limit function by $u(t)$, such that for any $f \in L_1(0, t_0)$,

$$\int_0^{t_0} f(s)u_i(s)\,ds \to \int_0^{t_0} f(s)u(s)\,ds, \qquad i \to \infty.$$

Therefore, from (A.1) with $k = k_i$, we obtain, letting $i \to \infty$,

$$x(t) = z(t) - \int_0^t h(t - s)u(s)\,ds, \qquad 0 < t \leqq t_0.$$

The remaining part we must show is that the function $u(t)$ determined above satisfies the relation (3). Let $\{x_k(t)\}$ be the subsequence converging uniformly to $x(t)$. For any $\delta > 0$ there exists a number $k(\delta)$ such that $|x_k(t) - x(t)| \leqq \delta/2$ for $0 < t \leqq t_0, k > k(\delta)$ and $0 < r_k < \delta/2$ for $k > k(\delta)$. Therefore, for any real number α, we obtain the relation

$$u_k(t)\alpha \leqq \operatorname*{ess\,sup}_{|z - x_k(t)| \leqq \delta/2} \phi(z)\alpha \leqq \operatorname*{ess\,sup}_{|z - x(t)| \leqq \delta} \phi(z)\alpha, \qquad 0 < t \leqq t_0, \quad k > k(\delta),$$

from which, furthermore, we obtain

$$\int_{t'}^{t''} u_k(t)\alpha\, dt \leqq \int_{t'}^{t''} \operatorname*{ess\,sup}_{|z - x(t)| \leqq \delta} \phi(z)\alpha\, dt, \qquad 0 < t' < t'' \leqq t_0, \quad k > k(\delta).$$

On dividing the inequality by $t'' - t'$ and letting $t'' \to t' + 0$ and after that letting $\alpha = +1$ and $\alpha = -1$, we obtain the desired form.

LEMMA A.2 (Continuation of the solution). *Let $x_1(t)$ $(u_1(t))$ be a solution of* (1) *on* $(0, t_0]$. *Let $x_2(t)$ $(u_2(t))$ be a solution of the equation*

$$\text{(A.2)} \qquad x(t) = \hat{z}(t) - \int_0^t h(t - s)\phi(x(s))\, ds, \qquad 0 < t \leqq \beta, \quad \beta > 0,$$

where $\hat{z}(t) = z(t + t_0) - \int_0^{t_0} h(t + t_0 - s)u_1(s)\, ds$. Then $x(t)$ defined by $x(t) = x_1(t)$, $0 < t \leqq t_0$ and $x(t) = x_2(t - t_0), t_0 < t \leqq t_0 + \beta$ is a solution on $(0, t_0 + \beta]$.

The proof is clear by direct calculation.

Proof of Theorem A.1. By Lemma A.1 there exists a solution $x_1(t)$ on $(0, t_0]$, $t_0 > 0$. Observing that $z(t)$ is continuous, (A.2) has a solution $x_2(t)$ on $(0, \beta]$, $\beta > 0$. Therefore, by Lemma A.2, the solution can be continued to $(0, t_0 + \beta]$. Repeating this process we can prove Theorem A.1.

Acknowledgment. The author is indebted to S. Kodama and H. Shirakawa for suggesting this area of research and for their help and encouragement during its development, and to the referee who made many useful comments throughout the paper.

REFERENCES

[1] A. KH. GELIG, *Investigation of stability of nonlinear discontinuous automatic control systems with a nonunique equilibrium state*, Avtomat. i Telemekh., 25 (1964), pp. 153–160 = Automat. Remote Control, 25 (1964), pp. 141–148.

[2] H. MAEDA AND S. KODAMA, *Analysis of the stability problem of relay servo-systems*, Trans. Inst. Electronics Comm. Engrs. Japan, 52-C (1969), pp. 143–150.

[3] V. A. YACUBOVICH, *Frequency conditions for the stability of a class of nonlinear integral control equations*, Vestnik. Leningrad Univ., (1967), pp. 109–125.

[4] YA. Z. TSYPKIN, *Theory of Relay Control Systems*, Gostekhizdat, Moscow, 1955.

[5] D. V. ANOSOV, *Stability of equilibrium position in relay system*, Automat. Remote Control, 20 (1959), pp. 130–143.

[6] R. K. MILLER, *On the linearization of Volterra integral equations*, J. Math. Anal. Appl., 23 (1968), pp. 198–208.

[7] N. V. AZBELEV, LI MUN SU AND R. E. RAGIMHANOV, *Defining the concept of a solution to an integral equation with discontinuous operator*, Soviet Math. Dokl., 7 (1966), pp. 1437–1440.

[8] A. F. FILIPPOV, *Differential equations with discontinuous right-hand side*, Math. Sb., 51 (1960), pp. 99–128. English transl., Amer. Math. Soc. Transl., 42 (1964), pp. 199–231.

[9] R. E. A. C. PALEY AND N. WIENER, *Fourier Transform in the Complex Domain*, American Mathematical Society, New York, 1934.

[10] C. A. DESOER, *A general formulation of the Nyquist criterion*, IEEE Trans. Circuit Theory, CT-12 (1965), pp. 230–234.

[11] E. HILLE, *Analytic Function Theory*, vol. I, Ginn, New York, 1959.

[12] J. C. WILLEMS, *The Analysis of the Feedback Systems*, MIT Press, Cambridge, Mass., 1971.

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

SIAM J. CONTROL
Vol. 11, No. 2, May 1973

STRUCTURE OF INDEX-INVARIANT SYSTEMS*

A. S. MORSE† AND L. M. SILVERMAN‡

Abstract. This paper introduces the concept of a controllability module and then uses it to study certain algebraic properties of index-invariant, time-varying, linear systems. It is shown that controllability modules possess a property similar to the pole placement property of controllability subspaces. Application of controllability modules to index-invariant systems leads to a new construction of Brunovský's decomposition which serves to clarify the algebraic relationships upon which the decomposition is based.

Introduction. In a recent article [1], controllability subspaces and related algebraic concepts were introduced and successfully applied to the problem of noninteracting control. Although these concepts have since proved useful in discussing other problems of system synthesis, they are effectively limited to time-invariant linear multivariable systems. The present article is motivated by a desire to extend these concepts to a broader class of physically significant processes; namely, time-varying linear systems.

To treat the time-varying case, it is useful to regard the entries in the system coefficient matrices as elements of a suitably defined ring of time functions (§ 1). Modules thus replace vector spaces and the controllability module becomes the principal algebraic object of interest (§ 2).

It is known that the state space of a controllable linear constant system admits a canonical decomposition into independent controllability subspaces, each subspace being generated by a single vector [2]. An analogous decomposition for a restricted class of time-varying systems is developed in § 3. This result and the corollary which follows it coincide with earlier results due to Brunovský [3] and thus we cannot claim originality. Nevertheless, the present approach is new and should be of interest in that it clearly exhibits the algebraic relationships upon which the constructions of [3] are based.

Our results, which are all of an algebraic nature, depend on a useful theorem due to Doležal [4] which characterizes the range and null spaces of constant rank time-varying matrices. In § 4 we summarize this theorem and provide an interpretation in module theoretic terms.

1. Preliminaries. *Notation.* Let $I \equiv [t_0, t_1]$ be a closed interval on the real line and let R denote the ring of continuous, infinitely differentiable, real-valued functions, $I \to$ reals, together with pointwise addition and multiplication; $\bar{R} \subset R$ denotes the subring (field) of functions in R which are constant on I. Vector spaces over $\bar{R}$ are denoted by script letters with overbars; $\bar{\mathscr{R}}^n$ is n-dimensional

* Received by the editors September 24, 1971, and in revised form May 4, 1972.

† Department of Engineering and Applied Science, Yale University, New Haven, Connecticut 06520. The research of this author was supported in part by the National Science Foundation under Grant GK 30098.

‡ Department of Electrical Engineering, University of Southern California, Los Angeles, California 90007. The research of this author was supported by the U.S. Army Research Office–Durham under Grant DA-ARO-D-31-124-71-637.

space and $\mathbf{e}_n \equiv \{e_1, e_2, \cdots, e_n\}$ is a basis for $\bar{\mathscr{R}}^n$. Modules over R are denoted by script letters; $\mathscr{R}^n$ is the free module generated by $\mathbf{e}_n$. If $x \in \mathscr{R}^n$, $\mathscr{x}$ denotes the free module generated by x. Whenever $\mathscr{S}$ is free, $d(\mathscr{S}) \equiv$ dimension $\mathscr{S}$; $\mathscr{S} \approx \mathscr{T}$ means $\mathscr{S}$ is isomorphic to $\mathscr{T}$. The symbol $\sum_n$ denotes the class of all free submodules of $\mathscr{R}^n$ which are also direct summands of $\mathscr{R}^n$. If k is an integer, $\mathbf{k} \equiv \{1, 2, \cdots, k\}$.

Each $n \times m$ matrix M over R is considered to be the unique representation of a morphism $\mathscr{R}^m \to \mathscr{R}^n$ in the bases $\mathbf{e}_m$ and $\mathbf{e}_n$; a matrix and its corresponding morphism are denoted by the same symbol. The function $\circ: R \to R$ is the ring derivation defined by $r \mapsto dr/dt$; $D: \mathscr{R}^n \to \mathscr{R}^n$ denotes the function $x \mapsto \sum_{i \in \mathbf{n}} e_i \dot{r}_i$, where $x = \sum_{i \in \mathbf{n}} e_i r_i$, $e_i \in \mathbf{e}_n$ and $r_i \in R$.

If $M: \mathscr{R}^m \to \mathscr{R}^n$ is a function, $\ker M \equiv \{x : Mx = 0, x \in R^m\}$; if $\mathscr{S} \subset \mathscr{R}^m$, $M\mathscr{S}$ denotes the image of $\mathscr{S}$ under M; if M is an R-morphism, $\mathscr{M} \equiv M\mathscr{R}^m$. The notation $D\mathscr{S} + \mathscr{S} \equiv \{x : x = Ds_1 + s_2, s_i \in \mathscr{S}\}$ is an R-module.

In this article we shall be specifically concerned with R-morphisms $A: \mathscr{R}^n \to \mathscr{R}^n$ and $B: \mathscr{R}^m \to \mathscr{R}^n$ which are associated with the system $\dot{x} = Ax + Bu$. We use the notation $\{A - D|\mathscr{B}\}_i \equiv \mathscr{B} + (A - D)\mathscr{B} + \cdots + (A - D)^{i-1}\mathscr{B}$. For simplicity we assume that $\{A - D|\mathscr{B}\}_n = \mathscr{R}^n$ (i.e., (A, B) is a uniformly controllable pair [5]) and $\ker B = 0$.

Let $T: \mathscr{R}^n \to \mathscr{R}^n$ be a fixed automorphism and define the morphism $\dot{T}: \mathscr{R}^n \to \mathscr{R}^n$ by $e_i \mapsto DTe_i$, $e_i \in \mathbf{e}_n$. It is a simple matter to show that the following diagram

$$\begin{tikzcd}
\mathscr{R}^n \arrow[r, "A - D"] \arrow[d, "T"'] & \mathscr{R}^n \arrow[d, "T"] \\
\mathscr{R}^n \arrow[r, dashed, "M"'] & \mathscr{R}^n
\end{tikzcd}$$

commutes if and only if $M = TAT^{-1} + \dot{T}T^{-1} - D$. Associated with the function M, one can define a morphism $H: \mathscr{R}^n \to \mathscr{R}^n$ by requiring that the restriction of H to $\bar{\mathscr{R}}^n$ coincide with the restriction of M to $\bar{\mathscr{R}}^n$; clearly $H = TAT^{-1} + \dot{T}T^{-1}$. When speaking of the morphism of M we shall always mean the morphism H. Clearly the automorphism T corresponds to the matrix transformation $A \mapsto TAT^{-1} + \dot{T}T^{-1}$; a transformation of this type corresponds to the coordinate change $y = Tx$ associated with the differential equation $\dot{x} = Ax$; i.e., $\dot{y} = (TAT^{-1} + \dot{T}T^{-1})y$.

Properties of R-modules and R-morphisms. In the following sections we shall be concerned with R-matrices of a special type. Let M be an $n \times m$ matrix with elements in R. By regarding M as a matrix-valued function of t, $M(t)$, one can define $\rho(t) \equiv \operatorname{rank} M(t)$ for each fixed $t \in I$. In general, there does not seem to be any way to interpret $\rho(t)$ in ring theoretic terms unless $\rho(t)$ is constant on I. We shall say that the R-matrix M and its corresponding R-morphism have *constant rank* ρ if and only if $\rho(t) \equiv \rho$ is independent of $t \in I$. Constant rank morphisms admit the following characterization.

PROPOSITION 1.1. *Let $M: \mathscr{R}^m \to \mathscr{R}^n$ be a fixed R-morphism. Then M has constant rank if and only if $\mathscr{M} \in \sum_n$.*

This proposition states that the image of any constant rank morphism possesses a basis which may be completed to $\mathscr{R}^n$. It also states that there is a one-to-one correspondence between elements of $\sum_n$ and constant rank morphisms with codomain $\mathscr{R}^n$. A proof of this proposition, which is a ring theoretic interpretation of Doležal's theorem [4], is given in § 4.

The second property we wish to emphasize here is stated below.

PROPOSITION 1.2. *Let* $\mathscr{S}$, $\mathscr{T}$, $\mathscr{U}$ *be fixed submodules of* $\mathscr{R}^n$ *satisfying* $\mathscr{S} \in \sum_n$ *and* $\mathscr{S} = \mathscr{T} \oplus \mathscr{U}$. *Then* $\mathscr{T}$ *is free if and only if* $\mathscr{U}$ *is free.*

Note that this proposition states a general property of R-modules. Our reason for presenting the proposition at this point is to emphasize that this is the *only* special property of R-modules which we require. Thus all results which follow are true over any commutative ring K with ring derivation $\delta: K \to K$, provided Proposition 1.2 holds for K-modules and $\ker \delta$ contains a subring $\bar{K}$ which is also a field. The set of continuous, infinitely differentiable, periodic functions on I is an example of such a ring. Proposition 1.2 is proved in § 4.

2. Controllability modules. Controllability subspaces have proved useful in describing the algebraic structure of linear time-invariant systems [1]. To treat time-varying systems, we introduce a similar concept. A submodule $\mathscr{S} \subset \mathscr{R}^n$ is called a *controllability module* of the pair (A, B) if for some morphism $F: \mathscr{R}^n \longrightarrow \mathscr{R}^m$,

$$\{A + BF - D|\mathscr{B} \cap \mathscr{S}\}_i \in \sum_n, \quad i \in \mathbf{n}, \tag{2.1}$$

and

$$\mathscr{S} = \{A + BF - D|\mathscr{B} \cap \mathscr{S}\}_n. \tag{2.2}$$

At present, the rather restrictive requirement (2.1) seems necessary in order to obtain useful results. In effect, (2.1) means that each $\mathscr{S}_i \equiv \{A + BF - D|\mathscr{B} \cap \mathscr{S}\}_i$ must be the image of a constant rank morphism.

Controllability modules are important for at least two reasons. First, each controllability module corresponds to a subsystem of the system $\dot{x} = Ax + Bu$ which can be completely controlled while not influencing the remainder of the system. Second, controllability modules possess a property similar to the pole placement property of controllability subspaces [1]. We shall elaborate on these points later.

Since the Cayley–Hamilton theorem does not hold over the ring R, one cannot conclude that modules satisfying (2.2) are $(A + BF - D)$-invariant. Nevertheless, controllability modules do possess this property.

LEMMA 2.1. *Let* $\mathscr{S}$ *be a fixed controllability module and suppose that* F *is chosen so that* (2.1) *and* (2.2) *hold. Then*

$$(A + BF - D)\mathscr{S} \subset \mathscr{S}. \tag{2.3}$$

Proof. Clearly (2.3) holds if $\mathscr{S}$ is either $\mathscr{R}^n$ or 0. If $\mathscr{S}$ is a proper submodule of $\mathscr{R}^n$, $d(\mathscr{S}) < n$ and $d(\mathscr{B} \cap \mathscr{S}) > 0$. With $\mathscr{S}_i \equiv \{A + BF - D|\mathscr{B} \cap \mathscr{S}\}_i$, $i \in \mathbf{n}$, there follows $d(\mathscr{S}_{i+1}) \geqq d(\mathscr{S}_i)$, $i \in \mathbf{n - 1}$, and $d(\mathscr{S}_1) > 0$. If inequality holds for all $i \in \mathbf{n - 1}$, then $d(\mathscr{S}_{i+1}) > i$, $i \in \mathbf{n - 1}$; thus $d(\mathscr{S}) = d(\mathscr{S}_n) > n - 1$ or $\mathscr{S} = \mathscr{R}^n$, a contradiction. Therefore $d(\mathscr{S}_{i+1}) = d(\mathscr{S}_i)$ for some $i \in \mathbf{n - 1}$; clearly $\mathscr{S}_{i+1} = \mathscr{S}_i$ so $(A + BF - D)\mathscr{S}_i \subset \mathscr{S}_i$. By induction, $\mathscr{S}_j = \mathscr{S}_i$ for $j > i$ and (2.3) follows.

The following lemma characterizes a class of submodules for which (2.3) is true.

LEMMA 2.2. *Let* $\mathscr{U} \in \sum_n$, *There exists a morphism* $F : \mathscr{R}^n \to \mathscr{R}^m$ *such that*

$$(A + BF - D)\mathscr{U} \subset \mathscr{U} \tag{2.4}$$

if and only if

$$(A - D)\mathscr{U} \subset \mathscr{B} + \mathscr{U}. \tag{2.5}$$

Proof. If (2.4) holds, $u \in \mathscr{U}$ implies $(A + BF - D)u \in \mathscr{U}$ or $(A - D)u \in BF\mathscr{U} + \mathscr{U} \subset \mathscr{B} + \mathscr{U}$ and (2.5) is true. To prove the converse, let $\{u_1, \cdots, u_l\}$ be a basis for $\mathscr{U}$; from (2.5) there follows $(A - D)u_i = Bx_i + \hat{u}_i$, $i \in \mathbf{l}$, for some $x_i \in \mathscr{R}^m$, $\hat{u}_i \in \mathscr{U}$. Define F on $\mathscr{U}$ by $u_i \mapsto -x_i$, $i \in \mathbf{l}$. Since $\mathscr{U} \in \sum_n$, by Proposition 1.2 there exists a free module $\mathscr{V}$ such that $\mathscr{R}^n = \mathscr{U} \oplus \mathscr{V}$; thus the domain of F can be extended to $\mathscr{R}^n$. Clearly $(A + BF - D)u_i = \hat{u}_i \in \mathscr{U}$, $i \in \mathbf{l}$; since for arbitrary $r_i \in R$,

$$(A + BF - D)\sum_{i \in \mathbf{l}} u_i r_i = \sum_{l \in \mathbf{l}} (r_i(A + BF - D)u_i - \dot{r}_i u_i) \in \mathscr{U},$$

(2.4) is true..

It is possible to completely characterize controllability modules.

PROPOSITION 2.1. *Let* $\mathscr{S} \subset \mathscr{R}^n$ *be fixed. Then* $\mathscr{S}$ *is a controllability module of the pair* (A, B) *if and only if*

$$(A - D)\mathscr{S} \subset \mathscr{B} + \mathscr{S},$$

$\mathscr{S} = \mathscr{S}_l$, *and* $\mathscr{S}_i \in \sum_n$, $i \in \mathbf{l}$; *the* $\mathscr{S}_i$ *are defined by* $\mathscr{S}_0 = 0$,

$$\mathscr{S}_i = ((A - D)\mathscr{S}_{i-1} + \mathscr{S}_{i-1} + \mathscr{B}) \cap \mathscr{S}, \quad i \in \mathbf{l},$$

where $l = d(\mathscr{S})$.

With Lemma 2.2 at hand, the proof of this proposition parallels exactly the corresponding proof of the theorem which characterizes controllability subspaces [1]; thus the present proof is omitted.

It is useful to note that the class of $F : \mathscr{R}^n \to \mathscr{R}^m$ for which (2.1) and (2.2) hold coincides with $\mathbf{F}(\mathscr{S})$, the class of F for which (2.3) is true. To establish this observe that if (2.1) and (2.2) hold for some F, then by Lemma 2.1, $F \in \mathbf{F}(\mathscr{S})$. Conversely, if $F \in \mathbf{F}(\mathscr{S})$, then

$$\begin{aligned}\mathscr{S}_i &\equiv ((A - D)\mathscr{S}_{i-1} + \mathscr{S}_{i-1} + \mathscr{B}) \cap \mathscr{S} = ((A + BF - D)\mathscr{S}_{i-1} + \mathscr{S}_{i-1} + \mathscr{B}) \cap \mathscr{S} \\ &= ((A + BF - D)\mathscr{S}_{i-1} + \mathscr{S}_{i-1}) \cap \mathscr{S} + \mathscr{B} \cap \mathscr{S} \\ &= (A + BF - D)\mathscr{S}_{i-1} + \mathscr{S}_{i-1} + \mathscr{B} \cap \mathscr{S}, \quad i \in \mathbf{l}.\end{aligned}$$

By induction on i there follows $\mathscr{S}_i = \{A + BF - D | \mathscr{B} \cap \mathscr{S}\}_i$, $i \in \mathbf{l}$, so (2.1) and (2.2) hold.

Controllability modules have a property similar to the pole assignment property of controllability subspaces. Below we describe this property for the special case when $d(\mathscr{B} \cap \mathscr{S}) = 1$.

PROPOSITION 2.2. *Let* $\mathscr{S} \subset \mathscr{R}^n$ *be a fixed controllability module with* $d(\mathscr{B} \cap \mathscr{S}) = 1$; *write* $\mathscr{b} \equiv \mathscr{B} \cap \mathscr{S}$ *and* $l \equiv d(\mathscr{S})$. *Let*

$$\alpha(\lambda) \equiv \lambda^l + \sum_{i \in \mathbf{l}} \bar{r}_i \lambda^{i-1}$$

be a fixed monic polynomial of degree l with coefficients $\bar{r}_i \in \bar{R}$. There exists an $F \in \mathbf{F}(\mathscr{S})$ such that

$$(2.6)\qquad \alpha(A + BF - D)b \equiv ((A + BF - D)^l + \sum_{i \in \mathbf{l}} \bar{r}_i(A + BF - D)^{i-1})b = 0.$$

Proof. Since $\mathscr{S}$ is a controllability module, $\mathbf{F}(\mathscr{S})$ is nonempty. Choose any fixed $F_0 \in \mathbf{F}(\mathscr{S})$ and write $A_0 \equiv A + BF_0$. Clearly $\{b, (A_0 - D)b, \cdots, (A_0 - D)^{l-1}b\}$ is a basis for $\mathscr{S}$. Thus

$$(2.7)\qquad (A_0 - D)^l b = \sum_{i \in \mathbf{l}} \tilde{r}_i(A_0 - D)^{i-1}b$$

for some $\tilde{r}_i \in R$. Since $r(A_0 - D)x = (A_0 - D)rx + \dot{r}x$ for all $x \in \mathscr{R}^n$ and $r \in R$, with suitable $r_i \in R$, (2.7) may be rewritten as

$$(2.8)\qquad (A_0 - D)^l b = \sum_{i \in \mathbf{l}} (A_0 - D)^{i-1} r_i b.$$

The computation of the r_i from the $\tilde{r}_i$ is tedious but straightforward. To proceed, define $s_1 \equiv b$ and $s_{i+1} \equiv (A_0 - D)s_i - r_{l+1-i}b$, $i \in \mathbf{l} - \mathbf{1}$. Clearly $\{s_1, \cdots, s_l\}$ is a basis for $\mathscr{S}$; in addition, by a simple computation,

$$(2.9)\qquad (A_0 - D)s_l = r_1 b.$$

Define $f_1 : \mathscr{S} \to \mathscr{R}^1$ so that $f_1 s_i = -r_{l+1-i}$, $i \in \mathbf{l}$; as in the proof of Lemma 2.2, the domain of f_1 may be extended to $\mathscr{R}^n$. Write $b = Bg$ and $A_1 = A_0 + Bgf_1$. It follows that

$$s_{i+1} = (A_1 - D)s_i = (A_1 - D)^i b, \quad i \in \mathbf{l} - \mathbf{1},$$

and

$$(2.10)\qquad (A_1 - D)^l b = 0.$$

Write $\bar{\mathscr{S}}$ for the $\bar{\mathscr{R}}$-vector space with basis $\{s_1, \cdots, s_l\}$ and let $\bar{A}$ be the restriction of $(A_1 - D)$ to $\bar{\mathscr{S}}$. In view of (2.10), $\bar{A}$ is an $\bar{R}$-endomorphism and $\bar{\mathscr{S}}$ is $\bar{A}$-cyclic with generator b. It follows from well-known results [1] that for suitable $f_2 : \bar{\mathscr{S}} \to \bar{\mathscr{R}}$, $\bar{\mathscr{S}}$ can be made $(\bar{A} + bf_2)$-cyclic with minimal polynomial $\alpha(\lambda)$ and generator b. By first extending the domain of f_2 to $\mathscr{S}$ in the obvious way and then extending the domain further to all of $\mathscr{R}^n$, the desired result is at hand. Clearly with $F = F_0 + g(f_1 + f_2)$, (2.6) is true and $F \in \mathbf{F}(\mathscr{S})$, thus completing the proof.

Remark. Proposition 2.2 can be generalized: If $\mathscr{S}$ is a controllability module with $d(\mathscr{B} \cap \mathscr{S}) > 1$, then there exists a $b \in \mathscr{B} \cap \mathscr{S}$ and an $F \in \mathbf{F}(\mathscr{S})$ such that (2.6) is true. At present there does not seem to be any direct way to establish this fact, perhaps because not every free generator of $\mathscr{B} \cap \mathscr{S}$ is an appropriate candidate for b. Nevertheless, such an (F, b) pair does indeed exist; this will be obvious from the results of the next section.

Remark. Consider again the controllability module of Proposition 2.2 and suppose that F has been chosen so that (2.6) is true. Let $T : \mathscr{S} \to \mathscr{R}^n$ be the morphism $(A + BF - D)^{i-1}b \mapsto e_i$, $i \in \mathbf{l}$, $e_i \in \mathbf{e}_n$. Extend the domain of T to $\mathscr{R}^n$ so that T is an automorphism. It should be clear that the matrix of the function $T(A + BF)T^{-1} + \dot{T}T^{-1} - D$ has the structure

$$\begin{bmatrix} \bar{A}_1 & A_2 \\ 0 & A_3 \end{bmatrix},$$

where $\bar{A}_1$ is an $l \times l$ $\bar{R}$-matrix with characteristic polynomial $\alpha(\lambda)$. In addition, the matrix representation of the morphism $TBg = Tb$ has the structure

$$\begin{bmatrix} \bar{b}_1 \\ 0 \end{bmatrix},$$

where $\bar{b}_1$ is an $l \times l$ $\bar{R}$-matrix; the pair $(\bar{A}_1, \bar{b}_1)$ is completely controllable. Thus by introducing suitable feedback, one can transform that portion of the system $\dot{x} = Ax + Bu$ associated with $\mathscr{S}$ into a time-invariant controllable system. In the next section, it will be shown that this can be done for any controllability module.

3. Index-invariant systems. In this section we study the invariant properties of the matrix pair (A, B) under transformations of the type

$$(A, B) \mapsto (T(A + BF)T^{-1} + \dot{T}T^{-1}, TBG), \tag{3.1}$$

where T and G are invertible in R and F; $\mathscr{R}^n \to \mathscr{R}^m$ is an R-morphism. Our results are only valid for a restricted class of systems: roughly speaking, all systems whose controllability properties are invariant with respect to time. To be more specific, define

$$m_i(t) \equiv \operatorname{rank}[B(t), (A - D)B(t), \cdots, (A - D)^{i-1}B(t)]$$

for $t \in I$ and $i \in \mathbf{n} + \mathbf{1}$. The pair (A, B) is called *index-invariant* on I if for all $i \in \mathbf{n} + \mathbf{1}, m_i(t) \equiv m_i =$ const. on I and $m_n = m_{n+1}$. The m_i are called the *structure indices* of the index-invariant pair (A, B). Since by hypothesis $\mathscr{R}^n = \{A - D|\mathscr{B}\}_n$, $m_n = m_{n+1} = n$. Clearly (A, B) is an index-invariant pair if and only if $\mathscr{R}^n$ is a controllability module.

Associated with the m_i one can define a second set of integers n_j as follows. Write $k_0 = 0$ and $k_i \equiv m + m_i - m_{i+1}$ for $i \in \mathbf{n}$. Define $n_j \equiv i$ for $j = k_{i-1} + 1, \cdots, k_i$ and $i \in \mathbf{n}$. Since $k_n = m + m_n - m_{n+1} = m$, there are exactly m integers n_j. Clearly $n_1 \leqq n_2 \leqq \cdots \leqq n_m$; in addition,

$$\begin{aligned}
\sum_{j \in \mathbf{m}} n_j &= (k_1 - k_0) + 2(k_2 - k_1) + \cdots + n(k_n - k_{n-1}) \\
&= nk_n - \sum_{i \in \mathbf{n-1}} k_i \\
&= nk_n - \sum_{i \in \mathbf{n-1}} (m + m_i - m_{i+1}) \\
&= k_n + (n - 1)(k_n - m) - \sum_{i \in \mathbf{n-1}} (m_i - m_{i+1}) \\
&= k_n + (n - 1)(k_n - m) - m_1 + m_n.
\end{aligned}$$

Since $k_n = m_1 = m$,

$$\sum_{j \in \mathbf{m}} n_j = m_n = n. \tag{3.2}$$

The n_i have been discussed elsewhere [2], [3] and are called the *controllability indices* of the index-invariant pair (A, B). Controllability indices are central to the structure of the pair (A, B).

THEOREM 3.1. *Let (A, B) be an index-invariant pair with controllability indices $\{n_1, \cdots, n_m\}$. There exist controllability modules $\mathcal{S}_1, \cdots, \mathcal{S}_m$ satisfying*

$$d(\mathcal{B} \cap \mathcal{S}_i) = 1, \quad i \in \mathbf{m}, \tag{3.3}$$

$$d(\mathcal{S}_i) = n_i, \quad i \in \mathbf{m}, \tag{3.4}$$

$$\mathcal{R}^n = \bigoplus_{i \in \mathbf{m}} \mathcal{S}_i. \tag{3.5}$$

This theorem states that any index-invariant system can be decomposed (with suitable F and G) into m independent subsystems, each completely controlled by a scalar input. To prove the theorem, use will be made of the following special result.

LEMMA 3.1. *Let (A, B) be an index-invariant pair with structure indices $\{m_1, \cdots, m_{n+1}\}$. There exist modules $\mathcal{V}_i \in \sum_n$, $i \in \mathbf{n}$, satisfying*

$$\mathcal{V}_1 \subset \mathcal{V}_2 \subset \cdots \subset \mathcal{V}_n = \mathcal{B}, \tag{3.6}$$

$$(A - D)^i \mathcal{V}_i \subset \{A - D|\mathcal{B}\}_i, \quad i \in \mathbf{n}, \tag{3.7}$$

$$d(\mathcal{V}_i) = m + m_i - m_{i+1}, \quad i \in \mathbf{n}. \tag{3.8}$$

Proof. Write $\mathcal{B}_0 \equiv 0$ and $\mathcal{B}_i \equiv \{A - D|\mathcal{B}\}_i$. Define $\mathcal{V}_n \equiv \mathcal{B}$ and note that.(3.7) and (3.8) hold for $i = n$. To define the remaining $\mathcal{V}_i$, $i \in \mathbf{n} - \mathbf{1}$, proceed as follows. Since $\mathcal{B}_i \in \sum_n$, $i \in \mathbf{n}$, by Proposition 1.2 there exist free $\mathcal{U}_i$ such that

$$\mathcal{B}_{i+1} = \mathcal{B}_i \oplus \mathcal{U}_i, \quad i \in \mathbf{n} - \mathbf{1}. \tag{3.9}$$

Let $P_i : \mathcal{B}_{i+1} \to \mathcal{U}_i$ be the projection $(u + b) \mapsto u$, $u \in \mathcal{U}_i$, $b \in \mathcal{B}_i$; write $Q : \mathcal{B} \to \mathcal{R}^n$ for the insertion of $\mathcal{B}$ in $\mathcal{R}^n$. Since $\mathcal{B}_{i+1} = (A - D)^i\mathcal{B} + \mathcal{B}_i$, $i \in \mathbf{n} - \mathbf{1}$, it follows from (3.9) that

$$P_i(A - D)^i Q\mathcal{B} = \mathcal{U}_i, \quad i \in \mathbf{n} - \mathbf{1}. \tag{3.10}$$

By direct verification, $P_i(A - D)^i Q : \mathcal{B} \to \mathcal{U}_i$ is a morphism; by (3.10) even an epimorphism.

Define $\mathcal{V}_i$ by

$$\mathcal{V}_i \equiv \ker P_i(A - D)^i Q, \quad i \in \mathbf{n} - \mathbf{1}. \tag{3.11}$$

Clearly $P_i(A - D)^i Q\mathcal{V}_i = P_i(A - D)^i\mathcal{V}_i = 0$ or $(A - D)^i\mathcal{V}_i \subset \ker P_i = \mathcal{B}_i$ and (3.7) is true. To establish (3.8), note that since $\mathcal{U}_i$ is free,

$$\mathcal{B} = \mathcal{V}_i \oplus \mathcal{W}_i, \quad i \in \mathbf{n} - \mathbf{1}, \tag{3.12}$$

for some $\mathcal{W}_i \approx \mathcal{U}_i$ (cf. [6, Chap. VI, Lemma 2]). Thus

$$\begin{aligned} d(\mathcal{V}_i) &= d(\mathcal{B}) - d(\mathcal{W}_i) = m - d(\mathcal{U}_i) = m - (d(\mathcal{B}_{i+1}) - d(\mathcal{B}_i)) \\ &= m + m_i - m_{i+1}, \quad i \in \mathbf{n} - \mathbf{1}, \end{aligned}$$

so (3.8) is true. In addition, since $\mathcal{B} \in \sum_n$ and $\mathcal{W}_i \approx \mathcal{U}_i$, by Proposition 1.2 and (3.12), $\mathcal{V}_i$ is also free. Since $\mathcal{B}$ is a direct summand of $\mathcal{R}^n$, so is $\mathcal{V}_i$; i.e., $\mathcal{V}_i \in \sum_n$, $i \in \mathbf{n}$.

To complete the proof, it is enough to verify (3.6). By (3.7), $(A - D)^{i+1}\mathcal{V}_i \subset (A - D)\mathcal{B}_i \subset \mathcal{B}_{i+1}$; clearly $P_{i+1}(A - D)^{i+1}\mathcal{V}_i = 0$, so $\mathcal{V}_i \subset \ker P_{i+1}(A - D)^{i+1}Q = \mathcal{V}_{i+1}$, $i \in \mathbf{n} - \mathbf{1}$, and (3.6) is true.

Proof of Theorem 3.1. Write $k_i \equiv m + m_i - m_{i+1}$, $i \in \mathbf{n}$. By Lemma 3.1, there exists a basis $\{b_1, \cdots, b_{k_1}\}$ for $\mathscr{V}_1$; it also follows from Lemma 3.1 that this basis may be extended sequentially so that $\{b_1, \cdots, b_{k_i}\}$ is a basis for $\mathscr{V}_i$, $i \in \mathbf{n}$. In this way, there results a basis $\{b_1, \cdots, b_m\}$ for $\mathscr{B}$ with special properties. From (3.7) and the definition of the n_i there follows

$$(A - D)^{n_i} b_i \in \mathscr{B}_{n_i}, \quad i \in \mathbf{m}.$$

Thus for each $i \in \mathbf{m}$, there exist $b^i_j \in \mathscr{B}$ such that

$$(A - D)^{n_i} b_i = \sum_{j \in \mathbf{n}_i} (A - D)^{j-1} b^i_{j-1}, \quad i \in \mathbf{m}. \tag{3.13}$$

Define $s^i_1 \equiv b_i$, $i \in \mathbf{m}$, and

$$s^i_{j+1} = (A - D)s^i_j - b^i_{n_i - j}, \quad j \in \mathbf{n}_i - \mathbf{1}, \quad i \in \mathbf{m}. \tag{3.14}$$

It follows that

$$(A - D)s^i_{n_i} = b^i_0, \quad i \in \mathbf{m}. \tag{3.15}$$

Write

$$\mathscr{S}_i \equiv \operatorname{span}\{s^i_1, \cdots, s^i_{n_i}\}, \quad i \in \mathbf{m}. \tag{3.16}$$

It follows from (3.14) and (3.15) that

$$(A - D)\mathscr{S}_i \subset \mathscr{B} + \mathscr{S}_i, \quad i \in \mathbf{m}. \tag{3.17}$$

Thus $\mathscr{S} \equiv \sum_{i \in \mathbf{m}} \mathscr{S}_i$ satisfies $(A - D)\mathscr{S} \subset \mathscr{S} + \mathscr{B}$; in addition, since $b_i \in \mathscr{S}_i$, $i \in \mathbf{m}$, $\mathscr{B} \subset \mathscr{S}$; thus $(A - D)\mathscr{S} \subset \mathscr{S}$. It follows that $(A - D)^i \mathscr{B} \subset \mathscr{S}$ for all $i \geqq 0$; thus $\mathscr{R}^n \subset \mathscr{S}$. But $\mathscr{S} \subset \mathscr{R}^n$, so $\mathscr{S} = \mathscr{R}^n$. From (3.16), $\mathscr{R}^n$ is spanned by at most $\sum_{i \in \mathbf{m}} n_i$ generators; but from (3.2), $\sum_{i \in \mathbf{m}} n_i = n$ which means that each $\mathscr{S}_i$ is free on the elements $\{s^i_1, \cdots, s^i_{n_i}\}$, $i \in \mathbf{m}$, and that the $\mathscr{S}_i$ are mutually independent. Thus (3.4) and (3.5) are both true.

To prove (3.3), fix i and suppose that $b \in \mathscr{B} \cap \mathscr{S}_i$; clearly $b = \sum_{j \in \mathbf{m}} r_j b_j$, $r_j \in R$; thus

$$b - r_i b_i \in \mathscr{S}_i \cap \sum_{\substack{j \in \mathbf{m} \\ j \neq i}} \mathscr{b}_j \subset \mathscr{S}_i \cap \sum_{\substack{j \in \mathbf{m} \\ j = i}} \mathscr{S}_j = 0;$$

thus $b = r_i b_i$ and b_i is a basis for $\mathscr{B} \cap \mathscr{S}_i$; therefore (3.3) is true.

To prove that $\mathscr{S}_i$ is a controllability module, write $\mathscr{T}^i_0 = 0$, $\mathscr{T}^i_j \equiv \operatorname{span}\{s^i_1, \cdots, s^i_j\}$ and note by (3.14) that

$$\mathscr{T}^i_j = ((A - D)\mathscr{T}^i_{j-1} - \mathscr{T}^i_{j-1} + \mathscr{B}) \cap \mathscr{S}, \quad j \in \mathbf{n}_i.$$

Since each $\mathscr{T}^i_j$ is a free direct summand of $\mathscr{S}_i \in \sum_n$, clearly $\mathscr{T}^i_j \in \sum_n$. Finally, since $\mathscr{S}_i = \mathscr{T}^i_{n_i}$ and (3.17) is true, application of Proposition 2.1 establishes the desired result.

Remark. Theorem 3.1 suggests a procedure for transforming any index-invariant pair (A, B) into a pair of (constant) $\bar{R}$-matrices $(\bar{A}, \bar{B})$. In particular, having determined a set of $\mathscr{S}_i$ for which (3.3)–(3.5) hold, define F on each $\mathscr{S}_i$ (using the construction of Proposition 2.2) so that $(A + BF - D)^{n_i} b_i = 0$. By the remark following Proposition 2.2, it is clear that for appropriate T and G, the pair $(\bar{A}, \bar{B})$ defined by

$$\begin{aligned} \bar{A} &\equiv T(A + BF)T^{-1} + \dot{T}T^{-1}, \\ \bar{B} &\equiv TBG \end{aligned} \tag{3.18}$$

can be made (constant) $\bar{R}$-matrices.

Clearly the structure indices of (A, B) are invariant under transformations of the type (3.1). In addition, every pair of completely controllable $\bar{R}$-matrices $(\bar{A}, \bar{B})$ is index-invariant. Thus we have established the following result.

COROLLARY 3.1. *Let (A, B) be a fixed pair of R-matrices. There exist transformations (T, F, G) such that the matrices $(\bar{A}, \bar{B})$ in* (3.18) *are constant $\bar{R}$-matrices if and only if (A, B) is an index-invariant pair.*

Remark. The original statement and proof of this corollary are due to Brunovský [3] who developed the result using matrix methods. Brunovský also showed that for fixed n and m there is a bijection between the class of ordered sets of positive integers $\{n_1, n_2, \cdots, n_m\}$ satisfying $n_1 \leqq n_2 \cdots \leqq n_m, \sum_{i \in \mathbf{m}} n_i = n$ and the set of equivalence classes $\{(A, B)\}$, where $(\bar{A}, \bar{B}) \in \{(A, B)\}$ if and only if $(\bar{A}, \bar{B})$ and (A, B) are related by (3.18) for some F and invertible T and G. Thus controllability indices uniquely specify index-invariant pairs (A, B) up to transformations (T, F, G).

Remark. Suppose (A, B) is index-invariant and T, F, G are chosen so that $(\bar{A}, \bar{B})$ in (3.18) are $\bar{R}$-matrices. It is easy to show that $(\bar{A}, \bar{B})$ is a controllable pair. Thus, using known methods [1], one can determine $\bar{R}$-matrices $\bar{F}$ and $\bar{x}$ such that $(\bar{A} + \bar{B}\bar{F}, \bar{B}\bar{x})$ is a controllable pair. It follows from (3.18) that the pair $(A + B(F + G\bar{F}T), BG\bar{x})$ is index-invariant and that $\mathscr{R}^n = \{A + B(F + G\bar{F}T) - D|\mathscr{b}\}_n$, where $b = BG\bar{x}$. Extending this argument, it is straightforward to show that for any fixed controllability module $\mathscr{S}$, there exists a $b \in \mathscr{B} \cap \mathscr{S}$ and an $F \in \mathbf{F}(\mathscr{S})$ such that $\mathscr{S} = \{A + BF - D|\mathscr{b}\}_{\tilde{n}}, \tilde{n} = d(\mathscr{S})$. Since $\mathscr{S}$ can be regarded as a controllability module of the pair $(A + BF, b)$, Proposition 2.2 applies thus showing that a pole-assignment-like construction is possible for any controllability module.

If $I = [0, \infty]$, invertible R-matrices correspond to Lyapunov transformations [7, pp. 116–118]; in this case, as noted in [8],[1] index-invariant systems may be stabilized with state feedback control.

4. Constant rank matrices.

In this section we discuss several properties of constant rank R-matrices and then conclude with proofs of Propositions 1.1 and 1.2. Consider first the following fundamental result.

PROPOSITION 4.1 (Doležal [4]). *Let M be an $n \times m$ R-matrix with constant rank ρ. There exists an $m \times m$ matrix S, invertible in R, and an $n \times \rho$ matrix W with constant rank ρ such that*

$$MS = [W, 0]. \tag{4.1}$$

This proposition may be restated in terms more suitable to our application.

[1] Actually the results of [8] are slightly more restrictive than this, since it is assumed there (in effect) that bases for the S_i can be constructed from the *columns* of $[B, (A - D)B, \cdots, (A - D)^{n-1}B]$ rather than from linear combinations of the columns.

PROPOSITION 4.1′. *Let $M:\mathscr{R}^m \to \mathscr{R}^n$ be an R-morphism with constant rank. Then both $\mathscr{M}$ and* ker *M are free R-modules.*

Proof. Clearly $\mathscr{M} = \mathscr{W}$; but $\mathscr{W}$ is free on the linearly independent columns of W; thus $\mathscr{M}$ is free. From (4.1) there follows $\ker M = \ker [W, 0]S^{-1} \approx \ker [W, 0]$ which is clearly free; thus $\ker M$ is free.

Remark 4.1. Let P be the canonical projection $P:\mathscr{R}^m \to \mathscr{R}^m/\ker M$. The matrix W in (4.1) may be regarded as a representation of the unique morphism $W: \mathscr{R}^m/\ker M \to \mathscr{R}^n$ for which $M = WP$. The existence of matrix representations for both W and P is assured by the fact that $\mathscr{R}^m$, $\mathscr{R}^n$ and $\mathscr{R}^m/\ker M$ are all free. In addition, P is epic and its right inverse is a morphism; the existence of a left inverse morphism for W (which is monic) may be established by dualizing Proposition 4.1.

To prove Proposition 1.1, use will be made of the following lemma.

LEMMA 4.1. *Let $H:\mathscr{U} \to \mathscr{V}$ be a fixed morphism of the R-modules $\mathscr{U}$ and $\mathscr{V}$ and suppose $\mathscr{V}$ is free. Then H possesses a left inverse morphism if and only if H is monic and $\mathscr{H} \equiv H\mathscr{U}$ is a direct summand of $\mathscr{V}$.*

This easily proved lemma is a slight variation of [6, Prop. 20, Chap. X], and thus a proof will not be given.

Proof of Proposition 1.1. If M has constant rank, by Proposition 4.1, $\mathscr{M}$ is free. By Remark 4.1, W is left-invertible and has codomain $\mathscr{R}^n$ which is free; thus by Lemma 4.1, $\mathscr{W}$ is a direct summand of $\mathscr{R}^n$. But $\mathscr{W} = \mathscr{M}$; therefore $\mathscr{M} \in \sum_n$.

To prove the converse, write (as in Remark 4.1) $M = WP$. Since $\mathscr{W} = \mathscr{M} \in \sum_n$, $\mathscr{W}$ is a free direct summand of $\mathscr{R}^n$. Since W is necessarily monic, Lemma 4.1 applies and W possesses a left inverse. Since $M = WP$, rank $M(t) \leqq$ rank $W(t) \leqq d(\mathscr{W}) \equiv \rho$. Since W is left-invertible and P is right-invertible, $W^{-1}MP^{-1} = E$, the $\rho \times \rho$ identity. Thus rank $M(t) \geqq \rho$, which with the previous inequality implies rank $M(t) = \rho$.

Proof of Proposition 1.2. Consider first the case when $\mathscr{S} = \mathscr{R}^n$. It is enough to show that if $\mathscr{T}$ is free then $\mathscr{U}$ is free. Suppose $\mathscr{T}$ is free and let $P: \mathscr{R}^n \to \mathscr{R}^n$ be the projection on $\mathscr{T}$ along $\mathscr{U}$. Clearly $\mathscr{P} = \mathscr{T} \in \sum_n$; by Proposition 1.1, P has constant rank. Thus by Proposition 4.1, $\ker P = \mathscr{U}$ is free.

For the case when $\mathscr{S}$ is a proper submodule of $\mathscr{R}^n$, write $\mathscr{R}^n = \mathscr{S} \oplus \mathscr{V}$; by the above argument, $\mathscr{V} \in \sum_n$. But $\mathscr{R}^n = (\mathscr{T} \oplus \mathscr{U}) \oplus \mathscr{V} = \mathscr{U} \oplus (\mathscr{T} \oplus \mathscr{V})$; since $\mathscr{T} \oplus \mathscr{V}$ is free, by the above argument $\mathscr{U}$ is free.

5. Concluding remarks. Application of the algebraic concepts discussed in this article to observer synthesis along lines paralleling [2], [9] should present no difficulties; the results of [10] can easily be obtained. Application of controllability modules to decoupling synthesis, however, is not so straightforward. For such an application one will probably need a module theoretic result similar to the vector space theorem which states that there is a unique maximal controllability subspace contained in a fixed subspace of state space [1]; an analogous result for controllability modules has not yet been established.

An interesting problem suggested by this study is to find conditions on the matrix pair (A, B) which insure the existence of transformations T and F such that $T(A + BF)T^{-1} + \dot{T}T^{-1}$ is an $\bar{R}$-matrix with preassigned characteristic polynomial. Although index-invariant systems possess this property, a complete solution to the problem is still outstanding.

The translation of the preceding algebraic results into matrix terms is not difficult. Although one can use the algorithm of [3] as a starting point for future research, the present coordinate-free approach appears more transparent and its application to other problems of system structure should prove rewarding.

REFERENCES

[1] 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.

[2]———, *Feedback invariants of linear multivariable systems*, Automatica, 8 (1972), pp. 93–100.

[3] P. Brunovský, *A classification of linear controllable systems*, Kybernetika, 3 (1970), pp. 173–188.

[4] V. Doležal, *The existence of a continuous basis of a certain linear subspace of E_r which depends on a parameter*, Casopis. Pěst. Mat., 89 (1964), pp. 466–469.

[5] L. M. Silverman and H. E. Meadows, *Controllability and observability in time-variable linear systems*, this Journal, 5 (1967), pp. 64–73.

[6] S. MacLane and G. Birkhoff, *Algebra*, Macmillan, New York, 1967.

[7] F. R. Gantmacher, *The Theory of Matrices*, vol. 2, Chelsea, New York, 1959.

[8] W. A. Wolovich, *On the stabilization of controllable systems*, IEEE Trans. Automatic Control, AC-13 (1968), pp. 569–572.

[9] W. M. Wonham, *Dynamic observers—geometric theory*, Ibid., AC-15 (1970), pp. 258–259.

[10] W. A. Wolovich, *On state estimation of observable systems*, Proc. 1968 JACC, Ann Arbor, Mich., 1968, pp. 210–220.

SIAM J. CONTROL
Vol. 11, No. 2, May 1973

DYNAMIC PROGRAMMING CONDITIONS FOR PARTIALLY OBSERVABLE STOCHASTIC SYSTEMS*

M. H. A. DAVIS† AND P. VARAIYA‡

Abstract. In this paper necessary and sufficient conditions for optimality are derived for systems described by stochastic differential equations with control based on partial observations. The solution of the system is defined in a way which permits a very wide class of admissible controls, and then Hamilton–Jacobi criteria for optimality are derived from a version of Bellman's "principle of optimality."

The method of solution is based on a result of Girsanov: Wiener measure is transformed for each admissible control to the measure appropriate to a solution of the system equation. The optimality criteria are derived for three kinds of information pattern: partial observations (control based on the past of only certain components of the state), complete observations, and "Markov" observations (observation of the current state). Markov controls are shown to be minimizing in the class of those based on complete observations for system models of a suitable type.

Finally, similar methods are applied to two-person zero-sum stochastic differential games and a version of Isaacs' equation is derived.

1. Introduction.

1.1. This paper concerns the control of a system represented by a stochastic differential equation of the form

$$dz_t = g(t, z, u)\,dt + \sigma(t, z)\,dB_t, \tag{1.1}$$

where z_t is the state at time t and the increments $\{dB_t\}$ are "Gaussian white noise." The control u is to be chosen so as to minimize the average cost

$$J(u) = E\int_0^T c(t, z, u)\,dt. \tag{1.2}$$

Here T is either a fixed time or a bounded random time. The solution of (1.1) is defined by the "Girsanov measure transformation" method (see § 1.3, § 2 below) which permits a wide class of admissible controls. Controls based on three types of information pattern (partial and complete observation of the past, observation of the current state) are considered. In each case a principle of optimality similar to that of Rishel [13] is proved, and criteria for optimality analogous to the Hamilton–Jacobi equation of dynamic programming established by using an Ito process representation of the value function. Controls based on observation of the current state are shown to be minimizing in the class of those based on complete observation for system models of a suitable type. Finally similar methods

* Received by the editors October 27, 1971, and in revised form May 1, 1972. This work was supported by the National Science Foundation under Grant GK-10656X and by the U.S. Army Research Office–Durham, under Contract DAGCO-67-C-0046.

† Department of Electrical Engineering and Computer Sciences and the Electronics Research Laboratory, University of California, Berkeley, California. Now at Imperial College of Science and Technology, London SW7 2BT, England.

‡ Department of Electrical Engineering and Computer Sciences and the Electronics Research Laboratory, University of California, Berkeley, California 94720.

are applied to two-person zero-sum differential games, and a version of Isaacs' equation derived.

The results presented here are closely related to those of Fleming on optimal control of diffusion processes. A brief outline of the latter is given in § 1.2 below in order to give the flavor of the former and for purposes of comparison. Some other possible approaches to stochastic control are mentioned; then, in the light of these, a more detailed statement of the contents of this paper will be found in § 1.3.

1.2. Control of diffusion processes. The results outlined here will be found in Fleming [9] and in the references there. Let $F \subset R^n$ be open and define the cylinder $Q \subset R^{n+1}$ by

$$Q = [0, 1] \times F.$$

The system equation, to be solved in Q, is

$$\begin{aligned} d\xi_t &= g(t, \xi_t, u_t)\,dt + \sigma(t, \xi_t)\,dB_t, \\ \xi_0 &= x \in F. \end{aligned} \tag{1.3}$$

$\{B_t\}$ is a separable n-vector Brownian motion process defined on some probability space $(\Omega, \mathscr{A}, P)$. σ is an $n \times n$-matrix-valued function on $[0, 1] \times R^n$, and $g:[0, 1] \times R^n \times R^l \rightarrow R^n$; both are of class C^2, with g, σ, g_x, σ_x bounded on $[0, 1] \times R^n \times K$, for K compact in R^l. Also there exists $c > 0$ such that

$$\sum_{i,j} a_{ij}(t, x)\mu_i\mu_j \geqq c|\mu|^2 \tag{1.4}$$

for each $\mu \in R^n$, where $a = \sigma\sigma'$ ($' =$ transpose).

The control u_t is given by $u_t = Y(t, \xi_t)$, where Y is Lipschitz and takes values in a compact set $K \subset R^l$. Thus the information pattern consists of complete observations of the current state. Under the above conditions (1.3) determines uniquely a diffusion process ξ on $[0, 1]$ with $E|\xi_t|^2 < \infty$ for each t. The objective is to choose Y so as to minimize

$$J(Y) = E\int_0^\tau c(t, \xi_t, Y[t, \xi_t])\,dt,$$

where τ is the first exit time from Q. Let $g^Y(t, x) = g(t, x, Y[t, x])$ and similarly for c^Y. Define the differential operator

$$\Lambda\phi = \phi_t + \tfrac{1}{2}\sum_{i,j} a_{ij}\phi_{x_ix_j} \tag{1.5}$$

and consider the boundary problem

$$\begin{aligned} &\Lambda\psi^Y + \psi^Y_x g^Y + c^Y = 0, \qquad (t, x) \in Q, \\ &\psi^Y = 0, \qquad (t, x) \in \partial' Q = \partial Q - \{0\} \times F. \end{aligned} \tag{1.6}$$

Under the stated conditions this has a unique solution with the required differentiability properties. Applying Ito's differential formula to the function $\psi^Y(t, \xi_t)$, where ξ_t is the solution of (1.3) with $u_t = Y(t, \xi_t)$, gives

$$\psi^Y(0, x) = E\int_0^\tau c^Y\,dt = J(Y).$$

One can now drop the probabilistic interpretation and regard the problem as that of choosing the coefficients of the partial differential equation (1.6) so as to minimize the initial value $\psi^Y(0, x)$.

Let $U(s, x)$ be the minimum cost, over the class of admissible controls, starting at $(s, x) \in Q$. Formal application of Bellman's principle of optimality leads to the Hamilton–Jacobi equation

$$\begin{aligned} &\Lambda U(t,x) + \min_Y \{U_x(t,x)g^Y(t,x) + c^Y(t,x)\} = 0 \quad \text{in } Q, \\ &\qquad\qquad\qquad\qquad\qquad\qquad U(t,x) = 0 \quad \text{on } \partial' Q. \end{aligned} \tag{1.7}$$

Fleming's "verification theorem" [9, Thm. 6.1] says that if:

(i) $\phi(t, x)$ is a suitably smooth solution of (1.7), and

(ii) $u^0 = Y^0(t, x)$ is characterized by the property that $[\phi_x(t, x)g(t, x, v) + c(t, x, v)]$ is minimized in Q by $v = Y^0(t, x)$, then

$$\phi(t, x) = U(t, x) = \psi^{Y^0}(t, x).$$

A unique solution of (1.7) satisfying the conditions of the verification theorem exists if a is bounded and Lipschitz and satisfies the uniform ellipticity condition (1.4), K is compact and convex, and the boundary ∂F has certain smoothness properties. (It is possible to relax these conditions somewhat.)

The above theory can be generalized in various ways. Let $C = C_n[0, 1]$ be the space of continuous functions on $[0, 1]$ with values in R^n. Let $\mathscr{F}_t$ be the σ-field in C generated by the cylinder sets $\{z \in C : z_s \in \Gamma\}$ where Γ is a Borel set in R^n and $s \leqq t$. For $z \in C$, $t \in [0, 1]$, define $\pi_t z \in C$ by

$$\pi_t z(s) = \begin{cases} z_s, & s \leqq t, \\ z_t, & s > t. \end{cases}$$

A function of the form $u_t = u(t, \pi_t z)$ is "nonanticipative" in that it is adapted to$\mathscr{F}_t$. A unique solution to (1.3) using a nonanticipative control u is obtained if u is Lipschitz, i.e.,

$$|u(t, \pi_t \eta) - u(t, \pi_t \xi)| \leqq \kappa \|\eta - \xi\|,$$

where $\|\cdot\|$ is the uniform norm in C. See [16]. Here the information pattern is complete observation of the past. It turns out that Markov controls are minimizing in the class of nonanticipative controls, so the Markov theory is the natural one in the case of complete observations. The partially observable Markov case where $u_t = Y(t, \xi_t)$ only depends on certain components of ξ_t can be considered though this is a somewhat artificial problem since the controller generally does better by using all the past observations, not just the current value.

1.3. Methods of the type outlined above suffer from two main drawbacks:

(i) The dependence of the admissible controls on the observations has to be "smooth" (e.g. Lipschitz) to insure the existence of a solution; for optimal control it is undesirable to be limited in this way.

(ii) The observation σ-fields $\{\mathscr{Y}_t^{(u)}\}$ depend on what control is being used. This tends to vitiate variational methods since varying the control at a certain time affects the admissibility of controls applied at subsequent times. There are two cases where this does not apply: (a) complete observations, as above, since then the observation σ-fields are those generated by the Brownian process; (b) linear systems of the form

$$\begin{aligned} dx_t &= A_t x_t\,dt + u_t\,dt + dB_1(t), \\ dy_t &= F_t x_t\,dt + dB_2(t). \end{aligned} \tag{1.8}$$

In this case $\mathscr{Y}_t^{(u)} = \mathscr{Y}_t^0$, where $\mathscr{Y}_t^0$ are the σ-fields generated by $\{y_t^0\}$, $\{x_t^0, y_t^0\}$ being the solution of (1.8) with $u = 0$. This is the basic fact behind the separation theorem [18], [19] which says that there is one optimal control of the form $u_t = u(t, \hat{x}_t)$, where $\hat{x}_t = E[x_t|\mathscr{Y}_t^0]$. Of course, one can define other problems where the observation σ-fields do not depend on the control (this amounts to observing some function of the noise). Then variational methods can be used; see for example [20]. But then the problem loses its feedback aspects.

The system equation (1.1) treated in this paper is more general than (1.3) in that the dependence of the matrices g and σ on the state is nonanticipative rather than "Markov." The method of solution—given in § 2—is designed to avoid (i) and (ii) above. In § 1.2 above one took a measurable space $(\Omega, \mathscr{A})$ and random variables $\{B_t\}$ constituting a Brownian motion under a measure P, and defined a transformation (1.3), $B \to \xi$, of the random variables. Here a transformation $P \to P_u$ of the *measure* is defined such that the *original* random variables generate under P_u the measure in their sample space which is appropriate for the solution of (1.1). This transformation is well-defined with only minimal restrictions on the class of admissible controls, and the observation σ-fields do not depend on the control since they are always generated by the same random variables. On the other hand, the most that is claimed for the "solution" is that it has the right distributions.

Skorokhod remarks in the introduction to his book [5] that the methods of probability theory fall into two distinct groups, analytic and probabilistic, the former having to do only with the distributions of random variables, the latter based on operations with the random variables themselves. The method used here is something of a half-way house in that, while it is the distributions one is concerned with, the techniques used to derive them are definitely probabilistic.

This method has previously been used in [7], [8] (on the existence of optimal controls in the case of complete observations) and [19] (on the "separation theorem" of (b) above).

With a solution defined for each admissible control, the objective is to derive Hamilton–Jacobi-type conditions for optimality for the system (1.1) analogous to (1.7).

In [13] Rishel developed dynamic programming conditions for a very general class of stochastic systems. In § 3 below, Rishel's "principle of optimality" is proved in the present context and in § 4 conditions for optimality close to Rishel's are established. Using the special structure available here these can be recast (Theorem 4.2) in a form close to that of (1.7) above.

In § 5 the same methods are applied to the (simpler) completely observable case.

Section 6 deals with Markov control of systems similar to (1.3) (but without the technical conditions). The results here are direct extensions of those of § 1.2 above, coinciding with the latter when the relevant conditions are satisfied.

Differential games are susceptible to attack by the same methods. The paper concludes with a brief section (§ 7) outlining some of the possibilities in this direction.

2. Preliminaries. In differential form, the system equation (1.1) is

$$dz_t = g(t, z, u_t)\,dt + \sigma(t, z)\,dw_t \tag{2.1}$$

with initial condition $z(0) = z_0 \in R^n$, a fixed value. Here $t \in [0, 1]$ and for each t, $z_t \in R^n$, $w_t \in R^n$, $u_t \in R^l$. When necessary g and σ will be written $g' = (g'_1, g'_2)$ and $\sigma' = (\sigma'_1, \sigma'_2)$ ($'$ = transpose) corresponding to $z'_t = (x'_t, y'_t)$, the "state" and "observation" processes with dimensions $n - m$, m, respectively.

Let $\{B_t, t \in [0, 1]\}$ be an n-dimensional separable Brownian motion process on some probability space $(\Omega, \mathscr{A}, \mu)$. For each t let $\mathscr{A}_t \subset \mathscr{A}$ be the σ-field generated by the random variables $\{B_s, 0 \leqq s \leqq t\}$. Consider the stochastic differential equation

$$\begin{aligned} dz_t &= \sigma(t, z)\,dB_t, \\ z(0) &= z_0. \end{aligned} \tag{2.2}$$

The following properties are assumed for the $n \times n$-matrix-valued function $\sigma = [\sigma_{ij}]$:

(2.3)

(i) The elements $\sigma_{ij}: [0, 1] \times C \to R$ are jointly measurable functions; $\sigma_{ij}(t, \cdot)$ is $\mathscr{F}_t$-measurable for each t.[1]

(ii) There exists a process z_t, $t \in [0, 1]$, adapted to $\mathscr{A}_t$, satisfying (2.2) and

$$\sum_{i,j} \int_0^1 \sigma_{ij}^2(t, z)\,dt < \infty.$$

(iii) $\sigma(t, z)$ is nonsingular; in particular, $\sigma_2(t, z)$ has rank m, for almost all (t, z).

In (ii), the process $\{z_t\}$ is assumed to be unique in the following sense: All solutions of (2.2), which must necessarily have continuous sample paths, generate the same measure in the sample space $(C, \mathscr{F})$. This is the definition used by Girsanov in [11]. Conditions (ii) and (iii) are given in the form in which they are required rather than in such a form as to be easily verified; any sufficient conditions insuring their satisfaction could be imposed; see, e.g., [16], [23].

Under the conditions (2.3), (2.2) defines a measure P on $(C, \mathscr{F})$ by

$$PF = \mu[z^{-1}(F)] \quad \text{for } F \in \mathscr{F}.$$

Observe that

$$\mathscr{A}_t = z^{-1}(\mathscr{F}_t) \tag{2.4}$$

[1] C and $\{\mathscr{F}_t\}$ were defined in § 1.2. We are assuming that σ is defined as a function over $[0, 1] \times C$ rather than over $[0, 1] \times \Omega$. Throughout this paper the elements of C are denoted by z.

for each t, since

$$w_t = \int_0^t \sigma^{-1}\, dz.$$

The following properties are assumed for the function g:

(2.5)

(i) $g:[0,1] \times C \times \Xi \to R^n$ is jointly measurable. (Here Ξ, the control set, is a fixed Borel set of R^l.)

(ii) For fixed (t, u), $g(t, \cdot, u)$ is adapted to $\mathscr{F}_t$.

(iii) For all (t, z, u),

$$|\sigma^{-1}(t, z)g(t, z, u)| \leqq g^0(\|z\|),$$

where $\|\cdot\|$ is the uniform norm in C and g^0 is an increasing real-valued function. Thus

$$\int_0^1 |\sigma^{-1}g|^2\, dt \leqq [g^0(\|z\|)]^2 < \infty \quad \text{a.s.}(P).$$

Admissible controls. The class of admissible controls is denoted by $\mathscr{U}$ and defined as follows. For $s, t \in [0, 1]$, $s < t$, let $\mathscr{U}_s^t$ be the class of functions satisfying (2.6) below. Let $\mathscr{Y}_t$ be the sub-σ-field of $\mathscr{F}_t$ generated by the sets $\{z = (x, y) \in C : y_s \in \Gamma\}$, where Γ is a Borel set in R^m and $s \leqq t$. Note that $\mathscr{Y}_t$ does not depend on u. Thus $\mathscr{Y}_t$ is the σ-field generated by the past of the observation process.

(2.6)

(i) $u:[s, t] \times C \to \Xi \subset R^l$ is jointly measurable in (t, z).

(ii) For each t, $u(t, \cdot)$ is adapted to $\mathscr{Y}_t$.

(iii) $E[\rho_s^t(u)|\mathscr{F}_s] = 1$ a.s.(P).

Here $\rho_s^t(u) = \exp[\zeta_s^t(g^{(u)})]$, where $g^{(u)}(t, z) = g(t, z, u[t, z])$ and $\zeta_s^t(g^{(u)})$ is defined by

$$\zeta_s^t(g^{(u)}) = \int_s^t \{\sigma^{-1}(\tau, z)g(\tau, z, u[\tau, z])\}'\, dB\tau$$

$$(2.7) \qquad -\frac{1}{2}\int_s^t |\sigma^{-1}(\tau, z)g(\tau, z, u[\tau, z])|^2\, d\tau$$

$$= \int_s^t (\sigma^{-1}g)'\sigma^{-1}\, dz - \frac{1}{2}\int_s^t |\sigma^{-1}g|^2\, d\tau.$$

Now define $\mathscr{U} = \mathscr{U}_0^1$.

From (2.7), $\zeta_s^t(u)$ can be computed directly from $\{z_\tau, 0 \leqq \tau \leqq t\}$. Thus $\rho_s^t(u)$ can be regarded as a random variable on the probability space $(C, \mathscr{F}, P)$; in fact *this is taken as the basic space from now on*, the symbol E referring, as in (2.6)(iii), to integration with respect to the measure P. It is shown in [11] that (2.5) and (2.6(i)), (2.6(ii)) imply

$$E[\rho_s^t(u)|\mathscr{F}_s] \leqq 1 \quad \text{a.s.}$$

There is no known criterion for equality, though various sufficient conditions have been derived; see [7], [8].

Remark 1. If $u' \in \mathscr{U}_r^s$ and $u'' \in \mathscr{U}_s^t$, where $r \leqq s \leqq t$, then $u \in \mathscr{U}_r^t$, where

$$u(\tau, z) = \begin{cases} u'(\tau, z), & \tau \in [r, s), \\ u''(\tau, z), & \tau \in [s, t]. \end{cases}$$

Indeed, u clearly satisfies (2.6(i)), (2.6(ii)), and

$$\begin{aligned} E[\rho_{t'}^{t''}(u)|\mathscr{F}_{t'}] &= E[\rho_{t'}^{s}(u')E\{\rho_s^{t''}(u'')|\mathscr{F}_s\}|\mathscr{F}_{t'}] \\ &= E[\rho_{t'}^{s}(u')|\mathscr{F}_{t'}] = 1 \quad \text{a.s.} \end{aligned}$$

for $r \leqq t' \leqq s \leqq t'' \leqq t$. The other cases work similarly.

Remark 2. If $u \in \mathscr{U}$, then u restricted to $[s, t]$ belongs to $\mathscr{U}_s^t$. This follows from Lemma 2 of [11].

THEOREM 2.1 (Girsanov). *For $u \in \mathscr{U}$ let the measure P_u on $(C, \mathscr{F})$ be defined by*

$$P_u F = \int_F \rho_0^1(u)\, dP, \qquad F \in \mathscr{F}.$$

Then: (a) *$dw = dB - \sigma^{-1} g\, dt$ is a Brownian motion process under the measure μ_u defined by*

$$\mu_u[z^{-1}(F)] = P_u F.$$

(This defines μ_u for each $A \in \mathscr{A}$ in view of (2.4).)

(b) *The process $\{z_t\}$ considered on the space $(C, \mathscr{F}, P_u)$ satisfies*

$$\begin{aligned} dz_t &= g(t, z, u[t, z])\, dt + \sigma(t, z)\, dw_t, \\ z(0) &= z_0. \end{aligned} \tag{2.8}$$

This result is immediate from Girsanov's Theorem 1 [11]. Lemma 6 of [11] states that if $\{\theta_t\}$ is adapted to $\mathscr{A}_t$ and $\int |\theta_t|^2\, dt < \infty$ a.s., then

$$\int_0^t \theta_s\, dB_s = \int_0^t \theta_s\, dw_s + \int_0^t \theta_s \sigma^{-1} g\, ds.$$

Putting $\theta_t = \sigma(t, z)$, we see that (2.8) follows from (a) above and (2.2).

Theorem 2.1 shows that the process $\{z_t\}$ is, with measure P_u, a solution of (2.1) in the sense that

$$dz = g\, dt + \sigma\, d(\text{Brownian motion}).$$

Remark. All measures arising in this paper are, by definition, mutually absolutely continuous with respect to the measure P; so when some property is stated to hold "almost surely" (a.s.), it is irrelevant which measure is referred to.

Let $c : [0, 1] \times C \times \Xi \to R$ be a nonnegative real-valued function satisfying (2.5(i)), (2.5(ii)) and

$$c(t, z, u) \leqq k \quad \text{for all } (t, z, u) \in [0, 1] \times C \times \Xi, \tag{2.9}$$

where k is a real constant. The cost ascribed to an admissible control u is

$$J(u) = E_u\left[\int_0^1 c(s, z, u[s, z])\, ds\right] = E\left[\rho_0^1(u)\int_0^1 c_s^{(u)}\, ds\right]. \tag{2.10}$$

Note that this allows for a random stopping time τ as long as $\tau \leqq 1$ a.s., for $c_t' = c_t I_{[\tau \geqq t]}$ is an admissible cost rate function and

$$E_u \int_0^\tau c\, ds = E_u \int_0^1 c'\, ds.$$

The following results will be required in subsequent sections.

2.1. Compactness of the set of densities. Let $\mathscr{G}$ be the set of measurable functions $\gamma : [0, 1] \times C \to R^n$ adapted to $\mathscr{F}_t$ and satisfying:

$$\begin{aligned} &\text{(i)} \quad |\sigma^{-1}(t, z)\gamma(t, z)| \leqq g^0(\|z\|), \\ &\text{(ii)} \quad E[\exp\{\zeta_0^1(\gamma)\}] = 1. \end{aligned} \tag{2.11}$$

Let $\mathscr{D} = \{\exp[\zeta_0^1(\gamma)] : \gamma \in \mathscr{G}\}$.

THEOREM 2.2. *$\mathscr{D}$ is a weakly compact subset of $L_1(C, \mathscr{F}, P)$.*

This result is contained in Theorem 2 of [8] for the case $\sigma = I$ (the identity matrix). Only minor modifications are required to establish the result as stated.

Note that $\rho_0^1(u) \in \mathscr{D}$ for each $u \in \mathscr{U}$. Thus for any sequence $u_n \in \mathscr{U}$ there is a subsequence $\{u_{n_k}\}$ and an element $h \in \mathscr{G}$ such that

$$\rho_0^1(u_{n_k}) \to \exp[\zeta_0^1(h)]$$

weakly in L_1 as $k \to \infty$.

2.2. Innovations process and representation of Martingales. The main result here is Theorem 2.3, which says that any Martingale adapted to $\mathscr{Y}_t$ has a representation as a stochastic integral with respect to the "innovations process" of $\{y_t\}$, defined below. This definition was given in [15]. The result is proved in [10] for the case $\sigma = I$; the following is a similar method of proof using also ideas from [15].

Let $\gamma \in \mathscr{G}$, $\gamma = (\bar{h}, h)$ with dimensions $m - n$, n, and define the measure P^* on $(C, \mathscr{F})$ by the Radon–Nikodym derivative

$$dP^* = \exp[\zeta_0^1(\gamma)]\, dP.$$

P^* is a probability measure in view of (2.11) and from Girsanov's theorem the process $\{z_t\}$ considered on the space $(C, \mathscr{F}, P^*)$ satisfies

$$dz_t = \gamma_t\, dt + \sigma_t\, dw_t, \tag{2.12}$$

where $(w_t, \mathscr{F}_t, P^*)$ is a Brownian motion. From (2.12), the observation process $\{y_t\}$ satisfies

$$dy_t = h_t\, dt + \sigma_2(t)\, dw_t \tag{2.13}$$

and

$$\int_0^1 |h_t|^2\, dt < \infty \quad \text{a.s.} \tag{2.14}$$

Choose any vector $\theta \in R^m$ and let $\xi_t = \theta' y_t$. Applying Ito's differential formula to the function $F(\xi) = \xi^2$ gives

$$\int_0^t \theta' \sigma_2 \sigma_2' \theta \, ds = \xi_t^2 - \xi_0^2 - 2 \int_0^t \xi_s \, d\xi_s.$$

This shows that the symmetric positive definite matrix $\sigma_2(t)\sigma_2'(t)$ is $\mathscr{Y}_t$-measurable for each t. Thus there exists a unitary matrix Q_t and a diagonal matrix L_t, both $\mathscr{Y}_t$-measurable, such that

$$\sigma_2(t)\sigma_2'(t) = Q_t L_t Q_t'. \tag{2.15}$$

Now define

$$\begin{aligned} T_t &= (L_t)^{-1/2} Q_t', \\ \hat{h}_t &= E^*[h_t | \mathscr{Y}_t], \\ \tilde{h}_t &= h_t - \hat{h}_t. \end{aligned} \tag{2.16}$$

($E^*[\cdot | \mathscr{Y}_t]$ denotes conditional expectation with respect to P^*.) The *innovations process* $\{v_t\}$ is defined by $v_0 = 0$ and

$$\begin{aligned} dv_t &= T_t(dy_t - \hat{h}_t \, dt) \\ &= T_t(\sigma_2(t) \, dw_t + \tilde{h}_t \, dt). \end{aligned} \tag{2.17}$$

LEMMA 2.1. *$(v_t, \mathscr{Y}_t, P^*)$ is a Brownian motion process.*

Proof. It is evident from the definition that $\{v_t\}$ is adapted to $\mathscr{Y}_t$ and has almost all sample paths continuous. Pick $\theta \in R^m$. In view of (2.15) and (2.16), for each t,

$$T_t \sigma_2(t) \sigma_2'(t) T_t' = I \tag{2.18}$$

so that applying Ito's differential formula to the function $f(v) = e^{i\theta' v}$ gives (using (2.17)),

$$e^{i\theta' v(t)} - e^{i\theta' v(s)} = \int_s^t i\theta' \, e^{i\theta' v(\tau)} T_\tau \tilde{h}_\tau \, d\tau + \int_s^t (-\tfrac{1}{2}|\theta|^2) \, e^{i\theta' v(\tau)} \, d\tau$$
$$+ \int_s^t i\theta' \, e^{i\theta' v(\tau)} T_\tau \sigma_2(\tau) \, dw_\tau.$$

Now,

$$\begin{aligned} E^*[i\theta' \, e^{i\theta' v(\tau)} T_\tau \tilde{h}_\tau | \mathscr{Y}_s] &= E^*\{i\theta' \, e^{i\theta' v(\tau)} T_\tau E^*[h_\tau - \hat{h}_\tau | \mathscr{Y}_\tau] | \mathscr{Y}_s\} \\ &= 0 \quad \text{a.s.}, \end{aligned}$$

and

$$E^*\left[\int_s^t i\theta' \, e^{i\theta' v(\tau)} T_\tau \sigma_2(\tau) \, dw_\tau | \mathscr{Y}_s \right] = 0 \quad \text{a.s.}$$

Thus,

$$E^*[e^{i\theta' v(t)} - e^{i\theta' v(s)} | \mathscr{Y}_s] = E^*\left[\int_s^t (-\tfrac{1}{2}|\theta|^2) \, e^{i\theta' v(\tau)} \, d_\tau | \mathscr{Y}_s \right],$$

or, alternatively,

$$E^*[\exp\{i\theta'\{v(t) - v(s)\}\} - 1|\mathscr{Y}_s]$$

$$= -\tfrac{1}{2}|\theta|^2 E^*\left[\int_s^t \exp\{i\theta'\{v(\tau) - v(s)\}\}\, d\tau|\mathscr{Y}_s\right]. \tag{2.19}$$

Pick $A \in Y_s$ and define

$$\eta_t = \int_A \exp\{i\theta'\{v(t) - v(s)\}\}\, dP^*.$$

Then from (2.19),

$$\eta_t = P^*A - \tfrac{1}{2}|\theta|^2 \int_A \int_s^t \exp[i\theta'\{v(\tau) - v(s)\}]\, d\tau\, dP^*$$

$$= P^*A - \tfrac{1}{2}|\theta|^2 \int_s^t \eta_\tau\, d\tau.$$

This integral equation has the unique solution

$$\eta_t = P^*A \exp[-\tfrac{1}{2}|\theta|^2(t - s)]$$

from which it is immediate that

$$E^*[\exp[i\theta'\{v(t) - v(s)\}]|\mathscr{Y}_s] = \exp[-\tfrac{1}{2}|\theta|^2(t - s)].$$

The statement of the lemma follows from this.

THEOREM 2.3. *Suppose $(M_t, \mathscr{Y}_t, P^*)$ is a martingale. Then there exists a process $\{\psi_t\}$ adapted to $\mathscr{Y}_t$ such that*

$$\int_0^1 |\psi_t|^2\, dt < \infty \quad a.s.$$

and

$$M_t = M_0 + \int_0^t \psi_s\, dv_s. \tag{2.20}$$

Proof. M_0 is a constant a.s. since $\mathscr{Y}_0 = \{C, \varnothing\}$. For convenience assume that $E^*M_t = M_0 = 0$. For $n = 1, 2, \cdots$, define

$$\tau_n = \min\left(1, \inf\left\{t : \int_0^t |T_s\hat{h}_s|^2\, ds \geqq n\right\}\right).$$

This is a stopping time of $\mathscr{Y}_t$, and $\tau_n \uparrow 1$ a.s. from (2.11(i)). Now define

$$\pi_t = \exp\left(\int_0^t (-T_s\hat{h}_s)\, dv_s - \frac{1}{2}\int_0^t |T_s\hat{h}_s|^2\, ds\right),$$

and define the measure $\tilde{P}_n$ by

$$d\tilde{P}_n = \pi_{t\wedge\tau_n}\, dP^*.$$

From Girsanov's theorem $\tilde{P}_n$ is a probability measure for each n, and the process

$$Y_t^n = v_t + \int_0^{t \wedge \tau_n} T_s \hat{h}_s \, ds \tag{2.21}$$

is a Brownian motion under $\tilde{P}_n$. Let

$$\mathscr{Y}_t^n = \sigma\{Y_s^n, 0 \leqq s \leqq t\}.$$

By Theorem 3 of [6], if $(\tilde{M}_t, \mathscr{Y}_t^n, \tilde{P}_n)$ is a separable martingale, then it has continuous sample paths and has the representation

$$\tilde{M}_t = \int_0^t \phi_s^n \, dY_s^n.$$

Observe from (2.17) and (2.21) that $Y_t^n = \int_0^t T_s \, dy_s$ for $t < \tau_n$ and hence that

$$\mathscr{Y}_t^n = \mathscr{Y}_{t \wedge \tau_n}.$$

Thus if $(\tilde{M}_t, \mathscr{Y}_{t \wedge \tau_n}, \tilde{P}_n)$ is a martingale,

$$\tilde{M}_{t \wedge \tau_n} = \int_0^{t \wedge \tau_n} \phi_s^n T_s \, dy_s. \tag{2.22}$$

Now suppose $(M_t, \mathscr{Y}_t, P^*)$ is a separable martingale. Then $(\tilde{M}_t, \mathscr{Y}_{t \wedge \tau_n}, \tilde{P}_n)$ is a martingale, where

$$\tilde{M}_t = M_{t \wedge \tau_n}(\pi_{t \wedge \tau_n})^{-1}. \tag{2.23}$$

Indeed, $(M_{t \wedge \tau_n}, \mathscr{Y}_{t \wedge \tau_n}, P^*)$ is a martingale by the optional sampling theorem, and, denoting integration with respect to $\tilde{P}_n$ by $\tilde{E}_n$, we have

$$\begin{aligned} \tilde{E}_n[\tilde{M}_t | \mathscr{Y}_{s \wedge \tau_n}] &= \tilde{E}_n[M_{t \wedge \tau_n}(\pi_{t \wedge \tau_n})^{-1} | \mathscr{Y}_{s \wedge \tau_n}] \\ &= \frac{E^*[M_{t \wedge \tau_n}(\pi_{t \wedge \tau_n})^{-1} \pi_{\tau_n} | \mathscr{Y}_{s \wedge \tau_n}]}{E^*[\pi_{\tau_n} | \mathscr{Y}_{s \wedge \tau_n}]} \\ &= \frac{E^*[M_{t \wedge \tau_n} | \mathscr{Y}_{s \wedge \tau_n}]}{\pi_{s \wedge \tau_n}} \\ &= M_{s \wedge \tau_n}(\pi_{s \wedge \tau_n})^{-1} = \tilde{M}_s. \end{aligned}$$

In this case $\tilde{M}_t = \tilde{M}_{t \wedge \tau_n}$ so that from (2.22),

$$\begin{aligned} \tilde{M}_t &= \int_0^{t \wedge \tau_n} \phi_s^n T_s \, dy_s \\ &= \int_0^{t \wedge \tau_n} \phi_s^n \, dv_s + \int_0^{t \wedge \tau_n} \phi_s^n T_s \hat{h}_s \, ds. \end{aligned} \tag{2.24}$$

Now $\pi_{t \wedge \tau_n}$ satisfies the Ito equation

$$\pi_{t \wedge \tau_n} = 1 - \int_0^{t \wedge \tau_n} \pi_s T_s \hat{h}_s \, dv_s. \tag{2.25}$$

Applying the Ito differential formula to the product in (2.23), using (2.24), (2.25), gives

$$M_{t\wedge\tau_n} = \int_0^{t\wedge\tau_n} \psi_s^n \, dv_s,$$

where

$$\psi_s^n = \pi_s(\phi_s^n - \tilde{M}_s T_s \hat{h}_s).$$

Such a representation is clearly unique, so that

$$\psi_s^n = \psi_s^{n'} \quad \text{for } n' \geqq n, \quad s \geqq \tau_n.$$

If ψ is the function which, for each n, agrees with ψ^n on $[s < \tau_n]$, then

$$M_{t\wedge\tau_n} = \int_0^{t\wedge\tau_n} \psi_s \, dv_s,$$

i.e.,

$$M_t = \int_0^t \psi_s \, dv_s \tag{2.26}$$

on $[t < \tau_n]$ for each n. Thus (2.26) holds a.s. for each t since $\tau_n \uparrow 1$ a.s.

3. Value function and principle of optimality. The results here are similar to those of Rishel [13]. The *value function* W_u is defined by (3.3) below and shown in Theorem 3.1 to satisfy a version of Bellman's principle of optimality. In Rishel's paper this depended on the class of controls satisfying a condition called "relative completeness." Here it turns out (Lemma 3.1) that this condition is always satisfied.

Suppose control $u \in \mathscr{U}_0^t$ is used on $[0, t]$ and $v \in \mathscr{U}_t^1$ on $(t, 1]$. Then the expected remaining cost at time t, given the observations up to that time, is

$$\begin{aligned} \psi_{uv}(t) &= E_{uv}\left[\int_t^1 c_s^{(v)} \, ds | \mathscr{Y}_t\right] \\ &= \frac{E[\rho_0^t(u)\rho_t^1(v)\int_t^1 c_s^{(v)} \, ds | \mathscr{Y}_t]}{E[\rho_0^1 | \mathscr{Y}_t]} \quad \text{a.s.} \end{aligned} \tag{3.1}$$

See [3, § 24.2]. We omit the superscript v if it is clear from the context. Define

$$f_{uv}(t) = E\left[\rho_0^t(u)\rho_t^1(v) \int_t^1 c_s \, ds | \mathscr{Y}_s\right].$$

The notations $\psi_u = \psi_{uu}$ and $f_u = f_{uu}$ for $u \in \mathscr{U}_0^1$ will also be used. Now $f_{uv}(t) \in L_1(C, \mathscr{F}, P)$ since $f_{uv} \geqq 0$ a.s. and

$$Ef_{uv}(t) = E\left[\rho_0^t(u)\rho_t^1(v) \int_t^1 c_s \, ds\right] \leqq k(1 - t)$$

from (2.9). L_1 is a complete lattice [2, p. 302] under the partial ordering $f_1 \prec f_2 \Leftrightarrow f_1(z) \leqq f_2(z)$ a.s. The set $\{f_{uv}(t) : v \in \mathscr{U}_t^1\}$ is bounded from below by the zero function, so the following infimum exists in L_1 for each t:

$$V(u, t) = \bigwedge_{v \in \mathscr{U}_t^1} f_{uv}(t). \tag{3.2}$$

Notice that the "normalizing factor" $E[\rho_0^1|\mathscr{Y}_t] = E[\rho_0^t|\mathscr{Y}_t]$ does not depend on v. The *value function* $W_u(t)$ is thus defined as

$$W_u(t) = \bigwedge_{v \in \mathscr{U}_t^1} E_{uv}\left[\int_t^1 c_s\, ds|\mathscr{Y}_t\right] = \frac{1}{E[\rho_0^t(u)|\mathscr{Y}_t]} V(u, t). \tag{3.3}$$

Thus $V(u, t)$ is an unnormalized version of the g.l.b. of the expected additional cost at time t. Suppose $\theta \in L_1$, $V(u, t) \prec \theta$; i.e., $V(u, t)(z) < \theta(z)$ for $z \in M$, $PM = 1$. Then there exists $v \in \mathscr{U}_t^1$ and a set M_v with $PM_v > 0$ such that $f_{uv}(t, z) < \theta(z)$ for $z \in M_v$. The class $\mathscr{U}$ is said to be *relatively complete* [13] if for any $t \in [0, 1]$ and $\varepsilon > 0$ there exists $v \in \mathscr{U}_t^1$ such that

$$f_{uv}(t) < V(u, t) + \varepsilon \quad \text{a.s.}$$

This amounts to saying, in the above, that for $\theta = V = \varepsilon$ there is a v with $PM_v = 1$. The fact (Lemma 3.1) that this is true is used in the proof of Theorem 3.1.

LEMMA 3.1. *$\mathscr{U}$ is relatively complete.*

Proof. Fix $\varepsilon > 0$, $t \in [0, 1]$ and $u \in \mathscr{U}_0^t$. Let $V(z) = V(t, u)(z)$, and for $v \in \mathscr{U}_t^1$ let

$$M_v = \{z : f_{uv}(z) < V(z) + \varepsilon\} \subset \mathscr{Y}_t.$$

A partial ordering is defined on the set $X = \{(v, M_v) : v \in \mathscr{U}_t^1\}$. Then Zorn's lemma is used to establish the existence of a maximal element (v^*, M_{v^*}) which has the property that $PM_{v^*} = 1$, proving the lemma.

The partial ordering $\succ$ on X is as follows:

$$(u, M) \succ (v, M_v) \quad \text{if and only if}$$

$$\begin{aligned} &\text{(i)} \quad M_u \supset M_v, \\ &\text{(ii)} \quad PM_u > PM_v, \\ &\text{(iii)} \quad u \text{ and } v \text{ agree on } M_v. \end{aligned} \tag{3.4}$$

Each chain in $(X, \succ)$ has an upper bound. Indeed, let $\{(v_\alpha, M_\alpha) : \alpha \in A\}$ be a chain in $(X, \succ)$.

1. If for some $\alpha_1 \in A$, $PM_1 = \sup_{\alpha \in A}\{PM_\alpha\}$, then $(u_{\alpha_1}, M_{\alpha_1})$ is an upper bound.

2. If $PM_{\alpha_1} < \sup_{\alpha \in A}\{PM_\alpha\} = m$ for each $\alpha_1 \in A$, then $PM_\alpha \uparrow m \leqq 1$. For each $n = 1, 2, \cdots$ pick α_n such that

$$PM_{\alpha n} > m - 1/n.$$

Let $M = \bigcup_{\alpha \in A} M_\alpha$. Then $M = \bigcup_{n=1}^\infty M_{\alpha_n}$; clearly $M \supset \bigcup_{n=1}^\infty M_{\alpha_n}$ and conversely, given $\alpha \in A$, $PM_\alpha < m - 1/n'$ for some integer n', and hence, $M_\alpha \subset M_{\alpha n'} \subset \bigcup_{n=1}^\infty M_{\alpha_n}$. Thus M is $\mathscr{Y}_t$-measurable and $PM = m$.

3. Define the control v on $t \leqq \tau \leqq 1$ as follows:

$$v(\tau, z) = v_{\alpha_n}(\tau, z), \qquad z \in M_{\alpha_n}.$$

This specifies v on M; on M^c let $v(\tau, z) = v_{\alpha_1}(\tau, z)$. v is clearly measurable; and

$$\{z : v(\tau, z) \in \Gamma\} = \bigcup_{i=1}^{\infty} M_{\alpha_i} \cap \{z : v_{\alpha_i}(\tau, z) \in \Gamma\} \in \mathscr{Y}_\tau$$

so that v is adapted to $\mathscr{Y}_t$. Let $M_1 = M_{\alpha_1} \cup M^c$, $M_i = M_{\alpha_i} - \bigcup_{j=1}^{i-1} M_j$ for $i = 1, 2, 3, \cdots$. Then $\{M_i\}$ is a partition of C into $\mathscr{Y}_t$-measurable sets. Hence,

$$\begin{aligned} E[\rho_t^1(v)|\mathscr{Y}_t] &= \sum_{i=1}^{\infty} E[I_{M_i}\rho_t^1(v_{\alpha_i})|\mathscr{Y}_t] \\ &= \sum_{i=1}^{\infty} I_{M_i}E[\rho_t^1(v_{\alpha_i})|\mathscr{Y}_t] \\ &= \sum_{i=1}^{\infty} I_{M_i} = 1 \quad \text{a.s.} \end{aligned}$$

Thus $v \in \mathscr{U}_t^1$.

4. (v, M) is an upper bound for $\{(v_\alpha, M_\alpha) : \alpha \in A\}$.

(a) $(v, M) \in X$; i.e., $M = M_v = \{z : f_{uv}(z) < V(z) + \varepsilon\}$, since $z \in M$ so $z \in M_{\alpha_i}$ for some i; hence $v = v_{\alpha_i}$ and $f_{uv}(z) < V(z) + \varepsilon$, while $z \in M^c$ implies $v(z) = v_{\alpha_1}(z)$ and $z \in M_v^c$ since $M_{\alpha_1} \subset M$.

(b) For any $\alpha \in A$, $(v, M) \succ (v_\alpha, M_\alpha)$. This is immediate from (3.4). Since each chain in X has an upper bound, X has a maximal element, i.e., an element (v^*, M^*) with the property that for each comparable $(v_\alpha, M_\alpha) \in X$,

$$(v^*, M^*) \succ (v_\alpha, M_\alpha).$$

It remains to show that $PM^* = 1$. Suppose $PM^* < 1$. Then $P(M^*)^c > 0$ so there exists $v' \in \mathscr{U}_t^1$ and a set $\Psi \subset (M^*)^c$ with $P\Psi > 0$ such that

$$f_{uv'}(z) < V(z) + \varepsilon, \qquad z \in \Psi.$$

Recall that Ψ, M are $\mathscr{Y}_t$-measurable. Define

$$v^0(t, z) = \begin{cases} v^*(t, z), & z \in M^*, \\ v'(t, z), & z \in (M^*)^c. \end{cases}$$

This is admissible and

$$M^0 = \{z : v^0(t, z) < V(z) + \varepsilon\} \supset M^* \cup \Psi.$$

Thus $PM^0 > PM$, and hence $(v^0, M^0) \succ (v^*, M^*)$, contradicting the maximality of (v^*, M^*). So $PM^* = 1$, as required.

THEOREM 3.1. *For each* $t \in [0, 1]$ *and* $u \in \mathscr{U}_0^t$, *the value function* $W_u(t)$ *satisfies the* "principle of optimality":

$$W_u(t) \leqq E_u\left[\int_t^{t+h} c_s^{(u)}\, ds|\mathscr{Y}_t\right] + E_u[W_u(t+h)|\mathscr{Y}_t] \quad a.s. \tag{3.5}$$

for each $h > 0$. *Furthermore* u *is optimal if and only if there is equality in* (3.5) *for all* t, *and* $h > 0$.

Proof.

$$V(u,t) = \bigwedge_{v\in\mathscr{U}_t^1} E\left[\rho_0^t(u)\rho_t^1(v)\int_t^{t+h} c_s^{(v)}\,ds + \rho_0^t(u)\rho_t^1(v)\int_{t+h}^1 c_s^{(v)}\,ds|\mathscr{Y}_t\right]$$

$$\leqq E\left[\rho_0^{t+h}(u)\int_t^{t+h} c_s^{(u)}\,ds|\mathscr{Y}_t\right] \tag{3.6}$$

$$+ \bigwedge_{v\in\mathscr{U}_{t+h}^1} E\left[\rho_0^{t+h}(u)\rho_{t+h}^1(v)\int_{t+h}^1 c_s^{(v)}\,ds|\mathscr{Y}_t\right].$$

(Otherwise there would be a $v\in\mathscr{U}_{t+h}^1$ and $M\in\mathscr{Y}_t$ with $PM > 0$ such that

$$V(u,t) - E\left[\rho_0^{t+h}(u)\int_t^{t+h} c_s\,ds|\mathscr{Y}_t\right] > E\left[\rho_0^{t+h}(u)\rho_{t+h}^1(v)\int_{t+h}^1 c_s\,ds|\mathscr{Y}_t\right]$$

for $z\in M$; i.e.

$$V(u,t) > E\left[\rho_0^{t+h}(u)\rho_{t+h}^1(v)\int_t^1 c_s\,ds|\mathscr{Y}_t\right] \quad \text{for } z\in M,$$

a contradiction.) The next step is to show that

$$E[V(u,t+h)|\mathscr{Y}_t] = \bigwedge_{v\in\mathscr{U}_{t+h}^1} E\left[\rho_0^{t+h}(u)\rho_{t+h}^1(v)\int_{t+h}^1 c_s^{(v)}\,ds|\mathscr{Y}_t\right]. \tag{3.7}$$

For any $v'\in\mathscr{U}_{t+h}^1$,

$$V(u,t+h) \leqq E\left[\rho_0^{t+h}(u)\rho_{t+h}^1(v')\int_{t+h}^1 c_s^{(v')}\,ds|\mathscr{Y}_{t+h}\right] \quad \text{a.s.}$$

Hence,

$$E[V(u,t+h)|\mathscr{Y}_t] \leqq E\left[\rho_0^{t+h}(u)\rho_{t+h}^1(v')\int_{t+h}^1 c_s^{(v')}\,ds|\mathscr{Y}_t\right] \quad \text{a.s.}$$

Therefore,

$$E[V(u,t+h)|\mathscr{Y}_t] \leqq \bigwedge_{v'\in\mathscr{U}_{t+h}^1} E\left[\rho_0^{t+h}(u)\rho_{t+h}^1(v')\int_{t+h}^1 c_s^{(v')}\,ds|\mathscr{Y}_t\right] \quad \text{a.s.}$$

Since the class $\mathscr{U}_t^1$ is relatively complete, given $t+h$, $u\in\mathscr{U}_0^{t+h}$, and $\varepsilon > 0$, there exists $u'\in\mathscr{U}_{t+h}^1$ such that

$$E\left[\rho_0^{t+h}(u)\rho_{t+h}^1(u')\int_{t+h}^1 c_s^{(u')}\,ds|\mathscr{Y}_{t+h}\right] \leqq V(u,t+h) + \varepsilon \quad \text{a.s.}$$

Then,

$$E\left[\rho_0^{t+h}(u)\rho_{t+h}^1(u')\int_{t+h}^1 c_s^{(u')}\,ds|\mathscr{Y}_t\right] \leqq E[V(u,t+h)|\mathscr{Y}_t] + \varepsilon \quad \text{a.s.,}$$

and thus,

$$\bigwedge_{u'\in\mathscr{U}_{t+h}^1} E\left[\rho_0^{t+h}(u)\rho_{t+h}^1(u')\int_{t+h}^1 c_s^{(u')}\,ds|\mathscr{Y}_t\right] \leqq E[V(u,t+h)|\mathscr{Y}_t] \quad \text{a.s.}$$

This establishes (3.7). From (3.6) and (3.7),

$$V(u,t) \leqq E\left[\rho_0^{t+h}(u)\int_t^{t+h} c_s^{(u)}\,ds|\mathscr{Y}_t\right] + E[V(u,t+h)|\mathscr{Y}_t] \quad \text{a.s.}$$

Dividing this through by $E[\rho_0^t(u)|\mathscr{Y}_t]$ gives (3.5) after noting that

$$\begin{aligned}\frac{E[V(u,t+h)|\mathscr{Y}_t]}{E[\rho_0^t(u)|\mathscr{Y}_t]} &= \frac{E[W_u(t+h)E\{\rho_0^{t+h}|\mathscr{Y}_{t+h}\}|\mathscr{Y}_t]}{E[\rho_0^t|\mathscr{Y}_t]} \\ &= \frac{E[W_u(t+h)\rho_0^{t+h}|\mathscr{Y}_t]}{E[\rho_0^t|\mathscr{Y}_t]} \\ &= E_u[W_u(t+h)|\mathscr{Y}_t].\end{aligned}$$

The last assertion is evident from the argument above.

4. Conditions for optimality. Two sets of criteria—Theorems 4.1 and 4.2—are presented in this section. Theorem 4.1. is included for two reasons: it is used later in establishing other criteria, and it is the equivalent in the present context of Rishel's results (Theorems 8 and 9 of [13]). The process "$W_u(t)$" of Rishel's Theorem 9 corresponds to the process $Z_u(t)$ defined below. Theorem 4.2 is the main result of this section.

The objective of this and the following sections is to get Hamilton–Jacobi criteria for optimality, i.e., a local characterization of the optimal policy in terms of the value function. The results are bound to be less than satisfactory in the case of partial observations as there is a different value function for each control: the expected remaining cost from a certain time on depends on what control was applied prior to that time. By restricting attention, as Rishel does, to "value decreasing" controls, in which class the optimal control, if it exists, must lie, one can get some way towards the above characterization. This is Theorem 4.2.

The case of "complete observations" is a great deal simpler as here there is only one value function. This case is treated in § 5.

THEOREM 4.1. *$u^* \in \mathscr{U}$ is optimal if and only if there exists a constant J^* and for each $u \in \mathscr{U}$ an integrable process $\{\alpha_u(t)\}$ adapted to $\mathscr{Y}_t$ and satisfying:*

$$\text{(i)} \qquad E_u\int_0^1 \alpha_u(s)\,ds = J^*,$$

$$\text{(ii)} \qquad \begin{aligned}E_{u^*}[c_t^{(u^*)}|\mathscr{Y}_t] - \alpha_{u^*}(t) &= 0,\\ E_u[c_t^{(u)}|\mathscr{Y}_t] - \alpha_u(t) &\geqq 0\end{aligned}$$

for almost all (t, z).

Then $J^* = J(u^*)$, the cost of the optimal policy.

Proof. Suppose u^* is optimal. Let $J^* = J(u^*) = \psi_{u^*}(0)$. Define $\kappa_u = J^*(\psi_u(0))^{-1}$; thus $\kappa_u \leqq 1$ and $\kappa_u = 1$ if u is optimal. Then the process

$$\alpha_u(t) = \kappa_u E_u[c_t^{(u)}|\mathscr{Y}_t]$$

is clearly integrable and in fact satisfies (i) and (ii). Indeed,

$$E_u\int_0^1 \alpha_u(s)\,ds = \kappa_u E_u\left[\int_0^1 c_s\,ds\right] = \kappa_u\psi_u(0) = J^*$$

and

$$E_u[c_t|\mathscr{Y}_t] - \alpha_u(t) = (1 - \kappa_u)E_u[c_t|\mathscr{Y}_t] \geqq 0.$$

Conversely, suppose there exists an integrable process $\{\alpha_u(t)\}$ satisfying (i) and (ii). Let $Z_u(t)$ be defined by

$$Z_u(t) = E_u[a_u(1)|\mathscr{Y}_t] - a_u(t),$$

where $a_u(t) = \int_0^t \alpha_u(s)\,ds$. It is easy to check that

$$\psi_u(t) = E_u\left[\int_0^t E_u(c_s^{(u)}|\mathscr{Y}_s)\,ds|\mathscr{Y}_t\right] - \int_0^t E_u(c_s^{(u)}|\mathscr{Y}_s)\,ds.$$

Thus

$$\begin{aligned}\psi_u(t) - Z_u(t) &= E_u\left[\int_0^1 E_u(c_s^{(u)}|\mathscr{Y}_s)\,ds - \int_0^1 \alpha_u(s)\,ds|\mathscr{Y}_t\right]\\ &\quad - \int_0^t E_u[c_s^{(u)}|\mathscr{Y}_s]\,ds + \int_0^t \alpha_u(s)\,ds\\ &= E_u\left[\int_t^1 \{E_u(c_s^{(u)}|\mathscr{Y}_s) - \alpha_u(s)\}\,ds|\mathscr{Y}_t\right]\\ &\geqq 0 \quad \text{a.s. from (ii).}\end{aligned}$$

It follows that

$$\psi_u \geqq Z_u \quad \text{a.e.} (d\lambda \times dP).$$

Similar steps using the equality in (ii) lead to

$$\psi_{u^*} = Z_{u^*} \quad \text{a.e.}(d\lambda \times dP).$$

Thus

$$E_u\psi_u(0) \geqq E_uZ_u(0),$$
$$E_{u^*}\psi_{u^*}(0) = E_{u^*}Z_{u^*}(0).$$

But $\psi_u(0) = J(u)$ and $E_uZ_u(0) = E_{u^*}Z_{u^*}(0) = J^*$ from (i). So from the relations above,

$$J(u) \geqq J^* = J(u^*)$$

for all $u \in \mathscr{U}$. This completes the proof.

Following Rishel [13], a control $u \in \mathscr{U}$ is called *value decreasing* if

$$W_u(t) \geqq E_u[W_u(t + h)|\mathscr{Y}_t] \quad \text{a.s. for each } t,$$

i.e., if $(W_u(t), \mathscr{Y}_t, P_u)$ is a supermartingale. Any optimal control is value decreasing: from Theorem 3.1, if u is optimal,

$$W_u(t) - E_u[W_u(t + h)|\mathscr{Y}_t] = E_u\left[\int_t^{t+h} c_s^{(u)}\,ds|\mathscr{Y}_t\right] \geqq 0 \quad \text{a.s.}$$

On the other hand, optimal controls could conceivably be the only value decreasing ones, though normally one would expect this class to be a good deal larger.

In the case of value decreasing controls the value function can be represented as an Ito process and the conditions for optimality restated in a more intuitively appealing way. The term "Ito process"is borrowed from [11, p. 286].

LEMMA 4.1. *Let $u \in \mathscr{U}$ be value decreasing. Then there exist processes $\{\Lambda W_u\}$, $\{\nabla W_u\}^2$ taking values in R, R^n respectively and adapted to $\mathscr{Y}_t$, such that*

$$\text{(i)} \quad \int_0^1 |\Lambda W_u|^2 \, dt < \infty \quad a.s.,$$

$$\text{(ii)} \quad E \int_0^1 |\nabla W_u| \, ds < \infty,$$

$$\text{(iii)} \quad W_u(t) = J^* + \int_0^t \Lambda W_u(s) \, ds + \int_0^t \nabla W_u(s) \, dy_s \quad a.s.$$

under measure P.

Proof. By assumption, $(W_u(t), \mathscr{Y}, P_u)$ is a supermartingale; thus from (3.5),

$$|E_u[W_u(t+h) - W_u(t)]| \leqq E_u\left[\int_t^{t+h} c_s^{(u)} \, ds\right] \leqq kh.$$

Thus the function $t \to E_u W_u(t)$ is right-continuous, and therefore (VI T4) (such a reference is to Meyer's book [4]) $\{W_u(t)\}$ admits a right-continuous modification, which is assumed to be the version chosen. It is clear from the definition that $W_u(t) \to 0$ as $t \uparrow 1$, a.s. and in $L_1(P_u)$ so that $\{W_u(t)\}$ is a potential and by (VII T29) there exists a unique integrable natural increasing process $\{A_t\}$ which generates $W_u(t)$; i.e. such that

$$W_u(t) = E_u[A_1 | \mathscr{Y}_t] - A_t. \tag{4.1}$$

Define, for $h > 0$,

$$\beta_t^h = \frac{1}{h}\{W_u(t) - E_u[W_u(t+h) | \mathscr{Y}_t]\}.$$

Then (VII T29) also states that

$$\int_0^t \beta_s^h \, ds \to A_t \tag{4.2}$$

weakly in $L_1(P_u)$ as $h \downarrow 0$ for each fixed t. Now from (3.5),

$$\beta_t^h \leqq \frac{1}{h} E_u\left[\int_t^{t+h} c_s^{(u)} \, ds | \mathscr{Y}_t\right] \leqq k \quad \text{a.s.}$$

Thus the subset $\mathscr{H} = \{\beta_t^h : h > 0\}$ is uniformly integrable and hence, from (II T23), weakly compact in $L_1(P_u)$. There therefore exists a sequence $h_n \downarrow 0$ and an element α_t of L_1 such that

$$\beta_t^{h_n} \xrightarrow{w} \alpha_t \quad \text{as } n \to \infty.$$

[2] So-called because they play a similar role to the functions $\Lambda\phi$ and $\phi_x = \nabla\phi$ in the Markov case (see § 1.2). This will become apparent.

It is then immediate that there is a sequence $h_n \downarrow 0$ and a subset $\{\alpha_t : t \in S\} \subset L_1$, where S is a countable dense subset of $[0, 1]$, such that

$$\beta_t^{h_n} \xrightarrow{w} \alpha_t \quad \text{as } n \to \infty \quad \text{for each } t \in S.$$

For $t \notin S$ define α_t by

$$\alpha_t = w\text{-}\lim_{\substack{s \downarrow t \\ s \in S}} \alpha_s. \tag{4.3}$$

To see that this limit exists, note that β_s^h is right-continuous in s for each fixed h. Let $\theta \in L_\infty$. For $t, t' \in S$, $t' > t$,

$$|E_u\theta(\alpha_t - \alpha_{t'})| \leqq |E_u\theta(\alpha_t - \beta_t^{h_n})| + |E_u\theta(\alpha_{t'} - \beta_{t'}^{h_n})| + |E_u\theta(\beta_{t'}^{h_n} - \beta_t^{h_n})|. \tag{4.4}$$

Now $\theta(\beta_{t'}^{h_n} - \beta_t^{h_n}) \to 0$ a.s. as $t' \downarrow t$, and hence also in L_1, in view of the uniform integrability. Choosing n such that the sum of the first two terms in (4.4) is less than $\frac{1}{2}\varepsilon$ and then t' such that $|E_u\theta(\beta_{t'}^{h_n} - \beta_t^{h_n})| < \frac{1}{2}\varepsilon$ gives

$$|E_u\theta(\alpha_{t'} - \alpha_t)| < \varepsilon.$$

Thus if $t_n \downarrow t$, $\{\alpha_{t_n}\}$ is a weak Cauchy sequence and the limit in (4.3) exists.

For $\theta \in L_\infty$,

$$E_u\theta\left(\int_0^t \alpha_s\, ds - A_t\right) \leqq E_u\theta\left(\int_0^t \alpha_s\, ds - \int_0^t \beta_s^h\, ds\right) + E_u\theta\left(\int_0^t \beta_s^h\, ds - A_t\right). \tag{4.5}$$

The last term converges to zero along $\{h_n\}$ from (4.2), and since the expectations $E_u\beta_s^h$ are uniformly bounded for $h > 0$, by Lebesgue's bounded convergence theorem,

$$\int_0^t E_u\theta(\alpha_s\, ds - \beta_s^{h_n})\, ds \to 0, \qquad n \to \infty.$$

Thus from (4.5),

$$E_u\theta\left(\int_0^t \alpha_s\, ds - A_t\right) = 0, \qquad \theta \in L_\infty, \quad t \in [0, 1].$$

It follows that

$$A_t = \int_0^t \alpha_s\, ds \quad \text{a.s. for each } t. \tag{4.6}$$

Recalling (4.2) and in view of (4.6), evidently

$$\alpha_t = w\text{-}\lim_{n \to \infty} \beta_t^{h_n}$$

for *every* subsequence $\{h_n\}$ such that the limit exists. Therefore

$$\alpha_t = w\text{-}\lim_{h \downarrow 0} \beta_t^h.$$

Now (4.1) says

$$W_u(t) = E_u[A_1 | \mathscr{Y}_t] - \int_0^t \alpha_s\, ds. \tag{4.7}$$

$Y_t = E_u[A_1|\mathscr{Y}_t]$ is a right-continuous, hence separable, uniformly integrable martingale on $(C, \mathscr{F}, P_u)$. Applying Theorem 2.3 with $\gamma = g^{(u)}$, $P^* = P_u$, shows that $\{Y_t\}$ has the representation

$$Y_t = Y_0 + \int_0^t \psi_s \, dv_s, \tag{4.8}$$

where $dv_t = T_t(dy_t - \hat{g}_2^{(u)} \, dt)$ is a Wiener process under P_u. Here

$$\hat{g}_2^{(u)} = E_u[g_2(t, z, u_t)|\mathscr{Y}_t]. \tag{4.9}$$

Thus from (4.7) and (4.8),

$$W_u(t) = Y_0 - \int_0^t [\alpha_s + \psi_s T_s \hat{g}_2^{(u)}(s)] \, ds + \int_0^t \psi_s T_s \, dy_s.$$

Now $W_u(0) = J^* = Y_0$; and defining

$$\Lambda W_u(t) = -\alpha_t - \psi_t T_t \hat{g}_2^{(u)}(t)$$

and

$$\nabla W_u(t) = \psi_t T_t$$

finally gives

$$W_u(t) = J^* + \int_0^t \Lambda W_u(s) \, ds + \int_0^t \nabla W_u(s) \, dy_s$$

as required.

THEOREM 4.2. *$u^* \in \mathscr{U}$ is optimal if and only if there exists a constant J^* and for each value decreasing control $u \in \mathscr{U}$ processes $\{\eta_t^{(u)}\}$, $\{\xi_t^{(u)}\}$, taking values in R, R^m respectively and adapted to $\mathscr{Y}_t$, and satisfying the following conditions*:

(i) $\int_0^1 |\xi_t^{(u)}|^2 \, dt < \infty \quad a.s., \qquad E\int_0^t \xi_t^{(u)} \, dy_t = 0;$

(ii) $\chi^{(u)}(1) = 0 \quad a.s.,$

where

$$\chi^{(u)}(t) = J^* + \int_0^t \eta_s^{(u)} \, ds + \int_0^t \xi_s^{(u)} \, dy_s;$$

(iii) $\eta_t^{(u)} + \xi_t^{(u)}\hat{g}_2^{(u)}(t) + \hat{c}_t^{(u)} \geqq 0 = \eta_t^{(u^*)} + \xi_t^{(u^*)}\hat{g}_2^{(u^*)}(t) + \hat{c}_t^{(u^*)}$

for almost all $(t, z) \in [0, 1] \times C$.

Then $\chi_t^{(u^*)} = W_{u^*}(t)$ a.s. and $J^* = J(u^*)$, the minimal cost. Here $\hat{g}_2^{(u)}(t)$ is defined by (4.9) *above, and $\hat{c}_t^{(u)}$ is defined similarly.*

Proof. Suppose $u \in \mathscr{U}$ is value decreasing. Then from (3.5),

$$W_u(t) - E_u[W_u(t + h)|\mathscr{Y}_t] \leqq E_u\left[\int_t^{t+h} c_s^{(u)} \, ds|\mathscr{Y}_t\right]. \tag{4.10}$$

Now from Lemma 4.1,

$$W_u(t) = J^* + \int_0^t \Lambda W_u(s)\,ds + \int_0^t \nabla W_u(s)\,dy_s.$$

Under measure P_u, $\{y_t\}$ has, from Lemma 2.1, the innovations process representation

$$dy_t = T_t^{-1}\,dv_t + \hat{g}_2^{(u)}(t)\,dt,$$

and thus

$$W_u(t) = J^* + \int_0^t [\Lambda W_u(s) + \nabla W_u(s)\hat{g}_2^{(u)}(s)]\,ds + \int_0^t \nabla W_u(s) T_s^{-1}\,dv_s.$$

Therefore,

$$W_u(t) - E_u[W_u(t+h)|\mathscr{Y}_t] = -E_u\left[\int_t^{t+h} [\Lambda W_u(s) + \nabla W_u(s)\hat{g}_2^{(u)}(s)]\,ds|\mathscr{Y}_t\right],$$

so (4.10) becomes

$$E_u\left[\int_t^{t+h} (\Lambda W_u(s) + \nabla W_u(s)\hat{g}_2^{(u)}(s) + c_s^{(u)})\,ds|\mathscr{Y}_t\right] \geqq 0 \quad \text{a.s.} \tag{4.11}$$

Denote the integrand in (4.11) by X_s and take $\theta \in L_\infty$. Then

$$\frac{1}{h}E_u\left[\theta E_u\left\{\int_t^{t+h} X_s\,ds|\mathscr{Y}_t\right\}\right] = \frac{1}{h}\int_t^{t+h} E_u\{E_u[\theta|\mathscr{Y}_t]X_s\}\,ds$$
$$\to E_u\{E_u[\theta|\mathscr{Y}_t]X_t\} = E_u\{\theta E_u[X_t|\mathscr{Y}_t]\}$$

as $h \downarrow 0$ for almost all t. Hence from (4.11),

$$\Lambda W_u(t) + \nabla W_u(t)\hat{g}_2^{(u)}(t) + \hat{c}_t^{(u)} \geqq 0 \tag{4.12}$$

for almost all (t, z). If u is optimal, then equality holds in (4.10), and hence in (4.12). Thus, identifying

$$\eta_t^{(u)} = \Lambda W_u(t), \quad \xi_t^{(u)} = \nabla W_u(t), \quad \chi_t^{(u)} = W_u(t),$$

properties (i)–(iii) are seen to hold.

Conversely, suppose J^*, $\{\eta_t^{(u)}\}$, $\{\xi_t^{(u)}\}$ exist and satisfy (i)–(iii), for each value decreasing control. Let $u \in \mathscr{U}$ be value decreasing. Then under measure P_u, $\chi_t^{(u)}$ satisfies

$$\chi_t^{(u)} = J^* + \int_0^t (\eta_s^{(u)} + \xi_s^{(u)}\hat{g}_2^{(u)}(s))\,ds + \int_0^t \xi_s^{(u)} T_s^{-1}\,dv_s, \tag{4.13}$$

where $\{dv_t\}$ is a Brownian motion. Define

$$\alpha_u(t) = -\eta_t^{(u)} - \xi_t^{(u)}\hat{g}_2^{(u)}(t).$$

Then from (4.13) and (ii),

$$E_u \int_0^1 \alpha_u(s)\,ds = J^*.$$

$\{\alpha_u(t)\}$ is adapted to $\mathscr{Y}_t$, and from (iii),

$$E_u[c_t^{(u)}|\mathscr{Y}_t] - \alpha_u(t) \geqq 0. \tag{4.14}$$

In the case $u = u^*$, (4.14) holds with equality. It now follows from Theorem 4.1 that u^* is optimal in the class of value decreasing controls. Since these are, as remarked earlier, the only candidates for the optimum, u^* must be optimal in $\mathscr{U}$.

Since u^* is optimal, $W_{u^*}(t) = \psi_{u^*}(t)$ from Theorem 3.1. Now

$$\begin{aligned}
\psi_{u^*}(t) &= E_{u^*}\left[\int_t^1 c_s^{(u^*)}\,ds|\mathscr{Y}_t\right] \\
&= E_{u^*}\left[\int_t^1 (-\eta_s^{(u^*)} - \xi_s^{(u^*)}\hat{g}_2^{(u^*)})\,ds - \int_t^1 \xi_s^{(u^*)}\,dv_s|\mathscr{Y}_t\right] \qquad \text{(from (iii))} \\
&= E_{u^*}\left[-\int_t^1 \eta_s^{(u^*)}\,ds - \int_t^1 \xi_s^{(u^*)}\,dy_s|\mathscr{Y}_t\right] \\
&= E_{u^*}\left[\int_0^t \eta_s^{(u^*)}\,ds + \int_0^t \xi_s^{(u^*)}\,dy_s|\mathscr{Y}_t\right] + J^* \qquad \text{(from (ii))} \\
&= \int_0^t \eta_s^{(u^*)}\,ds + \int_0^t \xi_s^{(u^*)}\,dy_s + J^* \\
&= \chi_t^{(u^*)}.
\end{aligned}$$

Thus $\chi_t^{(u^*)} = W_{u^*}(t)$, as claimed.

5. Completely observable systems. This section treats the case where the entire past of z is available for control; i.e. (in the definitions of § 2) $m = n$ and $\mathscr{Y}_t = \mathscr{F}_t$ for each t. Thus the admissible controls (denoted by $\mathscr{N}$) are functionals of the past of z, and are for that reason sometimes referred to as "nonanticipative controls" [9].

The considerable simplification that results in this case is due to the fact that there is now only one value function. In fact,

$$\begin{aligned}
\psi_{uv}(t) &= E_{uv}\left[\int_t^1 c_s^{(v)}\,ds|\mathscr{F}_t\right] \\
&= \frac{E[\rho_0^t(u)\rho_t^1(v)\int_t^1 c_s^{(v)}\,ds|\mathscr{F}_t]}{\rho_0^t(u)} \\
&= E\left[\rho_t^1(v)\int_t^1 c_s^{(v)}\,ds|\mathscr{F}_t\right]
\end{aligned} \tag{5.1}$$

does not depend on u; thus $W_u(t) = W(t)$ for all u, where

$$W(t) = \bigwedge_{v\in\mathscr{N}_t^1} E\left[\rho_t^1(v)\int_t^1 c_s^{(v)}\,ds|\mathscr{F}_t\right].$$

The principle of optimality (3.5) becomes

$$W(t) \leqq E_u\left[\int_t^{t+h} c_s^{(u)}\,ds|\mathscr{F}_t\right] + E_u[W(t+h)|\mathscr{F}_t]. \tag{5.2}$$

Using this, a genuine Hamilton–Jacobi criterion (Theorem 5.1) for optimality can be obtained, as a corollary to Theorem 4.2.

THEOREM 5.1 (Nonanticipative controls). *$u^* \in \mathcal{N}$ is optimal if and only if there exist a constant J^* and processes $\{\eta_t\}$, $\{\xi_t\}$ taking values in R, R^n respectively, adapted to $\mathscr{F}_t$, and satisfying the following conditions:*

$$\text{(i)}\quad \int_0^1 |\xi_t|^2\,dt < \infty \quad a.s., \qquad E\int_0^1 \xi_t\,dz_t = 0,$$

$$\text{(ii)}\quad \chi(1) = 0 \quad a.s.,$$

where

$$\chi(t) = J^* + \int_0^t \eta_s ds + \int_0^t \xi_s\,dz_s,$$

$$\text{(iii)}\quad \eta_t + \xi_t g_t^{(u)} + c_t^{(u)} \geqq 0 = \eta_t + \xi_t g_t^{(u^*)} + c_t^{(u^*)}$$

for almost all (t, z), for each $u \in \mathcal{N}$.

Then $\chi(t) = W_t$ a.s. and $J^* = J(u^*)$, the minimal cost.

Now the value function $W(t)$ is defined whether or not an optimal control exists. As a result of independent interest we obtain a representation for $W(t)$ as an Ito process (Theorem 5.2).

Recall the definitions of the sets $\mathscr{G}$ and $\mathscr{D}$ from § 2.

LEMMA 5.1. *There exists a process $h \in \mathscr{G}$ such that $(W_t, \mathscr{F}_t, P^*)$ is a supermartingale, where*

$$dP^*/dP = \exp[\zeta_0^1(h)].$$

Proof. Select a sequence $\{u_n\} \subset \mathcal{N}$ such that

$$J(u_n) = \psi_{u_n}(0) \downarrow W(0) = J^*.$$

Now $g^{(u_n)} \in \mathscr{G}$ and hence $\rho_0^1(u_n) \in \mathscr{D}$ for each n. From Theorem 2.2 there exists a subsequence, also denoted by $\{\rho(u_n)\}$, and an element $h \in \mathscr{G}$ such that

$$\rho_0^1(u_n) \to \rho^* \quad \text{weakly in } L_1(P), \tag{5.3}$$

where

$$\rho^* = \exp[\zeta_0^1(h)].$$

Evidently, from (5.3), for any $t \in [0, 1]$,

$$\rho_0^t(u_n) = E[\rho_0^1(u_u)|\mathscr{F}_t] \to E[\rho^*|\mathscr{F}_t] = \exp[\zeta_0^t(h)]. \tag{5.4}$$

Define the measure P^* by $dP^* = \rho^*\,dP$ and let

$$\rho^{*t}_0 = E[\rho^*|\mathscr{F}_t].$$

To show that $(W_t, \mathscr{F}_t, P^*)$ is a supermartingale it suffices to prove that for any $t, h, F \in \mathscr{F}_t$,

$$\int_F (W_{t+h} - W_t)\,dP^* = \int_F \rho^{*t+h}_0 (W_{t+h} - W_t)\,dP \leqq 0. \tag{5.5}$$

Let $\rho_* = \rho_0^{*t+h}$ and $\rho_n = \rho_0^{t+h}(u_n)$. Then

$$\begin{aligned}\int_F \rho_*(W_{t+h} - W_t) &= \int_F (\rho_* - \rho_n)(W_{t+h} - W_t) + \int_F \rho_n(\psi_{u_n}(t) - W_t)\\ &\quad + \int_F \rho_n(W_{t+h} - \psi_{u_n}[t+h]) + \int_F \rho_n(\psi_{u_n}(t+h) - \psi_{u_n}(t)).\end{aligned} \tag{5.6}$$

The third and fourth terms of (5.6) are nonpositive, the third because $\psi_{u_n}(t+h)$ majorizes W_{t+h} and the fourth because ψ_{u_n} is a supermartingale under P_{u_n}.

Fix $\varepsilon > 0$ and choose n' such that $\psi_{u_{n'}}(0) < W(0) + \varepsilon$ for $n \geqq n'$. From (5.2) (with $t = 0, h = t$),

$$E_{u_n}[\psi_{u_n}(t) - W_t] < \varepsilon$$

for each t. Hence,

$$\begin{aligned}\int_F \rho_n[\psi_{u_n}(t) - W_t] &\leqq \int_C \rho_n[\psi_{u_n}(t) - W_t]\\ &= E_{u_n}[\psi_{u_n}(t) - W_t] \leqq \varepsilon \quad \text{for } n \geqq n'.\end{aligned} \tag{5.7}$$

Now $[W_{t+h} - W_t]I_F \in L_\infty$, so there exists n'' such that for $n \geqq n''$,

$$\int_F (\rho_* - \rho_n)(W_{t+h} - W_t) < \varepsilon.$$

Thus for $n \geqq \max[n', n'']$, in (5.6),

$$\int_F \rho_*(W_{t+h} - W_t) < \varepsilon$$

which is equivalent to (5.5) since ε was arbitrary. This completes the proof.

THEOREM 5.2. *There exist processes $\{\Lambda W_t\}, \{\nabla W_t\}$ taking values in R, R^n, respectively, and adapted to $\mathscr{F}_t$, such that*

$$\text{(i)} \quad \int_0^1 |\nabla W|^2\, ds < \infty \quad a.s.,$$

$$\text{(ii)} \quad E\int_0^1 |\Lambda W|\, ds < \infty,$$

$$\text{(iii)} \quad W_t = J^* + \int_0^t \Lambda W_s\, ds + \int_0^t \nabla W_s\, dz_s$$

a.s. under measure P.

Proof. Choose a sequence $\{u_n\} \subset \mathscr{U}$ satisfying (5.3) and such that

$$J(u_n) \to W(0) \quad \text{as } n \to \infty.$$

Now

$$\begin{aligned}|E^*(W_{t+h} - W_t)| &= |E[\rho^*(W_{t+h} - W_t)]|\\ &\leqq |E[(\rho^* - \rho(u_n))(W_{t+h} - W_t)]|\\ &\quad + |E[\rho(u_n)(W_{t+h} - W_t)]|.\end{aligned} \tag{5.8}$$

The first term on the right goes to zero as $n \to \infty$ since $(W_{t+h} - W_t) \in L_\infty$ and since $\rho(u_n) \to \rho^*$ weakly in L_1 by (5.3). Also

$$E[\rho(u_n)(W_{t+h} - W_t)] = E[\rho(u_n)(W_{t+h} - \psi_{u_n}(t + h))] + E[\rho(u_n)(\psi_{u_n}(t) - W_t)] + E[\rho(u_n)(\psi_{u_n}(t + h) - \psi_{u_n}(t))],$$

and by (5.7) the first two terms on the right go to zero as $n \to \infty$. Finally from (5.1) it is easy to check that

$$E[\rho(u_n)(\psi_{u_n}(t) - \psi_{u_n}(t + h))] = E\left[\rho_t^{t+h}(u_n) \int_t^{t+h} c_s\, ds\right] \leqq kh.$$

Thus letting $n \to \infty$ in (5.8) we get

$$|E^*(W_{t+h} - W_t)| \leqq kh. \tag{5.9}$$

This implies, as in Lemma 4.1, the existence of a right-continuous modification of W_t; and since $(W_t, \mathscr{F}_t, P^*)$ is a potential, it follows again from (VII T 29) of [4] that

$$W_t = E^*[A_1|\mathscr{F}_t] - A_t,$$

where

$$A_t = \operatorname*{w-lim}_{h\downarrow 0} \int_0^t \beta_s^h\, ds$$

and

$$\beta_t^h = (1/h)(W_t - E^*[W_{t+h}|\mathscr{F}_t]).$$

The next stage is to show that $\alpha_t = dA_t/dt = \text{w-lim}_{h\downarrow 0}\beta_t^h$. It suffices to show that $\mathscr{H} = \{\beta_t^h : h > 0\}$ is uniformly integrable; then the rest of the proof is exactly as in the proof of Lemma 4.1. From (II T19) of [4], $\mathscr{H}$ is uniformly integrable if:

(i) $E^*\beta_t^h$ are uniformly bounded for $h > 0$, and

(ii) $\int_F|\beta_t^h|\, dP^* \to 0$ as $P^*F \to 0$, uniformly in h.

(i) follows from (5.9). Since β_t^h is $\mathscr{F}_t$-measurable, in proving (ii) we can restrict ourselves to $F \in \mathscr{F}_t$. Now

$$\begin{aligned}\int_F h\beta_t^h\, dP^* &= \int_F [W_t - W_{t+h}]\, dP^* \\ &= \int_F [W_t - W_{t+h}](\rho^* - \rho(u_n))\, dP + \int_F [W_t - W_{t+h}]\rho(u_n)\, dP.\end{aligned} \tag{5.10}$$

Once again since $(W_t - W_{t+h}) \in L_\infty$ and $\rho(u_n) \xrightarrow{w} \rho^*$, the first term on the right goes to zero as $n \to \infty$. Next,

$$\begin{aligned}\int_F [W_t - W_{t+h}]\rho(u_n)\, dP &= \int_F \rho(u_n)(W_t - \psi_{u_n}(t))\, dP \\ &+ \int_F \rho(u_n)(\psi_{u_n}(t + h) - W_{t+h})\, dP \\ &+ \int_F \rho(u_n)(\psi_{u_n}(t) - \psi_{u_n}(t + h))\, dP.\end{aligned} \tag{5.11}$$

From (5.7), the first two terms on the right go to zero as $n \to \infty$. On the other hand, from (5.1),

$$\psi_{u_n}(t) = E\left[\rho_t^{t+h}(u_n)\int_t^{t+h} c_s\, ds|\mathscr{F}_t\right] + E[\rho_t^{t+h}(u_n)\psi_{u_n}(t+h)|\mathscr{F}_t],$$

so that

$$\begin{aligned}\int_F \rho(u_n)\psi_{u_n}(t)\,dP &= \int_F \rho_0^t(u_n)\psi_{u_n}(t)\,dP\\ &= \int_F \rho_0^{t+h}(u_n)\left[\int_t^{t+h} c_s\,ds\right]dP + \int_F \rho_0^{t+h}(u_n)\psi_{u_n}(t+h)\,dP.\end{aligned}$$

Also,

$$\int_F \rho(u_n)\psi_{u_n}(t+h)\,dP = \int_F \rho_0^{t+h}(u_n)\psi_{u_n}(t+h)\,dP$$

so that the last term in (5.11) is equal to

$$\int_F \rho_0^{t+h}(u_n)\left[\int_t^{t+h} c_s\,ds\right]dP \leqq kh\int_F \rho_0^{t+h}(u_n)\,dP$$

and converges to khP^*F as $n \to \infty$. Thus letting $n \to \infty$ in (5.10) we conclude that

$$\int_F h\beta_t^h\,dP^* \leqq khP^*F,$$

and (ii) is established. Therefore,

$$W_t = E^*[A_1|\mathscr{F}_t] - \int_0^t \alpha_s\,ds. \tag{5.12}$$

To represent the separable martingale $E^*[A_1|\mathscr{F}_t]$, again Theorem 2.3 is used. Recall from Lemma 5.1 that

$$dP^* = \exp[\zeta_0^1(h)]\,dP.$$

Thus,

$$dw = \sigma^{-1}(dz - h_t\,dt) \tag{5.13}$$

is a Brownian motion under P^* and is in fact the innovations process for $\{z_t\}$ since it is adapted to $\mathscr{F}_t$. From Theorem 2.3 there exists a process $\{\phi_t\}$ such that

$$E^*[A_1|\mathscr{F}_t] = E^*[A_1] + \int_0^t \phi_s\,dw_s. \tag{5.14}$$

Combining (5.12)–(5.14) gives

$$W_t = J^* + \int_0^t \Lambda W_s\,ds + \int_0^t \nabla W_s\,dz_s,$$

where

$$\Lambda W_t = -\alpha_t - \phi_t\sigma_t^{-1}h_t, \qquad \nabla W_t = \phi_t\sigma_t^{-1}.$$

This is the desired result.

6. Markov controls. In this section a more restricted class of models is considered, namely those where the system matrices g and σ depend at a given time on the state only at that time. More precisely, let $\mathscr{B}_t$ be the σ-field generated by the single random variable z_t. The definitions (2.3) and (2.5) are unchanged except for (2.3(ii)) and (2.5(ii)) which now read:

(6.1) (ii) For each fixed (t, u), $g(t, \cdot, u)$ and $\sigma(t, \cdot)$ are $\mathscr{B}_t$-measurable.

In view of [3, § 35.1a] this amounts to saying that g and σ are functions on $[0, 1] \times R^n \times R^l$ taking on values $g(t, z_t, u)$ and $\sigma(t, z_t)$ at (t, z, u).

The class of *Markov controls* is denoted by $\mathscr{M} = \mathscr{M}_0^1$, where $\mathscr{M}_s^t$ is the class of functions u satisfying the following conditions:

(6.2)
(i) $u:[0, 1] \times R^n \to \Xi \subset R^l$ is jointly measurable,
(ii) $E[\rho_s^t(u)|\mathscr{F}_s] = 1$ a.s.,

where $\rho_s^t(u)$ is defined by (2.6) with

$$g_t^{(u)} = g(t, z_t, u[t, z_t]).$$

Let $u \in \mathscr{M}$. Then, from Theorem 2.1, under measure P_u the process $\{z_t\}$ satisfies

$$z_t = z_s + \int_s^t g_\tau^{(u)}\,d\tau + \int_s^t \sigma_\tau\,dw_\tau, \tag{6.3}$$

where $(w_t, \mathscr{F}_t, P_u)$ is a Brownian motion. From (6.3) it is evident that

$$E_u[z_t|\mathscr{F}_s] = E_u[z_t|\mathscr{B}_s] \quad \text{a.s.}$$

and it is easy to see that z_t is a Markov process under P_u; hence the term "Markov controls." $\{z_t\}$ is also Markov under the original measure P.

The cost rate function c is also assumed to satisfy a condition similar to (6.1), so that

$$c_t^{(u)}(z) = c(t, z_t, u[t, z_t]).$$

Stopping the process at the first exit time τ from a cylinder Q (as in § 1.2) can be accommodated within this framework. For, let $I(s, x) = 1$ for $(s, x) \in Q$ and $= 0$ elsewhere. Then the new system functions $g^\circ = Ig$, $\sigma^\circ = I\sigma$ and $c^\circ = Ic$ satisfy all the relevant conditions. If $u \in \mathscr{M}$ and E_u° denotes integration with respect to the measure corresponding to $g^{\circ(u)}$, σ°, then

$$E_u\left[\int_0^\tau c_s^{(u)}\,ds\right] = E_u^\circ\left[\int_0^1 c_s^{\circ(u)}\,ds\right].$$

The remaining cost function $\psi_u(t)$ is defined for $u \in \mathscr{M}$ as

$$\begin{aligned}\psi_u(t) &= E_u\left[\int_t^1 c_s^{(u)}\,ds|\mathscr{B}_t\right]\\ &= E_u\left[\int_t^1 c_s^{(u)}\,ds|\mathscr{F}_t\right]\\ &= E\left[\rho_t^1(u)\int_t^1 c_s^{(u)}\,ds|\mathscr{F}_t\right].\end{aligned}$$

This does not depend on u_s for $s \in [0, t]$; there is therefore, as in the case of complete observations, a single value function $U(t, z_t)$ defined by

$$U_t = U(t, z_t) = \bigwedge_{u \in \mathscr{M}_t^1} \psi_u(t).$$

Since $\mathscr{M}_t^1 \subset \mathscr{N}_t^1$ it is clear that $U_t \geqq W_t$ a.s. for each t. The main result of this section (Theorem 6.2) is that in fact $U_t = W_t$. This is intuitively clear: since the system's evolution from time t depends only on z_t the controller gains nothing by taking account of previous values z_s, $s < t$. The proof depends on a principle of optimality for the Markov case and results exactly analogous to Lemma 3.1 and Theorem 5.1 for the completely observable case. The proofs are almost identical here, the Markov property stepping in wherever the fact $\mathscr{F}_s \subset \mathscr{F}_t$ for $s < t$ was used in § 5. So in the following, complete details are provided only where there is significant deviation from the corresponding previous proofs.

LEMMA 6.1 (Markov principle of optimality). *Let $u \in \mathscr{M}$. Then for each t, h,*

$$U_t \leqq E_u\left[\int_t^{t+h} c_s^{(u)}\,ds|\mathscr{B}_t\right] + E_u[U_{t+h}|\mathscr{B}_t] \quad a.s. \tag{6.4}$$

The following result is the analogue of Theorem 5.2.

LEMMA 6.2. *There exist measurable functions $\Lambda U:[0, 1] \times R^n \to R$ and $U_x:[0, 1] \times R^n \to R^n$ such that:*

(i) $E\int_0^1 |\Lambda U(t, z_t)|\,dt < \infty$,

(ii) $\int_0^1 |U_x(t, z_t)|^2\,dt < \infty \quad a.s.$,

(iii) $U(t, z_t) = J_M + \int_0^t \Lambda U(s, z_s)\,ds + \int_0^t U_x(s, z_s)\,dz_s$,

where $J_M = \inf_{u \in \mathscr{M}} J(u)$, the minimum Markov cost.

Proof. The methods of Theorem 5.2 can be used to show that U_t has the representation:

$$U_t = J_M + \int_0^t \eta_s\,ds + \int_0^t \xi_s\,dz_s, \tag{6.5}$$

where $\{\eta_\tau\}$, $\{\xi_\tau\}$ are adapted to $\mathscr{F}_t$. It remains to show that η_t, ξ_t are $\mathscr{B}_t$-measurable for each t. For $n = 1, 2, \cdots$ let

$$\tau_n = \min\left(1, \inf\left\{t: \int_0^t |\xi_s|^2\, ds \geqq n\right\}\right).$$

τ_n is a stopping time of $\{\mathscr{F}_t\}$ and $\tau_n \uparrow 1$ a.s. since

$$\int_0^1 |\xi_s|^2\, ds < \infty \quad \text{a.s.}$$

Let

$$\xi_t^{(n)} = \begin{cases} \xi_t & \text{for } t \leqq \tau_n, \\ 0 & \text{for } \tau_n < t \leqq 1. \end{cases}$$

Let

$$M_t = \int_0^t \xi_s\, dz_s,$$

$$M_t^{(n)} = M_{t\wedge\tau_n} = \int_0^t \xi_n^{(n)}\, dz_s. \tag{6.6}$$

Now $E\int_0^1 |\xi_s^{(n)}|^2\, ds \leqq n$, so that $M_t^{(n)}$ is a second order (square integrable) martingale for each n; thus M_t is by definition a local second order martingale. The following results are proved in Kunita and Watanabe [12]. Let

$$\Upsilon = \{(a_1(t) - a_2(t)) \colon a_i(t \wedge \tau_n) \text{ is a natural, integrable increasing process adapted to } \mathscr{F}_t, i = 1, 2; n = 1, 2, \cdots\}.$$

If $(X_t, \mathscr{F}_t)$, $(Y_t, \mathscr{F}_t)$ are local second order martingales, there exists a unique process $\langle Y, X\rangle_t \in \Upsilon$ such that for $t > s$,

$$E[(X_{t\wedge\tau_n} - X_{s\wedge\tau_n})(Y_{t\wedge\tau_n} - Y_{s\wedge\tau_n})|\mathscr{F}_s] = E[\langle Y, X\rangle_{t\wedge\tau_n} - \langle Y, X\rangle_{s\wedge\tau_n}|\mathscr{F}_s] \quad \text{a.s.}$$

In addition,

$$\langle Y, X\rangle_t = \tfrac{1}{4}(\langle X + Y\rangle_t - \langle X - Y\rangle_t), \tag{6.7}$$

where

$$\langle X\rangle_t = \langle X, X\rangle_t.$$

$\langle X\rangle_t$ is known as the *quadratic variation* of X for the following reason: if X has continuous sample paths, then [12, Thm. 1.3] there exists a sequence of partitions $\{t_k^{(n)}, k = 1, 2, \cdots, k_n\}$ of $[0, t]$ such that

$$\max_k |t_k^{(n)} - t_{k-1}^{(n)}| \to 0, \qquad n \to \infty, \tag{6.8}$$

$$\sum_k (X_{t_{k^{(n)}}} - X_{t_{k-1}^{(n)}})^2 \to \langle X\rangle_t - \langle X\rangle_0 \quad \text{a.s. as } n \to \infty. \tag{6.9}$$

It is shown in [15] that for local martingales of the form (6.6),

$$\langle M\rangle_t = \int_0^t |\xi_s'\sigma_s|^2\, ds \quad \text{a.s.} \tag{6.10}$$

Also, referring to (6.5) and (6.9), we have

$$\sum_k [U_{t_k^{(n)}} - U_{t_{k-1}^{(n)}}]^2 \to \langle M\rangle_t - \langle M\rangle_0 \quad \text{a.s. as } n \to \infty. \tag{6.11}$$

(The sums corresponding to $\int \eta_s \, ds$ converge to zero a.s. since this term is of bounded variation.)

Let superscript i denote the ith component of a vector or row of a matrix. Then from (6.6),

$$M_t + z_t^i = \int_0^t (\xi_s' \sigma_s + \sigma_s^i) \, dB_s,$$

so that, using (6.10), we have

$$\langle M + z^i\rangle_t = \int_0^t (|\xi_s'\sigma_s|^2 + |\sigma_s^i|^2 + 2\xi_s'\sigma_s(\sigma_s^i)') \, ds,$$

$$\langle M - z^i\rangle_t = \int_0^t (|\xi_s'\sigma_s|^2 + |\sigma_s^i|^2 - 2\xi_s'\sigma_s(\sigma_s^i)') \, ds.$$

Therefore, using (6.7),

$$\langle M, z\rangle_t = \int_0^t \sigma_s \sigma_s' \xi_s \, ds, \tag{6.12}$$

i.e.,

$$\xi_t = (\sigma_t \sigma_t')^{-1} \frac{d}{dt} \langle M, z\rangle_t,$$

where $\langle M, z\rangle_t$ is the n-vector whose ith component is $\langle M, z^i\rangle_t$.

In view of (6.9) and (6.11), for each $h > 0$ there is a sequence of partitions $\{t_k^{(n)}\}$ of $[t, t + h]$ satisfying (6.8) and

$$\sum_k (Y_{t_k^{(n)}} - Y_{t_{k-1}^{(n)}})^2 \to \langle X\rangle_{t+h} - \langle X\rangle_t \quad \text{a.s.,} \qquad n \to \infty, \tag{6.13}$$

where in this case $X_t = M_t + z_t^i$ or $M_t - z_t^i$ and $Y_t = U_t + z_t^i$ or $U_t - z_t^i$. In either case, for any n the sum on the left of (6.13) is an $\mathscr{F}_t^{t+h}$-measurable random variable, where

$$\mathscr{F}_t^{t+h} = \sigma\{z_s, s \in [t, t + h]\}.$$

It follows from (6.12) that

$$\xi_t^{(h)} = \frac{1}{h} \int_t^{t+h} \xi_s^i \, ds$$

is $\mathscr{F}_t^{t+h}$-measurable. Now $\xi_t^{(h)} \to \xi_t^i \; w - L_1$ for almost all t. Hence a subsequence of a sequence of convex combinations converges a.s. and therefore ξ_t^i is $\mathscr{F}_t^{t+h}$-measurable for every h and hence measurable with respect to

$$\bigcap_{h>0} \mathscr{F}_t^{t+h} = \mathscr{B}_t.$$

There is thus a measurable function $U_x : [0, 1] \times R^n \to R^n$ such that

$$U_x(t, z_t) = \xi_t. \tag{6.14}$$

Referring back to (6.5) now gives

$$\int_t^{t+h} \eta_s\, ds = U_{t+h} - U_t - \int_t^{t+h} U_x(s, z_s)\, dz_s.$$

Thus $(1/h)\int_t^{t+h} \eta_s\, ds$ is $\mathscr{F}_t^{t+h}$-measurable, so η_t must be $\mathscr{B}_t$-measurable by the same reasoning as above. Defining

$$\Lambda U(t, z_t) = \eta_t$$

concludes the proof of the lemma.

COROLLARY. *Suppose the value function $U(t, x)$ has continuous first, and continuous first and second, partial derivatives respectively in t and in x; then*

$$(6.15) \qquad U_x(t, x) = \frac{\partial}{\partial x} U(t, x),$$

$$(6.16) \qquad \Lambda U(t, x) = \frac{\partial}{\partial t} U(t, x) + \frac{1}{2} \sum_{i,j} \frac{\partial^2}{\partial x_i\, \partial x_j} U(t, x)(\sigma\sigma')_{ij}.$$

Proof. Denote the right-hand sides of (6.15) and (6.16) by U_x', $\Lambda U'$ respectively. Under measure P,

$$dz_t = \sigma(t, z_t)\, dB_t,$$

so applying Ito's lemma to the function $U(t, z_t)$ gives

$$dU_t = U_x'(t)\, dz_t + \Lambda U'(t)\, dt.$$

Thus $\int_0^t (U_x - U_x')\, dz = \int_0^t (\Lambda U' - \Lambda U)\, dt$ and the left-hand member is a local martingale which must be of bounded variation. It follows that $\int_0^t (U_x - U_x')\, dz = 0$ a.s. for each t, and hence that $\Lambda U_t' = \Lambda U_t$, $U_x'(t) = U_x(t)$ a.s.

Remark. The corollary shows that the results of this section are precisely equivalent to those of Fleming mentioned in the Introduction, when the relevant conditions are satisfied. But in [9] conditions are given which guarantee that U is smooth.

Here is the analogue of Theorem 5.1.

THEOREM 6.1 (Markov controls). $u^* \in \mathscr{M}$ *is optimal if and only if there exists a constant J_M and measurable functions $\eta : [0, 1] \times R^n \to R$ and $\xi : [0, 1] \times R^n \to R^n$ satisfying:*

(i) $\displaystyle\int_0^1 |\xi(t, z_t)|^2\, dt < \infty$ *a.s.*, $\quad E \displaystyle\int_0^1 \xi(t, z_t)\, dz_t = 0,$

(ii) $\chi(1) = 0$ *a.s.*,

where

$$\chi(t) = J_M + \int_0^t \eta(s, z_s)\, ds + \int_0^t \xi(s, z_s)\, dz_s,$$

(iii) $\eta(t, z_t) + \xi(t, z_t)g(t, z_t, u[t, z_t]) + c(t, z_t, u[t, z_t]) \geqq 0$ *a.s.*,

$\eta(t, z_t) + \xi(t, z_t)g(t, z_t, u^*[t, z_t]) + c(t, z_t, u^*[t, z_t]) = 0$ *a.s.*

Then $\chi(t) = U_t$ *a.s. and* $J_M = J(u^*)$, *the cost of the optimal Markov policy.*

Proof. The proof is as for Theorem 4.2 by use of Lemma 6.2.

Notice that since $u(t, z_t)$ can take any value in Ξ, and in the restriction of Wiener measure to $\mathscr{B}_t$ is absolutely continuous with respect to Lebesgue measure, (iii) is equivalent to

$$\eta(t,x) + \min_{v\in\Xi}\{\xi(t,x)g(t,x,v) + c(t,x,v)\} = 0 \tag{6.17}$$

for all $(t, x) \in [0, 1] \times R^n$, and the optimal policy u^* is characterized by the property that $[U_x(t, x)g(t, x, v) + c(t, x, v)]$ is minimized by $v = u^*(t, x)$.

THEOREM 6.2. *For the system considered in this section* (*i.e., satisfying* (6.1)),

$$\inf_{u\in\mathscr{M}} J(u) = \inf_{u\in\mathscr{N}} J(u),$$

where $\mathscr{N}$ *is the class of nonanticipative controls.*

Proof. From Theorem 6.1 and (6.17),

$$\Lambda U(t,x) + U_x(t,x)g(t,x,v) + c(t,x,v) \geqq 0 \tag{6.18}$$

for all $(t, x, v) \in [0, 1] \times R^n \times \Xi$. Let $u \in \mathscr{N}$. Then the process $\{w_t\}$ defined by

$$dw_t = \sigma_t^{-1}(-g^{(u)}\,dt + dz_t)$$

is a Brownian motion under P_u and

$$U_t = J_M + \int_0^t (\Lambda U_s + U_x g^{(u)})\,ds + \int_0^t U_x\sigma\,dw.$$

Now $U(1) = 0$ a.s., so taking expectations at $t = 1$ gives

$$\begin{aligned} J_M &= E_u\int_0^1 (-\Lambda U - U_x g^{(u)})\,ds \\ &\leqq E_u\int_0^1 c^{(u)}\,ds \quad \text{(from (6.18))} \\ &= J(u). \end{aligned}$$

Since u was arbitrary,

$$J_M \leqq \inf_{u\in\mathscr{N}} J(u).$$

The reverse inequality is immediate from the inclusion $\mathscr{M} \subset \mathscr{N}$.

7. A note on two-person, zero-sum stochastic differential games. Stochastic differential games—control problems where there are several controllers with conflicting objectives—can also be treated by methods based at least implicitly on dynamic programming. For instance, Friedman in [22] has developed a theory using partial differential equations analogous to that of Fleming [9] for the optimal control problem. The "Girsanov" method of this paper can also be applied. The intention here is not to provide an exhaustive account but merely to indicate one or two of the possibilities; in particular, attention is restricted to two-person zero-sum games where complete information is available to both players. The method was first applied to games of this type by Varaiya [8], [21]. See Theorem 7.1 below.

The game (G) is defined as follows. The system dynamics are represented by

$$dz_t = g(t, z, u, v)\, dt + \sigma(t, z)\, dw_t,$$

where g and σ satisfy (2.5) with the obvious modifications. The control *strategies* u and v take values in $\Xi_1 \subset R^{l_1}$ and $\Xi_2 \subset R^{l_2}$ respectively and satisfy (2.6) with $\mathscr{Y}_t = \mathscr{F}_t$ (complete observations). The measure P_{uv} is defined for any admissible strategy (u, v) by

$$\frac{dP_{uv}}{dP} = \rho_0^1(uv) = \exp\left[\zeta_0^1(g^{(uv)})\right],$$

where

$$g^{(uv)}(t, z) = g(t, z, u[t, z], v[, z]).$$

The *payoff* is

$$J(u, v) = E_{uv}\left[\int_0^1 c_s^{(uv)}\, ds\right].$$

Here E_{uv} denotes expectation with respect to P_{uv} and $c_s^{(uv)} = c(s, z, u[s, z], v[s, z])$ is a bounded function satisfying conditions similar to those satisfied previously by the cost function. Player I (control u) is attempting to minimize the payoff while player II (control v) wants to maximize it. The game has a *saddle point* if there is a pair of strategies (the *equilibrium strategies*) (u^*, v^*) such that for all admissible (u, v),

$$J(u^*, v) \leqq J(u^*, v^*) \leqq J(u, v^*).$$

Assumption. There exist equilibrium strategies (u^*, v^*) for the game (G).

In [8], [21] it is shown that a saddle point does in fact exist under certain conditions.

THEOREM 7.1. *Suppose*:

(i) $\sigma = I$ (*the identity matrix*);

(ii) g *has the form*

$$g(t, z, u, v) = \begin{bmatrix} g_1(t, z, u) \\ g_2(t, z, v) \end{bmatrix};$$

(iii) *for fixed* (t, z), $g_1(t, z, \cdot)$ *and* $g_2(t, z, \cdot)$ *are continuous on* Ξ_1, Ξ_2 *respectively.*

(iv) $g_1(t, z, \Xi_1)$ *and* $g_2(t, z, \Xi_2)$ *are closed and convex for each* (t, z).

Then game (G) *has a saddle point.*

Let (u^*, v^*) be an equilibrium strategy for (G) and let $P^* = P_{u^*v^*}$, $E^* = E_{u^*v^*}$. For any admissible strategy, define the process ψ_t^{uv} by

$$\psi_t^{uv} = E_{uv}\left[\int_t^1 c_s^{(uv)}\, ds | \mathscr{F}_t\right].$$

Let

$$\phi_t = \psi_t^{u^*v^*}. \tag{7.1}$$

LEMMA 7.1. *For each* $t \in [0,1]$ *and* $h > 0$,

$$
\begin{aligned}
(7.2)\qquad & E_{u^*v}\left[\int_t^{t+h} c_s^{u^*v}\, ds \mid \mathscr{F}_t\right] + E_{u^*v}[\phi_{t+h} \mid \mathscr{F}_t] \leqq \phi_t \\
& \leqq E_{uv^*}\left[\int_t^{t+h} c_s^{uv^*}\, ds \mid \mathscr{F}_t\right] + E_{uv^*}[\phi_{t+h} \mid \mathscr{F}_t] \quad a.s.
\end{aligned}
$$

Proof. Suppose there is a strategy v for player II such that for some t, h,

$$
\phi_t < E_{u^*v}\left[\int_t^{t+h} c_s^{u^*v}\, ds \mid \mathscr{F}_t\right] + E_{u^*v}[\phi_{t+h} \mid \mathscr{F}_t]
$$

for $z \in M \subset \mathscr{F}_t$, $PM > 0$. Define the strategy v' for player II by

$$
v' = \begin{cases} v, & t \in [t, t+h], \quad z \in M, \\ v^*, & \text{elsewhere}. \end{cases}
$$

Then

$$
\begin{aligned}
(7.3)\qquad J(u^*, v') - J(u^*, v^*) &= E_{u^*v'}\left(I_M \int_t^{t+h} c_s^{u^*v'}\, ds\right) \\
&\quad + E_{u^*v'}\left(I_M \int_{t+h}^{1} c_s^{u^*v'}\, ds\right) - E^*\left(I_M \int_t^{1} c_s^{*}\, ds\right),
\end{aligned}
$$

where I_M is the indicator function of M. Now,

$$
\begin{aligned}
E_{u^*v'}\left(I_M \int_t^{t+h} c_s^{u^*v'}\, ds\right) &= E^*\left(I_M E\left[\rho_t^{t+h}(u^*v) \int_t^{t+h} c_s^{u^*v}\, ds \mid \mathscr{F}_t\right]\right) \\
&= E^*\left(I_M E_{u^*v}\left[\int_t^{t+h} c_s^{u^*v}\, ds \mid \mathscr{F}_t\right]\right) \\
&> \phi_t I_M - E^*(I_M E_{u^*v}[\phi_{t+h} \mid \mathscr{F}_t]) \\
&= \phi_t I_M - E^*(E[\rho_t^{t+h}(u^*v) I_M \phi_{t+h} \mid \mathscr{F}_t]) \\
(7.4)\qquad &= \phi_t I_M - E_{u^*v'}(I_M \phi_{t+h}).
\end{aligned}
$$

From (7.3) and (7.4),

$$
J(u^*, v') > J(u^*, v^*).
$$

So PM must be zero. The other inequality in (7.2) is proved similarly.

LEMMA 7.2. *There exist processes* $\Lambda\phi$, $\nabla\phi$ *such that*

$$
\phi_t = J^* + \int_0^t \Lambda\phi_s\, ds + \int_0^t \nabla\phi_s\, dz_s.
$$

Proof. It is easy to check that

$$\phi_t = E^*\left[\int_0^1 c_s^*\, ds | \mathscr{F}_t\right] - \int_0^t c_s^*\, ds.$$

Under measure P^* the innovations process of z is $dw = \sigma^{-1}(dz - g^*\, dt)$. Hence from Theorem 2.3 there is process $\{\gamma_t\}$ such that

$$E^*\left[\int_0^1 c_s^*\, ds | \mathscr{F}_t\right] = \int_0^t \gamma_s \sigma_s^{-1}(dz_s - g_s^*\, ds).$$

The result follows after defining

$$\nabla\phi = \gamma\sigma^{-1}, \qquad \Lambda\phi = c^* - \nabla\phi g^*.$$

THEOREM 7.2. *(u^*, v^*) is an equilibrium strategy if and only if there exist processes $\{\eta_t\}$, $\{\xi_t\}$ adapted to $\mathscr{F}_t$, and a constant J^* such that*:

(i) $\int_0^1 |\xi_t|^2\, dt < \infty$ *a.s. and* $E\int_0^1 \xi_t\, dz_t = 0$,

(ii) $\chi(1) = 0$ *a.s.*,

where

$$\chi(t) = J^* + \int_0^t \eta_s\, ds + \int_0^t \xi_s\, dz_s,$$

(iii) $$\eta_t + \min_u (g^{(uv^*)}\xi + c^{(uv^*)}) = \eta_t + (g^{(u^*v^*)}\xi + c^{(u^*v^*)})$$
$$= \eta_t + \max_v (g^{(u^*v)}\xi + c^{(u^*v)}) = 0.$$

Then $\chi_t = \phi_t$ a.s. for each t, and J^* is the value of the game.

Proof. Sufficiency is proved as in the proof of Theorem 4.2. Necessity is established by showing that $\eta_t = \Lambda\phi_t$ and $\xi_t = \nabla\phi_t$ satisfy (i)–(iii). Fixing $v = v^*$ and using precisely the methods of Theorem 4.2 together with Lemma 7.1 gives the result with the left-hand side of (iii), while fixing $u = u^*$ similarly gives the right-hand side. This completes the proof.

For $p \in R^n$, define

$$H(t, x, u, v, p) = p'g(t, x, u, v) + c(t, x, u, v).$$

Then, from (iii) above, we have

$$\min_u \max_v H(t, x, u, v, p) = \max_v \min_u H(t, x, u, v, p) \tag{7.5}$$

for all $(t, x, p) \in [0, 1] \times C \times R^n$.

The equality (iii) in Theorem 7.2 is a version of Isaacs' equation (the game equivalent of the Hamilton–Jacobi equation). The partial differential equation counterpart of this for the Markov (pure strategies) case was derived by Friedman in [22], and a solution shown to exist under certain conditions; notably, under the assumption that (7.5) is satisfied.

REFERENCES

[1] J. L. Doob, *Stochastic Processes*, John Wiley, New York, 1953.

[2] N. Dunford and J. T. Schwartz, *Linear Operators, Part I*, Interscience, New York, 1958.

[3] M. Loève, *Probability Theory*, 3rd ed., Van Nostrand, Princeton, N.J., 1963.

[4] P. A. Meyer, *Probability and Potentials*, Blaisdell, Waltham, Mass., 1966.

[5] A. V. Skorokhod, *Studies in the Theory of Random Processes*, Addison-Wesley, Reading, Mass., 1965.

[6] J. M. C. Clark, *The representation of functionals of Brownian motion by stochastic integrals*, Ann. Math. Statist., 41 (1970), pp. 1282–1295.

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

[8] T. E. Duncan and P. P. Varaiya, *On the solutions of a stochastic control system*, this Journal, 9 (1971), pp. 354–371.

[9] W. H. Fleming, *Optimal continuous-parameter stochastic control*, SIAM Rev., 11 (1969), pp. 470–509.

[10] M. Fujisaki, G. Kallianpur and H. Kunita, *Stochastic differential equations for the non-linear filtering problem*, Osaka J. Math., 9 (1972), pp. 19–40.

[11] 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.

[12] H. Kunita and S. Watanabe, *On square integrable martingales*, Nagoya Math. J., 30 (1967), pp. 209–245.

[13] R. Rishel, *Necessary and sufficient dynamic programming conditions for continuous-time stochastic optimal control*, this Journal, 8 (1970), pp. 559–571.

[14] H. S. Witsenhausen, *A counterexample in stochastic optimum control*, this Journal, 6 (1968), pp. 131–147.

[15] E. Wong, *Representation of martingales, quadratic variation and applications*, this Journal, 9 (1971), pp. 621–633.

[16] W. H. Fleming and M. Nisio, *On the existence of optimal stochastic controls*, J. Math. Mech., 15 (1966), pp. 777–794.

[17] W. H. Fleming, *Duality and a priori estimates in Markovian optimization problems*, J. Math. Anal. Appl., 16 (1966), pp. 254–279.

[18] W. M. Wonham, *On the separation theorem of stochastic control*, this Journal, 6 (1968), pp. 312–326.

[19] M. H. A. Davis and P. P. Varaiya, *Information states for linear stochastic systems*, J. Math. Anal. Appl., 37 (1972), pp. 384–402.

[20] H. J. Kushner, *On the stochastic maximum principle—fixed time of control*, J. Math. Anal. Appl., 11 (1965), pp. 78–92.

[21] P. P. Varaiya, *Differential games*, Proc. 6th Berkeley Symp. on Mathematical Statistics and Probability, to appear.

[22] A. Friedman, *Stochastic Differential Games*, J. Differential Equations, 11 (1972), pp. 79–108.

[23] D. W. Stroock and S. R. S. Varadhan, *Diffusion processes with continuous coefficients, II*, Comm. Pure Appl. Math., 22 (1969), pp. 479–530.

SIAM J. CONTROL
Vol. 11, No. 2, May 1973

THE STABILIZING SOLUTION OF THE ALGEBRAIC RICCATI EQUATION*

B. P. MOLINARI†

Abstract. This paper investigates solutions $\hat{X}$ of the algebraic Riccati equation $F'X + XF - XGG'X + Q = 0$ with the property $\operatorname{Re} \lambda(F - GG'\hat{X}) \leqq 0$. The uniqueness and existence of such a solution is completely characterized. In addition, the conditions are strengthened to characterize the important stabilizing solution ($\operatorname{Re} \lambda(F - GG'\hat{X}) < 0$). All standard existence results are recovered as special cases.

1. Introduction. This paper is concerned with the algebraic Riccati equation

$$F'X + XF - XGG'X + Q = 0, \tag{ARE}$$

where F and G are $n \times n$ and $n \times m$ real matrices respectively, and Q is an $n \times n$ real symmetric matrix. No a priori sign-definite condition is imposed on Q.

It is a standard property that solutions X of the ARE satisfy

$$\det(Is + F - XGG')\det(Is - F + GG'X) = \det(Is - M),$$

where M is the associated Hamiltonian matrix

$$M = \begin{bmatrix} F & GG' \\ Q & -F' \end{bmatrix}. \tag{1.1}$$

Hence it is meaningful to seek solutions of the ARE subject to eigenvalue restrictions on $F - GG'X$. This paper establishes existence and uniqueness properties of solutions $\hat{X}$ and $\check{X}$ satisfying

$$\operatorname{Re} \lambda(F - GG'\hat{X}) \leqq 0, \tag{1.2}$$

$$\operatorname{Re} \lambda(F - GG'\check{X}) \geqq 0. \tag{1.3}$$

It also clearly establishes when these conditions can be sharpened to

$$\operatorname{Re} \lambda(F - GG'\hat{X}) < 0, \tag{1.4}$$

$$\operatorname{Re} \lambda(F - GG'\check{X}) > 0. \tag{1.5}$$

Such a solution $\hat{X}$ is termed a stabilizing solution of the ARE.

These results have direct application. The ARE arises in least squares problems defined on stationary linear systems, namely:

(i) the regulator problem [1], [3], [17];
(ii) the dual filtering problem [4];
(iii) the stability theory of feedback systems [16].

In these applications the solution of the ARE involved is either $\hat{X}$ or $\check{X}$. Moreover, Willems [17] has recently shown that all real symmetric solutions of the ARE can be characterized in terms of $\hat{X}$ and $\check{X}$.

In a companion paper [12], the author has established a one-to-one equivalence between the solutions $\hat{X}$ and a spectral factorization of a related real rational

* Received by the editors August 3, 1971, and in revised form June 5, 1972.

† School of Electrical Engineering, University of New South Wales, Kensington, Australia 2033.

matrix. This paper uses this equivalence to transcribe well-known existence results for spectral factorization [19] into basic existence results for $\hat{X}$. Elementary transformations and manipulations extend these results into various necessary and sufficient conditions for the existence of $\hat{X}$ and $\check{X}$. All standard (sufficient) conditions for the existence of $\hat{X}$ are easily recovered.

Since the first draft of this paper, Willems [17] has obtained necessary and sufficient existence results for $\hat{X}$ and $\check{X}$ by a different method. This paper provides a wider range of equivalent conditions, several of which are easier to interpret.

2. Equivalence results. A companion paper [12] relates solutions $\hat{X}$ of the ARE to certain factorizations of the real rational matrix

$$\Phi_K(s) = Z'(-s)\Phi(s)Z(s), \tag{2.1}$$

where

$$\Phi(s) = I + G'(-Is - F')^{-1}Q(Is - F)^{-1}G, \tag{2.2}$$

$$Z(s) = I - K(Is - F + GK)^{-1}G, \tag{2.3}$$

$$K \text{ is a real } m \times n \text{ matrix}. \tag{2.4}$$

The matrix $\Phi(s)$ is related to M by

$$\det[\Phi(s)] = \frac{\det(Is - M)}{\det(Is + F)\det(Is - F)}.$$

Multiplication by $Z'(-s)$ and $Z(s)$ merely serves to cancel the poles of $\Phi(s)$ and to replace them by poles in a more acceptable position. This is indicated by

$$\det[\Phi_K(s)] = \frac{\det(Is - M)}{\det(Is + F - GK)\det(Is - F + GK)}.$$

These relations are established in [12].

Remark 1. The matrices $\Phi(s)$ and $\Phi_K(s)$ can be visualized with reference to the integral

$$J = \int_0^\infty x'Qx + u'u\,dt,$$

defined on trajectories of

$$\dot{x} = Fx + Gu, \qquad x(0) = 0.$$

Should J exist, then

$$J = \frac{1}{2\pi}\int_{-\infty}^{\infty} U'(-j\omega)\Phi(j\omega)U(j\omega)\,d\omega,$$

where $U(j\omega)$ is the Fourier transform of $u(t)$ [2, Chap. 5.6]. On the other hand, define $v = u + Kx$. Then

$$\dot{x} = (F - GK)x + Gv, \qquad x(0) = 0,$$

$$U(j\omega) = Z(j\omega)V(j\omega),$$

$$J = \frac{1}{2\pi}\int_{-\infty}^{\infty} V'(-j\omega)\Phi_K(j\omega)V(j\omega)\,d\omega.$$

The factorization of $\Phi_K(s)$ appropriate to this paper is given by the following.

DEFINITION 1. A matrix $\Delta(s)$ is a *spectral factorization of* $\Phi_K(s)$ if and only if $\Delta(s)$ is an $m \times m$ real rational matrix satisfying:

(a) $\Delta'(-s)\Delta(s) = \Phi_K(s)$;

(b) $\Delta(s)$ is analytic in $\operatorname{Re} s \geqq 0$;

(c) $\Delta^{-1}(s)$ is analytic in $\operatorname{Re} s > 0$;

(d) $\lim_{s \to \infty} \Delta(s) = I$.

This is a simple normalization of the standard spectral factorization of Youla [19].

The equivalence result used in this paper is a slight extension of Theorem 4 of [12] and may be stated as the following theorem.

THEOREM 1. *Assume that* (F, G) *is controllable and that* $\operatorname{Re} \lambda(F - GK) < 0$. *Then* $\Delta(s)$ *is a spectral factorization of* $\Phi_K(s)$ *if and only if*

(i) $\Delta(s)$ *admits a realization*

$$\Delta(s) = I + (G'\hat{X} - K)(Is - F + GK)^{-1}G$$

for some real symmetric $\hat{X}$, *and*

(ii) $\hat{X}$ *is a solution of the* ARE *satisfying* $\operatorname{Re} \lambda(F - GG'\hat{X}) \leqq 0$.

Further, the relation is one-to-one. Finally, $\Delta^{-1}(s)$ *is analytic for* $\operatorname{Re} s \geqq 0$ *if and only if* $\operatorname{Re} \lambda(F - GG'\hat{X}) < 0$.

Proof. Theorem 4 of [12] cannot be applied directly to $\Phi_K(s)$, since it may not be of full degree $2n$. However, a simple transformation gives a rational matrix to which the theorem is applicable. Consider the $m \times m$ nonsingular matrix $Z_L(s)$ defined on a real $m \times n$ matrix L by

$$Z_L(s) = I - (L - K)(Is - F + GL)^{-1}G,$$

where $\operatorname{Re} \lambda(F - GL) < 0$, and $(F - GL)$ and M have no common eigenvalues. Such matrices L exist since (F, G) is controllable [14, Thm. 5.4.2]. Now

$$Z(s)Z_L(s) = I - L(Is - F + GL)^{-1}G$$

and

$$Z_L'(-s)\Phi_K(s)Z_L(s) = \Phi_L(s),$$

where $\Phi_L(s)$ is defined on L exactly as $\Phi_K(s)$ is defined on K. Via Definition 1 and the definition of $Z_L(s)$ it follows that:

(i) Δ is a spectral factorization of $\Phi_K(s)$ if and only if ΔZ_L is a spectral factorization of $\Phi_L(s)$;

(ii) the relation between Δ and ΔZ_L is one-to-one;

(iii) Δ^{-1} is analytic in $\operatorname{Re} s \geqq 0$ if and only if $(\Delta Z_L)^{-1}$ is analytic in $\operatorname{Re} s \geqq 0$.

Now, Theorem 4 of [12] characterizes the solution $\hat{X}$ as claimed in Theorem 1, but in terms of spectral factorizations ΔZ_L of $\Phi_L(s)$, via the realization

$$\Delta Z_L = I - (G'X - L)(Is - F + GL)^{-1}G.$$

Theorem 2 follows by the equivalences just stated and the fact that the last realization can be written

$$\Delta Z_L = [I - (G'\hat{X} - K)(Is - F + GK)^{-1}G]Z_L(s).$$

Note that it is not claimed that the realization for $\Delta(s)$ is minimal. Clearly, it must be nonminimal if $\delta[\Phi_K] < 2n$.

3. Existence results. This section first uses the well-known theory [19] to give the following.

THEOREM 2. *Assume* $\operatorname{Re}\lambda(F - GK) < 0$. *Then* $\Phi_K(j\omega) \geqq 0$ *if and only if there exists a spectral factorization* $\Delta(s)$ *of* $\Phi_K(s)$. *Further*, $\Delta(s)$ *is unique. Finally*, $\Phi_K(j\omega) > 0$ *if and only if* $\Delta^{-1}(s)$ *is analytic in* $\operatorname{Re} s \geqq 0$.

Proof. First note that the assumption implies that poles p of $\Phi_K(s)$ satisfy $\operatorname{Re} p \neq 0$. Now given $\Phi_K(j\omega) \geqq 0$, the classic result of Youla [19] establishes a real rational matrix $H(s)$ satisfying conditions (a)–(c) of Definition 1. Moreover, from (a) of Definition 1,

$$\sum_{k=1}^{m} |H_{ik}(j\omega)|^2 = \Phi_{K_{ii}}(j\omega),$$

and it follows that $\lim_{\omega\to\infty} H_{ik}(j\omega)$ exists for all i, k. Since $H(s)$ is rational it follows that

$$\lim_{s\to\infty} H(s) = N,$$

and by (a) of Definition 1, N is an orthogonal matrix. It is easily checked that $\Delta = N'H(s)$ is a spectral factorization of $\Phi_K(s)$ according to Definition 1. That the existence of $\Delta(s)$ implies $\Phi_K(j\omega) \geqq 0$ is trivial from

$$\Delta(j\omega) * \Delta(j\omega) = \Phi_K(j\omega).$$

Assume two spectral factorizations Δ_1 and Δ_2. Then by Youla's result,

$$\Delta_1(s) = N\Delta_2(s)$$

for some constant orthogonal matrix N. Taking limits $s \to \infty$ shows that $N = I$, and that the factorization is unique. Finally consider $\Phi_K(j\omega) > 0$. From condition (a) of Definition 1 it is clear that $\det \Delta(j\omega) \neq 0$ for all ω. This plus condition (b) of Definition 1 implies that $\Delta^{-1}(s)$ has no poles on the imaginary axis $s = j\omega$, and hence is analytic for $\operatorname{Re} s \geqq 0$. The converse is easy and completes the proof.

The result immediately allows an existence and uniqueness statement to be given for the required solutions of the ARE.

THEOREM 3. *Assume that* (F, G) *is controllable. Then the following conditions are equivalent*:

(A.1) *the* ARE *has a real symmetric solution* $\hat{X}$ *satisfying* $\operatorname{Re}\lambda(F - GG'\hat{X}) \leqq 0$;

(A.2) *the* ARE *has a real symmetric solution* $\check{X}$ *satisfying* $\operatorname{Re}\lambda(F - GG'\check{X}) \geqq 0$;

(A.3) $\Phi_K(j\omega) \geqq 0$, *for any* K *satisfying* $\operatorname{Re}\lambda(F - GK) < 0$;

(A.4) $\Phi(j\omega) \geqq 0$.

Moreover, $\hat{X}$ *and* $\check{X}$ *are unique in the set of all real symmetric solutions of the* ARE.

Proof. (A.1) ↔ (A.3). Merely combine Theorems 1 and 2, using the intermediate notion of spectral factorization.

(A.1) → (A.2). Consider $Y = -X$. The ARE becomes

$$\text{(ARE}') \qquad (-F)'Y + Y(-F) - YGG'Y + Q = 0,$$

where $(-F, G)$ is controllable. This implies a real symmetric solution $\hat{Y}$ of the ARE′ satisfying $\operatorname{Re}\lambda(-F - GG'\hat{Y}) \leqq 0$. That is, $\operatorname{Re}\lambda(F + GG'\hat{Y}) \geqq 0$. Denoting $\hat{Y} = -\check{X}$ gives (A.2).

(A.1) ← (A.2). This is exactly analogous to the last step.

(A.3) ↔ (A.4). This is immediate from (2.1).

Finally $\hat{X}$ is unique in the set of all real symmetric solutions of the ARE since it has a one-to-one relation to the unique spectral factorization. The solution $\check{X}$ is unique by the equivalence (A.1) ↔ (A.2).

The equivalence (A.1)–(A.2)–(A.4) has recently been given by Willems [17], using a different method. The proof uses a novel representation of Q to separate the ARE into two special ARE equations, both of which are amenable to standard theory [3, Thms. 23.4 and 25.2]. It is fair to point out that the proof of these underlying results is not trivial.

Finally, it is of particular interest to the applications of $\hat{X}$ and $\check{X}$ to sharpen the eigenvalue conditions to strict inequalities. This is done in the following theorem.

THEOREM 4. *Assume that* (F, G) *is controllable. Then following conditions are equivalent*:

(B.1) *the* ARE *has a real symmetric solution* $\hat{X}$ *satisfying* $\operatorname{Re} \lambda(F - GG'\hat{X}) < 0$;

(B.2) *the* ARE *has a real symmetric solution* $\check{X}$ *satisfying* $\operatorname{Re} \lambda(F - GG'\check{X} > 0$;

(B.3) $\Phi_K(j\omega) > 0$, *for any* K *satisfying* $\operatorname{Re} \lambda(F - GK) < 0$;

(B.4) $\Phi(j\omega) > \varepsilon G'(-Ij\omega - F')^{-1}(Ij\omega - F)^{-1}G$ *for some* $\varepsilon > 0$;

(B.5) $\operatorname{Re} \lambda(M) \neq 0$.

Moreover, $\hat{X}$ *and* $\check{X}$ *are unique in the set of all solutions of the* ARE.

Proof. (B.1) ↔ (B.3). Again by Theorems 1 and 2, through the intermediate concept of spectral factorization.

(B.1) ↔ (B.2). This follows exactly as before.

(B.3) → (B.4). Since $(Ij\omega - F + GK)^{-1}G$ is bounded for all ω, it follows that

$$\Phi_K(j\omega) > \varepsilon G'(-Ij\omega - F' + K'G')^{-1}(Ij\omega - F + GK)^{-1}G, \quad \text{for some } \varepsilon > 0.$$

Condition (B.4) then follows from (2.1), from the identity

$$(Is - F + GK)^{-1}G \cdot Z^{-1}(s) = (Is - F)^{-1}G,$$

and from the fact that $\det [Z^{-1}(j\omega)] \neq 0$.

(B.3) ← (B.4). Reversing the last step shows that (B.4) implies

$$\Phi_K(j\omega) \geqq \varepsilon G'(-Ij\omega - F' + K'G')^{-1}(Ij\omega - F + GK)^{-1}G.$$

The inequality $\geqq$ is applicable since it is not guaranteed that $\det [Z(j\omega)] \neq 0$ for all ω.

Consider $\alpha^*\Phi_K(j\omega)\alpha$, for $\alpha \neq 0$. Two cases are possible:

(a) $G\alpha \neq 0$. Then $(Ij\omega - F + GK)^{-1}G\alpha \neq 0$, and the above inequality implies that $\alpha^*\Phi_K(j\omega)\alpha > 0$.

(b) $G\alpha = 0$. Then it is easily checked that $Z(s)\alpha = \alpha$, and that $\alpha^*\Phi_K(j\omega)\alpha = \alpha^*\alpha > 0$.

In other words, $\Phi_K(j\omega)$ is positive definite for all ω.

(B.3) ↔ (B.5). This is shown independently in the following lemma. Finally, uniqueness is shown by assuming two stabilizing solutions X_1 and X_2 (not necessarily real symmetric). These both satisfy the ARE, which by rearrangement may be written

$$(F' - X_2GG')X_1 + X_1(F - GG'X_1) + X_2GG'X_1 + Q = 0,$$

$$(F' - X_2GG')X_2 + X_2(F - GG'X_1) + X_2GG'X_1 + Q = 0.$$

Subtraction gives

$$(F' - X_2GG')(X_1 - X_2) + (X_1 - X_2)(F - GG'X_1) = 0.$$

But

$$\det(Is + F' - X_2GG')\det(Is - F + GG'X_2) = \det(Is - M) = \det(-Is - M).$$

Hence $\operatorname{Re}\lambda(-F' + X_2GG') > 0$ (since the other n eigenvalues of M satisfy $\operatorname{Re}\lambda < 0$). By standard theory [7, Chap. VIII] this implies that $X_1 - X_2 = 0$ is the unique solution of the matrix linear equation, and the stabilizing solution is unique.

It remains to establish the following lemma.

LEMMA 1. *Assume* $\operatorname{Re}\lambda(F - GK) < 0$. *Then* $\Phi_K(j\omega) > 0$ *if and only if* $\operatorname{Re}\lambda(M) \neq 0$.

Proof. The matrices $\Phi_K(s)$ and M are related by

$$\det\Phi_K(s) = \frac{\det(Is - M)}{\det(Is + F - GK)\det(Is - F + GK)}.$$

See, for example, [12, Lemma 3]. Any factors cancelled on the right-hand side are certainly not on the imaginary axis. Hence $\Phi_K(j\omega) > 0$ immediately implies $\operatorname{Re}\lambda(M) \neq 0$. Now denote the minimum eigenvalue (necessarily real) of the Hermitian matrix $\Phi_K(j\omega)$ by $\lambda_{\min}(\omega)$. By standard theory $\lambda_{\min}$ is a continuous function of the components of $\Phi_K(j\omega)$ [15, Thm. 4.6], which in turn are continuous functions of ω. Hence $\lambda_{\min}$ is a continuous function of ω. Now assume

$$\lambda_{\min}(\omega_1) \leqq 0, \quad \text{for some } \omega_1.$$

Since $\lambda_{\min} \to 1$ as $\omega \to \infty$, it follows from the intermediate value theorem of elementary calculus [6, p. 66] that

$$\lambda_{\min}(\omega_2) = 0 \quad \text{for some} \quad \omega_1 \leqq \omega_2 < \infty.$$

Hence M has an eigenvalue $j\omega_2$. In other words $\operatorname{Re}\lambda(M) \neq 0$ implies that $\Phi_K(j\omega) > 0$, and the proof is complete.

The equivalence (B.1)–(B.2)–(B.4) has been given by Willems [17], by considering the solutions guaranteed by Theorem 3 and investigating the range of the inequality $\hat{X} > \check{X}$. The additional conditions given in Theorem 4 are interesting in several respects. First, the sharpening of condition (A.3) to (B.3) is somewhat neater than the relation of (A.4) to (B.4). Condition (B.5) is of considerable importance since, although it is a little hard to visualize, it provides by far the most direct test in terms of the data F, G and Q. Finally, it can be noted that the equivalence (B.1)–(B.5) has long been an article of faith in linear regulator theory [9], [10], but to the author's knowledge a proof has never appeared in the literature.

Remark 2. Using the methods of this paper, it does not seem easy to provide a condition (A.5) corresponding to (B.5). A likely candidate is that the elementary divisors of M on the imaginary axis are of even multiplicity.

4. Extended existence result. For the stabilizing solution the controllability condition can be weakened to give the following complete characterization.

THEOREM 5. *The* ARE *has a real symmetric solution* $\hat{X}$ *satisfying* $\operatorname{Re}\lambda(F - GG'\hat{X}) < 0$ *if and only if*:

(i) (F, G) *is stabilizable*;

(ii) $\operatorname{Re}\lambda(M) \neq 0$.

Proof. The necessary conditions are immediate. That the conditions are also sufficient is shown via Theorem 4. Now if (F, G) is stabilizable [18] there exists a nonsingular real matrix T such that [5, p. 276]

$$T^{-1}F = \begin{bmatrix} F_{11} & F_{12} \\ 0 & F_{22} \end{bmatrix}, \qquad T^{-1}G = \begin{bmatrix} G_1 \\ 0 \end{bmatrix},$$

where (F_{11}, G_1) is controllable and $\operatorname{Re}\lambda(F_{22}) < 0$. Defining

$$T'XT = \begin{bmatrix} Y_{11} & Y_{12} \\ Y_{21} & Y_{22} \end{bmatrix}, \qquad T'QT = \begin{bmatrix} R_{11} & R_{12} \\ R_{21} & R_{22} \end{bmatrix},$$

the ARE can be transformed and partitioned into

$$F'_{11}Y_{11} + Y_{11}F_{11} - Y_{11}G_1G'_1Y_{11} + R_{11} = 0,$$

$$(F_{11} - G_1G'_1Y'_{11})'Y_{12} + Y_{12}F_{22} + (R_{12} + Y_{11}F_{12}) = 0,$$

$$F'_{22}Y_{21} + Y_{21}(F_{11} - G_1G'_1Y_{11}) + (R_{21} + F'_{12}Y_{11}) = 0,$$

$$F'_{22}Y_{22} + Y_{22}F_{22} + (R_{22} - Y_{21}G_1G'_1Y_{12} + F'_{12}Y_{12} + Y_{21}F_{12}) = 0.$$

These equations can be solved sequentially for Y_{11}, Y_{12}, Y_{21} and Y_{22}. Now, by Theorem 4, the reduced algebraic Riccati equation has a real symmetric solution $\hat{Y}_{11}$ satisfying $\operatorname{Re}\lambda(F_{11} - G_1G'_1\hat{Y}_{11}) < 0$. This follows since the reduced Hamiltonian matrix

$$M_{11} = \begin{bmatrix} F_{11} & G_1G'_1 \\ R_{11} & -F'_{11} \end{bmatrix}$$

satisfies

$$\det(Is + F_{22})\det(Is - M_{11})\det(Is - F_{22}) = \det(Is - M),$$

and hence $\operatorname{Re}\lambda(M_{11}) \neq 0$. By the standard theory of matrix linear equations the remaining equations have unique solutions, denoted $\hat{Y}_{12}$, $\hat{Y}_{21}$ and $\hat{Y}_{22}$. Denoting the entire solution by $\hat{Y}$, it can be checked to be a real symmetric matrix. Consider the corresponding solution $\hat{X} = (T')^{-1}\hat{Y}T^{-1}$ of the ARE. Now

$$\begin{aligned} T^{-1}(F - GG'\hat{X})T &= T^{-1}FT - (T^{-1}G)(T^{-1}G)'\hat{Y} \\ &= \begin{bmatrix} F_{11} - G_1G'_1\hat{Y}_{11} & F_{12} - G_1G'_1\hat{Y}_{12} \\ 0 & F_{22} \end{bmatrix}. \end{aligned}$$

Hence $\operatorname{Re}\lambda(F - GG'\hat{X}) < 0$. In other words, a real symmetric stabilizing solution of the ARE has been constructed. Exactly as in Theorem 4 it can be shown to be unique in the set of all solutions of the ARE.

COROLLARY. *In Theorem* 5, *condition* (ii) *can be replaced by*
(ii′) $\Phi_K(j\omega) > 0$ *for any* K *satisfying* $\operatorname{Re}\lambda(F - GK) < 0$.

5. Special cases. This last section recovers from Theorem 5 various standard sufficient conditions for the existence of a stabilizing solution $\hat{X}$ of the ARE. These conditions are applicable for special cases of F, G and Q.

So far no conditions have been imposed on the square matrix F (apart from the implicit conditions of Theorem 5). A weak eigenvalue restriction allows the following.

THEOREM 6. *Assume that* $\operatorname{Re}\lambda(F) \neq 0$. *Then the* ARE *has a real symmetric stabilizing solution* $\hat{X}$ *if and only if*:
(i) (F, G) *is stabilizable*;
(ii) $\Phi(j\omega) > 0$.

Proof. Directly from the realization (2.3),

$$\det Z(s) = \frac{\det(Is - F)}{\det(Is - F + GK)}$$

and $Z(s)$ is nonsingular for $\operatorname{Re} s = 0$. Hence $\Phi(j\omega) > 0$ is equivalent to $\Phi_K(j\omega) > 0$ for any K satisfying $\operatorname{Re}\lambda(F - GK) < 0$. The result now follows from the corollary to Theorem 5.

COROLLARY. *Assume*
(i) $\operatorname{Re}\lambda(F) < 0$ *and* (F, G) *is controllable*;
(ii) $Q = -H'H$ *and* (H, F) *is observable*;
(iii) $\Phi(j\omega) > 0$.
Then the ARE *has a real symmetric stabilizing solution* $\hat{X}$.

Proof. The proof is immediate. For a standard proof, see [3, Thm. 25.2].

In the regulator problem of linear optimal control, the main interest is with the case $Q = H'H \geqq 0$. It is clear that this immediately guarantees condition (A.4) of Theorem 3, and hence the existence of a real symmetric solution $\hat{X}$ of the ARE satisfying $\operatorname{Re}\lambda(F - GG'\hat{X}) \leqq 0$. However, the sharpening of the result to stabilizing solutions of the ARE is not automatic.

The essential problem is to isolate the occurrence of imaginary eigenvalues in M. This is done via the following lemmas.

LEMMA 2. *The matrix pair* (F, G) *is controllable if and only if no nonzero vectors* ξ *satisfy*

$$\xi^*(I\lambda - F) = 0, \qquad \xi^*G = 0,$$

where, clearly, λ *is an eigenvalue of* F.

Proof. See [14, Thm. 2.6.2].

LEMMA 3. *Assume* $Q = H'H$. *Then* $\operatorname{Re}\lambda(M) \neq 0$ *if and only if all eigenvalues of* F *satisfying* $\operatorname{Re}\lambda = 0$ *are controllable modes of* (F, G) *and observable modes of* (H, F).

Proof. Assume that $j\omega$ is an eigenvalue of M. That is

$$\begin{bmatrix} M \end{bmatrix}\begin{bmatrix} \alpha \\ \beta \end{bmatrix} = j\omega\begin{bmatrix} \alpha \\ \beta \end{bmatrix}. \tag{5.1}$$

Multiplication gives

$$[\beta^* \quad \alpha^*]\begin{bmatrix} F & GG' \\ H'H & -F' \end{bmatrix}\begin{bmatrix}\alpha \\ \beta\end{bmatrix} = j\omega [\beta^* \quad \alpha^*]\begin{bmatrix}\alpha \\ \beta\end{bmatrix},$$

$$(\beta^*F\alpha - \alpha^*F'\beta) + (\beta^*GG'\beta + \alpha^*H'H\alpha) = j\omega(\beta^*\alpha + \alpha^*\beta).$$

Equating real parts gives

$$\beta^*GG'\beta + \alpha^*H'H\alpha = 0.$$

That is,

$$\beta^*G = 0, \qquad H\alpha = 0.$$

Further, substituting this into (5.1) allows

$$\beta^*(Ij\omega - F) = 0, \qquad (Ij\omega - F)\alpha = 0.$$

In other words, $\beta \neq 0$ implies that $j\omega$ is an uncontrollable mode of (F, G), and $\alpha \neq 0$ implies that $j\omega$ is an unobservable mode of (H, F). The algebra can be reversed, which establishes the lemma.

This allows the following.

THEOREM 7. *Assume that* $Q = H'H$. *Then the* ARE *has a real symmetric stabilizing solution* $\hat{X}$ *if and only if*:

(i) (F, G) *is stabilizable*;

(ii) *all eigenvalues of* F *satisfying* $\operatorname{Re}\lambda = 0$ *are observable modes of* (H, F).

Overspecifying the conditions (i) and (ii) gives the following.

COROLLARY 1. *Assume that* $Q = H'H$ *and* $\operatorname{Re}\lambda(F) \neq 0$. *Then the* ARE *has a real symmetric stabilizing solution* $\hat{X}$ *if and only if* (F, G) *is stabilizable.*

COROLLARY 2. *Assume that*

(i) (F, G) *is stabilizable*;

(ii) $Q = H'H$, *where* (H, F) *is detectable.*

Then the ARE *has a real symmetric stabilizing solution* $\hat{X}$.

Proof. For alternative proofs, see Potter [13] and Wonham [18].

Of course, these last conditions can be further strengthened to controllability and observability respectively, to give the classic result of Kalman [8].

Finally, it is of interest to repeat a standard link between the stabilizing solution $\hat{X}$ (which is unique) and nonnegative solutions of the ARE (which are not necessarily unique [11]).

LEMMA 4. *Assume that* $Q = H'H$ *and consider* X *a real symmetric solution of the* ARE.

(i) *If* (H, F) *is observable, then* $X > 0$ *if and only if* $\operatorname{Re}\lambda(F - GG'X) < 0$.

(ii) *If* (H, F) *is detectable, then* $X \geqq 0$ *if and only if* $\operatorname{Re}\lambda(F - GG'X) < 0$.

Proof. See Wonham [18, Thm. 4.1].

Note added in proof. A recent paper by Kucera [20] establishes the theory of nonnegative stabilizing solutions by direct arguments.

REFERENCES

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

[2] E. Apslund and L. Bungart, *A First Course in Integration*, Holt, Rinehart and Winston, New York, 1966.

[3] R. W. Brockett, *Finite-dimensional Linear Systems*, John Wiley, New York, 1970.

[4] R. S. Bucy and P. D. Joseph, *Filtering for Stochastic Processes with Applications to Guidance*, Interscience, New York, 1968.

[5] C. T. Chen, *Introduction to Linear System Theory*, Holt, Rinehart and Winston, New York, 1970.

[6] R. Courant, *Differential and Integral Calculus*, Blackie, London, 1937.

[7] F. R. Gantmacher, *Theory of Matrices*, vol. 1, Chelsea, New York, 1959.

[8] R. E. Kalman, *Contributions to the theory of optimal control*, Bol. Soc. Mat. Mexicana, 5 (1960), pp. 102–119.

[9] A. G. MacFarlane, *An eigenvector solution of the optimal linear regulator problem*, J. Electron Contr., 14 (1963), pp. 643–654.

[10] S. A. Marshall and H. Nicholson, *Optimal control of linear multivariable systems with quadratic performance criteria*, Proc. IEE, 117 (1970), pp. 1705–1713.

[11] K. Martensson, *On the matrix Riccati equation*, Information Sci., 3 (1971), pp. 17–49.

[12] B. P. Molinari, *Equivalence relations for the algebraic Riccati equation*, this Journal, 11 (1973), pp. 272–285.

[13] J. E. Potter, *A matrix equation arising in statistical filter theory*, Rep. NASA CR-270, M.I.T., Cambridge, Mass., 1965.

[14] H. H. Rosenbrock, *State Space and Multivariable Theory*, Nelson, London, 1970.

[15] B. Wendroff, *Theoretical Numerical Analysis*, Academic Press, New York, 1966.

[16] J. C. Williams, *The generation of Lyapunov functions for input–output stable systems*, this Journal, 9 (1971), pp. 105–134.

[17] ———, *Least-squares stationary optimal control and the algebraic Riccati equation*, IEEE Trans. Automatic Control, AC-16 (1971), pp. 621–634.

[18] W. M. Wonham, *On a matrix Riccati equation of stochastic control*, this Journal, 6 (1968), pp. 681–697.

[19] D. C. Youla, *On the factorization of rational matrices*, IRE Trans. Information Theory, IT-7 (1961), pp. 172–189.

[20] V. Kucera, *A contribution to matrix quadratic equations*, IEEE Trans. Automatic Control, AC-17 (1972), pp. 344–347.

SIAM J. Control
Vol. 11, No. 2, May 1973

EQUIVALENCE RELATIONS FOR THE ALGEBRAIC RICCATI EQUATION*

B. P. MOLINARI†

Abstract. By generalizing the notion of spectral factorization, solutions X of the matrix quadratic equation $BX + XA - XCDX + Q = 0$ are shown to have a one-to-one relation with factorizations of a rational matrix. By progressive specialization of this factorization, equivalence results are obtained in turn for symmetric solutions, Hermitian solutions, stabilizing solutions and positive definite solutions of the special case of the algebraic Riccati equation $F'X + XF - XGG'X + Q = 0$.

1. Introduction. The normalized algebraic Riccati equation

$$\text{(ARE)} \qquad F'X + XF - XGG'X + Q = 0$$

is fundamentally associated with least squares problems defined on trajectories of stationary linear systems. Problems of this type include:

(i) the regulator problem of linear optimal control [5], [7], [18],

(ii) the dual filtering problem of estimation theory [8] and

(iii) Lyapunov stability theory of feedback systems [17].

These least squares problems involve only particular solutions of the ARE, and the theory of these solutions is adequate for the problem application.

However, the fundamental problem of characterizing all solutions of the ARE is not so well developed, particularly when no a priori sign-definite assumption is made on Q. One approach characterizes solutions in terms of eigenvalues and eigenvectors of the Hamiltonian matrix:

$$M = \begin{bmatrix} F & GG' \\ Q & -F' \end{bmatrix}. \tag{1.1}$$

Unfortunately the equivalence statement involves the nonsingularity of an intermediate matrix [13], [15], which has hindered the generation of existence results via this approach. A second approach, recently published [18], relates general solutions of the ARE to two extremal solutions encountered in least squares problems. As the existence of these extremal solutions has been established, a proper picture of the complete set of solutions of the ARE emerges for the first time.

This paper establishes a third characterization of solutions of the ARE. It is motivated by the solution of the linear regulator problem by Weiner–Hopf methods [6], [16], which involves a factorization of the general form $\Delta'(-s)\Delta(s)$ (spectral factorization) of the rational matrix

$$\Phi(s) = I + G'(-Is - F')^{-1}Q(Is - F)^{-1}G. \tag{1.2}$$

This paper relates solutions of the ARE not to $\Phi(s)$ but to a closely related rational matrix $\Phi_K(s)$. In particular, it establishes a one-to-one relation between solutions

* Received by the editors August 3, 1971, and in revised form May 26, 1972.

† School of Engineering, University of New South Wales, Kensington, Australia 2033.

of the ARE factorizations of the general form $V(s)W(s)$. Specializing this notion of factorization in turn characterizes the symmetric, Hermitian and the stabilizing solutions of the ARE.

This work is very closely related to the results of Popov [14] and of Anderson [1]–[4]. In particular, Anderson establishes [3] an equivalence between certain factorizations of the form $W'(-s)W(s) = \Phi(s)$ and real symmetric solutions of a matrix quadratic equation. However, for $\Phi(s)$ defined by (1.2) this equation in general is not the ARE (see later comments). This paper, on the other hand, carefully preserves the relation to the ARE by modifying $\Phi(s)$.

It must be stressed that this paper is involved only with equivalence results and no attempt is made to provide existence results. However, as the theory of rational matrices is reasonably well established [21], [10], [11], these follow readily enough and will be dealt with separately. One such application to special solutions of the ARE is given in a companion paper.

The methods used are essentially those introduced by Anderson in [1]–[4]. These in turn rest squarely on the realization theory of constant finite-dimensional linear systems and the related notion of the degree $\delta[Z]$ of a rational matrix $Z(s)$ [12], [7, Chap. 2], [9, Chap. 6].

2. Problem statement. It must be realized at the outset that the general solution of the ARE is a complex matrix, neither symmetric nor Hermitian. Because of this fact there is no additional difficulty in studying the slightly more general matrix quadratic equation

$$\text{(MQE)} \qquad BX + XA - XCDX + Q = 0,$$

where A, B and Q are $n \times n$ real matrices and C and D' are $n \times m$ real matrices. No symmetry is claimed for Q.

The matrices M and Φ are logically redefined as

$$(2.1) \qquad M = \begin{bmatrix} A & CD \\ Q & -B \end{bmatrix},$$

$$(2.2) \qquad \Phi(s) = I + D(-Is - B)^{-1}Q(Is - A)^{-1}C.$$

There are several easy connections between the MQE and the M and Φ matrices. Before these are given it is convenient to state the following aids to matrix manipulation.

LEMMA 1. *The following two conditions are equivalent*:

$$BY + YA + Q = 0,$$

$$(-Is - B)^{-1}Y + Y(Is - A)^{-1} = (-Is - B)^{-1}Q(Is - A)^{-1}.$$

If $Z(s) = I + K(Is - A)^{-1}C$, *then*

$$\det(Z) = \frac{\det(Is - A + CK)}{\det(Is - A)},$$

$$Z^{-1}(s) = I - K(Is - A + CK)^{-1}C.$$

Proof. The determinant relation is given by Kalman in [12, Prop. 6]. The others are trivial.

The matrices M and Φ and the MQE are related according to the following.

LEMMA 2.

$$\Phi^{-1}(s) = I + [0 \quad D]\begin{bmatrix} Is - M \end{bmatrix}^{-1}\begin{bmatrix} C \\ 0 \end{bmatrix},$$

$$\det(\Phi) = \frac{\det(Is - M)}{\det(Is - B)\det(Is - A)}.$$

If X is a solution of the MQE, *then*

$$[I + D(-Is - B)^{-1}XC][I + DX(Is - A)^{-1}C] = \Phi(s), \tag{2.3}$$

$$\det(Is + B - XCD)\det(Is - A + CDX) = \det(Is - M). \tag{2.4}$$

Proof. By immediate calculation

$$I - [0 \quad D]\begin{bmatrix} Is - A & 0 \\ -Q & Is + B \end{bmatrix}^{-1}\begin{bmatrix} C \\ 0 \end{bmatrix} = \Phi(s),$$

and the application of Lemma 1 provides the first two identities. Equation (2.3) is obtained by expanding the product using Lemma 1, and (2.4) follows by taking determinants of (2.3).

Relation (2.3) shows that solutions of the MQE imply certain factorizations of $\Phi(s)$ of the general form $V(s)W(s) = \Phi(s)$, where V and W are square rational matrices. This paper determines a situation where a converse result holds.

It turns out that the arguments used in this paper break down when there is a cancellation of factors in the rational expression for $\det(\Phi)$ (for example, when $\delta[\Phi] < 2n$). This paper avoids the impasse by considering rather a matrix $\Phi_K(s)$ where the troublesome poles of $\Phi(s)$ (if any) have been cancelled and replaced by poles in a more acceptable position. The cancellation technique is critical to the method of the paper, and utilizes the standard identity

$$(Is - A)^{-1}C[I - K(Is - A + CK)^{-1}C] = (Is - A + CK)^{-1}C. \tag{2.5}$$

This corresponds to the situation of system dynamics $\dot{x} = Ax + Cu$, where state feedback $u = v - Kx$ is used to modify the dynamics to $\dot{x} = (A - CK)x + Cv$. Taking Laplace transforms provides (2.5).

With this background, define

$$Z_A(s) = I - K_A(Is - A + CK_A)^{-1}C,$$

$$Z_B(s) = I - D(-Is - B + K_BD)^{-1}K_B$$

for real matrices K_A and K_B of appropriate dimension, and for convenience denote

$$A_K = A - CK_A, \qquad B_K = B - K_BD.$$

The rational matrix of interest is defined by

$$\Phi_K(s) = Z_B(s)\Phi(s)Z_A(s). \tag{2.6}$$

Its essential properties are collected in the following.

LEMMA 3.

$$\Phi_K(s) = I - [K_A \quad D]\begin{bmatrix} Is - A_K & 0 \\ -Q - K_BK_A & Is + B_K \end{bmatrix}^{-1}\begin{bmatrix} C \\ -K_B \end{bmatrix},$$

$$\Phi_K^{-1}(s) = I + [K_A \quad D]\begin{bmatrix} Is - M \end{bmatrix}^{-1}\begin{bmatrix} C \\ -K_B \end{bmatrix},$$

$$\det(\Phi_K) = \frac{\det(Is - M)}{\det(Is + B_K)\det(Is - A_K)}.$$

If X is a solution of the MQE, *then*

$$[I + D(-Is - B_K)^{-1}(XC - K_D)][I + (DX - K_C)(Is - A_K)^{-1}C] = \Phi_K(s).$$

Proof. Using (2.5) to expand (2.6) gives

$$\begin{aligned}\Phi_K(s) &= I - D(-Is - B_K)^{-1}K_B - K_A(Is - A_K)^{-1}C \\ &\quad + D(-Is - B_K)^{-1}(Q + K_BK_A)(Is - A_K)^{-1}C.\end{aligned} \tag{2.7}$$

By direct expansion this can be shown equivalent to the realization in the lemma. The expressions for $\Phi_K^{-1}(s)$ and $\det(\Phi_K)$ follow exactly as in Lemma 2. Finally the factorization identity is obtained by premultiplying (2.3) by $Z_B(s)$ and postmultiplying it by $Z_A(s)$. In other words, a solution of the MQE now implies a factorization of the general form $V(s)W(s) = \Phi_K(s)$. This paper seeks a converse result where a factorization of $\Phi_K(s)$ implies the existence of a solution of the MQE.

To establish this result the poles of $\Phi_K(s)$ have to be fixed via the following assumptions.

(A.1) (a) (A, C) is controllable;
(b) $\operatorname{Re}\lambda(A_K) < 0$;
(c) A_K and M have no common eigenvalues.

(A.2) (a) (D, B) is observable;
(b) $\operatorname{Re}\lambda(B_K) < 0$;
(c) $-B_K$ and M have no common eigenvalues.

Remark 1. The essential assumptions are (A.1a) and (A.2a). Given these conditions, standard results [9, Thm. 7.6] imply the existence of suitable K_A and K_B.

Remark 2. Should (A.1a) and (A.2a) not hold, then a canonical structure result [9, Thm. 5.17] can be used to reduce the MQE to a set of one smaller MQE where (A.1a) and (A.2a) do hold, and three matrix linear equations that can be solved sequentially.

The essential consequences of (A.1) and (A.2) are collected in the following.

LEMMA 4. *Assume* (A.1) *and* (A.2). *Then*

$$\delta[\Phi_K] = 2n, \tag{2.8}$$

$$\Phi_K(s) = I + D(-Is - B_K)^{-1}(LC - K_B) + (DL - K_A)(Is - A_K)^{-1}C, \tag{2.9}$$

where L is the unique solution of

$$B_KL + LA_K + K_BK_A + Q = 0. \tag{2.10}$$

Moreover, the realizations in (2.9) *are both minimal.*

Proof. Directly from Lemma 3, $\delta[\Phi_K] \leqq 2n$. By standard theory, $\delta[\Phi_K] = \delta[\Phi_K - I]$, and applying the Duffin–Hazony concept of degree [12, Def. 6] gives

$$\delta[\Phi_K] = \max_R \delta[\pi_R],$$

where

$$\det(R + (\Phi_K - I)) = \frac{\pi_R(s)}{\rho_R(s)}$$

and $\pi_R(s)$ and $\rho_R(s)$ are relatively prime. But for $R = I$ the conditions (A.1c) and (A.2c) applied to Lemma 3 give $\delta[\pi_1] = 2n$.

Hence

$$\delta[\Phi_K] \geqq 2n,$$

which immediately gives (2.8).

Now by (A.1b) and (A.2b), (2.10) certainly has a unique solution. Lemma 1 gives an equivalent condition which when substituted into (2.7) directly gives (2.9).

Finally denote

$$\delta[D(-Is - B_K)^{-1}(LC - K_B)] = m_B,$$

$$\delta[(DL - K_A)(Is - A_K)^{-1}C] = m_A.$$

By inspection,

$$m_B \leqq n, \qquad m_A \leqq n.$$

But from (2.9) the degrees are related by [12, Thm. B]

$$2n \leqq m_A + m_B.$$

Hence $m_A = m_B = n$, which completes the lemma.

3. Main equivalence result. The factorizations studied in this paper are precisely defined by the following.

DEFINITION 1. The pair (V, W) is a *factorization* of $\Phi_K(s)$ if and only if $V(s)$ and $W(s)$ are $m \times m$ rational matrices satisfying:

(a) $V(s)W(s) = \Phi_K(s)$;

(b) V is analytic in $\operatorname{Re} s \leqq 0$ and W is analytic in $\operatorname{Re} s \geqq 0$;

(c) V^{-1} has no poles in common with W, and W^{-1} has no poles in common with V;

(d) $\lim_{s\to\infty} V(s) = \lim_{s\to\infty} W(s) = I$.

Remark 3. This notion of factorization is a logical generalization of those involved in Weiner–Hopf methods [21], [10], [6], covariance generation [3] and feedback systems stability theory [22], [19].

Remark 4. The matrices $V(s)$ and $W(s)$ are "minimal" in several respects. By assumption, V has the minimum number of columns and W has the minimum number of rows. Condition (c) excludes the possibility of "inflating" a factorization (V, W) into

$$(VZ^{-1})(ZW) = \Phi_K(s),$$

where $Z(s)$ is analytic for $\operatorname{Re} s \geqq 0$ and $Z^{-1}(s)$ is analytic for $\operatorname{Re} s \leqq 0$. Condition

(d) similarly excludes the extension of a factorization (V, W) to

$$(VN^{-1})(NW) = \Phi_K(s),$$

where N is a constant nonsingular matrix.

Remark 5. It is not claimed that $V(s)$ or $W(s)$ is a real rational matrix. In fact, it is essential to allow the possibility of complex coefficients in V and W.

The essential result of this paper is now stated.

THEOREM 1. *Assume* (A.1) *and* (A.2). *Then the matrix pair* (V, W) *is a factorization of* $\Phi_K(s)$ *if and only if*

(i) *V and W admit the following minimal realizations for some X*:

$$V(s) = I + D(-Is - B_K)^{-1}(XC - K_B),$$

$$W(s) = I + (DX - K_A)(Is - A_K)^{-1}C,$$

and

(ii) *X is a solution of the* MQE.

Further, the relation between (V, W) and X is one-to-one. Finally, X is a real matrix if and only if both $V(s)$ and $W(s)$ are real rational matrices.

It is convenient to prove the theorem in several easy stages.

LEMMA 5. *Assume* (A.1) *and* (A.2). *If (V, W) is a factorization of $\Phi_K(s)$, then*

$$\delta[V] = \delta[W] = n.$$

Proof. Condition (d) of Definition 1 implies that $V(s)$ and $W(s)$ possess minimal realizations of the general form $V(s)$, $W(s) = I + H(Is - F)^{-1}G$. Lemma 1 then implies that

$$\det(V) = \frac{\pi_V(s)}{\rho_V(s)}, \qquad \det(W) = \frac{\pi_W(s)}{\rho_W(s)},$$

where π_V and ρ_V (relatively prime) are of the same degree, as are π_W and ρ_W. Moreover, all zeros of $\pi_V(s)$ are poles of $V^{-1}(s)$, and all zeros of $\pi_W(s)$ are poles of $W^{-1}(s)$.

Lemma 3 provides

$$\frac{\det(Is - M)}{\det(Is + B_K)\det(Is - A_K)} = \frac{\pi_V\pi_W}{\rho_V\rho_W}.$$

It follows from condition (c) of Definition 1 that $(\pi_V\pi_W)$ and $(\rho_V\rho_W)$ are relatively prime. Hence

$$\det(Is - M) = \pi_V\pi_W, \tag{3.1}$$

$$\det(Is + B_K)\det(Is - A_K) = \rho_V\rho_W, \tag{3.2}$$

and conditions (A.1b) and (A.2b) then imply that

$$\det(Is + B_K) = \rho_V(s), \qquad \det(Is - A_K) = \rho_W(s).$$

Hence exactly as in Lemma 4,

$$\delta[V] \geqq n, \qquad \delta[W] \geqq n.$$

Conversely, conditions (b) and (c) of Definition 1 imply by standard theory [1, Lemma 4] that

$$2n = \delta[\Phi_K] = \delta[V] + \delta[W].$$

The lemma follows immediately.

Equation (3.1) provides the basis for the following.

COROLLARY. *Assume* (A.1) *and* (A.2). *Then all the poles of* $V^{-1}(s)$ *and all the poles of* $W^{-1}(s)$ *are eigenvalues of* M.

This partitioning of the eigenvalues of M between $V(s)$ and $W(s)$ provides a formal link between this characterization of solutions of the MQE and the eigenvector characterization of Potter [15] and Martensson [13].

The next step establishes the structure of the minimal realizations.

LEMMA 6. *Assume* (A.1) *and* (A.2). *If* (V, W) *is a factorization of* $\Phi_K(s)$, *then* $V(s)$ *and* $W(s)$ *admit minimal realizations*

$$V(s) = I + D(-Is - B_K)^{-1}H_B,$$
$$W(s) = I + H_A(Is - A_K)^{-1}C,$$

for some H_B *and* H_A *of appropriate dimension.*

Proof. From Lemma 5, V and W admit minimal realizations [9, Chap. 6] of the form

$$V(s) = I + G_V(-Is - F_V)^{-1}H_V,$$
$$W(s) = I + H_W(Is - F_W)^{-1}G_W,$$

where F_V and F_W are $n \times n$ matrices, and G_V', H_V, H_W' and G_W are all $n \times m$ matrices (not necessarily real matrices). By direct expansion using Lemma 1

$$V(s)W(s) = I + G_V(-Is - F_V)^{-1}(H_V + YG_W)$$
$$+ (H_W + G_VY)(Is - F_W)^{-1}G_W,$$

where Y is the solution (unique by condition (b) of Definition 1) of the matrix linear equation

$$F_VY + YF_W + H_VH_W = 0.$$

Equating this expression to (2.9) and using a partial fraction argument gives

$$G_V(-Is - F_V)^{-1}(H_V + YG_W) = D(-Is - B_K)^{-1}(LC - K_B),$$
$$(H_W + G_VY)(Is - F_W)^{-1}G_W = (DL - K_A)(Is - A_K)^{-1}C.$$

Now all these realizations are minimal. The isomorphism property [7, Thm. 18.2] of minimal realizations implies

(a) there exists T_V such that $B_K = T_V^{-1}F_VT_V$, $D = G_VT_V$;

(b) there exists T_W such that $A_K = T_WF_WT_W^{-1}$, $C = T_WG_W$.

Using these matrices to transform the above realizations provides the lemma.

Remark 6. Although the realization theory just used was developed for real rational matrices, it holds equally well for complex rational matrices.

The last lemma provides the basis for the following.

Proof of Theorem 1. *Sufficient conditions.* For $V(s)$ and $W(s)$ defined in the statement of Theorem 1:

condition (d) (of Definition 1) holds by inspection;

condition (b) holds by (A.1b) and (A.2b);

condition (a) holds since X is a solution of the MQE (Lemma 3), and condition (c) follows from relation (2.4), conditions (A.1c) and (A.2c), and

$$V^{-1}(s) = I - D(-Is - B + XCD)^{-1}(XC - K_B),$$
$$W^{-1}(s) = I - (DX - K_A)(Is - A + CDX)^{-1}C.$$

Hence (V, W) is a factorization of $\Phi_K(s)$.

Necessary conditions. Consider the realizations of Lemma 6 and evaluate the product $V(s)W(s)$ exactly as before. Then

$$V(s)W(s) = I + D(-Is - B_K)^{-1}(YC + H_B) + (H_A + DY)(Is - A_K)^{-1}C,$$

where Y is the unique solution of

$$B_K Y + Y A_K + H_B H_A = 0. \tag{3.3}$$

Again equating this expression to (2.9) and using a partial fraction argument gives

$$D(-Is - B_K)^{-1}(YC + H_B) = D(-Is - B_K)^{-1}(LC - K_B),$$
$$(H_A + DY)(Is - A_K)^{-1}C = (DL - K_A)(Is - A_K)^{-1}C.$$

The observability of (D, B_K) and the controllability of (A_K, C) then give

$$YC + H_B = LC - K_B,$$
$$H_A + DY = DL - K_A.$$

Consider the well-defined matrix X defined by $X = L - Y$. Then

$$H_B = XC - K_B, \qquad H_A = DX - K_A,$$

which provides the desired realizations. It now remains to show that X is a solution of the MQE. Subtracting (3.3) from (2.10) gives

$$B_K(L - Y) + (L - Y)A_K - H_B H_A + K_B K_A + Q = 0. \tag{3.4}$$

Substituting gives

$$B_K X + X A_K - (XC - K_A)(DX - K_A) + K_B K_A + Q = 0.$$

Simple removal of brackets gives

$$BX + XA - XCDX + Q = 0, \quad \text{as required.}$$

One-to-one relation. Consider two factorizations (V_1, W_1) and (V_2, W_2), and assume that they are equal. In terms of the realizations just established this means that

$$X_1 - K_B = X_2 C - K_B,$$
$$DX_1 - K_A = DX_2 - K_A,$$

where X_1 and X_2 satisfy

$$B_K X_1 + X_1 A_K - (X_1 C - K_B)(DX_1 - K_A) + K_B K_A + Q = 0,$$
$$B_K X_2 + X_2 A_K - (X_2 C - K_B)(DX_2 - K_A) + K_B K_A + Q = 0.$$

Subtracting these equations and noting the last relations gives

$$B_K(X_1 - X_2) + (X_1 - X_2)A_K = 0.$$

Conditions (A.1b) and (A.2b) then imply that $X_1 - X_2 = 0$. Hence the corresponding solutions X_1 and X_2 are equal. The converse is trivial.

Real property. Consider $V(s)$ and $W(s)$ to be real rational matrices. Now

$$V(s) = I + D(-Is - B_K)^{-1}H_B.$$

Since D and B_K are both real matrices and (D, B_K) is observable it follows that H_B must be a real matrix. Similarly H_A must be a real matrix. Then (3.4) is a matrix linear equation with real matrix coefficients, and it must follow that $L - Y = X$ is a real matrix. Again the converse is trivial. This completes the proof.

Alternatively the factorization (V, W) can be characterized in terms of the solutions Y of (3.3). For the special case where $K_A = K_B = 0$ is allowable in conditions (A.1) and (A.2), this equation can be written as

$$BY + YA + (YC - LC)(DY - DL) = 0.$$

This is the approach taken by Anderson [3, Eq. (12)] for special (symmetric) cases of (V, W).

4. The algebraic Riccati equation.

This section specializes the results of § 3 to the algebraic Ricatti equation

$$F'X + XF - XGG'X + Q = 0. \tag{4.1}$$

This of course is the MQE for the special case

$$A = B' = F, \qquad C = D' = G,$$

and the associated matrices return to

$$M = \begin{bmatrix} F & GG' \\ Q & -F' \end{bmatrix}, \tag{4.2}$$

$$\Phi(s) = I + G'(-Is - F')^{-1}Q(Is - F)^{-1}G. \tag{4.3}$$

In the applications of the ARE no loss of generality is involved in the assumption:

(A.3) Q is symmetric.

Immediate consequences are

$$\Phi'(-s) = \Phi(s), \tag{4.4}$$

$$\det(-Is - M) = \det(Is - M). \tag{4.5}$$

The latter result follows immediately from Lemma 2.

It is essential to preserve the symmetry property (4.4). To this end denote $F_K = F - GK$, where K is a real matrix of appropriate dimensions, and consider $K_A = K'_B = K$. Then (2.6) specializes to

$$\Phi_K(s) = Z'(-s)\Phi(s)Z(s), \tag{4.6}$$

where

$$Z(s) = I - K(Is - F_K)^{-1}G = Z_A(s) = Z_B'(-s).$$

Clearly this gives the property

$$\Phi_K'(-s) = \Phi_K(s). \tag{4.7}$$

To establish pole-placement, assume

(A.4) (a) (F, G) is controllable;

(b) Re $(F_K) < 0$;

(c) F_K and M have no common eigenvalues.

Again (A.4a) implies the existence of matrices K satisfying conditions (b) and (c).

Note that (A.1) holds directly, and by the symmetry property (4.5) condition (A.2) also holds. Hence (A.3) and (A.4) are sufficient conditions for Theorem 1 to characterize the general complex solution of the ARE. However, the symmetry of the ARE admits the possibility of symmetric and Hermitian solutions. These are now characterized in turn by specializing the notion of factorization.

4.1. Symmetric solutions. Definition 1 is specialized to the following definition.

DEFINITION 2. A matrix $S(s)$ is a *symmetric factorization* of $\Phi_K(s)$ if and only if $S(s)$ is an $m \times m$ rational matrix satisfying:

(a) $S'(-s)S(s) = \Phi_K(s)$;

(b) $S(s)$ is analytic in Re $s \geqq 0$;

(c) $S^{-1}(-s)$ and $S(s)$ have no common poles;

(d) $\lim_{s\to\infty} S(s) = I$.

If, in addition, $S(s)$ is a real rational matrix, then such a factorization has application in covariance generation and has been studied by Anderson [3].

Clearly $(S'(-s), S(s))$ is a factorization of $\Phi_K(s)$ in the sense of Definition 1. This provides the basis for the following theorem.

THEOREM 2. *Assume* (A.3) *and* (A.4). *Then $S(s)$ is a symmetric factorization of $\Phi_K(s)$ if and only if*

(i) *$S(s)$ admits the minimal realization*

$$S(s) = I + (G'X - K)(Is - F_K)^{-1}G,$$

for some symmetric matrix X, and

(ii) *X is a solution of the* ARE.

Further, the relation is one-to-one. Finally, X is a real symmetric matrix if and only if $S(s)$ is a real rational matrix.

Proof. Define rational matrices V and W by

$$V(s) = S'(-s), \qquad W(s) = S(s).$$

Then (V, W) is a factorization of $\Phi_K(s)$, and Theorem 1 gives the existence of a solution X of the ARE (not yet symmetric) such that

$$V(s) = I + G'(-Is - F_K')^{-1}(XG - K'),$$

$$W(s) = I + (G'X - K)(Is - F_K)^{-1}G.$$

Now by (4.7), $\Phi_K(s) = W'(-s)V'(-s)$, and it is easily checked that $(W'(-s), V'(-s))$ is a factorization of $\Phi_K(s)$. By inspection, it is related to X', again a solution of the ARE. But by definition,

$$(V(s), W(s)) = (W'(-s), V'(-s)).$$

The one-to-one property of Theorem 1 then gives $X = X'$. The rest follows directly from Theorem 1.

4.2. Hermitian solutions. A parallel characterization can be given for Hermitian solutions. It is convenient to use the following notation introduced by Youla [21],

$$Z_*(s) = [Z(-\bar{s})]^*,$$

where $\bar{x}$ denotes the complex conjugate and x^* denotes the adjoint (complex conjugate transpose) of x respectively. If $Z(s)$ is a real rational matrix, $Z_*(s) = Z'(-s)$.

An appropriate notion of factorization turns out to be the following.

DEFINITION 3. A matrix $H(s)$ is a *Hermitian factorization* of $\Phi_K(s)$ if and only if $H(s)$ is an $m \times m$ rational matrix satisfying:

(a) $H_*(s)H(s) = \Phi_K(s)$;

(b) $H(s)$ is analytic for $\operatorname{Re} s \geqq 0$;

(c) $H_*^{-1}(s)$ and $H(s)$ have no common poles;

(d) $\lim_{s\to\infty} H(s) = I$.

Note that if $H(s)$ is a real rational matrix, it reduces to a real symmetric factorization of $\Phi_K(s)$.

Exactly as for the last theorem the following can be shown.

THEOREM 3. *Assume* (A.3) *and* (A.4). *Then $H(s)$ is a Hermitian factorization of $\Phi_K(s)$ if and only if*

(i) *$H(s)$ admits the minimal realization*

$$H(s) = I + (G'X - K)(Is - F_K)^{-1}G,$$

for some Hermitian matrix X, and

(ii) *X is a solution of the* ARE.

Further, the relation is one-to-one.

4.3. Stabilizing solutions. In view of relation (2.4), it is valid to consider eigenvalue restrictions on the matrix $(F - GG'X)$ for solutions of the ARE, especially the restriction of the eigenvalues to a half-plane. An appropriate factorization is given by the following.

DEFINITION 4. A matrix $\Delta(s)$ is a *spectral factorization of* $\Phi_K(s)$ if and only if $\Delta(s)$ is a real $m \times m$ rational matrix satisfying:

(a) $\Delta'(-s)\Delta(s) = \Phi_K(s)$;

(b) $\Delta(s)$ is analytic for $\operatorname{Re} s \geqq 0$;

(c) $\Delta^{-1}(s)$ is analytic for $\operatorname{Re} s > 0$;

(d) $\lim_{s\to\infty} \Delta(s) = I$.

Clearly condition (d) merely normalizes the standard concept of spectral factorization [21], [10].

A spectral factorization $\Delta(s)$ satisfies the definition of a real symmetric factorization. Theorem 2 provides the basis for the following.

THEOREM 4. *Assume* (A.3) *and* (A.4). *Then $\Delta(s)$ is a spectral factorization of*

$\Phi_K(s)$ *if and only if*

(i) $\Delta(s)$ *admits a minimal realization*

$$\Delta(s) = I + (G'\hat{X} - K)(Is - F_K)^{-1}G,$$

for some real symmetric matrix $\hat{X}$, *and*

(ii) $\hat{X}$ *is a solution of the* ARE *satisfying* $\operatorname{Re} \lambda(F - GG'\hat{X}) \leqq 0$.

Proof. It remains to show that the eigenvalue condition is equivalent to condition (c) of Definition 4. By inspection,

$$\Delta^{-1}(s) = 1 - (G'\hat{X} - K)(Is - F + GG'\hat{X})^{-1}G.$$

Since this realization is minimal, it follows that $\Delta^{-1}(s)$ and $(F - GG'\hat{X})$ have the same invariant factors [12, Thm. 1], which is precisely what is required.

COROLLARY. *Assume* (A.3) *and* (A.4). *Consider* $\Delta(s)$ *and* $\hat{X}$ *involved in Theorem* 4. *Then* $\Delta^{-1}(s)$ *is analytic for* $\operatorname{Re} s \geqq 0$ *if and only if* $\operatorname{Re} \lambda(F - GG'\hat{X}) < 0$.

A solution of the ARE satisfying this last condition has been called a stabilizing solution in the sense that a feedback control $u = -G'\hat{X}x$ will stabilize the differential equation $\dot{x} = Fx + Gu$.

4.4. Positive definite solutions. Finally, it is worth noting a further restriction that is standard for the regulator problem:

(A.5) $Q = H'H$.

For this case the following equivalence is well known.

THEOREM 5. *Assume* (A.5) *and consider* X *a real symmetric solution of the* ARE.

(i) *If* (H, F) *is observable, then* $X > 0$ *if and only if* $\operatorname{Re} \lambda(F - GG'X) < 0$.

(ii) *If* (H, F) *is detectable, then* $X \geqq 0$ *if and only if* $\operatorname{Re} \lambda(F - GG'X) < 0$.

Proof. See Wonham [20, Thm 4.1].

Under the additional assumptions of $Q = H'H$ and (H, F) observable, Theorem 4 and its corollary can be written in terms of positive definite solutions of the ARE rather than stabilizing solutions. This then provides a precise statement of the equivalence between positive definite solutions of the ARE and spectral factorizations. For the special case $m = 1$ (G is an $n \times 1$ vector) this equivalence is well known [7, § 27]. For the general case Anderson has pointed out [2] that a positive definite solution of the ARE implies a spectral factorization. The converse result has apparently not been published but has long been an article of faith in optimal linear regulator theory. It was an interest in this converse problem that motivated the more general results of this paper.

5. Example. The following simple example may help to illuminate the nature of the results of this paper. Consider

$$F'X + XF - XGG'X + Q = 0,$$

where

$$F\begin{bmatrix} 0 & 1 \\ -2 & -3 \end{bmatrix}, \qquad G = \begin{bmatrix} 0 \\ 1 \end{bmatrix}, \qquad Q = \begin{bmatrix} 21 & 4 \\ 4 & 1 \end{bmatrix}.$$

Note that the pair (F, G) is in the controllable canonical form. The special properties of $(Is - F)^{-1}G$ then easily provide

$$\Phi(s) = 1 + \frac{-s^2 + 21}{(s^2 - 3s + 1)(s^2 + 3s + 2)}$$

$$= \frac{s^4 - 6s^2 + 25}{(s^2 - 3s + 2)(s^2 + 3s + 2)}.$$

It can be checked that (A.4) holds for $K = 0$. All possible factorizations of $\Phi(s)$ are now listed, along with the corresponding complete set of solutions to the ARE.

V, W	X
$\frac{s^2 + 4s + 5}{s^2 - 3s + 2}, \frac{s^2 - 4s + 5}{s^2 + 3s + 2}$	$\begin{bmatrix} -30 & 3 \\ 3 & -7 \end{bmatrix}$
$\frac{s^2 - 2js - 5}{s^2 - 3s + 2}, \frac{s^2 + 2js - 5}{s^2 + 3s + 2}$	$\begin{bmatrix} -10 - 10j & -7 \\ -7 & -3 + 2j \end{bmatrix}$
$\frac{s^2 - (3 + 4j)}{s^2 - 3s + 2}, \frac{s^2 - (3 - 4j)}{s^2 + 3s + 2}$	$\begin{bmatrix} -10 & -5 - 4j \\ -5 + 4j & -3 \end{bmatrix}$
$\frac{s^2 - (3 - 4j)}{s^2 - 3s + 2}, \frac{s^2 - (3 + 4j)}{s^2 + 3s + 2}$	$\begin{bmatrix} -10 & -5 + 4j \\ -5 - 4j & -3 \end{bmatrix}$
$\frac{s^2 + 2js - 5}{s^2 - 3s + 2}, \frac{s^2 - 2js - 5}{s^2 + 3s + 2}$	$\begin{bmatrix} -10 + 10j & -7 \\ -7 & -3 - 2j \end{bmatrix}$
$\frac{s^2 - 4s + 5}{s^2 - 3s + 2}, \frac{s^2 + 4s + 5}{s^2 + 3s + 2}$	$\begin{bmatrix} 10 & 3 \\ 3 & 1 \end{bmatrix}$

For this simple case all solutions are either symmetric or Hermitian matrices. This is not true in general.

6. Conclusions. This paper has precisely established, in quite general terms, an equivalence that has long been assumed to hold. Solutions of the algebraic Riccati equation have been shown to have a one-to-one relation with certain "minimal factorizations" of a real rational matrix. Apart from being of intrinsic interest this equivalence has considerable utility as results from the theory of rational matrices can be simply transcribed into existence results for solutions of the ARE. One such application to stabilizing solutions is considered in a companion paper.

REFERENCES

[1] B. D. O. ANDERSON, *A system theory criterion for positive real matrices*, this Journal, 5 (1967), pp. 171–182.

[2] ———, *An algebraic solution to the spectral factorization problem*, IEEE Trans. Automatic Control, AC-12 (1967), pp. 410–414.

[3] ———, *The Inverse Problem of stationary covariance generation*, J. Statistical Phys., 1 (1969), pp. 133–147.

[4] B. D. O. ANDERSON AND J. B. MOORE, *Algebraic structure of generalized positive real matrices*, this Journal, 6 (1968), pp. 615–624.

[5] B. D. O. ANDERSON, *Linear Optimal Control*, Prentice-Hall, Englewood Cliffs, N.J., 1971.

[6] J. Bongiorno, *Minimum sensitivity design of linear multivariable feedback control systems by matrix spectral factorization*, IEEE Trans. Automatic Control, AC-14 (1969), pp. 665–673.

[7] R. W. Brockett, *Finite-dimensional Linear Systems*, John Wiley, New York, 1970.

[8] R. S. Bucy and P. D. Joseph, *Filtering for Stochastic Processes with Applications to Guidance*, Interscience, New York, 1968.

[9] C. T. Chen, *Introduction to Linear System Theory*, Holt, Rinehart and Winston, New York, 1970.

[10] M. C. Davis, *Factoring the spectral matrix*, IEEE Trans. Automatic Control, AC-8 (1963), pp. 296–305.

[11] V. A. Jakubovic, *Factorization of symmetric matrix polynomials*, Soviet Math. Dokl., 11 (1970), pp. 1261–1264.

[12] R. E. Kalman, *Irreducible realizations and the degree of a rational matrix*, J. Soc. Indust. Appl. Math., 13 (1965), pp. 520–544.

[13] K. Martensson, *On the matrix Riccati equation*, Information Sci., 3 (1971), pp. 17–49.

[14] V. M. Popov, *Hyperstability and optimality of automatic systems with several control functions*, Rev. Roumaine Sci. Tech. Sér. Electrotech. Energ., 9 (1964), pp. 629–690.

[15] J. E. Potter, *Matrix quadratic solutions*, SIAM J. Appl. Math., 14 (1966), pp. 496–501.

[16] E. G. Rynaski and R. F. Whitbeck, *The theory and applications of linear optimum control*, Rep. AD 632 553, Cornell Aeronautics Laboratory, Buffalo, N.Y., 1966.

[17] J. C. Willems, *The generation of Lyapunov functions for input–output stable systems*, this Journal, 9 (1971), pp. 105–134.

[18] ———, *Least squares stationary optimal control and the algebraic Riccati equation*, IEEE Trans. Automatic Control, AC-16 (1971), pp. 621–634.

[19] J. C. Willems and R. W. Brockett, *Some new rearrangement inequalities having application in stability analysis*, Ibid., AC-13 (1968), pp. 539–549.

[20] W. M. Wonham, *On a matrix Riccati equation of stochastic control*, this Journal, 6 (1968), pp. 681–697.

[21] D. C. Youla, *On the factorization of rational matrices*, IRE Trans. Information Theory, IT-7 (1961), pp. 172–189.

[22] G. Zames and P. L. Falb, *Stability conditions for systems with monotone and slope-restricted nonlinearities*, this Journal, 6 (1968), pp. 89–108.

SIAM J. CONTROL
Vol. 11, No. 2, May 1973

A CRITERION FOR THE BOUNDED-INPUT, BOUNDED-OUTPUT STABILITY OF TIME-VARYING NONLINEAR SYSTEMS*

EMILE K. HADDAD†

Abstract. A new approach is used to derive sufficient conditions for the boundedness of a broad class of time-varying nonlinear feedback systems. The stability problem is defined and treated independently from the problem of well-posedness of the feedback system. The a priori restrictions stipulated on the nature of the system are relatively unrestrictive. In the general time-dependent system considered, time-variation may be exhibited by linear as well as the nonlinear portions of the system. The type of linear subsystem admitted includes the class of systems with feed-forward differentiators. The stability criterion is derived via a time-domain method of analysis that makes no reference to Lyapunov functions, transform relationships, or normed spaces. A number of examples demonstrate that the present criterion can predict stability information which is not readily obtainable for other comparable criteria.

1. Introduction. In the past few years the stability properties of nonlinear and time-varying systems have been extensively studied. A broad class of such systems can be represented by the feedback format shown in Fig. 1, which consists of a linear subsystem H interacting with a nonlinear characteristic N. Most of the available criteria dealing with the stability of time-varying nonlinear fe̶edback systems are applicable only to cases where the linear subsystem H is stationary [1]–[7]. This paper presents a new criterion for the bounded-input, bounded-output stability of feedback systems in which the nonlinearity, or the linear subsystem, or both may be time-varying.

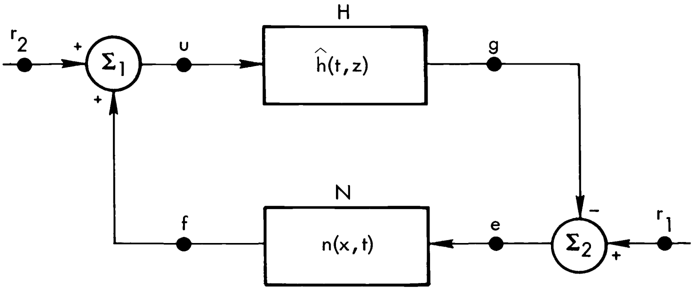


FIG. 1. *The time-varying nonlinear feedback system S*

The approach adopted in this paper is based on the point of view that the stability problem can be divorced from the questions of existence, uniqueness, and causality of solutions, and the related concept of well-posedness of the feedback model.[1] The practical motivation for such an approach can be justified by the fact that there are situations where the stability properties of a physical process can be adequately predicted from an "ill-posed" model of the process. Specific illustrations of this point will be given in subsequent sections of the paper.

* Received by the editors June 22, 1971, and in revised form May 3, 1972.

† Lockheed–Georgia Company, Marietta, Georgia 30061.

[1] For a discussion of the concept of well-posedness see J. C. Willems [12].

The stability criterion presented in this paper is applicable to a broad class of systems. In the derivation of the criterion, the a priori restrictions imposed on the nature of the system are kept to a minimum. For instance, no sector restriction is initially imposed on the graph of the nonlinear characteristic $n(x, t)$, and continuity in x and t is not stipulated. Furthermore, the nonlinear relationship $n(x, t)$ is allowed to be multivalued in x, as in the case of hysteresis. The time-varying linear subsystem H is allowed to be an improper system [8], such that its impulse response may contain any number of nth order delta functions which may be considered to represent feed-forward differentiators. Such systems are not uncommonly encountered, notably in circuit theory. Moreover, H is allowed to be a noncausal, or anticipative, system.

The method of analysis presented in this paper is rather distinct from the types of analyses to be found in previous investigations of the stability problem. In the course of the derivation, the system behavior is examined entirely in the time domain, with no reference to Lyapunov functions, frequency-domain relationships, or normed spaces.

The interpretation of the results of this paper and their application to specific systems is fairly straightforward. The conditions for stability stated in the criterion lend themselves to a simple graphical interpretation. Application to various specific systems is illustrated by a number of examples. In one of the examples, a stationary subsystem H is considered, and bounded-input, bounded-output stability is established for a nonlinearity that occupies a sector considerably larger than the sector obtained by the circle criterion.

2. System descriptions, assumptions, and problem definition. In this section, the nature of the feedback system under consideration is described in detail. The a priori assumptions about the system behavior are stated. The specific stability problem to be studied is formulated. In the interest of admitting a broad class of systems, and enhancing the generality of the results, the initial restrictions on the nature of the system will be kept to a minimum.

2.1. The feedback system. The schematic shown in Fig. 1 is a *model* for an *observed system or process S*. The observation *terminals*, represented by dots in the figure, are associated with the observation variables or *signals*, r_1, r_2, e, f, u, and g. These signals are represented by real-valued functions of the real variable t (time). It is assumed that there is a value t_0 such that

$$r_1(t) = r_2(t) = e(t) = f(t) = u(t) = g(t) = 0 \quad \text{for all } t < t_0. \tag{1}$$

Thus, t_0 may be described as the starting instant of the process or the *initial instant* of excitation of the system S. The observation of S takes place over the interval $[t_0, \infty)$, and it is therefore assumed that the signals are defined for all $t \in [t_0, \infty)$. The system signals are assumed to satisfy a number of specific relationships or *constraints*. The three-terminal junctions Σ_1 and Σ_2 induce the following constraints between their terminal signals:

$$r_1 - g - e = 0 \quad \text{for all } t \in [t_0, \infty), \tag{2}$$

$$r_2 + f - u = 0 \quad \text{for all } t \in [t_0, \infty). \tag{3}$$

Similarly, the blocks H and N induce specific constraints between their respective terminal signals. The exact description of these constraints will be given in subsequent sections. For the time being, let these constraints be symbolically represented by

(4) $$H[u, g] = 0 \quad \text{for all } t \in [t_0, \infty),$$

(5) $$N[e, f] = 0 \quad \text{for all } t \in [t_0, \infty).$$

The constraints (2) through (5) will be referred to as the process (system) *dynamics*. A specific set of signals $s = (r_1, r_2, e, f, u, g)$ is said to be an admissible process, or *a solution* to the system S, if the given signals r_1, r_2, e, f, u, and g satisfy the system dynamics, i.e. satisfy the constraints (2) through (5). The constituent signals of a given solution will be referred to as the *components* of the solution. The stability problem to be investigated relates to the question of boundedness of the components of system solutions. This is expounded in the following section.

2.2. Stability problem. Let Z represent the set of *all* possible solutions to the given system S. Let W be the set of all functions defined on $[t_0, \infty)$ which are bounded over finite intervals, i.e.,

(6) $$W \equiv \{w(t) \mid \sup_{[t_0, T]} |w(t)| < \infty \text{ for all } T \in [t_0, \infty)\}.$$

If $w(t) \in W$, then $w(t)$ is said to exhibit no "finite escape time." Let Y denote the set of all system solutions whose component signals exhibit no finite escape time, i.e.,

(7) $$Y \equiv \{(r_1, r_2, e, f, u, g) \in Z \mid r_1, r_2, e, f, u, g \in W\}.$$

In this paper, we shall be mainly concerned with the class of solutions represented by Y.

The designations of "inputs" and "outputs" will be used in conjunction with the signals of the system S, not to imply a cause-and-effect relationship, but rather to distinguish between signals for which some information is known a priori and signals for which information is being sought (a posteriori). In the study of the set of system solutions Y, it will be assumed that some information is given about some of the component signals, say r_1 and r_2, which are therefore designated as inputs, and that information is being sought for the other signals, which are then designated as outputs. The information to be obtained on the outputs is to be derived from the given information on the inputs and the system constraints (dynamics).

This brings us to the stability problem to be investigated in this paper, which may be stated as follows: given a solution $(r_1, r_2, e, f, u, g) \in Y$, what a priori information on r_1 and r_2, and on the system constraints, would guarantee that the signals e, f, u, and g are bounded over $[t_0, \infty)$? In this problem setting, the signals r_1 and r_2 are regarded as inputs and the other signals as outputs, and the information to be sought about the output signals is their boundedness over $[t_0, \infty)$.

In this formulation of the stability problem, two points should be emphasized. First, attention is being given only to solutions that belong to Y, i.e. solutions whose component signals do not exhibit finite escape time. Second, the question of stability of solutions is being divorced from the question of uniqueness of solutions. With regard to the first point, it should be mentioned that, for a given system, it might be possible to have admissible solutions which do not belong to Y. In this respect, a legitimate stability problem would be the following: given a solution $(r_1, r_2, e, f, u, g) \in Z$, given that $r_1, r_2 \in W$, what conditions on the system constraints would guarantee that $e, f, u, g \in W$, i.e. $(r_1, r_2, e, f, u, g) \in Y$? This question will not be tackled in this paper.[2]

As to the question of uniqueness, it should be noted that, for some given (r_1, r_2), there might exist more than one solution $s_n = (r_1, r_2, e_n, f_n, u_n, g_n) \in Y$. Nevertheless, the stability problem stated above is still meaningful: what conditions would guarantee the boundedness of e_n, f_n, u_n, g_n for all n? If the *model* of a physical process does not establish unique correspondence between the inputs (r_1, r_2) and the other signals, then the model is said to be "ill-posed" *with respect to uniqueness*. This means that, for a given input (r_1, r_2), the model *by itself* is not sufficient for the determination of the actual output observed in the physical process, and that some additional information (data) about the physical process is needed for such a determination. Nonetheless, such a model may still be adequate for the study of the stability of the process, i.e. "well-posed" *with respect to stability*. In this regard, one can give concrete examples of realizable physical processes whose stability properties can be adequately predicted from a model that is ill-posed with respect to uniqueness. In such instances, it can be shown that the additional information about the model which may be needed to establish uniqueness need not invalidate the predictions about the stability of the process. To recapitulate: the problem of stability of solutions can be conceptually separated from the problem of uniqueness of solutions, and there is legitimate practical justification for doing so.

2.3. The linear subsystem H. In subsequent discussion, it will be assumed that the terminal constraint induced by the block H can be expressed in the form,

$$(8)\qquad g(t) = \int_{t_0}^{\infty} u(z)h(t, z)\, dz + \sum_{n=0}^{N} \sum_{m=0}^{M} a_{nm}(t)u^{(n)}(t - b_{nm}(t)) \quad \text{for all } t \in [t_0, \infty),$$

where $h(t, z)$, $a_{nm}(t)$, $b_{nm}(t)$ are given functions; $N, M \in [0, \infty]$ are given integers; and $u^{(n)}(\cdot)$ is the nth derivative of $u(\cdot)$. Note that the relationship in (8) qualifies the subsystem H as a linear operator (mapping), and one may write

$$(9)\qquad g = H(u) = Hu.$$

If $N \geqq 1$, then $g(t)$ depends explicitly on derivatives of $u(t)$, and the system H is said to be *improper*. Such would be the case, for example, if H represents a two-terminal input impedance circuit having an inductance in series with its terminals. If $N = 0$, H is said to be *proper*. The functions $b_{nm}(t)$ represent time-varying displacement (delay or advance) of $u(\cdot)$. If $b_{nm}(t) \geqq 0$ and if $h(t, z) = 0$ whenever $z > t$, then H is said to be *nonanticipative*. Under these conditions the value of $g(t)$

[2] In a forthcoming paper, a sufficient condition for the system of Fig. 1 to exhibit no finite escape time, will be presented. That condition is in fact weaker than the condition stipulated by the stability criterion presented here. Consequently, the a priori requirement of no finite escape time is automatically satisfied.

at any given t does not depend on future values of $u(t)$. Otherwise, H would be described as being *anticipative.*

Remark 1. Ordinarily, the constraint in (8) would include an additive term $g_0(t)$ which solely depends on the initial conditions existing in the system at t_0. It is assumed here that such a term is combined with the input $r_1(t)$. This entails no loss of generality, yet it allows us to treat H as a linear mapping between u and g.

Remark 2. By using generalized functions (distributions), one can define for the system H an equivalent impulse response kernel $\hat{h}(t, z)$, such that the relationship in (8) can be expressed in the compact form,

$$(10) \qquad g(t) = \int_{t_0}^{\infty} u(z)\hat{h}(t, z)\, dz.$$

The expression for $\hat{h}(t, z)$ would then be

$$(11) \qquad \hat{h}(t, z) \equiv \hat{h}_t(z) \equiv h(t, z) + \sum_{n=0}^{N} \sum_{m=0}^{M} (-1)^n a_{nm}(t)\, \delta^{(n)}(z - t + b_{nm}(t)),$$

where $\delta^{(n)}$ is the delta function of order n. The designation of $\hat{h}(t, z)$ as $\hat{h}_t(z)$ is intended to mean that $\hat{h}$ should be regarded as a function of z with t as a parameter.

2.4. The nonlinear characteristic N. The constraint induced by the block N is assumed to be a nonlinear time-varying relationship between $e(t)$ and $f(t)$ of the form

$$(12) \qquad f(t) = n(e(t), t) \quad \text{for all } t \in [t_0, \infty).$$

The relationship in (12) need not represent a unique correspondence between the functions e and f. In other words, the relationship $y = n(x, t)$ is allowed to be multivalued in x. Thus, we are admitting a class of nonlinear characteristics that exhibit hysteresis. We assume that for any fixed t and any $x \in (-\infty, \infty)$ there is at least one value of $y \in (-\infty, \infty)$ that satisfies the given relationship $y = n(x, t)$. Note that under these conditions, the constraint $y = n(x, t)$ does not always qualify as an operator or mapping from $(-\infty, \infty)$ into $(-\infty, \infty)$. No conditions of continuity are stipulated on the relationship $y = n(x, t)$. Furthermore, no sector restriction is imposed on the graph of $y = n(x, t)$, thereby admitting cases where $n(x, t)$ is discontinuous at $x = 0$ and $n(0, t) \neq 0$.

3. Definitions and preliminaries. In this section we introduce the special definitions and symbols to be used in the statement and derivation of the main results. A transformation involving two parameter functions, $k(t)$ and $\beta(t)$, is introduced. The significance and implications of this transformation will be elaborated throughout the rest of the paper.

Let $k(t)$ and $\beta(t)$ be arbitrary real-valued functions defined for all t, with $\beta^{-1}(t)$ well-defined for all t. Consider the systems $N_{k\beta}$, $H_{k\beta}$ and R_k described in Fig. 2. The blocks designated $k(t)$, $\beta(t)$ and $1/\beta(t)$ represent multiplicative time-varying gain constraints. The constraint induced by $N_{k\beta}$ between its terminal signals is

$$(13) \qquad y = \beta[n(e, t) - ke] \quad \text{for all } t \in [t_0, \infty).$$

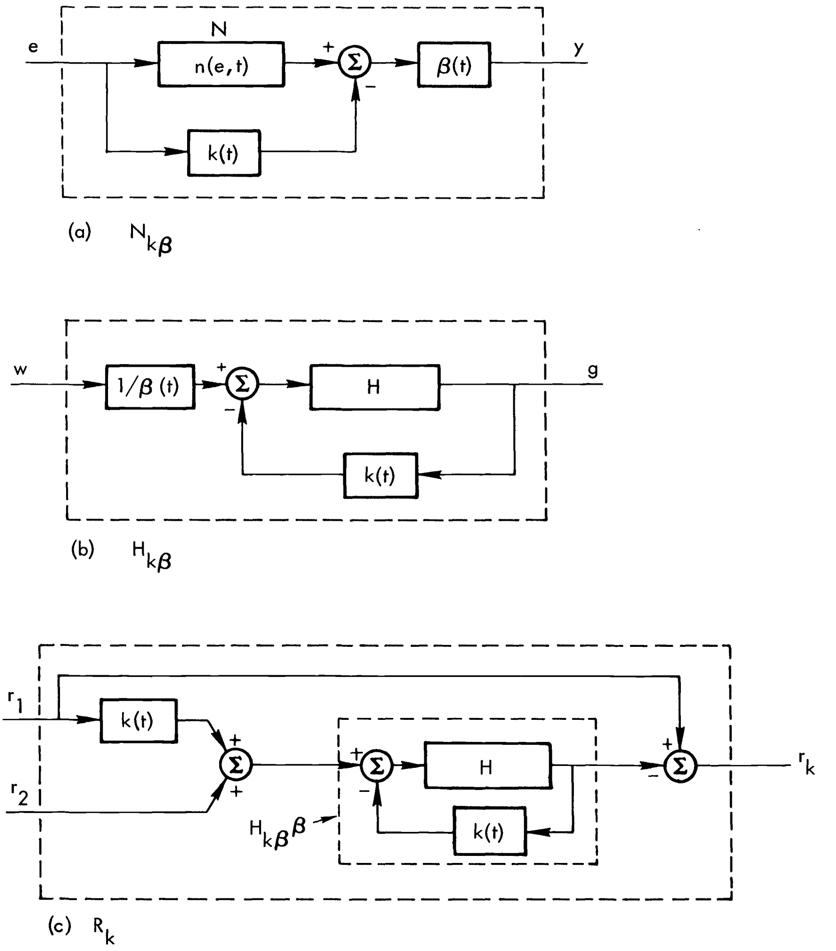


FIG. 2. *The systems* $N_{k\beta}$, $H_{k\beta}$ *and* R_k

Hence, $N_{k\beta}$ may be characterized by a nonlinear relationship $n_{k\beta}$ defined as

$$(14) \qquad n_{k\beta}(x, t) \equiv \beta[n(x, t) - kx].$$

The constraint induced by $H_{k\beta}$ is

$$(15) \qquad g = H\left(\frac{w}{\beta} - kg\right) \quad \text{for all } t \in [t_0, \infty),$$

where H is the linear operator described in § 2.3 and specified by (8) and (9). The constraint induced by R_k among its terminal signals r_1, r_2, and r_k is

$$(16) \qquad r_k = r_1 - H(r_2 + kr_k).$$

The stability conditions, to be discussed in the following section, will be expressed in terms of simple parameters obtained from the systems $H_{k\beta}$, $N_{k\beta}$ and R_k.

Although the constraints for $H_{k\beta}$, $N_{k\beta}$, and R_k are well-defined for an arbitrary $k(t)$ and an arbitrary invertible $\beta(t)$, we shall find it necessary for our purposes to restrict the choice of the pair (k, β) by a number of conditions. These conditions are designed to force the constraints of $H_{k\beta}$ and $N_{k\beta}$ to have certain properties which will be essential in the derivation of the stability results. The restrictions on $k(t)$ and $\beta(t)$ are stated in the following definition.

Definition. A pair of functions (k, β) is said to *belong to the class* C, viz. $(k, \beta) \in C$, if the following conditions C_1, C_2, and C_3 are satisfied:

(C_1) If one can find functions $h_{k\beta}(t, z)$, $a_{k\beta m}(t)$, $b_{k\beta m}(t) \geqq 0$, such that the constraint (15) for $H_{k\beta}$ can be expressed as

$$g(t) = H_{k\beta}w = \int_{t_0}^{t} w(z)h_{k\beta}(t, z)\,dz + \sum_m a_{k\beta m}(t)w(t - b_{k\beta m}(t)) \qquad \text{for all } t \in [t_0, \infty), \tag{17}$$

i.e., such that every pair (w, g) that satisfies (15) also satisfies (17).

(C_2) If the function $A_{k\beta}(t)$, defined by

$$A_{k\beta}(t) \equiv \int_{t_0}^{t} |h_{k\beta}(t, z)|\,dz + \sum_m |a_{k\beta m}(t)| \quad \text{for all } t \in [t_0, \infty), \tag{18}$$

is bounded.

(C_3) If the nonlinear characteristic $n_{k\beta}(x, t)$, defined in (14), satisfies the following property:

$$\bar{n}_{k\beta}(\lambda) \equiv \sup_{\substack{|x| \leqq \lambda \\ t \in [t_0, \infty)}} |n_{k\beta}(x, t)| < \infty \quad \text{for all } 0 \leqq \lambda < \infty. \tag{19}$$

In subsequent discussions, the condition $(k, \beta) \in C$ will be used to imply the three conditions C_1, C_2, and C_3. The significance of each of these conditions will be thoroughly discussed in conjunction with the stability results to be presented in § 4. For the time being, note that condition C_1 implies that the system $H_{k\beta}$ is a linear operator, and the constraint in (17) may be symbolically written as

$$g = H_{k\beta}(w) = H_{k\beta}w \quad \text{for all } t \in [t_0, \infty). \tag{20}$$

This, in turn, implies that the system R_k is a linear operator which maps any given (r_1, r_2) into a unique r_k according to the relationship

$$r_k = r_1 - H_{k\beta}\beta(kr_1 + r_2) \quad \text{for all } t \in [t_0, \infty). \tag{21}$$

This can be verified by comparing R_k with $H_{k\beta}$ in Fig. 2, and noting that $H_{k\beta}\beta$ is in fact identical to the feedback subsystem appearing in R_k. Note that $H_{k\beta}\beta$ is independent of β, and therefore r_k is independent of β.

Finally, we define $G_{k\beta}(x, t)$ in terms of $n_{k\beta}(x, t)$ as follows:

$$G_{k\beta}(x, t) \equiv |n_{k\beta}(x, t)/x| \quad \text{for all } x \neq 0. \tag{22}$$

Note that if $n_{k\beta}$ is multivalued in x, then $G_{k\beta}$ would also be multivalued in x.

The functions $A_{k\beta}(t)$, $G_{k\beta}(x, t)$ and $r_k(t)$, as defined in (18), (22), and (21) respectively, will be the key parameters in the stability criteria to be presented in the following section. For a given (k, β), the functions $A_{k\beta}(t)$ and $G_{k\beta}(x, t)$ may be described as the *instantaneous gains* for $H_{k\beta}$ and $N_{k\beta}$ respectively. The signal $r_k(t)$, which depends on the inputs $r_1(t)$, $r_2(t)$, and on $k(t)$, may be described as the system "*effective input*" corresponding to the specific choice of $k(t)$.

Remark 3. The function $A_{k\beta}(t)$, defined in (18), depends on t_0, and one may appropriately write $A_{k\beta}(t, t_0)$.

4. Main results. The theorem which follows provides a sufficient condition for the boundedness of the signal $e(t)$ in any solution $s = (r_1, r_2, e, f, u, g) \in Y$. The condition involves an inequality in terms of the functions $A_{k\beta}(t)$ and $G_{k\beta}(x, t)$ which must hold for *large values* of x and t. Conditions for the boundedness of the other output signals, namely, $f(t)$, $u(t)$, and $g(t)$, are also discussed. Proofs of the results will be given in a later section on derivations.

4.1. Boundedness of $e(t)$.

THEOREM. *Let* $s = (r_1, r_2, e, f, u, g) \in Y$ *be a solution to* S. *Let* $(k, \beta) \in C$ *such that*

$$r_k(t) \quad \textit{is bounded}, \tag{23}$$

$$A_{k\beta}(t)G_{k\beta}(x, t) \leqq 1 - \varepsilon \quad \textit{for all } t \geqq T \quad \textit{and} \quad |x| > X, \tag{24}$$

where $\varepsilon > 0$, $X \geqq 0$, *and* $T \geqq t_0$ *are some (finite) numbers. Then the signal* $e(t)$ *is bounded.*

The theorem states that the boundedness of $e(t)$ can be inferred from two bits of a priori information: condition (23) which relates to the "input" signals r_1 and r_2, and condition (24) which relates to the system constraints. This is consistent with our formulation of the stability problem discussed in § 2.2. Note that the functions $A_{k\beta}(t)$, $r_k(t)$, and $G_{k\beta}(x, t)$, as defined in (18), (21), and (22), depend on the specific choice of the parameters (k, β). Therefore, in applying the theorem, one should choose $(k, \beta) \in C$ first and then test conditions (23) and (24). It should be emphasized that the boundedness of $r_k(t)$ in condition (23) does not necessarily imply the boundedness of r_1 and r_2. In other words, the output signal $e(t)$ may be bounded even though r_1 or r_2 may be unbounded. This will be illustrated by a specific example. On the other hand, one may consider the function $r_k(t)$, as given in (21), to be the system "effective input" relative to the specific choice of $k(t)$. With this terminology, and with $e(t)$ considered as output, the theorem may be described as a criterion for "bounded-input, bounded-output" stability.

The inequality in condition (24) should be satisfied with some specific choice for the numbers $\varepsilon > 0$, $T \geqq t_0$, and $X \geqq 0$. Note that if condition (24) is satisfied with some specific $\varepsilon = \varepsilon_1$, $T = T_1$, $X = X_1$, then the condition would also be satisfied with *any* $\varepsilon < \varepsilon_1$, any $T > T_1$, and any $X > X_1$. In other words, if the condition is to be satisfied at all, it should be satisfied with arbitrarily small values of $\varepsilon > 0$, and with arbitrarily large values of T and X. This observation leads to the following result.

COROLLARY. *Let* $s = (r_1, r_2, e, f, u, g) \in Y$ *be a solution to* S. *Let* $(k, \beta) \in C$ *such that*

$$r_k(t) \quad \textit{is bounded}, \tag{25}$$

$$F_{k\beta} \equiv \max\left\{\overline{\lim_{x\to\infty}}\ \overline{\lim_{t\to\infty}}\ A_{k\beta}(t)G_{k\beta}(x, t),\ \overline{\lim_{x\to-\infty}}\ \overline{\lim_{t\to\infty}}\ A_{k\beta}(t)G_{k\beta}(x, t)\right\} < 1. \tag{26}$$

Then $e(t)$ *is bounded.*

The symbol $\overline{\lim}$ denotes the limit superior. Condition (26) does not require the determination of specific numbers such as ε, T and X of condition (24), and therefore may be easier to apply in those cases where the evaluation of the limits is not complicated.

Remark 4. If $G_{k\beta}$ is multivalued at some $x = \hat{x}$, then $G_{k\beta}(\hat{x}, t)$ in (24) and (26) should be understood to represent the largest value taken by $G_{k\beta}$ at $(\hat{x}, t)$.

4.2. Graphical interpretation. The condition of stability stated in (24) may be given a simple graphical interpretation or visualization in terms of the graph of the given nonlinearity $n(x, t)$. By substituting into (24) the expressions for $G_{k\beta}(x, t)$ and $n_{k\beta}(x, t)$, as given in (22) and (14) respectively, one obtains,

$$|\beta(t)[n(x,t) - k(t)x]| \leqq [(1 - \varepsilon)/A_{k\beta}(t)]|x| \quad \text{for all } t \geqq T, \quad |x| > X. \tag{27}$$

Since $|x| > X$ implies $x > X$ or $x < -X$, one has,

$$\begin{aligned}[k(t) - (1 - \varepsilon)/|\beta(t)|A_{k\beta}(t)]x &\leqq n(x,t) \\ &\leqq [k(t) + (1 - \varepsilon)/|\beta(t)|A_{k\beta}(t)]x \quad \text{for all } t \geqq T, \quad x > X,\end{aligned} \tag{28}$$

$$\begin{aligned}[k(t) + (1 - \varepsilon)/|\beta(t)|A_{k\beta}(t)]x &\leqq n(x,t) \\ &\leqq [k(t) - (1 - \varepsilon)/|\beta(t)|A_{k\beta}(t)]x \quad \text{for all } t \geqq T, \quad x < -X.\end{aligned} \tag{29}$$

The conditions (28) and (29) are graphically represented in Fig. 3. The nonlinear characteristic $n(x, t)$ is shown plotted as a function of x for some specific value of $t \geqq T$. In effect, the condition of the theorem stipulates that the outer portions of the graph of the nonlinearity (i.e. the portions of $n(x, t)$ for $|x| > X$) should belong to the shaded region $S_{k\beta}$. Note that, in general, the region $S_{k\beta}$ and the nonlinearity $n(x, t)$ are both time-varying, and that the conditions depicted in Fig. 3 are required

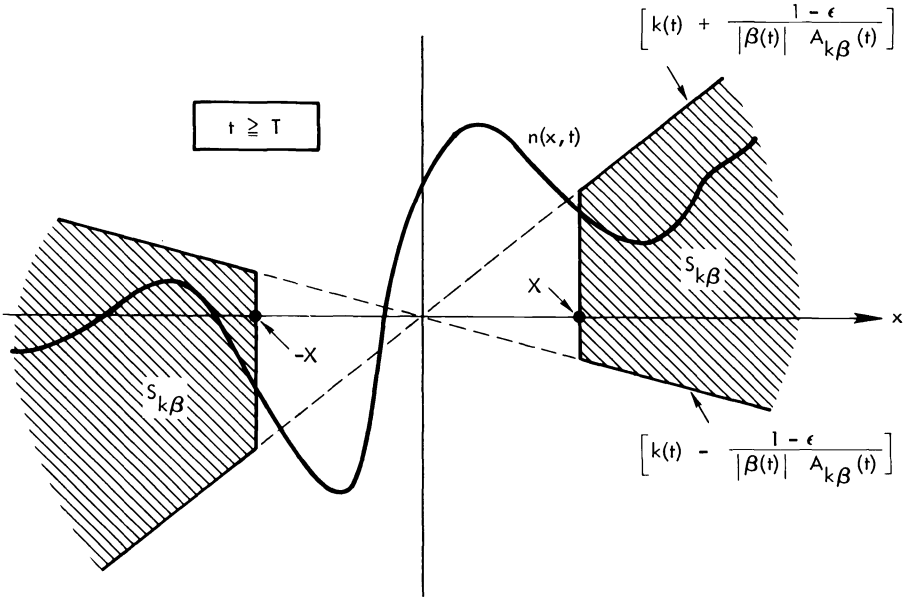

$\left[k(t) + \dfrac{1-\epsilon}{|\beta(t)|\ A_{k\beta}(t)}\right] x$
$t \geqq T$
$n(x,t)$
$S_{k\beta}$
X
x
$-X$
$S_{k\beta}$
$\left[k(t) - \dfrac{1-\epsilon}{|\beta(t)|\ A_{k\beta}(t)}\right] x$

FIG. 3. *Graphical representation of the criterion*

to hold only for $t \geqq T$. Recall also that T and X can be arbitrarily large. Thus, the theorem states that the waveform $e(t)$ is bounded if the *outer* portions of the graph of the nonlinearity *eventually* enter the region $S_{k\beta}$.

4.3. The condition $(k,\beta) \in C$. The three conditions which define the allowable class of the parameter functions $k(t)$ and $\beta(t)$, namely, C_1, C_2, and C_3, were presented in § 3 in a rather ad hoc fashion. In this section we shall elaborate on the implications of these conditions.

The first condition, namely C_1, stipulates that (k, β) should be chosen such that the resulting system $H_{k\beta}$ is characterized by the constraint

$$g(t) = \int_{t_0}^{t} w(z)h_{k\beta}(t, z)\,dz + \sum_m a_{k\beta m}(t)w(t - b_{k\beta m}(t)), \tag{30}$$

with $b_{km}(t) \geqq 0$. This implies that $H_{k\beta}$ should qualify as a *proper, causal, linear operator.* It should be emphasized that this requirement does not necessarily restrict the given linear system H, which is characterized by (8), to be proper or causal.

In the second condition, namely C_2, the stipulation that $A_{k\beta}(t)$ be bounded is equivalent to the requirement that $H_{k\beta}$ be "bounded-input, bounded-output stable," i.e. if $w(t)$ is bounded then $g(t)$ is bounded, and conversely [8]. This does not necessarily restrict the given system H to be bounded-input, bounded-output stable.

The third condition, as expressed in (19), implies that the nonlinear characteristic $n_{k\beta}(x, t)$ should be a bounded function of t for any finite value of x. This condition need not be satisfied by the given nonlinear characteristic $n(x, t)$.

4.4. Boundedness of $f(t)$, $u(t)$, **and** $g(t)$. Having established the boundedness of $e(t)$, the boundedness of the other "output" signals may be verified by examining the relationship of these signals to $e(t)$ and the input signals r_1 and r_2:

$$f(t) = n(e(t), t), \tag{31}$$

$$u(t) = r_2(t) + f(t) = r_2 + n(e(t), t), \tag{32}$$

$$g(t) = r_1(t) - e(t). \tag{33}$$

The boundedness of $f(t)$ may be deduced from (31). If, in addition, $r_2(t)$ is known to be bounded, then $u(t)$ is bounded. If $r_1(t)$ is bounded, then (33) implies g is bounded. It should be emphasized that any such additional conditions that might be needed to establish the boundedness of $f(t)$, $u(t)$ or $g(t)$ are not necessary for the boundedness of $e(t)$. It is possible for $e(t)$ to be bounded when other signals in the system are unbounded.

5. Examples. In this section, the application of the results is illustrated by a number of examples. The constraint for the linear system H or $H_{k\beta}$ will be specified in the form of a differential equation. The standard methods for obtaining the impulse response (Green's function) can then be employed to express the constraint in the integral form of (17). As mentioned earlier, the unforced (zero-input) response in H or H_k due to initial conditions will be lumped with the input r_1. Recall that we consider solutions $s \in Y$, i.e., we assume that the signals exhibit no finite escape time.[2] In some of the examples, the Bellman–Gronwall inequality [9] may be used to *verify* that $s \in Y$ if r_1 and $r_2 \in W$. (See also [10] and [11].)

In Example 1, the system H is specified, and the results are used to obtain general conditions on the nonlinearity under which the boundedness of $e(t)$ can be guaranteed. In Example 2, a *stationary* linear system H is considered, and the present criteria give better results than the circle criterion. In Example 3, an improper system H is considered. In Example 4, a system with periodically time-varying coefficients is considered.

Example 1. Let H be characterized by the differential equation

$$t^2\ddot{g} + 4t\dot{g} + (t^2 + 1)g = u, \qquad t \geqq t_0 > 0. \tag{34}$$

The nonlinear characteristic $y = n(x, t)$ is not specified, but is assumed to satisfy the following "sector" restriction:

$$|n(x,t)| \leqq (a + bt)|x|^p + ct + d; \quad a, b, c, d > 0, \quad 0 < p \leqq 1. \tag{35}$$

Figure 4 shows a graphical representation of this condition, which is assumed to hold for all $t \geqq t_0 > 0$. For $p = 1$, the symmetrical time-varying sector would

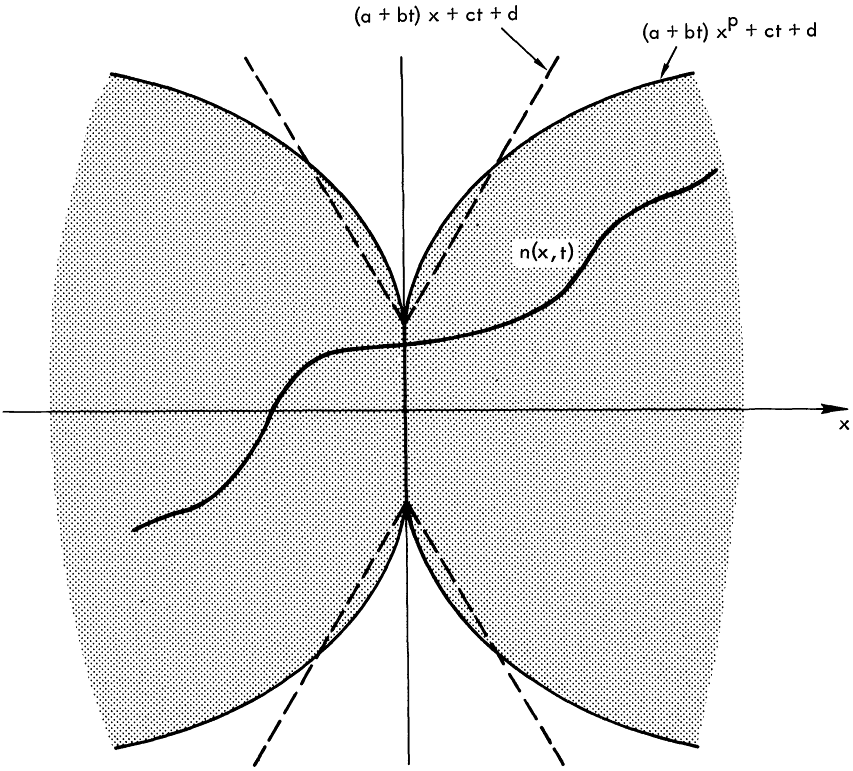


FIG. 4. *Condition on nonlinearity of Example* 1

have straight boundaries. The inputs r_1 and r_2 are assumed to satisfy the condition

$$(36) \qquad |r_1(t)| < B_1, \qquad |r_2(t)| < B_0 + B_2 t, \qquad B_0, B_1, B_2 > 0.$$

It is required to specify conditions on the magnitudes of the constants a, b, c, d, p, B_0, B_1, and B_2, under which the boundedness of $e(t)$ can be guaranteed.

In the corollary of § 4.1, let $(k, \beta) = (1, 1/t)$.[3] For this choice of (k, β) the system $H_{k\beta}$ is governed by the following constraint:

$$t^2\ddot{g} + 4t\dot{g} + (t^2 + 1)g = .w/\beta - kg = tw - g,$$

$$(37) \qquad t\ddot{g} + 4\dot{g} + \left(\frac{t^2 + 2}{t}\right)g = w, \qquad t \geqq t_0 > 0.$$

$$(38) \qquad g(t) = \int_{t_0}^{t} w(z)t^{-2}z \sin(t - z)\, dz, \qquad t \geqq t_0 > 0.$$

The step from (37) to (38) can be verified by the standard methods of obtaining the impulse response (Green's function) for a differential equation [8]. Now, we verify that the condition $(k, \beta) = (1, 1/t) \in C$ is satisfied. Equation (38) implies that condition C_1 is satisfied with

$$h_{k\beta}(t, z) = t^{-2}z \sin(t - z), \qquad a_{k\beta m}(t) = 0, \quad b_{k\beta m}(t) = 0.$$

Furthermore, condition C_2 is satisfied since $A_{k\beta}(t)$ is bounded:

$$(39) \qquad A_{k\beta}(t) = \int_{t_0}^{t} |t^{-2}z \sin(t - z)|\, dz < \frac{1}{2} \quad \text{for all } t \geqq t_0 > 0.$$

To verify condition C_3, note that from (14) and (35) one obtains,

$$(40) \quad |n_{k\beta}(x, t)| = \left|\frac{n(x, t) - x}{t}\right| \leqq \frac{a + bt}{t}|x|^p + c + \frac{d}{t} + \frac{|x|}{t} \quad \text{for all } t \geqq t_0 > 0,$$

and condition C_3, as expressed in (19), is satisfied.

Next, we examine condition (25). From (21) and from (38) which defines the operator $H_{k\beta}(\cdot)$, one has,

$$(41) \quad r_k = r_1 - H_{k\beta}\left(\frac{r_1 + r_2}{t}\right) = r_1(t) - \int_{t_0}^{t} \frac{r_1(z) + r_2(z)}{z} \cdot t^{-2}z \sin(t - z)\, dz.$$

Combining (36) with (41), one obtains,

$$|r_k(t)| < B_1 + B_2 + \frac{(B_0 + B_1)}{t}, \qquad t \geqq t_0 > 0.$$

[3] For β to be well-defined for all t, let $\beta(t) = 1$ for $t = 0 < t_0$. Since the system at hand involves no delay, the exact nature of the function $\beta(t)$ for $t < t_0$ is inconsequential.

Thus, $r_k(t)$ is bounded for *arbitrary* values of B_0, B_1, and B_2. Finally, we examine condition (26). From (39) and (40), one has,

$$A_{k\beta}(t)G_{k\beta}(x,t) < \frac{1}{2}\left[\frac{a+bt}{t}|x|^{p-1} + \frac{c}{|x|} + \frac{d}{t|x|} + \frac{1}{t}\right], \qquad x \neq 0,$$

$$\overline{\lim_{t \to \propto}}\, A_{k\beta}(t)G_{k\beta}(x,t) \leqq \frac{1}{2}\frac{b}{|x|^{1-p}} + \frac{c}{|x|}.$$

Thus, $F_{k\beta} = 0$ if $p < 1$, $F_{k\beta} \leqq b/2$ if $p = 1$.

This leads us to the following conclusion: if in (35), $p < 1$, $e(t)$ is bounded for arbitrary values of a, b, c, d, B_0, B_1, and B_2; if $p = 1$, $e(t)$ is bounded for $b < 2$ and arbitrary values of a, c, d, B_0, B_1, and B_2.

Example 2. Let H be the stationary system governed by the equation

$$\dddot{g} + 3\ddot{g} + 3\dot{g} + g = \dot{u} - u, \qquad t \geqq t_0 = 0.$$

Let $n(x,t)$ be given as

$$n(x,t) = K\frac{x(1+\cos x)(\sin t)}{10|\cos t| + 0.1}, \qquad K > 0.$$

The inputs r_1 and r_2 are assumed to be bounded. It is required to find the largest value of K for which the boundedness of $e(t)$ is guaranteed. Since the subsystem H is stationary, the circle criterion is applicable to the system at hand [9]. The circle criterion gives $K = 0.05$ as the maximum allowable value for which boundedness is guaranteed. The results of this paper give the substantially larger value of $K = 1.10$.

Letting $k(t) = 0$, $\beta(t) = 10|\cos t| + 0.1$, the nonlinear characteristic $n_{k\beta}$ and the impulse response $h_{k\beta}$ of $H_{k\beta}$ are given by

(42) $$n_{k\beta}(x,t) = Kx(1+\cos x)\sin t,$$

(43) $$h_{k\beta}(t,z) = \frac{(t-z)(1+z-t)\,e^{z-t}}{10|\cos z| + 0.1}.$$

Figure 5 shows a plot of $A_{k\beta}(t)$ versus t as evaluated on the digital computer. Observe that[4]

(44) $$A_{k\beta}(t) < 0.45, \qquad t \geqq 50.$$

Referring to the conditions stated in the theorem of § 4.1, it is easily verified that $(k, \beta) \in C$ and that r_k is bounded. Furthermore, from (42) and (44) one has,

$$A_{k\beta}(t)G_{k\beta}(x,t) < 0.45K(1+\cos x)|\sin t| < 0.9K$$

for all $|x| > 0$, $t \geqq 50$. Thus, if $K \leqq 1.1$, one has,

$$A_{k\beta}(t)G_{k\beta}(x,t) < 1 - 0.01 \quad \text{for all } |x| > 0, \quad t \geqq 50.$$

Thus, condition (24) is satisfied with $\varepsilon = 0.01$, $X = 0$, $T = 50$. Hence, all the conditions of the theorem are satisfied, and $e(t)$ is bounded.

[4] The use of the digital computer to verify (44) is not intended as a method of proof. The result in (44) can be *proven* by rather cumbersome manipulations of the expression for $A_{k\beta}(t)$.

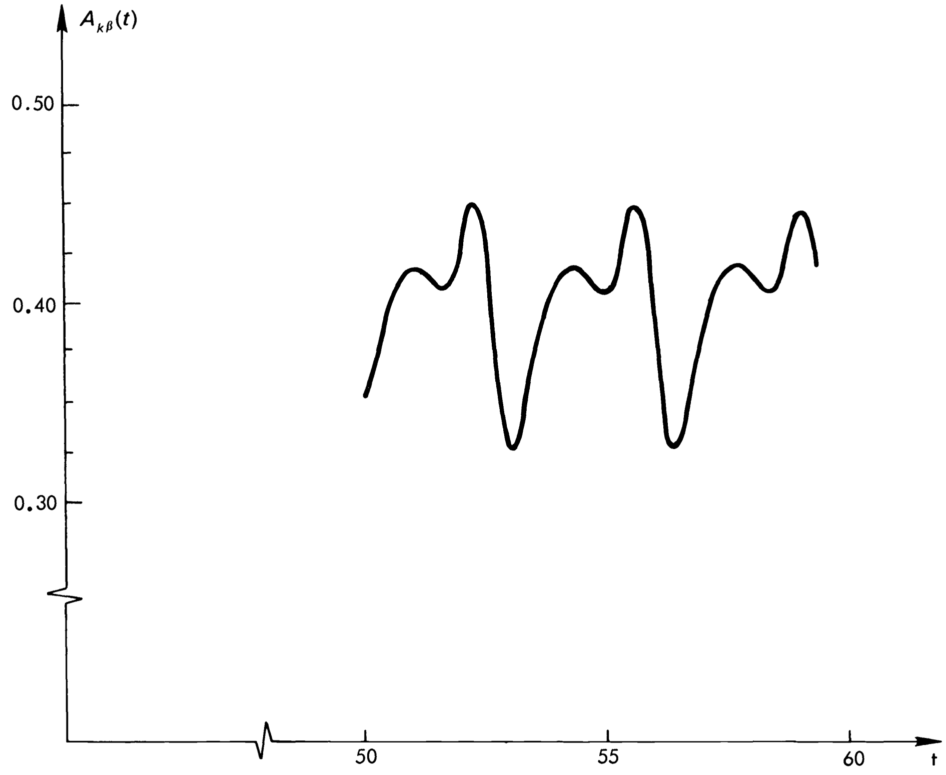


FIG. 5. *The function $A_{k\beta}(t)$ of Example* 2

Example 3. Let H be the *improper* system governed by

$$4t\dot{g} + 2g = t^2\ddot{u}, \qquad t \geqq t_0 > 0.$$

Letting $(k, \beta) = (1, 1)$, the constraint for $H_{k\beta}$ is

$$t^2\ddot{g} + 4t\dot{g} + 2g = t^2\ddot{w}, \qquad t \geqq t_0 > 0,$$

$$g(t) = \int_{t_0}^{t} w(z)\frac{2t - 6z}{t^2}\,dz + w(t).$$

Thus, condition C_1 is satisfied with $h_{k\beta}(t, z) = (2t - 6z)/t^2$, $m = 1$, $a_{k\beta 1}(t) = 1$, $b_{k\beta 1} = 0$. Condition C_2 is satisfied since

$$A_{k\beta}(t) = \int_{t_0}^{t} \frac{(2t - 6z)}{t^2}\,dz + 1 < \frac{8}{3}, \qquad t \geqq \tfrac{3}{2}t_0 > 0. \tag{45}$$

Assume that the nonlinearity $n(x, t)$ satisfies condition C_3. Assuming that r_1 and r_2 are bounded, (45) implies that r_k is bounded. Furthermore, (45) implies that condition (24) would be satisfied for any $n(x, t)$ such that

$$G_{k\beta}(x, t) = \left|\frac{n(x, t) - x}{x}\right| \leqq \frac{3}{8} - \varepsilon \quad \text{for all } t \geqq T \geqq \tfrac{3}{2}t_0, \quad |x| > X. \tag{46}$$

Thus, for any $n(x, t)$ that satisfies (46), $e(t)$ is bounded. Recall that T and X can be arbitrarily large, i.e. condition (46) places a restriction on the behavior of $n(x, t)$ for large values of x and t.

Example 4. Let H be governed by the integro-differential constraint

$$\dot{g} + g = e^{-t}(\sin^2 t) \int_{t_0}^{t} u(x)\, e^x\, dx \quad \text{for all } t \geqq t_0. \tag{47}$$

With $(k, \beta) = (0, 1)$, one obtains

$$h_{k\beta}(t, z) = e^{-t}e^{z} \int_{z}^{t} \sin^2 x\, dx, \qquad a_{k\beta m}(t) = 0,$$

$$A_{k\beta}(t) < (1/5)(\sin^2 t + 2 - \sin 2t) \quad \text{for all } t \geqq t_0.$$

Thus, $e(t)$ would be bounded if r_1 and r_2 are bounded, and if the nonlinearity satisfies condition C_3 and the constraint

$$|n(x, t)| < \frac{5(1 - \varepsilon)}{\sin^2 t + 2 - \sin 2t}|x| \quad \text{for all } t \geqq T, \quad |x| > X,$$

for some $T \geqq t_0$ and $X > 0$. It should be noted that the constraint for H in (47) can be expressed as a second order differential equation with periodic coefficients:

$$(\sin t)\ddot{g} + 2(\sin t - \cos t)\dot{g} + (\sin t - 2\cos t)g = (\sin^3 t)u. \tag{48}$$

Any pair (u, g) which satisfies (47) also satisfies (48).

6. Derivations. Let $s = (r_1, r_2, e, f, u, g) \in Y$ be a solution to S. Let (k, β) be a pair of functions for which the hypothesis of the theorem is satisfied. Recall that the condition $(k, \beta) \in C$ implies that conditions C_1, C_2, and C_3 are satisfied. Define the function $y(t)$ as

$$y \equiv \beta(f - ke). \tag{49}$$

The function $y(t)$ will play a primary role in proving the boundedness of $e(t)$.

The proof of the theorem will be organized and presented in four steps:

(a) *Step* 1. We will show that the functions $e(t)$ and $y(t)$ satisfy the following constraints:

$$y = n_{k\beta}(e, t) \quad \text{for all } t \geqq t_0, \tag{50}$$

$$e = r_k - H_{k\beta}(y) \quad t \geqq t_0. \tag{51}$$

(b) *Step* 2. We will assume that the function $y(t)$ is unbounded. Based on this assumption, a basic property of the function $y(t)$ will be derived. This property is true for a broad class of unbounded functions, and will therefore be stated and proven as a general result in the form of a lemma.

(c) *Step* 3. The results of Steps 1 and 2 are then used to deduce a contradiction to the hypothesis of the theorem. This implies that the assumption of unboundedness of $y(t)$ is not valid.

(d) *Step* 4. The boundedness of $e(t)$ is then deduced from the boundedness of $y(t)$ and the constraint (51) of Step 1.

Step 1. The validity of (50) follows directly from the definition of $n_{k\beta}$ in (14):

$$n_{k\beta}(e, t) = \beta[n(e, t) - ke] = \beta(f - ke) = y.$$

To prove (51), note first that

$$u = (y + \beta kr_1 + \beta r_2)/(\beta) - kg. \tag{52}$$

This follows from (49), (2), and (3). Letting $w \equiv (y + \beta kr_1 + \beta r_2)$, one has,

$$g = H(u) = H\left(\frac{w}{\beta} - kg\right), \tag{53}$$

$$g = H_{k\beta}(w). \tag{54}$$

The step from (53) to (54) is a consequence of condition C_1 implied by the hypothesis $(k, \beta) \in C$. Finally, from (2), (54), and (21), we obtain the required result:

$$e = r_1 - g = r_1 - H_{k\beta}(w) = r_1 - H_{k\beta}(\beta kr_1 + \beta r_2 + y)$$
$$= r_1 - H_{k\beta}\beta(kr_1 + r_2) - H_{k\beta}(y),$$

$$e = r_k - H_{k\beta}(y). \tag{55}$$

Step 2. We now show that $y \in W$. Given any $T \in [t_0, \infty)$, define $\lambda \equiv \sup_{[t_0,T]} |e(t)|$. Note that $s \in Y$ implies $e \in W$, and therefore $\lambda < \infty$. From (50), one has,

$$\sup_{[t_0,T]} |y(t)| = \sup_{t\in[t_0,T]} |n_{k\beta}(e(t), t)| \leqq \sup_{\substack{t\in[t_0,T]\\ |x|\leqq\lambda}} |n_{k\beta}(x, t)| \leqq \bar{n}_{k\beta}(\lambda) < \infty, \tag{56}$$

where the last two inequalities in (56) follow from condition C_3 stated in (19). Hence $y \in W$. Having established that $y(t)$ exhibits no finite escape time, we now *assume* $y(t)$ to be unbounded on $[t_0, \infty)$. This assumption leads to a contradiction, as will be demonstrated in Step 3. But first, we need the following result, which applies to $y(t)$, and to any unbounded function in W.

LEMMA. *Let $y(t) \in W$ be an unbounded function defined over $[t_0, \infty)$. Given any $T \geqq t_0$, any $\delta > 0$, and any M, one can find an instant $\tau \geqq T$ such that*

$$|y(\tau)| > \sup_{[t_0,\tau]} |y(t)| - \delta, \qquad \tau \geqq T, \tag{57}$$

$$|y(\tau)| > M, \qquad \tau \geqq T. \tag{58}$$

Proof. For convenience, we introduce the notation

$$\bar{y}(a, b) \equiv \sup_{[a,b]} |y(t)|, \qquad a, b \in [t_0, \infty).$$

Note that the hypothesis $y \in W$ implies that $\bar{y}(a, b) < \infty$ whenever $a, b \in [t_0, \infty)$. Since $y(t)$ is unbounded, one can find an instant $t_1 > T$ such that $|y(t_1)|$ is larger than $M + \delta$ and larger than $\bar{y}(t_0, T)$:

$$|y(t_1)| > \bar{y}(t_0, T), \qquad t_1 > T, \tag{59}$$

$$|y(t_1)| > M + \delta, \qquad t_1 > T. \tag{60}$$

From the definition of supremum, it follows that,

$$\bar{y}(t_0, t_1) = \max\{\bar{y}(t_0, T), \bar{y}(T, t_1)\}. \tag{61}$$

From (59) one has,

$$\bar{y}(t_0, T) < |y(t_1)| \leqq \bar{y}(T, t_1). \tag{62}$$

Combining (62) with (61), one obtains,

$$\bar{y}(t_0, t_1) = \bar{y}(T, t_1). \tag{63}$$

Since $\bar{y}(T, t_1)$ is the supremum of $|y(t)|$ over $[T, t_1]$, one can find a $\tau \in [T, t_1]$ such that,

$$|y(\tau)| > \bar{y}(T, t_1) - \delta, \qquad \tau \in [T, t_1]. \tag{64}$$

Substituting (63) into (64), and noting that $\bar{y}(t_0, t_1) \geqq \bar{y}(t_0, \tau)$, one obtains the first required relationship,

$$|y(\tau)| > \bar{y}(t_0, \tau) - \delta.$$

From (64), (62), and (60), one obtains the second required relationship,

$$|y(\tau)| > |y(t_1)| - \delta > M + \delta - \delta = M.$$

Step 3. It is now shown that the assumption in Step 2, namely, "$y(t)$ is unbounded," entails a contradiction to the hypothesis of the theorem in (24). This will be demonstrated by producing *specific* values $t = \tau$ and $x \equiv \hat{x}$, such that

$$A_{k\beta}(\tau)G_{k\beta}(\hat{x}, \tau) > 1 - \varepsilon, \qquad \tau \geqq T, \quad |\hat{x}| > X,$$

which is contradictory to (24).

From (51) and (17) one has,

$$e(t) = r_k(t) - \int_{t_0}^{t} y(z)h_{k\beta}(t, z)\,dz - \sum_m a_{k\beta m}(t)y(t - b_{k\beta m}(t)), \qquad t \geqq t_0;$$

$$\begin{aligned} |e(t)| \leqq |r_k(t)| &+ (\sup_{[t_0,t]} |y(z)|) \int_{t_0}^{t} |h_{k\beta}(t, z)|\,dz \\ &+ \sum_m |a_{k\beta m}(t)|\,|y(t - b_{k\beta m}(t))|, \qquad t \geqq t_0. \end{aligned} \tag{65}$$

We now show that for any $t \geqq t_0$, one has,

$$|y(t - b_{k\beta m}(t))| \leqq \sup_{[t_0,t]} |y(z)| \quad \text{for all } t \geqq t_0. \tag{66}$$

Recall from condition C_1 that $b_{k\beta m}(t) \geqq 0$, therefore $(t - b_{k\beta m}(t)) \in [-\infty, t]$. Furthermore, from (49) and (1), $y(z) = 0$ for all $z < t_0$. Consequently,

$$|y(t - b_{k\beta m}(t))| \leqq \sup_{(-\infty,t]} |y(z)| = \sup_{[t_0,t]} |y(z)| \quad \text{for all } t \geqq t_0,$$

and (66) is valid. Substituting (66) into (65), and using the expression of $A_{k\beta}(t)$ from (18), one obtains,

$$|e(t)| \leqq |r_k(t)| + (\sup_{[t_0,t]} |y(z)|)A_{k\beta}(t) \quad \text{for all } t \geqq t_0. \tag{67}$$

By hypothesis, $r_k(t)$ and $A_{k\beta}(t)$ are bounded functions. Let $R > 0$ and $A > 0$ represent some finite bounds on r_k and $A_{k\beta}$ respectively, i.e.,

$$|r_k(t)| < R, \qquad A_{k\beta}(t) < A \quad \text{for all } t \geqq t_0. \tag{68}$$

Consider the (finite) numbers $\varepsilon > 0$ and $X \geqq 0$ specified in the hypothesis of the theorem, and define new numbers L, δ, and M, as follows:

$$L \equiv \max(X, 2R/\varepsilon); \tag{69}$$

$$\delta \equiv (\varepsilon L)/(2A); \tag{70}$$

$$M \equiv \bar{n}_{k\beta}(L) \equiv \sup_{\substack{|x| \leqq L \\ t \in [t_0, \infty)}} |n_{k\beta}(x, t)|. \tag{71}$$

Note that $L < \infty$, $\delta > 0$, and $M < \infty$ as a consequence of condition C_3 expressed in (19). Now consider the results of the lemma in Step 2, with T being the number specified in the hypothesis of the theorem, δ and M being the numbers specified in (70) and (71) above: one can find an instant τ such that,

$$\tau \geqq T \geqq t_0, \tag{72}$$

$$|y(\tau)| > M, \tag{73}$$

$$|y(\tau)| > \sup_{[t_0, \tau]} |y(z)| - \delta. \tag{74}$$

The inequality (67), which holds for all $t \geqq t_0$, is now written for the *specific* value $t = \tau \geqq t_0$:

$$|e(\tau)| \leqq |r_k(\tau)| + (\sup_{[t_0,\tau]} |y(z)|) A_{k\beta}(\tau),$$

$$|e(\tau)| < R + (|y(\tau)| + \delta) A_{k\beta}(\tau), \tag{75}$$

where the last step follows from (68) and (74). Recalling from (50) that $y(\tau) = n_{k\beta}(e(\tau), \tau)$, and introducing for convenience the notation $\hat{x} \equiv e(\tau)$, one has,

$$|\hat{x}| < R + (|n_{k\beta}(\hat{x}, \tau)| + \delta) A_{k\beta}(\tau). \tag{76}$$

In order to divide both sides of (76) by $|\hat{x}|$, we now show that $|\hat{x}| > L > 0$. *Assume*, to the contrary, that $|\hat{x}| \leqq L$; then it follows from (71) that

$$|y(\tau)| = |n_{k\beta}(e(\tau), \tau)| = |n_{k\beta}(\hat{x}, \tau)| \leqq M,$$

which is contradictory to (73). Therefore,

$$|\hat{x}| > L. \tag{77}$$

Recalling that $L = \max(X, 2R/\varepsilon) > 0$, it follows that $|\hat{x}| > 0$. Divide both sides of (76) by $\hat{x}$:

$$1 < \frac{R}{|\hat{x}|} + A_{k\beta}(\tau)\frac{|n_{k\beta}(\hat{x}, \tau)|}{|\hat{x}|} + A_{k\beta}(\tau)\frac{\delta}{|\hat{x}|}. \tag{78}$$

Rearranging terms, and recalling that $G_{k\beta}(\hat{x}, \tau) = n_{k\beta}(\hat{x}, \tau)/\hat{x}$,[5]

$$A_{k\beta}(\tau) G_{k\beta}(\hat{x}, \tau) > 1 - R/|\hat{x}| - A_{k\beta}(\tau)\delta/|\hat{x}|. \tag{79}$$

[5] If, for the specific value $\hat{x}$, $n_{k\beta}(\hat{x}, \tau)$ is multivalued, then $G_{k\beta}(\hat{x}, \tau) \geqq n_{k\beta}(\hat{x}, \tau)/\hat{x}$ as indicated in Remark 4, and (79) is still valid.

Combining (77) and (68) with (79), one has,

$$A_{k\beta}(\tau)G_{k\beta}(\hat{x}, \tau) > 1 - R/L - A\delta/L. \tag{80}$$

From the expressions for δ and L in (69) and (70), one obtains,

$$A\delta/L = \varepsilon/2, \qquad R/L \leqq \varepsilon/2. \tag{81}$$

Substituting (81) into (80), one has

$$A_{k\beta}(\tau)G_{k\beta}(\hat{x}, \tau) > 1 - \varepsilon.$$

Recalling from (77) and (69) that $\hat{x} > L \geqq X$, and from (72) that $\tau \geqq T$, one has,

$$A_{k\beta}(\tau)G_{k\beta}(\hat{x}, \tau) > 1 - \varepsilon, \qquad \tau \geqq T, \quad |\hat{x}| > X. \tag{82}$$

The resulting statement in (82) is a contradiction to the hypothesis of the theorem as stated in (24) with the specific values $x = \hat{x}$, $t = \tau$. Because of this contradiction, one concludes that the starting assumption, namely, "$y(t)$ is unbounded," cannot be valid. Hence $y(t)$ is bounded.

Step 4. Let Y be a bound on $y(t)$,

$$|y(t)| < Y \quad \text{for all } t \geqq t_0. \tag{83}$$

From (67) and (68), one has,

$$|e(t)| < R + YA \quad \text{for all } t \geqq t_0.$$

Hence, $e(t)$ is bounded, and the proof of the theorem is completed.

The proof of the corollary follows immediately from the theorem. Note that the difference between the hypotheses of the corollary and the theorem lies in conditions (26) and (24). It can be shown in a straightforward fashion that if condition (26) is satisfied, one can find an $\varepsilon > 0$, a $T \geqq t_0$, and an $X \geqq 0$, such that condition (24) is satisfied, and the boundedness of $e(t)$ follows from the theorem.

REFERENCES

[1] I. W. Sandberg, *A frequency-domain condition for the stability of feedback systems containing a single time-varying element*, Bell System Tech. J., 43 (1964), pp. 1601–1608.

[2] G. Zames, *On the input-output stability of time-varying nonlinear feedback systems—part II*, IEEE Trans. Automatic Control, AC-11 (1966), pp. 465–476.

[3] A. R. Bergen, R. P. Iwens and A. J. Rault, *On the input-output stability of nonlinear feedback systems*, Ibid., AC-11 (1966), pp. 742–744.

[4] C. A. Desoer and M. Y. Wu, *L^p-stability of nonlinear time-varying feedback systems*, this Journal, 7 (1969), pp. 356–364.

[5] J. Willems, *On generalizations of the Popov criterion*, Internat. J. Nonlinear Mechanics, 5 (1970), pp. 131–141.

[6] G. N. Sarma and R. A. Haddad, *Stability of discrete-time systems with periodic coefficients*, 1970 Joint Automatic Control Conference, preprints, 1970, pp. 125–128.

[7] E. K. Haddad, *New criteria for bounded-input-bounded-output and asymptotic stability of nonlinear systems*, Proc. Fifth World Congress of the International Federation of Automatic Control, Paris, 1972.

[8] L. A. Zadeh and C. A. Desoer, *Linear System Theory*, McGraw-Hill, New York, 1963.

[9] J. C. Hsu and A. U. Meyer, *Modern Control Principles and Applications*, McGraw-Hill, New York, 1968.

[10] J. LaSalle and S. Lefschetz, *Stability by Liapunov's Direct Method with Applications*, Academic Press, New York.

[11] L. Markus, *Escape times for ordinary differential equations*, Rend. Sem. Mat., Univ. di Torino, 11 (1952), no. 2, pp. 271–276.

[12] J. C. Willems, *The Analysis of Feedback Systems*, M.I.T. Press, Cambridge, Mass., 1970.

SIAM J. CONTROL
Vol. 11, No. 2, May 1973

EQUATIONS DESCRIBING MULTIDIMENSIONAL CAUSAL SYSTEMS*

VACLAV DOLEZAL†

Abstract. This paper contains an abstract treatment of functional equations which involve nonlinear causal operators. We consider equations on linear spaces that are certain extensions of Banach or Hilbert spaces. The causality is defined in the traditional way, but by a "past" we mean a set from a given family of subsets of a generalized "time domain."

An operator A is said to have a resolvent operator Q, if for each u from a certain space, the equation $x = A(x, u)$ possesses a unique solution x; then $x = Qu$. We prove a theorem giving conditions for the existence of Q, and for continuity and boundedness of Q in a certain sense.

Further results are obtained for the case $A(x, u) = Bx + u$, i.e., for an equation of Hammerstein's type. In particular, it is shown that some monotonicity properties of operators involved guarantee the existence and continuity of the resolvent Q. The theory is illustrated by several concrete examples.

1. Introduction. In this paper we develop an abstract treatment of functional equations which involve nonlinear causal operators. In fact, we deal with equations of Hammerstein's type in linear spaces which are certain extensions of Banach or Hilbert spaces. There is an extensive literature concerning Hammerstein's equation on a Hilbert space; we mention only papers [3] and [5], which are referred to in the text. However, the present theory appears rather as a generalization of results by Sandberg [2], [4] and Zames [6] than the theory of Hammerstein's equation. On the other hand, some present results can be reduced to known results on Hammerstein's equation by using causality of the operators involved.

According to the standard definition, an operator is called causal, if two elements equal on a past P, which is an interval of the form $(-\infty, T)$ or $[0, T]$, have images which coincide on P. In the present context, however, P will be any set from a given family. It turns out that the scale, i.e., the set of all pasts, can be any family of subsets of a fixed "time domain," provided the intersection of two pasts is a past and all pasts fill out the time domain.

In the first section of the paper we shall consider the equation $x = A(x, u)$, where the operator A is assumed to be causal in x. The operator A is said to have a resolvent operator Q if, for each u from a certain space, the equation $x = A(x, u)$ possesses a unique solution $x = Qu$. We shall prove a theorem giving conditions for the existence of Q, and for continuity of Q in a certain sense. Further results are obtained for the case $A(x, u) = Bx + u$.

In § 2 we consider again the above particular form of A under the additional assumption that the space involved is an appropriate extension of a Hilbert space. It turns out that certain monotonicity properties of operators involved then guarantee the existence and continuity of Q.

1.1. We begin with several basic concepts.

Let Ω be a nonempty set. The collection $\mathscr{T} = \{T_\alpha\}$ will be called a scale on Ω, if all $T_\alpha \subset \Omega$ and

1. $T_\alpha \cap T_\beta \in \mathscr{T}$ whenever $T_\alpha, T_\beta \in \mathscr{T}$,
2. $\bigcup_\alpha T_\alpha = \Omega$.

* Received by the editors December 15, 1970, and in revised form February 15, 1972.

† Department of Applied Mathematics and Statistics, State University of New York at Stony Brook, Stony Brook, Long Island, New York 11790.

Let $\mathscr{C}$ be a linear space, and let $\tilde{F}$ be the collection of all mappings x from Ω into $\mathscr{C}$. Clearly, $\tilde{F}$ becomes a linear space, if we define addition and multiplication by scalars pointwise. Let F and F^* be linear spaces such that $F^* \subset F \subset \tilde{F}$, and F^* is a Banach space with a norm $\|\cdot\|$.

Finally, for each $T \in \mathscr{T}$, let S_T be a linear mapping from $\tilde{F}$ into itself such that the following axioms are satisfied:

(i) $S_{T_1}S_{T_2} = S_{T_1 \cap T_2}$ for any $T_1, T_2 \in \mathscr{T}$.

(ii) If $x, y \in \tilde{F}$, then $x(t) = y(t)$ on $T \in \mathscr{T}$, if and only if $S_T x = S_T y$.

(iii) If $x \in \tilde{F}$, then $x \in F$ if and only if $S_T x \in F^*$ for every $T \in \mathscr{T}$.

(iv) If $x \in F^*$, then $\|S_T x\| \leqq \|x\|$ for any $T \in \mathscr{T}$.

(v) If $x \in F$ and there exists a number $a > 0$ such that $\|S_T x\| \leqq a$ for every $T \in \mathscr{T}$, then $x \in F^*$ and $\|x\| \leqq a$.

On occasion we shall use the shorthand notation $x_T = S_T x$. Observe that the above axioms imply the following facts:

(a) $S_{T_1}S_{T_2} = S_{T_2}S_{T_1}$ for any $T_1, T_2 \in \mathscr{T}$.

(b) $S_T^2 = S_T$ for every $T \in \mathscr{T}$ (i.e., S_T is a projection).

(c) $x(t) = x_T(t)$ on T for any $x \in \tilde{F}$ and $T \in \mathscr{T}$.

(d) $\|S_{T_1}x\| \leqq \|S_{T_2}x\|$ whenever $T_1 \subset T_2$, $T_1, T_2 \in \mathscr{T}$ and $x \in F$.

Facts (a) and (b) follow trivially from axiom (i). If $T \in \mathscr{T}$ and $x \in \tilde{F}$, we have by (b), $S_T(S_T x) = S_T x$, and consequently, by axiom (ii), $x_T(t) = x(t)$ on T. Finally, if $T_1 \subset T_2$ and $x \in F$, then $S_{T_2}x \in F^*$ by (iii); hence, by (iv), $\|S_{T_1}S_{T_2}x\| \leqq \|S_{T_2}x\|$, and now (d) follows from axiom (i) since $T_1 = T_1 \cap T_2$.

A trivial example of $\tilde{F}$, F, F^* and S_T which satisfy all the axioms arises as follows:

Let Ω be a Hausdorff space, let $\mathscr{T}$ be the family of all compact subsets of Ω, and put $\mathscr{C} = R^1$. Then $\tilde{F}$ is the set of all functions on Ω, F is the set of all functions in $\tilde{F}$ whose restrictions to compact sets are bounded and F^* is the set of all bounded functions with norm $\|x\| = \sup_{t\in\Omega} |x(t)|$. Finally, we let S_T be defined by

$$(S_T x)(t) = \begin{cases} x(t) & \text{for } t \in T, \\ 0 & \text{for } t \notin T. \end{cases}$$

Another simple but important example is as follows:

Let R^n be n-dimensional Euclidean space; for $t \in R^n$, let $|t|$ denote some norm of t. Also, let $R^n_+ = [0, \infty)^n$; if $t = (t_1, t_2, \cdots, t_n) \in R^n_+$, let $[t] = [0, t_1] \times [0, t_2] \times \cdots \times [0, t_n]$. With this notation, put $\Omega = R^n$ and $\mathscr{T} = \{[t] : t \in R^n_+, t_i > 0, i = 1, 2, \cdots, n\}$ and let $\mathscr{C} = R^m$,

$$F = \{x : x \text{ measurable}, \int_{[t]} |x(\tau)|^p \, d\tau < \infty \text{ for any } t \in R^n_+\},$$

$$F^* = \{x : x \text{ measurable}, \int_{R^n_+} |x(\tau)|^p \, d\tau < \infty\};$$

here $1 \leqq p < \infty$ and F^* is equipped with norm $\|x\| = (\int_{R^n_+} |x(\tau)|^p \, d\tau)^{1/p}$. If, for every $[\tau] \in \mathscr{T}$, we define $S_{[\tau]}$ by

$$(S_{[\tau]}x)(t) = \begin{cases} x(t) & \text{for } t \in [\tau], \\ 0 & \text{for } t \notin [\tau], \end{cases}$$

then we can easily verify that all axioms (i)–(v) are satisfied. Other examples will be discussed later.

Now, we define the concept of an operator causal with respect to a scale.

DEFINITION. The (not necessarily linear) operator $A: F \to F$ will be called *causal with respect to the scale* $\mathscr{T}$, if

$$(1.1) \qquad S_T A = S_T A S_T \quad \text{for every } T \in \mathscr{T}.$$

This concept of causality can be characterized in the same fashion as the usual causality in one dimension.

LEMMA 1. *An operator $A: F \to F$ is causal with respect to $\mathscr{T}$ if and only if*

$$(1.2) \quad x_1, x_2 \in F, \quad x_1(t) = x_2(t) \quad \text{on } T \in \mathscr{T} \Rightarrow (Ax_1)(t) = (Ax_2)(t) \quad \text{on } T.$$

The proof of this lemma is obvious.

From the definition it follows immediately that all operators from F into itself which are causal with respect to $\mathscr{T}$ constitute a ring.

We now prove a fundamental proposition about causal operators.

LEMMA 2. *Let $A: F \to F$ be causal with respect to the scale $\mathscr{T}$. If, for every $T \in \mathscr{T}$, the operator $S_T A$ has a unique fixed point $x^{(T)}$ in F^*, then A has a unique fixed point x in F. Moreover, we have $S_T x = x^{(T)}$ for every $T \in \mathscr{T}$.*

Proof. First, axiom (iii) shows that for any $T \in \mathscr{T}$, $S_T A$ maps F^* into itself. Next, we verify that $x_{T^*}^{(T)} = x_{T^*}^{(T')}$ with $T, T' \in \mathscr{T}$, $T^* = T \cap T'$. Indeed, we have

$$(1.3) \qquad x^{(T)} = S_T A x^{(T)}, \quad x^{(T')} = S_{T'} A x^{(T')};$$

thus, by axiom (i) and causality of A,

$$x_{T^*}^{(T)} = S_{T^*} S_T A x^{(T)} = S_{T^* \cap T} A x^{(T)} = S_{T^*} A x^{(T)} = S_{T^*} A x_{T^*}^{(T)}.$$

Similarly, $x_{T^*}^{(T')} = S_{T^*} A x_{T^*}^{(T')}$. Thus, both $x_{T^*}^{(T)}$ and $x_{T^*}^{(T')}$ are fixed points of $S_{T^*} A$. Since $T^* \in \mathscr{T}$ and both $x_{T^*}^{(T)}$ and $x_{T^*}^{(T')}$ are in F^*, we have by uniqueness that $x_{T^*}^{(T)} = x_{T^*}^{(T')}$, i.e., $x^{(T)}(t) = x^{(T')}(t)$ on T^* by (ii).

Now, define the element $x \in \tilde{F}$ as follows: if $t \in \Omega$, let $x(t) = x^{(T)}(t)$ for some $T \in \mathscr{T}$ with $t \in T$; such a T exists by assumption 2 on $\mathscr{T}$. This definition is unambiguous, since if $T' \in \mathscr{T}$ and $t \in T'$, then, as we have just shown, $x^{(T)}(t) = x^{(T')}(t)$.

On the other hand, if $T \in \mathscr{T}$, our construction of x implies that $x(t) = x^{(T)}(t)$ for every $t \in T$. Thus, we have $S_T x = S_T x^{(T)}$ by (ii), and since $S_T x^{(T)} \in F^*$ by (iii), we have $S_T x \in F^*$; consequently, again by (iii), $x \in F$.

In order to show that x satisfies the equation $x = Ax$, i.e., $x(t) = (Ax)(t)$ for any $t \in \Omega$, pick some $T \in \mathscr{T}$ with $t \in T$; by the above $x_T = x_T^{(T)} = x^{(T)}$, the last following from the fact that the fixed point of $S_T A$ is in the range of S_T. Causality of A yields $(Ax)_T = (Ax_T)_T = (Ax^{(T)})_T = x^{(T)} = x_T$, so that, by axiom (ii), $x(t') = (Ax)(t')$ for every $t' \in T$. Consequently, $x(t) = (Ax)(t)$ as desired.

To prove uniqueness, suppose that $y \in F$ exists such that $y = Ay$. Choosing $t \in \Omega$ arbitrarily, pick a $T \in \mathscr{T}$ with $t \in T$. Then, $y_T \in F^*$ by (iii) and $y_T = S_T Ay = S_T Ay_T$ by causality; consequently, by uniqueness of the fixed point $x^{(T)}$, necessarily $y_T = x^{(T)} = x_T$, which implies that $y = x$. This completes the proof.

We now introduce some further concepts. First we extend the notion of continuity.

DEFINITION. Let $\tilde{G}$ be a linear space, let $G^* \subset \tilde{G}$ be a normed linear space and let G be a subset of $\tilde{G}$ such that $G \cap G^* \neq \varnothing$. If $P: G \to F$ and $u \in G$, then the operator P is called *continuous at the point u*, if for every $\varepsilon > 0$ there exists a $\delta > 0$

such that for any $\tilde{u} \in G$ with $\tilde{u} - u \in G^*$ and $\|\tilde{u} - u\| < \delta$ we have $P\tilde{u} - Pu \in F^*$ and $\|P\tilde{u} - Pu\| < \varepsilon$. If the operator P is continuous at every point, it is called *continuous.*

Analogously, if, for $u \in G$, there exist constants $\mu > 0$ and $a > 0$ such that $P\tilde{u} - Pu \in F^*$ and $\|P\tilde{u} - Pu\| \leqq \mu\|\tilde{u} - u\|$ whenever $\tilde{u} \in G$, $\tilde{u} - u \in G^*$ and $\|\tilde{u} - u\| \leqq a$, then P is called *L-continuous at u.* If P is L-continuous at every point $u \in G$, it is called *L-continuous.*

Clearly, the above notion of continuity reduces to the usual one when $F = F^*$ and $G = G^*$. Furthermore, from axiom (iv) it follows that for any $T \in \mathscr{T}$ the operator $S_T : F \to F^*$ is L-continuous.

Now, let A be an operator mapping $F \times G$ into F. A will be said to have a resolvent operator $Q : G \to F$, if for every $u \in G$, the equation

$$x = A(x, u) \tag{1.4}$$

possesses a unique solution x in F. Then Q is defined by $Qu = x$.

We can now state the first result concerning equation (1.4).

THEOREM 1.1. *Let $A : F \times G \to F$. Assume that, for every $u \in G$, the operator $A_u = A(\cdot, u)$ is causal with respect to $\mathscr{T}$, and that for every $u \in G$ and $T \in \mathscr{T}$ there exist an integer $m_{u,T} \geqq 1$ and a number $0 \leqq \lambda_{u,T} < 1$ such that*

$$\|S_T(A_u^{m_{u,T}}x_1 - A_u^{m_{u,T}}x_2)\| \leqq \lambda_{u,T}\|S_T(x_1 - x_2)\| \tag{1.5}$$

for all $x_1, x_2 \in F$. Then A has a resolvent operator Q.

If, in particular, $G = F$ and A is causal in u with respect to $\mathscr{T}$, i.e.,

$$S_T A(x, \cdot) = S_T A(x, S_T \cdot) \tag{1.6}$$

for every $x \in F$ and $T \in \mathscr{T}$, then Q is also causal with respect to $\mathscr{T}$.

If, furthermore,

(i) *a fixed integer $m \geqq 1$ and a fixed number $0 \leqq \lambda < 1$ exist such that* (1.5) *is true for every $u \in G$, $T \in \mathscr{T}$ and $x_1, x_2 \in F$,*

(ii) *the operator $A_u^m x$ is continuous* (*L-continuous*) *in u for any $x \in F$, then Q is continuous* (*L-continuous*).

Finally, if also $A_u^m\theta \in F^*$ (θ denotes the zero of F), then $Qu \in F^*$ if $u \in G \cap G^*$.

Proof. We begin by observing that if $u \in G$, $k \geqq 1$ is an integer and $T \in \mathscr{T}$, then

$$S_T A_u^k = (S_T A_u)^k; \tag{1.7}$$

this follows easily by causality of A_u.

Choose $u \in G$ and $T \in \mathscr{T}$. Then (1.7) and (1.5) show that $(S_T A_u)^{m_{u,T}}$ is a contraction on F^*; indeed, if x_1, $x_2 \in F^*$, we have (dropping the indices of m and λ for brevity),

$$\begin{aligned} \|(S_T A_u)^m x_1 - (S_T A_u)^m x_2\| &= \|S_T(A_u^m x_1 - A_u^m x_2)\| \\ &\leqq \lambda\|S_T(x_1 - x_2)\| \leqq \lambda\|x_1 - x_2\| \end{aligned} \tag{1.8}$$

by axion (iv). Consequently, there exists a unique $x^{(T)} \in F^*$ such that

$$x^{(T)} = (S_T A_u)^m x^{(T)}, \tag{1.9}$$

and it follows that $x^{(T)}$ is also a unique fixed point (in F^*) of $S_T A_u$, i.e.,

$$x^{(T)} = S_T A_u x^{(T)}. \tag{1.10}$$

(See [1].) From Lemma 2 it follows that A_u has a unique fixed point x_u, so A has a resolvent Q given by $Qu = x_u$, for $u \in G$.

Now, assume that $G = F$ and (1.6) holds. Choose a fixed $u \in F$ and let $T \in \mathscr{T}$; putting $x = Qu$, $\tilde{x} = Qu_T$, we have by definition of Q,

$$x = A(x, u), \qquad \tilde{x} = A(\tilde{x}, u_T). \tag{1.11}$$

Thus, by causality of A_u and (1.6),

$$x_T = S_T A(x, u) = S_T A(x_T, u) = S_T A(x_T, u_T), \tag{1.12}$$

and similarly,

$$\tilde{x}_T = S_T A(\tilde{x}, u_T) = S_T A(\tilde{x}_T, u_T),$$

i.e., $x_T = S_T A_{u_T} x_T$ and $\tilde{x}_T = S_T A_{u_T} \tilde{x}_T$. Since both x_T and $\tilde{x}_T$ are in F^*, it follows by the uniqueness of the fixed point of $S_T A_{u_T}$ that $x_T = \tilde{x}_T$, i.e., $(Qu)_T = (Qu_T)_T$, so Q is, indeed, causal.

Next, assume that (i) and (ii) are satisfied, and choose a fixed $u \in G$ and $\varepsilon > 0$. If $T \in \mathscr{T}$ and $\tilde{u} \in G$ is such that $\tilde{u} - u \in G^*$, let $x^{(T)}$, $\tilde{x}^{(T)}$ satisfy

$$x^{(T)} = (S_T A_u)^m x^{(T)}, \qquad \tilde{x}^{(T)} = (S_T A_{\tilde{u}})^m \tilde{x}^{(T)}. \tag{1.13}$$

Then, by a standard argument we get from (1.8),

$$\|\tilde{x}^{(T)} - x^{(T)}\| \leqq (1 - \lambda)^{-1} \|S_T(A^m_{\tilde{u}} x^{(T)} - A^m_u x^{(T)})\|. \tag{1.14}$$

On the other hand, putting $x = Qu$, $\tilde{x} = Q\tilde{u}$, we have, by Lemma 2, (1.13) and the equivalence of (1.9) and (1.10), $x^{(T)} = S_T x$ and $\tilde{x}^{(T)} = S_T \tilde{x}$. Using this and the fact that operators A^m_u and $A^m_{\tilde{u}}$ are causal, we can write (1.14) as

$$\|S_T(\tilde{x} - x)\| \leqq (1 - \lambda)^{-1} \|S_T(A^m_{\tilde{u}} x - A^m_u x)\|. \tag{1.15}$$

By the continuity of $A^m_u x$ at the given $u \in G$, there exists a $\delta > 0$ such that for $\|\tilde{u} - u\| < \delta$ we have $A^m_{\tilde{u}} x - A^m_u x \in F^*$ and $\|A^m_{\tilde{u}} x - A^m_u x\| < (1 - \lambda)\varepsilon$. Introducing this into (1.15), we obtain by axiom (iv),

$$\|S_T(\tilde{x} - x)\| \leqq (1 - \lambda)^{-1} \|A^m_{\tilde{u}} x - A^m_u x\| < \varepsilon. \tag{1.16}$$

The right-hand side of (1.16) is independent of T; consequently, by axiom (v), $\tilde{x} - x \in F^*$ and $\|\tilde{x} - x\| \leqq \varepsilon$, i.e., $Q\tilde{u} - Qu \in F^*$ and $\|Q\tilde{u} - Qu\| \leqq \varepsilon$. Hence, Q is continuous at $u \in G$. The proof of L-continuity follows the same pattern.

As for the last assertion of the theorem, choose $u \in G^* \cap G$. If $T \in \mathscr{T}$, we have by (1.9) and (1.5),

$$\begin{aligned} \|x^{(T)}\| &\leqq \|(S_T A_u)^m x^{(T)} - (S_T A_u)^m \theta\| + \|(S_T A_u)^m \theta\| \\ &= \|S_T(A^m_u x^{(T)} - A^m_u \theta)\| + \|S_T A^m_u \theta\| \leqq \lambda \|S_T x^{(T)}\| + \|S_T A^m_u \theta\|. \end{aligned}$$

Since both $x^{(T)}$ and $A^m_u \theta$ are in F^*, we have by axiom (iv),

$$\|x^{(T)}\| \leqq \lambda \|x^{(T)}\| + \|A^m_u \theta\|,$$

i.e.,

$$\|x^{(T)}\| \leqq (1 - \lambda)^{-1}\|A_u^m\theta\|.$$

Finally, $x^{(T)} = S_T x = S_T Qu$, so that

$$\|S_T Qu\| \leqq (1 - \lambda)^{-1}\|A_u^m\theta\|.$$

Thus, by axiom (v), $Qu \in F^*$ and the proof is now complete.

In many concrete situations the operator A has the form $A(x, u) = Bx + u$, where B is a product of a linear and nonlinear operator; of course, $G = F$. Having this in mind, we shall now prove a theorem similar to one stated by Sandberg [2]. In order to simplify its proof, several observations are in order.

For $T \in \mathscr{T}$, let $F_T = \{x : x = S_T y, y \in F\} = S_T F$. Due to axiom (iii), F_T is a linear subspace of F^*, and by fact (b) in the beginning of this section, S_T is the identity operator on F_T.

LEMMA 3. *If F_T is equipped with the norm of F^*, then it is a Banach space.*

Actually, F_T is closed in F^*, since if $x_n \in F_T$ and $x_n \to x_0 \in F^*$, the continuity of S_T implies that $S_T x_n = x_n \to S_T x_0 = x_0 \in F_T$.

LEMMA 4. *Let $A : F \to F$ and $T \in \mathscr{T}$; then the operator $S_T A$ has a unique fixed point in F^* if and only if it has a unique fixed point in F_T.*

LEMMA 5. *Let K be a one-to-one operator from F onto F such that both K and K^{-1} are causal with respect to $\mathscr{T}$; then, for any $T \in \mathscr{T}$, the operator $S_T K$ is one-to-one from F_T onto F_T and $(S_T K)^{-1} = S_T K^{-1}$.*

The proof of these lemmas is obvious.

THEOREM 1.2. *Let $C, N : F \to F$ be causal with respect to $\mathscr{T}$, and let C be linear. Assume that a scalar λ exists such that*

(i) *$I - \lambda C$ is one-to-one from F onto F and $(I - \lambda C)^{-1}$ is causal with respect to $\mathscr{T}$;*

(ii) *$(I - \lambda C)^{-1}C$ maps F^* into itself and is bounded;*

(iii) *$\|S_T\{Nx_1 - Nx_2 - \lambda(x_1 - x_2)\}\| \leqq \mu\|S_T(x_1 - x_2)\|$ for all $x_1, x_2 \in F$ and $T \in \mathscr{T}$ and some $\mu > 0$;*

(iv) *$\|(I - \lambda C)^{-1}C\|\mu < 1$.*

Then the operator $A(x, u) = CNx + u$ has a resolvent operator Q and Q is causal with respect to $\mathscr{T}$.

Moreover, if

(v) *$(I - \lambda C)^{-1}$ maps F^* into itself and is bounded, then Q is L-continuous. If, in addition, $CN\theta \in F^*$, then Q maps F^* into F^*.*

Proof. We can prove this theorem either by using results of [3] or by employing methods as in the proof of Theorem 1.1. We indicate the main steps.

Choose $u \in F$; referring to Lemmas 2 and 4, consider the fixed point in F_T of the equation

$$x^{(T)} = S_T(u + CNx^{(T)}). \tag{1.17}$$

By Lemma 5, (1.17) is equivalent to

$$x^{(T)} = R_u x^{(T)}, \tag{1.18}$$

where the operator $R_u : F \to F_T$ is defined by

$$R_u x = S_T(I - \lambda C)^{-1} S_T\{u + C(N - \lambda I)x\}. \tag{1.19}$$

Using assumptions (i)–(iii), we can easily verify that

$$\|R_u x_1 - R_u x_2\| \leqq \tilde{\lambda}\|x_1 - x_2\| \tag{1.20}$$

for every $x_1, x_2 \in F_T$, where $\tilde{\lambda} = \|(I - \lambda C)^{-1} C\|\mu < 1$; hence, R_u is a contraction on F_T. Consequently, A has a resolvent operator Q. The causality of Q follows by the same argument as in the proof of Theorem 1.1.

Next, assume that (v) is satisfied and let $u \in F$. Choose $\tilde{u} \in F$ such that $\tilde{u} - u \in F^*$, and let $Qu = x = CNx + u$, $Q\tilde{u} = \tilde{x} = CN\tilde{x} + \tilde{u}$. Using Lemma 2 and (1.19), (1.20) it follows that

$$\|Q\tilde{u} - Qu\| \leqq \lambda^* \|\tilde{u} - u\| \tag{1.21}$$

with $\lambda^* = (1 - \tilde{\lambda})^{-1} \|(I - \lambda C)^{-1}\|$, i.e., Q is L-continuous.

The proof of the last assertion follows the same pattern as in the proof of Theorem 1.1.

The condition (i) in Theorem 1.2 may be replaced by another assumption (i)*, which will prove more convenient than (i) in our considerations in § 2, and which also can be tested more easily in specific situations.

To this purpose we consider causal operators on F^* (rather than on F).

LEMMA 6. *Let $K : F \to F$ be causal with respect to $\mathscr{T}$, and let $\tilde{K}$ be its restriction to F^*. If $\tilde{K}$ is one-to-one from F^* onto F^* and $\tilde{K}^{-1}$ is causal on F^* with respect to $\mathscr{T}$, then K is one-to-one from F onto F, and K^{-1} is causal with respect to $\mathscr{T}$.*

Proof. First observe that $\tilde{K}$ is causal on F^*. Next, if $T \in \mathscr{T}$, then $S_T\tilde{K}$ is one-to-one from F_T onto F_T, and $(S_T\tilde{K})^{-1} = S_T\tilde{K}^{-1}$. Indeed, both $S_T\tilde{K}$ and $S_T\tilde{K}^{-1}$ map F_T into itself, while on F_T we have $S_T\tilde{K}^{-1}S_T\tilde{K} = S_T\tilde{K}^{-1}\tilde{K} = S_T$, and $S_T\tilde{K}S_T\tilde{K}^{-1} = S_T\tilde{K}\tilde{K}^{-1} = S_T$. The statement follows on noting that S_T is the identity on F_T.

Next, let $y \in F$, and let $T \in \mathscr{T}$. By the above, there exists a unique $x^{(T)} \in F_T$ such that $S_T\tilde{K}x^{(T)} = S_T y$. Since $\tilde{K} = K$ on F^*, we have $S_T K x^{(T)} = S_T y$ and note that this equation is equivalent to $x^{(T)} = S_T\{(I - K)x^{(T)} + y\}$. Recalling Lemma 2, it follows that there exists a unique $x \in F$ such that $x = (I - K)x + y$, i.e., $Kx = y$. Since $y \in F$ can be chosen arbitrarily, K possesses an inverse K^{-1}.

In order to prove causality of K^{-1}, pick $y \in F$ and let $Kx = y$; choosing $T \in \mathscr{T}$, find $\tilde{x} \in F$ such that $K\tilde{x} = y_T$. Then $S_T K x = y_T$ and $S_T K\tilde{x} = y_T$; by the causality of K, however, $S_T K x_T = y_T$ and $S_T K\tilde{x}_T = y_T$. Since both x_T and $\tilde{x}_T$ are in F_T and $K = \tilde{K}$ on F^*, we have $x_T = \tilde{x}_T$, i.e., $(K^{-1}y)_T = (K^{-1}y_T)_T$, and the causality of K^{-1} is proved.

Using this result, we can deduce the following without further work.

THEOREM 1.3. *The condition* (i) *in Theorem* 1.2 *can be replaced by the condition* (i)*; *the restriction $\tilde{K}$ of $I - \lambda C$ to F^* is one-to-one and onto F^*, and $\tilde{K}^{-1}$ is causal on F^* with respect to $\mathscr{T}$.*

Note that the converse of Lemma 6 is not true: causality of K and K^{-1} need not imply that $\tilde{K}$ is one-to-one from F^* onto F^*.

We shall now discuss several specific examples which illustrate the application of the above theorems.

1.2. Let $\Omega = R^n_+$, $\mathscr{T} = \{[t]: t \in R^n_+,\ t_i > 0,\ i = 1, 2, \cdots, n\}$, where the symbols R^n_+, $[t]$ have the same meaning as in the example of § 1.1. If $t, t' \in R^n$, we shall write $t \leqq t'$ whenever $t_i \leqq t'_i$ for $i = 1, 2, \cdots, n$. Next, let $\mathscr{C} = R^1$ and let F be the set of all continuous functions on R^n_+. Finally, let F^* be the Banach space of all continuous bounded functions on R^n_+ endowed with the sup norm.

Define S_T for every $T \in \mathscr{T}$ as follows:

$$(1.22) \qquad (S_{[\tau]}x)(t) = \begin{cases} x(t) & \text{for } t \in [\tau], \\ x(\lambda_t t) & \text{for } t \notin [\tau], \end{cases}$$

where the number λ_t is such that $\lambda_t t$ is in the boundary of $[\tau]$.

It can be easily verified that all axioms (i)–(v) are satisfied. To construct an operator A, let $K(t, \xi, \mu, \eta_0, \eta_1, \cdots, \eta_k)$, a function of $2n + k + 2$ variables, be defined and continuous for $0 \leqq \xi \leqq t$, $t \in R^n_+$ and μ, $\eta_i \in R^1$, $i = 0, 1, 2, \cdots, k$. Assume that for every $\tau \in R^n_+$ and $a > 0$, there are constants $C_i^{\tau,a} \geqq 0$, $i = 0, 1, 2, \cdots, k$, such that

$$(1.23) \quad |K(t, \xi, \mu, \eta_0, \eta_1, \cdots, \eta_k) - K(t, \xi, \mu, \bar{\eta}_0, \bar{\eta}_1, \cdots, \bar{\eta}_k)| \leqq \sum_{i=0}^{k} C_i^{\tau,a} |\eta_i - \bar{\eta}_i|$$

whenever $0 \leqq \xi \leqq t \leqq \tau$, $|\mu| \leqq a$ and $\eta_i, \bar{\eta}_i \in R^1$, $i = 0, 1, \cdots, k$.

Next, let $\varphi_i : R^n_+ \to R^n_+$ be real continuous functions such that $0 \leqq \varphi_i(\xi) \leqq \xi$ for any $\xi \in R^n_+$, $i = 1, 2, \cdots, k$. For $x, u \in F$ and $t \in R^n_+$, $A(x, u)$ is defined by

$$(1.24) \quad [A(x, u)](t) = \int_{[t]} K(t, \xi, u(t), x(\xi), x(\varphi_1(\xi)), x(\varphi_2(\xi)), \cdots, x(\varphi_k(\xi)))\, d\xi.$$

It is clear that $A : F \times F \to F$. Applying Lemma 1, we see immediately that the operator $A_u = A(\cdot, u)$ is causal with respect to the scale $\mathscr{T}$ for any $u \in F$, and the same is true for $A(x, \cdot)$ with any $x \in F$. We are going to show that A possesses a resolvent operator Q and Q is causal with respect to $\mathscr{T}$, by applying Theorem 1.1.

To do this, choose some $u \in F$, $[t] \in \mathscr{T}$, and let $x_1, x_2 \in F$, $a = \sup_{t' \in [\tau]} |u(t')|$. If $t \in [\tau]$, we have by (1.24) and (1.23),

$$|\{A(x_1, u) - A(x_2, u)\}(t)|$$
$$\leqq \int_{[t]} \{C_0^{\tau,a} |x_1(\xi) - x_2(\xi)| + \sum_{i=1}^{k} C_i^{\tau,a} |x_1(\varphi_i(\xi)) - x_2(\varphi_i(\xi))|\}\, d\xi.$$

Using (1.22) and the property of the φ_i's, it follows immediately that

$$|\{A(x_1, u) - A(x_2, u)\}(t)| \leqq t_1 t_2 \cdots t_n C^{\tau,a} \|S_{[\tau]}(x_1 - x_2)\|,$$

where the t_i's are coordinates of t and $C^{\tau,a} = \sum_{i=0}^{k} C_i^{\tau,a}$.

For $m \geqq 1$ we have

$$(1.25) \qquad |(A_u^m x_1 - A_u^m x_2)(t)| \leqq \frac{(t_1 t_2 \cdots t_n)^m}{(m!)^n} (C^{\tau,a})^m \|S_{[\tau]}(x_1 - x_2)\|.$$

This is proven by induction in precisely the same way as the analogous inequality for nonlinear Volterra operators. Recalling the definition of $S_{[\tau]}$ we conclude from

(1.25) that

$$\|S_{[\tau]}(A_u^m x_1 - A_u^m x_2)\| \leqq \lambda_m \|S_{[\tau]}(x_1 - x_2)\|, \tag{1.26}$$

where $\lambda_m = (C^{\tau,a}\tau_1\tau_2 \cdots \tau_n)^m/(m!)^n$.

On taking m sufficiently large we achieve $\lambda_m < 1$, so that the condition (1.5) in Theorem 1.1 is satisfied. Consequently, A has a resolvent operator and this is causal.

1.3. Now let $\Omega = R^2$; if $t \in R^2$, put $|t| = (t_1^2 + t_2^2)^{1/2}$ and define

$$\mathscr{T} = \{\{t : t \in R^2, |t| \leqq \tau\} : \tau \geqq 1\},$$

the set of (closed) discs of radius $\geqq 1$. Furthermore, let $\mathscr{C} = R^1$, let F be the set of all continuous functions on R^2, and let F^* be the set of all continuous bounded functions on R^2 with uniform norm. If $T_\tau = \{t : t \in R^2, |t| \leqq \tau\} \in \mathscr{T}$, define S_{T_τ} by

$$(S_{T_\tau}x)(t) = \begin{cases} x(t) & \text{for } |t| \leqq \tau, \\ x\left(\dfrac{\tau}{|t|}t\right) & \text{for } |t| > \tau. \end{cases}$$

We can easily verify that axioms (i)–(v) are satisfied.

The following functions are defined in order to construct the operator A. Let $f_i : R^1 \to R^1$, $i = 1, 2$, be Lipschitz continuous functions, i.e.,

$$|f_i(\xi_1) - f_i(\xi_2)| \leqq \lambda_i|\xi_1 - \xi_2| \tag{1.27}$$

for all $\xi_1, \xi_2 \in R^1$ and some λ_i's, $i = 1, 2$. Let $g \in F^*$, and let $K(t, \xi)$ be a continuous function on $R^2 \times R^2$ such that

$$k = \sup_{t \in R^2} \int_{|\xi| \leqq |t|} |K(t, \xi)|\, d\xi < \infty.$$

Finally, for each $x, u \in F$, define $A(x, u)$ by

$$\{A(x, u)\}(t) = g(t) \int_{|\xi| \leqq 1} f_1(x(\xi))\, d\xi + \int_{|\xi| \leqq |t|} K(t, \xi) f_2(x(\xi))\, d\xi + u(t). \tag{1.28}$$

Clearly, A maps $F \times F$ into F. Recalling again Lemma 1 we see that the operator $A(\cdot, u)$ is causal with respect to the above scale $\mathscr{T}$ and also $A(x, \cdot)$ has the same property. We shall show that the operator A possesses a resolvent Q which is L-continuous and maps F^* into itself, provided that $\pi\lambda_1\|g\| + k\lambda_2 < 1$. Of course, it will suffice to verify the appropriate hypotheses of Theorem 1.1.

In order to prove it, choose u, x_1, $x_2 \in F$ and $\tau \geqq 1$; if $t \in T_\tau$, we have by (1.28) and (1.27),

$$|\{A(x_1, u) - A(x_2, u)\}(t)| \leqq |g(t)| \int_{|\xi| \leqq 1} \lambda_1 |x_1(\xi) - x_2(\xi)|\, d\xi + \int_{|\xi| \leqq |t|} |K(t, \xi)| \cdot \lambda_2 |x_1(\xi) - x_2(\xi)|\, d\xi.$$

Using the definition of S_{T_τ}, we easily obtain

$$\|S_{T_\tau}(A(x_1, u) - A(x_2, u))\| \leqq (\|g\|\lambda_1\pi + k\lambda_2)\|S_{T_\tau}(x_1 - x_2)\|,$$

so that (1.5) is satisfied with $m = 1$, irrespective of u. Also, (1.28) shows that $A(x, \cdot)$ is L-continuous. Hence, by Theorem 1.1, an L-continuous resolvent exists for A.

Finally, for $u \in F^*$,

$$\{A(\theta, u)\}(t) = g(t)\pi f_1(0) + f_2(0) \int_{|\xi| \leqq |t|} K(t, \xi)\, d\xi + u(t).$$

Our assumption on $K(t, \xi)$ shows that

$$\int_{|\xi| \leqq |t|} K(t, \xi)\, d\xi \in F^*.$$

Hence, $A(\theta, u) \in F^*$ and consequently, $Qu \in F^*$.

Another example is obtained by modifying the previous one as follows: Let all quantities have the same meaning as above, but assume that u is a fixed element in F. Let $G^* = R^1$, $G = (-1, 1)$, and define the operator $\tilde{A}: F \times G \to F$ by

$$\{\tilde{A}(x, v)\}(t) = vg(t) \int_{|\xi| \leqq 1} f_1(x(\xi))\, d\xi + \int_{|\xi| \leqq |t|} K(t, \xi) f_2(x(\xi))\, d\xi + u(t). \tag{1.29}$$

As above, it follows that $\tilde{A}$ possesses a resolvent $\tilde{Q}: F \to F$, provided $(\pi\lambda_1 \|g\| + k\lambda_2) < 1$.

Next, let $x \in F$ and let $v, \tilde{v} \in (-1, 1)$. From (1.29) we get

$$\{\tilde{A}(x, \tilde{v}) - \tilde{A}(x, v)\}(t) = (\tilde{v} - v) g(t) \int_{|\xi| \leqq 1} f_1(x(\xi))\, d\xi \in F^*,$$

whence

$$\|\tilde{A}(x, \tilde{v}) - \tilde{A}(x, v)\| \leqq |\tilde{v} - v| \cdot \|g\| \cdot \left| \int_{|\xi| \leqq 1} f_1(x(\xi))\, d\xi \right|.$$

Hence, $\tilde{A}(x, \cdot)$ is L-continuous on $(-1, 1)$, and consequently, by Theorem 1.1, $\tilde{Q}$ is also L-continuous.

2.1. In this section we assume that F^* is a real inner-product space, i.e., an inner product $\langle \cdot, \cdot \rangle$ is defined on $F^* \times F^*$ such that $\langle x, x \rangle^{1/2} = \|x\|$ for every $x \in F^*$; naturally, it is understood that the norm $\| \cdot \|$ satisfies the axioms (iv) and (v). We shall show that, if F^* is endowed with a more specific structure of an inner-product space, requirements on a quantity to be less than one can be traded for certain positivity properties of operators involved.

First, we prove the following proposition.

THEOREM 2.1. *Let $A: F \to F$ be causal with respect to $\mathscr{T}$, and assume that it satisfies the following conditions*:

(i) *there exists a $c > -1$ such that*

$$\langle S_T(x_1 - x_2), S_T(Ax_1 - Ax_2) \rangle \geqq c \|S_T(x_1 - x_2)\|^2 \quad \textit{for all } x_1, x_2 \in F \textit{ and } T \in \mathscr{T}, \tag{2.1}$$

(ii) *for every $T \in \mathscr{T}$ there exists a $\lambda_T > 0$ such that*

$$\|S_T(Ax_1 - Ax_2)\| \leqq \lambda_T \|S_T(x_1 - x_2)\| \quad \textit{for all } x_1, x_2 \in F. \tag{2.2}$$

Then the operator $\tilde{A}$ defined by $\tilde{A}(x, u) = -\tilde{A}x + u$ possesses a resolvent Q which is causal with respect to $\mathscr{T}$.

If, in addition, there exists a $\lambda > 0$ such that (2.2) *holds for all $x_1, x_2 \in F$ and $T \in \mathscr{T}$, then Q is L-continuous. Finally, if $A\theta \in F^*$, then Q maps F^* into F^*.*

Proof. Choose some $u \in F$. Referring to Lemma 2, choose a $T \in \mathscr{T}$ and consider the equation

$$x^{(T)} = S_T(-Ax^{(T)} + u) \tag{2.3}$$

on F^*. If $k > 0$, then (2.3) is clearly equivalent to

$$x^{(T)} = R_u^T x^{(T)}, \tag{2.4}$$

where the operator $R_u^T : F \to F^*$ is defined by

$$R_u^T x = (1 + k)^{-1}\{S_T(kI - A)x + u_T\}. \tag{2.5}$$

We are going to show that for a suitable choice of $k > 0$, R_u^T becomes a contraction on F^*. Thus, let $x_1, x_2 \in F^*$; then we have by (2.5), (2.1) and (2.2),

$$\begin{aligned} \|R_u^T x_1 - R_u^T x_2\|^2 &= (1 + k)^{-2}\|kS_T(x_1 - x_2) - S_T(Ax_1 - Ax_2)\|^2 \\ &= (1 + k)^{-2}\{k^2\|S_T(x_1 - x_2)\|^2 + \|S_T(Ax_1 - Ax_2)\|^2 \\ &\qquad - 2k\langle S_T(x_1 - x_2), S_T(Ax_1 - Ax_2)\rangle\} \\ &\leqq \mu_T(k)\|S_T(x_1 - x_2)\|^2 \leqq \mu_T(k)\|x_1 - x_2\|^2, \end{aligned} \tag{2.6}$$

where $\mu_T(k) = (1 + k)^{-2}\{k^2 - 2kc + \lambda_T^2\}$.

An obvious argument shows that there are values of k such that $\mu_T(k) < 1$; hence, there exists a unique $x \in F$ such that $x = -Ax + u$, i.e., $\tilde{A}$ has a resolvent Q. To prove causality of Q is already a matter of routine.

Next, if a $\lambda > 0$ exists such that (2.2) holds for all $x_1, x_2 \in F$ and $T \in \mathscr{T}$, we find $k_0 > 0$ as above so that $\|R_u^T x_1 - R_u^T x_2\| \leqq \sqrt{\mu_0}\|x_1 - x_2\|$ for all $x_1, x_2 \in F^*$ and $T \in \mathscr{T}$, where $\mu_0 = \mu(k_0) < 1$. Choose a $\tilde{u} \in F$ with $\tilde{u} - u \in F^*$ and let $T \in \mathscr{T}$. Then the same argument as in the proof of Theorem 1.1 shows that

$$\|Q\tilde{u} - Qu\| \leqq (1 - \sqrt{\mu_0})^{-1}(1 + k_0)^{-1}\|\tilde{u} - u\|,$$

i.e., Q is L-continuous.

The proof of the last statement is obvious.

Note also that the proof of Theorem 2.1 follows immediately from results in [3] and Lemma 2.

Our next theorem is similar to a result of Sandberg given in [4] for the case that $F^* = L_2$. As in [4], the proof of our theorem is based on three lemmas: 7, 8 and 10 below. While the lemmas in [4] deal with L_2, here we shall have to prove analogous results for an arbitrary inner-product space F^*.

First we have the following assertion.

LEMMA 7. *Let $M : F^* \to F^*$ be linear and bounded, and assume that there exists a $c > 0$ such that*

$$\langle Mx, x\rangle \geqq c\|x\|^2 \quad \textit{for all } x \in F^*. \tag{2.7}$$

Then the operator M is one-to-one from F^* onto F^*, and M^{-1} is bounded; *in fact*, $\|M^{-1}\| \leqq c^{-1}$.

Proof. The operator M is one-to-one, since $Mx_0 = 0$, $x_0 \in F^*$ implies by (2.7) that $x_0 = 0$. Also, M maps F^* onto F^*. To show this, note first that MF^* is closed. Indeed, suppose that $y_n \in MF^*$ and $y_n \to y$; then, for each n, there exists a (unique) $x_n \in F^*$ such that $Mx_n = y_n$. However, (2.7) implies by the Schwarz inequality that $\|Mu\| \geqq c\|u\|$. Consequently, for any integers n, m,

$$\|M(x_n - x_m)\| = \|y_n - y_m\| \geqq c\|x_n - x_m\|;$$

hence, x_n is a Cauchy sequence, i.e., there exists an $x \in F^*$ such that $x_n \to x$. Thus, $Mx_n = y_n \to Mx \in MF^*$ and $y = Mx$, i.e., MF^* is closed.

Next, suppose that $MF^* \neq F^*$; then there exists a $z \neq 0$ such that $z \perp MF^*$. By (2.7), however,

$$0 = \langle Mz, z\rangle \geqq c\|z\|^2 > 0,$$

a contradiction; hence, M is onto and M^{-1} exists. Finally, setting $u = M^{-1}x$ in the above inequality, we get $\|x\| \geqq c\|M^{-1}x\|$, which completes the proof.

Note that Lemma 7 is true even without the assumption that M is bounded; see [5].

LEMMA 8. *For every $T \in \mathscr{T}$ the operator S_T is self-adjoint, i.e.*,

$$\langle S_Tx, y\rangle = \langle x, S_Ty\rangle \quad \textit{for any } x, y \in F^*.$$

Proof. If λ is any real number, we have by axioms (iv) and (i),

$$\|S_Tx - \lambda y\|^2 - \|S_Tx - \lambda S_Ty\|^2 \geqq 0.$$

Hence,

$$\lambda^2(\|y\|^2 - \|y_T\|^2) - 2\lambda\{\langle x_T, y\rangle - \langle x_T, y_T\rangle\} \geqq 0$$

for all λ, and consequently, $\langle x_T, y\rangle - \langle x_T, y_T\rangle = 0$. Interchanging x and y, we get $\langle y_T, x\rangle - \langle y_T, x_T\rangle = 0$, which completes the proof.

LEMMA 9. *Let P be a linear operator from F^* onto F^* which is causal on F^* with respect to $\mathscr{T}$. Assume that there exists a $\mu > 0$ such that*

$$\langle Px, x\rangle \geqq \mu\|x\|^2 \quad \textit{for all } x \in E^*. \tag{2.8}$$

Then P^{-1} exists and is causal on F^* with respect to $\mathscr{T}$.

Proof. Inequality (2.8) shows that P is one-to-one so that P^{-1} exists. For any $y \in F^*$ and $T \in \mathscr{T}$ we have $\langle PS_Ty, S_Ty\rangle \geqq \mu\|S_Ty\|^2$ in view of axiom (iii). Thus, by Lemma 8 and the causality of P,

$$\mu\|S_Ty\|^2 \leqq \langle S_TPS_Ty, S_Ty\rangle = \langle S_TPy, S_Ty\rangle. \tag{2.9}$$

Setting $y = P^{-1}x \in F^*$ with $x \in F^*$ yields

$$\langle S_Tx, S_TP^{-1}x\rangle \geqq \mu\|S_TP^{-1}x\|^2. \tag{2.10}$$

This implies that $S_TP^{-1}x = 0$ if $S_Tx = 0$ for some $x \in F^*$ and $T \in \mathscr{T}$. The proof follows now from the linearity of P^{-1}, axiom (ii) and Lemma 1.

Note that, due to a theorem by Phillips (see [5]), the assumption that P is onto may be dropped.

Remark. If we assume that the operator P is one-to-one, onto and bounded, then Lemma 9 follows even if $\mu = 0$. To prove this, note first that P^{-1} is also bounded by the closed graph theorem. If $0 < \varepsilon < \frac{1}{2}\|P^{-1}\|^{-1}$, define $P_\varepsilon : F^* \to F^*$ by $P_\varepsilon = P + \varepsilon I = P(I + \varepsilon P^{-1})$. Clearly, P_ε possesses the inverse

$$P_\varepsilon^{-1} = \left\{ I + \sum_{k=1}^{\infty} (-1)^k \varepsilon^k (P^{-1})^k \right\} P^{-1}.$$

Consequently,

$$\|P_\varepsilon^{-1} - P^{-1}\| \leqq \|P^{-1}\| \sum_{k=1}^{\infty} \varepsilon^k \|P^{-1}\|^k \leqq 2\varepsilon \|P^{-1}\|^2. \tag{2.11}$$

On the other hand, P_ε satisfies the hypotheses of Lemma 9, i.e., P_ε^{-1} is causal on F^*. Thus, choosing $x \in F^*$ and $T \in \mathscr{T}$, we have by axiom (iv) and (2.11),

$$\begin{aligned} \|(S_T P^{-1} - S_T P^{-1} S_T)x\| &\leqq \|S_T(P^{-1} - P_\varepsilon^{-1})x\| \\ &\quad + \|S_T(P^{-1} - P_\varepsilon^{-1})S_T x\| + \|(S_T P_\varepsilon^{-1} - S_T P^{-1} S_T)x\| \\ &\leqq 4\varepsilon \|P^{-1}\|^2 \cdot \|x\|. \end{aligned}$$

Since ε may be chosen arbitrarily small and x is arbitrary, it follows that $S_T P^{-1} - S_T P^{-1} S_T = 0$.

LEMMA 10. *Let $M : F^* \to F^*$ be a linear bounded operator, and let a $c \geqq -\frac{1}{2}$ exist such that*

$$\langle Mx, x \rangle \geqq c\|x\|^2 \quad \text{for all } x \in F^*. \tag{2.12}$$

Then $(I + M)^{-1}$ exists and

$$\|(I + M)^{-1} M\| \leqq \{1 - (2c + 1)(1 + \|M\|)^{-2}\}^{1/2}.$$

Proof. The existence and boundedness of $(I + M)^{-1}$ follows from Lemma 7. Next, choose $x \in F^*$ and let $y = (I + M)^{-1} Mx$; then $y = x - (I + M)^{-1}x$. Thus, setting $z = (I + M)^{-1}x$, we have $y = x - z$, and consequently,

$$\begin{aligned} \|y\|^2 &= \|x\|^2 - 2\langle x, z \rangle + \|z\|^2 = \|x\|^2 - 2\langle z + Mz, z \rangle + \|z\|^2 \\ &= \|x\|^2 - 2\langle Mz, z \rangle - \|z\|^2. \end{aligned} \tag{2.13}$$

However, by (2.12),

$$2\langle Mz, z \rangle + \|z\|^2 \geqq (2c + 1)\|z\|^2;$$

moreover, from $x = (I + M)z$ we get $\|x\| \leqq (1 + \|M\|)\|z\|$, i.e.,

$$\|z\| \geqq (1 + \|M\|)^{-1}\|x\|.$$

Introducing this into (2.13), we obtain the desired result.

We are now ready to state the main result of this section.

THEOREM 2.2. *Let K, $N : F \to F$ be operators causal with respect to $\mathscr{T}$, let K be linear and assume that the following conditions are satisfied*:

(i) *there exist constants $k_1 \geqq 0$ and $k_2 \geqq 0$ such that*

$$\langle S_T(x_1 - x_2), S_T(Nx_1 - Nx_2) \rangle \geqq k_1 \|S_T(x_1 - x_2)\|^2 \tag{2.14}$$

and

$$\|S_T(Nx_1 - Nx_2)\| \leqq k_2\|S_T(x_1 - x_2)\| \tag{2.15}$$

for all $x_1, x_2 \in F$ *and* $T \in \mathscr{T}$;

(ii) *the restriction* $\tilde{K}$ *of* K *to* F^* *maps* F^* *into itself, is bounded and there exists a* $c \geqq 0$ *such that*

$$\langle \tilde{K}x, x\rangle \geqq c\|x\|^2 \quad \textit{for all } x \in F^*; \tag{2.16}$$

(iii) $k_1 + c > 0$.

Then the operator $A: F \times F \to F$ *defined by* $A(x, u) = -KNx + u$ *possesses an L-continuous resolvent* Q, *which is causal with respect to* $\mathscr{T}$.

If, in addition, $N\theta \in F^*$, *then* Q *maps* F^* *into* F^*.

Proof. Referring to Theorems 1.2 and 1.3, we are going to show that, with $C = -K$, there exists a $\lambda > 0$ such that conditions (i)* and (i)–(v) are satisfied.

Let $\lambda > 0$. Then the operator $\lambda\tilde{K}$ is linear and bounded on F^*, causal on F^* due to causality of K, and satisfies the condition $\langle \lambda\tilde{K}x, x\rangle \geqq \lambda c\|x\|^2$ by (2.16). Hence, according to Lemma 7, $I + \lambda\tilde{K}$ is one-to-one from F^* onto F^* and has a bounded inverse. Since $I + \lambda\tilde{K}$ is also causal on F^*, and $\langle (I + \lambda\tilde{K})x, x\rangle \geqq (1 + \lambda c)\|x\|^2$ for all $x \in F^*$, Lemma 9 shows that $(I + \lambda\tilde{K})^{-1}$ is causal on F^*. Finally, invoking Lemma 10 it follows that

$$\|(I + \lambda\tilde{K})^{-1}\tilde{K}\|^2 \leqq \frac{2(\|\tilde{K}\| - c) + \lambda\|\tilde{K}\|^2}{\lambda(1 + \lambda\|\tilde{K}\|)^2}. \tag{2.17}$$

Hence, condition (i)* in Theorem 1.3 is satisfied.

On the other hand, by (2.14) and (2.15) we have for every $x_1, x_2 \in F$ and $T \in \mathscr{T}$,

$$\begin{aligned} \|S_T\{Nx_1 - Nx_2 - \lambda(x_1 - x_2)\}\|^2 &= \|S_T(Nx_1 - Nx_2)\|^2 \\ &\quad + \lambda^2\|S_T(x_1 - x_2)\|^2 \\ &\quad - 2\lambda\langle S_T(Nx_1 - Nx_2), S_T(x_1 - x_2)\rangle \\ &\leqq \mu^2(\lambda)\|S_T(x_1 - x_2)\|^2, \end{aligned} \tag{2.18}$$

where $\mu^2(\lambda) = k_2^2 - 2\lambda k_1 + \lambda^2$. Thus,

$$\|(I + \lambda\tilde{K})^{-1}\tilde{K}\|^2\mu^2(\lambda) \leqq \varkappa(\lambda), \tag{2.19}$$

where

$$\begin{aligned} \varkappa(\lambda) &= \frac{(\lambda^2 - 2\lambda k_1 + k_2^2)[\lambda\|\tilde{K}\|^2 + 2(\|\tilde{K}\| - c)]}{\lambda(1 + \lambda\|\tilde{K}\|)^2} \\ &= 1 - \frac{\lambda^2\rho + p(\lambda)}{\lambda(1 + \lambda\|\tilde{K}\|)^2}, \end{aligned}$$

and $p(\lambda)$ is a linear polynomial while $\rho = 2(k_1\|\tilde{K}\|^2 + c) > 0$ by assumption (iii). Hence, $\varkappa(\lambda) < 1$ for sufficiently large λ. Thus, condition (iv) in Theorem 1.2 is satisfied, and the same is true for (ii), (iii) and (v). The proof of the fact $Q: F^* \to F^*$ under the condition $N\theta \in F^*$ is obvious.

We now proceed to discuss some specific examples which employ the above results.

2.2. Let $\Omega = [0, \infty)$ and put $\mathscr{T} = \{[0, \tau] : \tau > 0\}$, $\mathscr{C} = R^n$. If $a \in R^n$, let $|a|$ denote the Euclidean norm of a. Furthermore, let

$$F = \left\{ x : x \text{ measurable}, \int_0^\tau |x(t)|^2 \, dt < \infty \text{ for every } 0 < \tau < \infty \right\},$$

$$F^* = \left\{ x : x \text{ measurable}, \int_0^\infty |x(t)|^2 \, dt < \infty \right\},$$

and assume that F^* is equipped with the inner product

$$\langle x, y \rangle = \int_0^\infty x^T(t) y(t) \, dt$$

(the superscript T signifies the transposition of a vector). If $S_{[0,\tau]}$ is defined by

$$(S_{[0,\tau]}x)(t) = \begin{cases} x(t) & \text{for } 0 \leqq t \leqq \tau, \\ 0 & \text{elsewhere}, \end{cases}$$

then all the axioms on S_T are satisfied.

Next, let $M(t)$, $N(t)$ be continuous $n \times n$ matrix-valued functions on $[0, \infty)$ and let $N(t)$ be symmetric, with a continuous derivative $N'(t)$ on $[0, \infty)$. Assume that the matrices $M(t)$, $N(t)$ and $-N'(t)$ are positive semidefinite for every $t \geqq 0$.

The operator A is defined on F by

$$(Ax)(t) = M(t)x(t) + N(t) \int_0^t x(\xi) \, d\xi. \tag{2.20}$$

It is obvious that A maps F into itself and is causal with respect to $\mathscr{T}$. Furthermore, if $T = [0, \tau] \in \mathscr{T}$ and $x \in F$, then

$$\langle S_T x, S_T A x \rangle = \int_0^\tau x^T(t) M(t) x(t) \, dt + \int_0^\tau x^T(t) N(t) \int_0^t x(\xi) \, d\xi \, dt.$$

Setting $v(t) = \int_0^t x(\xi) \, d\xi$, it follows due to the symmetry of $N(t)$ that

$$\langle S_T x, S_T A x \rangle = \int_0^\tau x^T(t) M(t) x(t) \, dt + \tfrac{1}{2} v^T(\tau) N(\tau) v(\tau) - \frac{1}{2} \int_0^\tau v^T(\xi) N'(\xi) v(\xi) \, d\xi \geqq 0.$$

Denoting by $|P|$ the norm of (the $n \times n$ matrix) P associated with the Euclidean norm of an n-vector, we then have

$$\|S_T A x\| \leqq \left(\int_0^\tau x^T M^T M x \, dt \right)^{1/2} + \left(\int_0^\tau \left[\left(\int_0^t x \, d\xi \right)^T N^T N \left(\int_0^t x \, d\xi \right) \right] dt \right)^{1/2}.$$

Using the Schwarz inequality, we obtain

$$\|S_T A x\| \leqq \left\{ \sup_{[0,\tau]} |M(t)| + \left(\int_0^\tau t |N(t)|^2 \, dt \right)^{1/2} \right\} \|S_T x\|. \tag{2.21}$$

Since A is linear, inequalities (2.20), (2.21) show that the conditions (i) and (ii) in Theorem 2.1 are met; hence, the operator $\tilde{A}$ defined by $\tilde{A}(x, u) = -Ax + u$ possesses a causal resolvent Q.

Moreover, if matrices $M(t)$ and $N(t)$ satisfy the condition

$$\sup_{[0,\infty]} |M(t)| + \left(\int_0^\infty t|N(t)|^2\, dt\right)^{1/2} < \infty,$$

then clearly Q is L-continuous and maps F^* into itself.

2.3. Now let $\Omega = R_+^2$, $\mathscr{T} = \{[0, \tau_1] \times [0, \tau_2] : \tau_1 > 0, \tau_2 > 0\}$, and $\mathscr{C} = R^1$. Also, let

$$F = \left\{x : x \text{ measurable}, \int_0^{\tau_1}\int_0^{\tau_2} x^2(\xi_1, \xi_2)\, d\xi_1\, d\xi_2 < \infty \text{ for } 0 < \tau_1, \tau_2 < \infty\right\},$$

$$F^* = \left\{x : x \text{ measurable}, \int_0^\infty\int_0^\infty x^2(\xi_1, \xi_2)\, d\xi_1\, d\xi_2 < \infty\right\},$$

and let the inner product on F^* be defined in the obvious way. If $T = [0, \tau_1] \times [0, \tau_2] \in \mathscr{T}$ and if we define

$$(S_T x)(t_1, t_2) = \begin{cases} x(t_1, t_2) & \text{for } (t_1, t_2) \in T, \\ 0 & \text{elsewhere,} \end{cases}$$

then S_T satisfies the axioms.

Next, let $w(t_1, t_2)$ be a real measurable function on R_+^2 such that

$$W = \int_0^\infty\int_0^\infty |w(t_1, t_2)|\, d\xi_1\, d\xi_2 < \infty.$$

Also, let $f(\xi)$ be a real function on R^1 and suppose that $0 \leqq \alpha_1 \leqq \alpha_2$ exist such that

$$\alpha_1(\xi_1 - \xi_2)^2 \leqq (f(\xi_1) - f(\xi_2))(\xi_1 - \xi_2) \leqq \alpha_2(\xi_1 - \xi_2)^2 \tag{2.22}$$

for all $\xi_1, \xi_2 \in R^1$; also, assume that $f(0) = 0$. The operator A on F is now defined by

$$(Ax)(t_1, t_2) = \int_0^{t_1}\int_0^{t_2} w(t_1 - \xi_1, t_2 - \xi_2) f(x(\xi_1, \xi_2))\, d\xi_1\, d\xi_2. \tag{2.23}$$

Referring to Theorem 2.2, let

$$\begin{aligned} (Kx)(t_1, t_2) &= \int_0^{t_1}\int_0^{t_2} w(t_1 - \xi_1, t_2 - \xi_2) x(\xi_1, \xi_2)\, d\xi_1\, d\xi_2, \\ (Nx)(t_1, t_2) &= f(x(t_1, t_2)). \end{aligned} \tag{2.24}$$

From (2.22) it follows that $|f(\xi)| \leqq \alpha_2|\xi|$, and consequently N maps F into itself, and F^* into itself; moreover, we easily conclude that

$$\langle S_T(x_1 - x_2), S_T(Nx_1 - Nx_2)\rangle \geqq \alpha_1\|S_T(x_1 - x_2)\|^2$$

for all $x_1, x_2 \in F$ and $T \in \mathscr{T}$. Also, N is clearly causal with respect to $\mathscr{T}$.

On the other hand, a routine application of the Schwarz inequality shows that K maps F^* into itself and that $\|Kx\| \leqq W\|x\|$ for all $x \in F^*$; since K is clearly causal, it follows that $K:F \to F$.

Next, let $\hat{w}(i\omega_1, i\omega_2)$ be the Fourier transform of $w(t_1, t_2)$. Choosing an $x \in F^*$, let $\tilde{x}(i\omega_1, i\omega_2)$ stand for the Fourier–Plancherel transform of x. Then we have, by (2.24), the theorem on transforms of the convolution, and Parseval's equality (the bar denotes the complex conjugate),

$$\begin{aligned}
\langle x, Kx \rangle &= \int_0^\infty \int_0^\infty x(t_1, t_2)(Kx)(t_1, t_2)\, dt_1\, dt_2 \\
&= (2\pi)^{-2} \int_{-\infty}^\infty \int_{-\infty}^\infty \bar{\hat{x}}(i\omega_1, i\omega_2)(\widehat{Kx})(i\omega_1, i\omega_2)\, d\omega_1\, d\omega_2 \\
&= (2\pi)^{-2} \int_{-\infty}^\infty \int_{-\infty}^\infty \operatorname{Re} \hat{x}(\widehat{Kx})\, d\omega_1\, d\omega_2 \\
&= (2\pi)^{-2} \int_{-\infty}^\infty \int_{-\infty}^\infty \operatorname{Re} |\hat{x}|^2 \hat{w}\, d\omega_1\, d\omega_2 \\
&\geqq \varkappa (2\pi)^{-2} \int_{-\infty}^\infty \int_{-\infty}^\infty |\hat{x}|^2\, d\omega_1\, d\omega_2 \\
&= \varkappa \int_0^\infty \int_0^\infty x^2(t_1, t_2)\, dt_1\, dt_2 \\
&= \varkappa \|x\|^2,
\end{aligned}$$

where we have set

$$\varkappa = \inf_{(\omega_1, \omega_2) \in R^2} \operatorname{Re} \hat{w}(i\omega_1, i\omega_2).$$

Hence, if $\varkappa \geqq 0$ and $\alpha_1 + \varkappa > 0$, Theorem 2.2 shows that the operator $\tilde{A}(x, u) = -Ax + u$ possesses an L-continuous resolvent Q, which is causal with respect to the scale $\mathscr{T}$ and is such that $Q:F^* \to F^*$.

REFERENCES

[1] I. I. Kolodner, *Fixed points*, Amer. Math. Monthly, 71 (1964), p. 906.

[2] I. W. Sandberg, *Some results on the theory of physical systems governed by nonlinear functional equations*, Bell System Tech. J., 44 (1965), pp. 871–898.

[3] I. I. Kolodner, *Equations of Hammerstein's type in Hilbert spaces*, J. Math. Mech., 13 (1964), pp. 701–750.

[4] I. W. Sandberg, *On the L_2-boundedness of solutions of nonlinear functional equations*, Bell System Tech. J., 43 (1964), pp. 1581–1599.

[5] R. S. Phillips, *Dissipative operators and hyperbolic systems of partial differential equations*, Trans. Amer. Math. Soc., 90 (1959), pp. 193–254.

[6] G. Zames, *Functional analysis applied to nonlinear feedback systems*, IEEE Trans. Circuit Theory, CT-10 (1963), pp. 392–404.

SIAM J. CONTROL
Vol. 11, No. 2, May 1973

ON FEEDBACK CONTROL OF LINEAR STOCHASTIC SYSTEMS*

ANDERS LINDQUIST†

Abstract. Feedback control of linear continuous-time stochastic systems of general type is discussed. Various types of (classical) information patterns with both complete and partial observations (white and colored measurement noise) are considered. The cost functional is quadratic. A class of admissible control laws is defined which includes all linear and nonlinear control policies for which our problem makes sense, i.e., existence, uniqueness etc. are secured. Then, we determine the optimal control law by an imbedding procedure which amounts to solving a problem without a feedback loop. We investigate under what conditions the optimal control law is linear in the data.

1. Introduction. In recent years there has been a considerable interest in feedback control of linear continuous-time stochastic systems. However, as pointed out by Witsenhausen [20], the difficulties created by the feedback loop have frequently been overlooked, and therefore many results have appeared which as yet have not been rigorously justified. On the other hand, as one might expect, many rigorous proofs suffer from undesired technical restrictions.

The most well-known problem of this type is the stochastic linear-quadratic regulator problem with noisy measurements, for which various versions of the "separation theorem" hold. These versions usually differ in the way in which the set of admissible control laws is defined. By confining ourselves to control laws which are linear in the data, we can easily avoid the difficulties mentioned above. However, we usually want to compare them with nonlinear control laws even when such a comparison rules in favor of a linear one. To the author's knowledge the first fully rigorous proof along these lines appeared in the book [16] by Kushner, where only control laws satisfying a uniform Lipschitz condition in a certain state estimate are admitted. The state estimate is assumed to be generated by a linear Kalman filter to which the nonlinear feedback loop is added, but it is shown that this estimate is indeed the expected value of the current state given past observations, as long as we confine ourselves to admissible control laws. In the well-known paper [21] by Wonham the class of admissible control laws is defined in a more straightforward way, first excluding the possibility that the "information" carried by the observation process is control-dependent by requiring the control to be Lipschitz in this process. Then, the admissible control laws are defined to be Lipschitz continuous functions of the conditional expectation of the state given past observations. Moreover, the separation theorem is generalized to hold for nonquadratic cost functionals. (Also see [22].)

Of course the Lipschitz conditions are imposed to insure that there exist *unique* solutions of the feedback equations. Otherwise the problem would not

* Received by the editors May 2, 1972.

† Institute for Optimization and Systems Theory, Royal Institute of Technology, Stockholm 70, Sweden, and Department of Mathematics, University of Florida, Gainesville, Florida 32601. This research was supported by the Swedish Board for Technical Development under Grant 71-196/U126.

make sense. Unfortunately, these conditions exclude many control laws which are not sufficiently smooth but for other reasons are natural to admit. (See § 3.) In order to get rid of these technical restrictions, Davis and Varaiya [5] defined a new concept of solution using a theorem of Girsanov [11] to eliminate the control dependence. For other contributions in this spirit see Beneš [2] and Davis and Varaiya [6]. In a recent paper by Lindquist [17] on optimal control of linear stochastic systems (primarily devoted to stochastic functional differential equations) most technical restrictions mentioned above are dispensed with without renouncing the usual concept of solution. Also in a paper by Bensoussan [3] on the separation principle for distributed parameter systems the set of control laws is defined so as to avoid undesirable restrictions. However, contrary to [5] and [21], in both these papers the cost functional is quadratic.

In this paper we consider feedback control of linear stochastic systems of general type. Various types of information patterns with both complete and partial observations are considered. The cost functional is quadratic, but it is of a more general type than usually encountered in the literature. The approach is the same as that of [17], but the objective of this paper is somewhat different. In [17] our prime purpose was to determine explicit feedback solutions for linear stochastic time-lag systems. But, since for such systems the conditional expectation of the current state given past observations is no longer a sufficient statistic, we could not adhere to the approach of [21]. Thus we had to define our set of admissible control laws with a minimum of technical restrictions. However, rather than to discuss the problems of feedback, our main effort was to demonstrate that stochastic time-lag problems of the most general type can be handled in a rigorous way. Therefore, in this paper we shall present a more detailed discussion of our feedback approach, and at the same time we shall be able to present some extensions. In order to avoid obscuring our exposition, we have used technically less complicated examples than in [17] to illustrate our basic ideas. Nevertheless, in certain aspects they will be more general.

In § 3 we discuss the problems of feedback in a general context. We define the concepts of *stochastic open loop* (SOL) problem and *feedback* (FB) problem. A SOL problem is usually easy to solve but what we want is a solution of a FB problem. Therefore, our basic method is to imbed our FB problem in a suitable SOL problem, and to this end, in § 4, we derive an identity for the cost functional. In § 5, we investigate what conditions we have to impose *on the system* in order that the *optimal* control law be linear. This is to simplify the imbedding procedure and also to enable the practical implementation of the optimal control law. Thus we define our system so that among all nonlinear control laws which make sense (conditions of existence, uniqueness, etc. are fulfilled) the optimal one is linear. For stochastic systems of the type discussed above, this amounts to requiring the perturbing noise process to be a martingale in the case of complete observations and a Wiener process for partial observations. We may well be able to solve a SOL problem without these conditions, but the solution is usually of limited interest to us, since we do not know of any method to decide whether an arbitrary nonlinear control law is admissible for our FB problem. Finally, in § 6 we give some simple examples to illustrate our method. For instance, we prove the separation theorem for colored measurement noise and for time delay in the control.

For more explicit control and filtering solutions of systems with delay in the state process we refer the reader to [17] and [18].

2. Preliminaries. Let $x_0(t)$ be a (fixed) measurable[1] n-dimensional stochastic process with bounded second order moments, and let $K(t, s)$ be an $n \times m$ matrix function such that $\int_0^T |K(t, s)|^2\, ds$ is bounded. ($|\cdot|$ is the Euclidean norm.) We shall define three vector functions taking values in R^m, R^n and R^k, respectively, namely the *input* or *control* $u(t)$, the *state* $x(t)$, and the *output* or *observation* $z(t)$. These functions are related to each other in the following way:

$$x(t) = x_0(t) + \int_0^t K(t, s)u(s)\, ds, \tag{2.1}$$

$$z(t) = Hx(t), \tag{2.2}$$

where H is a constant $k \times n$ matrix. In the sequel, we shall often use the following shorthand notation:

$$x = x_0 + Ku, \tag{2.1$'$}$$

$$z = Hx. \tag{2.2$'$}$$

Therefore whenever u is a measurable stochastic process such that $E|u(t)|^2 < \infty$ is integrable, x and z are also measurable processes, and they have bounded second order moments.

Our object, however, will be to construct a feedback system. At each time t, $u(t)$ should be formed as a functional of observations received so far: $\{z(s);\ 0 \leqq s \leqq t\}$ in such a way as to minimize

$$EV_0(x, u), \tag{2.3}$$

where

$$V_s(x, u) = \int_s^T x'(t)Q_1(t)x(t)\, d\alpha(t) + \int_s^T u'(t)Q_2(t)u(t)\, dt. \tag{2.4}$$

Here Q_1 and Q_2 are bounded matrix functions which are nonnegative definite and positive definite respectively, denotes transpose and E expectation, and α is a monotone nondecreasing bounded function which is continuous on the right and thus defines a finite Borel measure μ_α. Moreover, Q_2 has a bounded inverse Q_2^{-1}.

In order to facilitate the formulation of this problem in more precise mathematical terms, we shall define a few concepts: Let P^k be the set of all measurable k-dimensional stochastic processes, and S^m the set of m-dimensional stochastic variables. Then the function

$$\pi : [0, T] \times P^k \to S^m$$

is a *nonanticipative* function of z if $\pi(t, z)$ is a function of $\{z(s);\ 0 \leqq s \leqq t\}$ only for

[1] In this paper a *measurable* n-dimensional stochastic process will be a $\mathscr{B} \times \mathfrak{S}$-measurable function $[0, T] \times \Omega \to R^n$, where $\mathscr{B}$ and $\mathfrak{S}$ are the sigma fields of Borel sets and events respectively. Then we have assumed an underlying complete probability space $(\Omega, \mathfrak{S}, P)$, where as usual Ω is the sample space with elements ω and P is the probability measure. As usual we shall write $x_0(t)$ instead of $x_0(t, \omega)$. All deterministic functions defined in this paper are Borel measurable.

each t and defines an element in P^m. The measurable process x is a *stochastic B-solution* of the equation

$$x(t) = x_0(t) + \int_0^t K(t,s)\pi(s, Hx)\, ds \tag{2.5}$$

if for each $t \in [0, T]$ it satisfies (2.5) with probability 1 and $E|x(t)|^2$ is *bounded*. In this paper we shall make no distinction between equivalent processes, i.e., processes which for each t are equal with probability 1.

Our model (2.1) of the controlled system is sufficiently general to include linear dynamic systems such as stochastic differential equations and stochastic functional differential equations. Since our prime interest is in differential systems of this type, the technical assumptions of boundedness imposed above are natural and convenient, but it should be pointed out that they are in no way crucial.

3. Feedback in linear stochastic systems. Let $\{\mathfrak{S}_t \subset \mathfrak{S}; 0 \leqq t \leqq T\}$ be a family of sigma fields and let $\mathscr{U}$ be the set of all m-dimensional stochastic processes such that:

(i) $u(t, \omega)$ is measurable (t, ω);

(ii) $\int_0^T E|u|^2\, dt < \infty$;

(iii) $u(t)$ is $\mathfrak{S}_t$-measurable for almost all t.

Consider the problem of finding a $u^* \in \mathscr{U}$ so as to minimize

$$EV_0(x_0 + Ku, u).$$

It will be shown in the Appendix that there indeed exists a unique $u^* \in \mathscr{U}$ for which the minimum is attained. Following [17], such a problem will be called a *stochastic open loop* (SOL) problem and $\mathscr{U}$ a SOL class. If all $\mathfrak{S}_t \equiv \mathfrak{S}_0$, we have an *open loop* (OL) problem, which is essentially an ordinary variational problem, but in general the SOL problem corresponds to the situation where the available amount of "information" (given by $\mathfrak{S}_t$) varies (usually increases) with time but is unaffected by the choice of u.

However, we are primarily interested in problems where information about the state process is provided by the observation process

$$z = Hx. \tag{3.1}$$

The problem is to determine a control law, that is, to design a "black box" in which the observations received so far are filtered and fed back into the system as a control signal (Fig. 1). Then we have a *feedback* (FB) problem. The "black box" will be described mathematically by a nonanticipative function $\pi:(t, z) \rightarrow u(t) = \pi(t, z)$, and we shall use the shorthand notation:

$$u = \pi z. \tag{3.2}$$

Of course, we have to define the set of admissible π in such a way that there exists a unique solution of the stochastic functional equation created by the feedback loop. (To this end, Wonham [21] only admitted π for which a certain Lipschitz condition is fulfilled. However, for technical reasons which will be revealed below we do not choose to formulate our problem in this way.) To avoid these rather intricate problems of existence, we could instead of our FB problem solve the

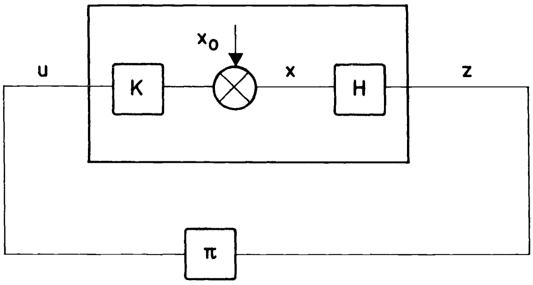


FIG. 1. *The FB problem*

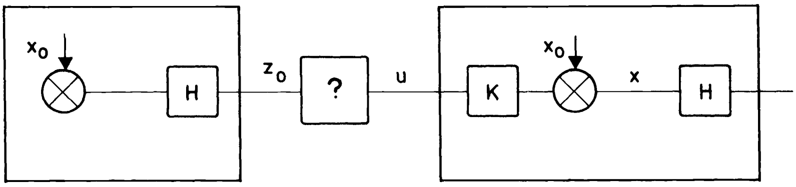


FIG. 2. *A SOL problem*

SOL problem (Fig. 2):

$$\min_{u\in\mathscr{U}_0} EV_0(x_0 + Ku, u), \tag{3.3a}$$

where $\mathscr{U}_0$ is the SOL class defined by

$$\mathfrak{S}_t = \sigma\{z_0(s); 0 \leqq s \leqq t\}; \tag{3.3b}$$

that is, the family of sigma fields generated by

$$z_0 = Hx_0. \tag{3.4}$$

Methods along these lines have been proposed [23], [24]. In fact, z_0 can be determined by subtracting HKu from z. However, a control system designed in this way will not be a proper closed loop system, and it will obviously lack some desirable properties associated with the concept of feedback. Also, the reader is warned against exchanging z_0 for z in (3.3), for then we cannot a priori assume that $\{\mathfrak{S}_t\}$ is constant with respect to variations of the control, and moreover questions of existence have to be settled.

Now, we define our class of admissible control laws in the following way: Π is the class of all nonanticipative functions

$$\pi:[0, T] \times P^k \rightarrow S^m$$

which are *measurable* in the sense that $\pi(t, y)$ is $\sigma\{y(s); 0 \leqq s \leqq t\}$-measurable for all (t, y) for which π is defined, and which fulfill the following conditions:

(i) there exists a unique stochastic B-solution x_π of

$$x = x_0 + K\pi Hx; \tag{3.5}$$

(ii) $u = \pi Hx_\pi \in \mathscr{U}_0$.

Then our problem is to determine a $\pi^* \in \Pi$ which minimizes

$$EV_0(x_\pi, \pi H x_\pi),$$

but in general we do not know whether there really exists an optimal π. However, if we define $\mathcal{U}_\Pi$ to be the set of *stochastic processes*

$$\mathcal{U}_\Pi = \{\pi H x_\pi : \pi \in \Pi\}, \tag{3.6}$$

it is clear that $\mathcal{U}_\Pi \subset \mathcal{U}_0$ and that

$$\begin{aligned} \inf_{\pi \in \Pi} EV_0(x_\pi, \pi H x_\pi) &= \inf_{u \in \mathcal{U}_\Pi} EV_0(x_0 + Ku, u) \\ &\geqq \min_{u \in \mathcal{U}_0} EV_0(x_0 + Ku, u). \end{aligned} \tag{3.7}$$

So if we can find an optimal u^* for the problem (3.3) so that $u^* \in \mathcal{U}_\Pi$, then we have found a solution of our FB problem provided that we can also determine a $\pi^* \in \Pi$ such that $u^* = \pi^* H x_{\pi^*}$.

At first sight it seems quite reasonable to assume that the class Π of admissible control laws includes all π for which our problem makes sense. The only point on which this claim could be questioned is the condition that $u(t) = \pi(t, Hx_\pi)$ be $\sigma\{z_0(s); 0 \leqq s \leqq t\}$-measurable for almost all t. However, it should be noted that this condition is true whenever the solution x_π of (3.5) is such that $z_\pi = Hx_\pi$ can be constructed as the limit in probability of a sequence of (measurable) nonanticipative functions of z_0. Therefore, for all practical purposes we can safely ignore all π which do not belong to Π.

As an *example* let us consider the following stochastic functional differential equation:

$$\begin{aligned} dx = [A_1(t)x(t) + A_2(t)x(t-h) &+ \int_{t-h}^{t} A_0(t,s)x(s)\,ds \\ &+ B_1(t)u(t) + B_2(t)u(t-h)]\,dt + C(t)\,dv \quad \text{for } t \geqq 0; \end{aligned} \tag{3.8}$$

$$x(t) = \xi(t) \quad \text{for } t \leqq 0,$$

where A_0, A_1, A_2, B_1, B_2 and C are bounded matrix functions, $B_2 \equiv 0$ for $t < h$, the delay $h > 0$, v is a stochastic vector process with orthogonal stationary increment such that

$$Ev(t) = 0; \qquad E\{v(s)v'(t)\} = I \min(s,t) \tag{3.9}$$

and ξ is a process with bounded second order moments. The processes ξ and v are independent.

Problems of this type have been studied under more general conditions in [17], where it was shown that (3.8) can be written in the following equivalent form:

$$x(t) = x_0(t) + \int_0^t K(t,s)u(s)\,ds, \tag{3.10}$$

where

$$\begin{aligned} x_0(t) = \Phi(t,0)\xi(0) &+ \int_{-h}^{0} \left[\Phi(t, s+h)A_2(s+h) + \int_0^h \Phi(t,\tau)A_0(\tau,s)\,d\tau\right]\xi(s)\,ds \\ &+ \int_0^t \Phi(t,s)C(s)\,dv(s), \end{aligned} \tag{3.11}$$

$$K(t,s) = \Phi(t,s)B_1(s) + \Phi(t,s+h)B_2(s+h) \tag{3.12}$$

and Φ is the transition matrix function:

$$\begin{gathered}\frac{\partial\Phi}{\partial t}(t,s) = A_1(t)\Phi(t,s) + A_2(t)\Phi(t-h,s) + \int_{t-h}^{t} A_0(t,\tau)\Phi(\tau,s)\,d\tau \quad \text{for } t \geqq 0; \\ \Phi(s,s) = I; \quad \Phi(t,s) = 0 \quad \text{for } t < 0.\end{gathered} \tag{3.13}$$

Now we have transformed our problem into the type discussed above, and it is not hard to see that the unique solution x_π of $x = x_0 + K\pi Hx$ ($\pi \in \Pi$) is also the unique solution of the feedback equation obtained when $u(t) = \pi(t, Hx)$ is inserted in (3.8). In fact, it is demonstrated in [17, Theorem 5.3] that the two equations can be transformed into each other, so that any (stochastic B-) solution of one is also a solution of the other.

Now, let Y^k be the space of all k-dimensional stochastic processes y which can be represented in the following form:

$$y(t) = \int_0^t q(s)\,ds + \int_0^t D(s)\,dw(s), \tag{3.14}$$

where q is a measurable stochastic process such that $E|q(t)|^2$ is integrable, D is a matrix function with square-integrable elements, and w is a stochastic vector process of type (3.9) with orthogonal increments. Then, putting $u = 0$ in (3.8), it is clear that $x_0 \in Y^n$ and $z_0 \in Y^k$. Define $\mathscr{L}$ to be the class of all functions

$$\varphi : [0,T] \times Y^k \to S^m$$

such that

$$\varphi(t,y) = f(t) + \int_0^t F(t,s)\,dy(s), \tag{$\mathscr{L}$}$$

where f is an L_2 vector function and F is an L_2 matrix kernel ($\iint |F|^2\,ds\,dt < \infty$). If there is a stochastic B-solution x_φ of (3.8) with $u = \varphi z$, we must clearly have $x_\varphi \in Y^n$ and consequently $z_\varphi \in Y^k$.

LEMMA 3.1. *For the dynamic system* (3.8) *we have*: $\mathscr{L} \subset \Pi$.

Proof (cf. [17]). First observe that if $\psi(t)$ is an L_2 matrix function,

$$\int_0^t \psi(s)\varphi(s,y)\,ds = \int_0^t \psi(s)f(s)\,ds + \int_0^t \int_\tau^t \psi(s)F(s,\tau)\,ds\,dy(\tau) \tag{3.15}$$

a.s. for all y for which φ is defined, that is, we can change the order of integration. In fact, considering (3.14) we can divide the last term of (3.15) into two and apply the usual Fubini theorem to the first term (for $\int |q|^2\,ds < \infty$ a.s.) and the stochastic Fubini theorem ([7, p. 431], [10, p. 197]) to the second. Now, inserting

$$K(t,s) = B_1(s) + B_2(s+h)\theta(t-s-h) + \int_s^t \Gamma(\tau,s)\,d\tau$$

(where Γ is an L_2 matrix kernel such that $\Gamma(t,s) = (\partial K/\partial t)(t,s)$ for $t \neq s + h$ and θ is the unit step function) into (3.10) and changing the order of integration, we obtain

$$x(t) = x_0(t) + \int_0^t \left[B_1(s)u(s) + B_2(s)u(s-h) + \int_0^s \Gamma(s,\tau)u(\tau)\,d\tau\right] ds. \tag{3.16}$$

Then after inserting $u(t) = \varphi(t, z)$ into (3.16), applying (3.15) and multiplying by H we have an expression of the following type:

$$dz = dz_0 + \left[\int_0^t G(t, s)\, dz(s) + g(t)\right] dt, \tag{3.17}$$

where G is an L_2 matrix kernel and g is an L_2 vector function.

Now, let the L_2 matrix kernel R be defined by the Volterra resolvent equation

$$G(t, s) = R(t, s) - \int_s^t R(t, \tau)G(\tau, s)\, d\tau, \tag{3.18}$$

exchange G in (3.17) for the right member of (3.18), and change the order of integration. Then we obtain

$$dz = dz_0 + \left[\int_0^t R(t, s)\, dz_0 + \int_0^t R(t, s)g(s)\, ds + g(t)\right] dt$$

which inserted into (3.16) with $u = \varphi z$ yields the unique solution x_φ. Evidently $\varphi H x_\varphi \in \mathscr{U}_0$. (We consider a measurable version of $\varphi(t, z)$. See [7, p. 430] or [10, p. 196].) This concludes the proof.

By prescribing some conditions of regularity on the sample functions of z such as continuity or boundedness, we could define π as a function of individual sample functions of z rather than the whole stochastic process. For example, if z has continuous sample functions, following Wonham [21] we could define the class Ψ of all functions

$$\psi : [0, T] \times C \to R^m$$

such that $\psi(t, \zeta)$ is a function of $\{\zeta(s) ; 0 \leqq s \leqq t\}$ and satisfies a Lipschitz condition

$$|\psi(t, \zeta_1) - \psi(t, \zeta_2)| \leqq \gamma\|\zeta_1 - \zeta_2\|,$$

where $\|\cdot\|$ denotes the sup norm in the space C of continuous functions on $[0, T]$ with values in R^k. Let C^k be the space of all k-dimensional stochastic processes with continuous sample functions, and define Π_{LIP} to be the class of all functions

$$(t, z) \in [0, T] \times C^k \to \psi(t, z) \in S^m,$$

where $\psi \in \Psi$. Then it can be shown that $\Pi_{\text{LIP}} \subset \Pi$. (See [21] and [22].) However, for the control of systems of type (3.8), $\mathscr{L}$ is often a very natural class of control laws. Indeed, below we shall introduce some further conditions so that the optimal control law $\pi^* \in \mathscr{L}$, but in general we have to impose still further conditions in order that $\pi^* \in \Pi_{\text{LIP}}$. In fact, $\mathscr{L} \not\subset \Pi_{\text{LIP}}$, for usually a stochastic integral cannot be defined samplewise. This is only possible if the functions $s \to F(t, s)$ are of bounded variation (and z has continuous sample functions). Then we can integrate by parts to obtain

$$\varphi(t, z) = F(t, t)z(t) - F(t, 0)z(0) - \int_0^t d_s F(t, s)z(s).$$

With a few additional conditions on F this control law will belong to Π_{LIP}.

4. An identity. Let $w(t)$ be a p-dimensional martingale[2] with finite second order moments, zero mean ($w(0) = 0$) and incremental covariances:

$$E\{dw_i(t)\,dw_j(t)\} = \begin{cases} d\beta_i(t), & j = i, \\ 0, & j \neq i, \end{cases} \tag{4.1}$$

where w_i ($i = 1, 2, \cdots, p$) are the components of w, and β_i are monotone nondecreasing bounded functions which are continuous from the right. In the Hilbert space H, with inner product $(\xi, \eta) = E\{\xi, \eta\}$, of all (real-valued) stochastic variables with finite second order moments, define H_t to be the subspace of stochastic variables which are measurable with respect to

$$\mathfrak{S}_t = \sigma\{w(s); 0 \leqq s \leqq t\} \tag{4.2}$$

and $\hat{H}_t$ to be the closed linear hull of $\{w_i(s); 0 \leqq s \leqq t, i = 1, 2, \cdots, p\}$ together with all constants, i.e., the set of all stochastic variables ξ which can be represented in the following way:

$$\xi = \bar{\xi} + \int_0^t f'(s)\,dw(s),$$

where $\bar{\xi}$ is a constant, f is an L_2 vector function, and integration is with respect to the stochastic measure

$$\mu((t_1, t_2]) = w(t_2) - w(t_1)$$

(cf. [10, p. 194]). Let $E_t\xi$ and $\hat{E}_t\xi$ denote the projections of $\xi \in H$ onto H_t and $\hat{H}_t$ respectively, i.e., the conditional and wide sense conditional expectations of ξ given $\{w(s); 0 \leqq s \leqq t\}$. Since $\hat{H}_t \subset H_t$, we can form the orthogonal complement $H_t \ominus \hat{H}_t$ of $\hat{H}_t$ in H_t. Finally, let $\mathscr{U}_w$ be the SOL class with $\{\mathfrak{S}_t\}$ given by (4.2).

LEMMA 4.1. *If* $u \in \mathscr{U}_w$ *is given by*

$$u(t) = \bar{u}(t) + \sum_{i=1}^{p} \int_0^t u_i(t, s)\,dw_i(s) + \tilde{u}(t) \tag{4.3}$$

where $\int |\bar{u}(t)|^2\,dt < \infty$, $\iint |u_i(t, s)|^2\,d\beta_i(s)\,dt < \infty$ ($i = 1, 2, \cdots, p$) *and* $\tilde{u} \in \mathscr{U}_w$ *is a stochastic process such that* $\tilde{u}_i(t) \in H_t \ominus \hat{H}_t$ *for almost all* t ($i = 1, 2, \cdots, m$), *then*

$$EV_0(x_0 + Ku, u) = V_0(\bar{x}, \bar{u}) + \sum_{i=1}^{k} \int_0^T V_s(x_i(\cdot, s), u_i(\cdot, s))\,d\beta_i(s) + EV_0(\tilde{x}, \tilde{u}),$$

$\bar{x}_0, x_i(\cdot, s)$ *and* $\tilde{x}$ *being defined in the following way*:

$$\bar{x}(t) = \bar{x}_0(t) + \int_0^t K(t, \tau)\bar{u}(\tau)\,d\tau,$$

$$x_i(t, s) = m_i(t, s) + \int_s^t K(t, \tau)u_i(\tau, s)\,d\tau,$$

$$\tilde{x}(t) = \tilde{x}_0(t) + \int_0^t K(t, \tau)\tilde{u}(\tau)\,d\tau,$$

[2] $E\{w(s)|w(\tau); 0 \leqq \tau \leqq t\} = w(t)$ for $t < s$. Since $E|w(t)|^2 < \infty$, $w(t)$ has *orthogonal increments*.

where[3] $\bar{x}_0(t) = Ex_0(t)$, $m_i(t, s) = (\partial/\partial\beta_i)E\{x_0(t)w_i(s)\}$, *and* $\tilde{x}_0(t) = x_0(t) - \hat{E}_t x_0(t)$.

Proof (cf. [17]). According to Lemma B.1 (see Appendix),

$$\hat{x}_0(t) = \bar{x}_0(t) + \sum_{i=1}^{k} \int_0^t m_i(t, s)\, dw_i(s)$$

is (a version of) the wide sense conditional mean $\hat{E}_t x_0(t)$. Then, inserting $x_0(t) = \hat{x}_0(t) + \tilde{x}_0(t)$ and (4.3) into $x_0 + Ku$, we have:

$$\begin{aligned} x(t) &= \bar{x}_0(t) + \int_0^t K(t, \tau)\bar{u}(\tau)\, d\tau \\ &\quad + \sum_i \left[\int_0^t m_i(t, s)\, dw_i(s) + \int_0^t K(t, \tau) \int_0^\tau u_i(\tau, s)\, dw_i(s)\, d\tau \right] \\ &\quad + \tilde{x}_0(t) + \int_0^t K(t, \tau)\tilde{u}(\tau)\, d\tau \\ &= \bar{x}(t) + \sum_i \int_0^t x_i(t, s)\, dw_i(s) + \tilde{x}(t), \end{aligned} \tag{4.4}$$

where we have used the stochastic Fubini theorem.

Now due to the martingale property,[4] $\tilde{u}(s) \perp \hat{H}_t$ for almost all $s \leqq t$, and therefore (since $\tilde{x}_0(t) \perp \hat{H}_t$), $\tilde{x}(t) \perp \hat{H}_t$. In fact, if

$$\xi_n = \int_s^t f_n'(s)\, dw(s), \tag{4.5}$$

where f_n is a vector step function, $E_s\xi_n = 0$, and therefore

$$E\{\tilde{u}(s)\xi_n\} = E\{\tilde{u}(s)E_s\xi_n\} = 0$$

for almost all $s \leqq t$, for $E = EE_s$ and $\tilde{u} \in \mathscr{U}_w$. But each $\xi \in \hat{H}_t \ominus \hat{H}_s$ can be represented as the limit in H of a fundamental $\{\xi_n\}$ of type (4.5), and therefore (for almost all $s \leqq t$) $E\{\tilde{u}(s)\xi\} = 0$ for all such ξ, and consequently $\tilde{u}(s) \perp \hat{H}_t \ominus \hat{H}_s$. But by definition, $\tilde{u}(s) \perp \hat{H}_s$ and hence $\tilde{u}(s) \perp \hat{H}_t$ (for almost all $s \leqq t$).

Since the terms of (4.4) are mutually orthogonal and the same is true for (4.3) for almost all t, we have

$$EV_0(x, u) = V_0(\bar{x}, \bar{u}) + \sum_i EV_0(x_i w_i, u_i w_i) + EV_0(\tilde{x}, \tilde{u}),$$

where $(x_i w_i)(t) = \int_0^t x_i(t, s)\, dw_i(s)$ and $u_i w_i$ are defined analogously. However,

$$\begin{aligned} EV_0(x_i w_i, u_i w_i) &= \int_0^T \int_0^t x_i'(t, s)Q_1(t)x_i(t, s)\, d\beta_i(s)\, d\alpha(t) \\ &\quad + \int_0^T \int_0^t u_i'(t, s)Q_2(t)u_i(t, s)\, d\beta_i(s)\, dt \\ &= \int_0^T V_s(x_i(\cdot, s), u_i(\cdot, s))\, d\beta_i(s), \end{aligned}$$

where we have used (4.1) and Fubini's theorem. This concludes the proof.

[3] $(\partial/\partial\beta_i)E\{x_0(t)w_i(s)\} = [(\partial/\partial\sigma)E\{x_0(t)w_i(\beta_i^{-1}(\sigma))\}]_{\sigma=\beta_i(s)}$ (see Appendix).

[4] $\tilde{u}(s) \perp \hat{H}_t$ means "all components of $\tilde{u}(s)$ are orthogonal to $\hat{H}_t$."

Remark. If w is a process of type (3.9), i.e., $\beta_i(t) = t$, we have

$$m_i(t,s) = \frac{\partial}{\partial s} E\{x_0(t)w_i(s)\}.$$

If $x_0(t) \in \hat{H}_t$, we have

$$x_0(t) = \bar{x}_0(t) + \sum_i \int_0^t m_i(t,s)\,dw_i(s).$$

5. The imbedding procedure. In order to exploit the identity derived in the previous section, we note that for many stochastic systems of interest it is possible to represent the uncontrolled observation process z_0 defined in § 3 in the following way:

$$z_0(t) = \bar{z}_0(t) + \int_0^t N(t,s)\,dw(s), \tag{5.1}$$

where w is an r-dimensional martingale defined as in § 4, $\bar{z}_0$ is a bounded deterministic function, and N is a $k \times r$ matrix function with columns N_i such that $\int |N_i(t,s)|^2\,d\beta_i(s)$ are bounded. With $\mathscr{U}_\Pi$, $\mathscr{U}_0$ and $\mathscr{U}_w$ defined as in §§ 3 and 4, we obtain

$$\mathscr{U}_\Pi \subset \mathscr{U}_0 \subset \mathscr{U}_w \tag{5.2}$$

and therefore we have imbedded the set $\mathscr{U}_\Pi$ of all control processes generated by admissible control laws in the SOL class defined by a martingale process. The following lemma will explain our basic method to construct optimal control laws.

LEMMA 5.1. *Let u^* be the optimal solution of the SOL problem*:

$$\min_{u \in \mathscr{U}_w} EV_0(x_0 + Ku, u). \tag{5.3}$$

If there is a $\pi^* \in \Pi$ such that

$$u^*(t) = \pi^*(t, z_0 + HKu^*), \tag{5.4}$$

then there exists an optimal solution of the FB problem

$$\min_{\pi \in \Pi} EV_0(x_\pi, \pi H x_\pi)$$

and it is provided by π^*.

Proof. Let $z^* = z_0 + HKu^*$. Then, according to (5.4), $u^* = \pi^* z^*$ and therefore $z^* = z_0 + HK\pi^* z^*$. But then z^* must be the unique solution z_{π^*} of $z = z_0 + HK\pi^* z$, and consequently $u^* = \pi^* z_{\pi^*}$. Now, due to (5.2),

$$\inf_{u \in \mathscr{U}_\Pi} EV_0(x_0 + Ku, u) \geqq EV_0(x_0 + Ku^*, u^*);$$

but since $u^* \in \mathscr{U}_\Pi$, equality holds. This concludes the proof of the lemma.

Of course this procedure can only be successful if w is so defined that $u^* \in \mathscr{U}_\Pi$. It is therefore desirable that $\mathscr{U}_w$ be as small as possible. If the transformation (5.1) is causally invertible, i.e., w can be represented as a measurable nonanticipative function of z_0, we have $\mathscr{U}_w = \mathscr{U}_0$. Then (5.1) is a *canonical representation* and w will be called an *innovation process.* (See, e.g., [4], [12], [13], [9], [14].) However, in general it is no trivial problem to decide whether u^* really belongs to $\mathscr{U}_\Pi$. Often a nonanticipative function expressing u^* in terms of $z^* = z_0 + HKu^*$ is quite easily determined (e.g., in the form of conditional expectations), but it still remains to show that this function belongs to Π, i.e., that *there exists a unique solution of the corresponding feedback system.* In order to reduce these difficulties and also to make full use of Lemma 4.1, we shall investigate under what conditions on the pair (x_0, w), the optimal SOL control u^* is *linear* in w.

THEOREM 5.2. *If the pair* (x_0, w) *fulfills the condition*

$$E_t x_0(t) = \hat{E}_t x_0(t) \quad \mu_\alpha\text{-a.e. on } [0, T], \tag{5.5}$$

i.e., the conditional and wide sense conditional expectations of $x_0(t)$ *given* $\{w(s);\ 0 \leqq s \leqq t\}$ *coincide for* μ_α*-almost all* t *on* $[0, T]$, *then the optimal solution* u^* *of the SOL problem* (5.3) *is given by*

$$u^*(t) = \bar{u}^*(t) + \sum_{i=1}^{r} \int_0^t u_i^*(t, s)\, dw_i(s), \tag{5.6}$$

where $\bar{u}^*$ *is the optimal* L_2 *solution of the problem*

$$\min_{\bar{u}} V_0(\bar{x}_0 + K\bar{u}, \bar{u}) \tag{5.7}$$

and $u_i^*(\cdot, s)$ $(i = 1, 2, \cdots, r; 0 \leqq s \leqq T)$ *are the optimal* L_2 *solutions of*

$$\min_{u_i(\cdot, s)} V_s(m_i(\cdot, s) + Ku_i(\cdot, s), u_i(\cdot, s)) \tag{5.8}$$

subject to the constraints $u_i(t, s) \equiv 0$ *for* $t < s$. *Here,*

$$\bar{x}_0(t) = Ex_0(t) \quad \text{and} \quad m_i(t, s) = \frac{\partial}{\partial \beta_i} E\{x_0(t) w_i(s)\}.$$

Moreover, if $x^* = x_0 + Ku^*$, *for all* $(\tau, t) \in [0, T] \times [0, T]$ *we have*

$$\hat{E}_\tau x^*(t) = \bar{x}^*(t) + \sum_{i=1}^{r} \int_0^\tau x_i^*(t, s)\, dw_i(s), \tag{5.9}$$

where $\bar{x}^* = \bar{x}_0 + K\bar{u}^*$ *and* $x_i^*(\cdot, s) = m_i(\cdot, s) + Ku_i^*(\cdot, s)$.

Proof. Let $\tilde{u}$ and $\tilde{x}_0$ be defined as in Lemma 4.1. Then, $\tilde{u}_i(s) \in H_s \subset H_t$ for almost all $s \leqq t$ and therefore $(K\tilde{u})_i(t) \in H_t$. Now, condition (5.5) implies that

$$E_t \tilde{x}_0(t) = E_t\{x_0(t) - \hat{E}_t x_0(t)\} = E_t x_0(t) - \hat{E}_t x_0(t) = 0 \quad \mu_\alpha\text{-a.e.},$$

and consequently $\tilde{x}_0(t) \perp H_t$ μ_α-a.e. Therefore $\tilde{x}_0(t)$ and $(K\tilde{u})(t)$ are orthogonal for μ_α-almost all t on $[0, T]$, and hence

$$EV_0(\tilde{x}, \tilde{u}) = EV_0(\tilde{x}_0, 0) + EV_0(K\tilde{u}, \tilde{u}) \geqq EV_0(\tilde{x}_0, 0),$$

where equality holds for $\tilde{u} = 0$. Therefore our assertion (5.6) follows from Lemma 4.1, for problems (5.7) and (5.8) indeed have unique solutions such that u^* defined by (5.6) is a measurable stochastic process (see [17]). To obtain (5.9), insert (5.6) into $x^* = x_0 + Ku^*$, change the order of integration (stochastic Fubini theorem) and apply Lemma B.1. This concludes the proof.

Note that we do not a priori assume that u^* is linear in w, but the linearity is a consequence of condition (5.5) and the martingale property. It should be clear from the proof of Lemma 4.1 that if we confine our set of admissible controls to those which are linear in w, we only need to assume that w have orthogonal increments to insure that (5.6) is optimal.

COROLLARY 5.2. *Either one of the following two conditions is sufficient for* (5.6) *to be the optimal solution of problem* (5.3):

(i) $x_0(t) \in \hat{H}_t$ *on* $[0, T]$;

(ii) x_0 *and* w *are jointly Gaussian.*

When condition (i) *holds,* $x^*(t)$ *can be obtained from* (5.9) *by putting* $\tau = t$, *and when condition* (ii) *is fulfilled,* $\hat{x}^*(t|\tau) = E_\tau x^*(t)$ *is given by* (5.9).

Proof. Both conditions are sufficient for (5.5) to hold. As for condition (ii), see, e.g., [10, pp. 228–229].

Remark. Conditions (i) and (ii) can be weakened so as to exploit the fact that (5.5) only needs to hold for μ_α-almost all t.

Now, if u^* is given by (5.6), we have

$$z^* = Nw + HKu^*, \tag{5.10}$$

$$u^* = U^*w, \tag{5.6'}$$

where N and U^* denote the affine transformations of (5.1) and (5.6) respectively. Then, if we can eliminate w from this system to obtain a nonanticipative function (5.4) expressing u^* in terms of z^*, there should be no problem in establishing whether this control law really belongs to Π, for it is linear in z^*. Once the optimal control law π^* has been determined, there remains the problem of implementing it. We shall refer to this problem as the *filtering problem*, since it often amounts to constructing a linear filter whose transfer property is described by π^*.

As an example, let us return to the dynamic system defined by (3.8). It is clear from (3.11) that if ξ is a known (deterministic) function and v is a martingale, we have a representation of type (5.1). Also, with $w = v$, condition (i) of Corollary 5.2 holds and consequently u^* is given by (5.6). Indeed, in § 6, we shall use this representation for the case with complete state information ($H = I$), but in general the SOL class $\mathscr{U}_v$ is too large. However, if v is a Wiener-process, ξ is Gaussian, $\{z(t)\,; t \leqq 0\}$ is deterministic, and $HC(t)$ is a square matrix with a bounded inverse, we have an innovation process

$$\begin{aligned} dw(t) &= (HC(t))^{-1}[dz_0(t) - H\hat{q}(t)\,dt], \\ w(0) &= 0, \end{aligned} \tag{5.11}$$

where

$$\hat{q}(t) = E\{q(t)|z_0(s)\,; 0 \leqq s \leqq t\}, \tag{5.12}$$

$$q(t) = A_1(t)x_0(t) + A_2(t)x_0(t-h) + \int_{t-h}^{t} A_0(t,s)x_0(s)\,ds. \tag{5.13}$$

The innovation process w is a Wiener-process, w and z_0 are related to each other by invertible and nonanticipative linear transformations, and the families of sigma fields generated by the two processes are identical. (See, e.g., [14], where other references are also given, or [17, Lemma 3.2].) Therefore,

$$\hat{q}(t) = \bar{q}(t) + \int_0^t Q(t,s)\,dw, \tag{5.14}$$

where the L_2 vector function $\bar{q}$ and L_2 matrix kernel Q are given by Lemma A.2, and hence z_0 can be represented in the form (5.1). Moreover, since x_0 and w are jointly Gaussian, condition (ii) of Corollary 5.2 holds, and u^* is given by (5.6). Now, from (3.16) we have

$$(5.15)\quad dz^* = dz_0 + \left[HB_1u^* + HB_2u^*(t-h) + \int_0^t H\Gamma(t,s)u^*(s)\,ds\right]dt.$$

Then, inserting z_0 given by (5.11) and (5.14) and u^* given by (5.6) into (5.15) and changing the order of integration, we obtain an expression of type

$$(5.16)\qquad (HC)^{-1}\,dz^* = dw + \int_0^t P(t,s)\,dw\,dt + p(t)$$

(where p and P are functions of the same type as $\bar{q}$ and Q). The resolvent technique previously used for equation (3.17) can now be applied to solve (5.16) for dw, which inserted into (5.6) yields

$$u^*(t) = \pi^*(t, z^*),$$

where $\pi^* \in \mathscr{L}$. Therefore, since $\mathscr{L}$ is contained in Π (Lemma 3.1), Lemma 5.1 implies that π^* is the optimal control law of our FB problem. We have not bothered to determine π^* explicitly but have only described how this can be done. The reason for this is that π^* is often more easily implemented by a linear filter in which w is formed as an intermediate process. Such a filter contains linear feedback loops, and we may ask whether the existence and uniqueness for the complete feedback system is preserved. However, this is the case, for mathematically these loops correspond to linear Volterra integral equations of either ordinary type or type (5.16), and therefore they can be resolved by reformulation using the resolvent equations. (Note that in this paper we allow no stochastic processes whose sample functions are not a.s. square integrable.)

6. Examples. In order to illustrate our basic technique, we shall apply the results of this paper to some simple and well-known problems, all of which will concern systems of type (3.8). However, we shall not consider delay in the state process since this would only introduce complications in notation without exposing any new ideas which cannot be found in [17].

Example 1. *Complete state information.* Consider the system:

$$(6.1)\qquad dx(t) = [A(t)x(t) + B_1(t)u(t) + B_2(t)u(t-h)]\,dt + C(t)\,dw,$$

where $x(0) = a$ is a deterministic vector, w is a martingale of type (4.1), and the matrix functions are defined as in § 3. The observation $z(t)$ is the state process $x(t)$ itself, and the problem is to determine a control law $\pi:(t,x) \to u(t) = \pi(t,x)$ in the class Π which minimizes

$$(6.2)\qquad E\left\{\int_0^T (x'Q_1x + u'Q_2u)\,dt + x'(T)Q_1(T)x(T)\right\}.$$

Here (6.2) is the cost functional (2.3) with $\alpha(t) = t$ for $t < T$ and $\alpha(t) = t + 1$ for $t \geqq T$.

Now, if Φ is defined as in (3.13) with $A_1 = A$ and $A_0 = A_2 = 0$, we have

$$x_0(t) = \Phi(t, 0)a + \int_0^t \Phi(t, s)C(s)\,dw(s) \tag{6.3}$$

and hence condition (i) of Corollary 5.2 holds. Therefore, (5.6) is the optimal solution of the SOL problem (5.3). The functions $\bar{x}_0$, m_i and K in Lemma 5.2 are given by

$$\bar{x}_0(t) = \Phi(t, 0)a, \tag{6.4}$$

$$m_i(t, s) = \Phi(t, s)c_i(s) \quad \text{for } t \geqq s, \tag{6.5}$$

$$K(t, s) = \Phi(t, s)B_1(s) + \Phi(t, s + h)B_2(s + h), \tag{3.12$'$}$$

where c_i is the ith column of C, and therefore problems (5.7) and (5.8) belong to the family of problems

$$\min \int_s^T (x'Q_1x + u'Q_2u)\,dt + x(T)Q_1(T)x(T) \tag{6.6}$$

when

$$\frac{dx}{dt} = Ax + B_1u(t) + B_2u(t - h) \quad \text{for } t > s, \qquad u(t) = 0 \quad \text{for } t < s,$$

where for (5.7), $s = 0$ and $x(0) = a$, and for (5.8), $x(s) = c_i(s)$. Now, according to Appendix C, we have the following feedback solutions:

$$\bar{u}^*(t) = P_0(t)\bar{x}^*(t) + \int_{t-h}^t P_1(t, \tau)\bar{u}^*(\tau)\,d\tau,$$

$$u_i^*(t, s) = P_0(t)x_i^*(t, s) + \int_{t-h}^t P_1(t, \tau)u_i^*(\tau, s)\,d\tau, \qquad t \geqq s,$$

which inserted into (5.6) yields, after applying the stochastic Fubini theorem,

$$u^*(t) = P_0(t)x^*(t) + \int_{t-h}^t P_1(t, \tau)u^*(\tau)\,d\tau, \tag{6.7}$$

for $x^*(t)$ is given by (5.9) with $\tau = t$ (Corollary 5.2), and $u_i^*(t, s) \equiv 0$ for $t < s$. Since P_1 is an L_2 matrix kernel and almost all sample functions of P_0x^* are square integrable, according to standard Volterra theory (see, e.g., [19]) there is an L_2 resolvent kernel P_2 such that

$$u^*(t) = P_0(t)x^*(t) + \int_0^t P_2(t, s)P_0(s)x^*(s)\,ds \tag{6.8}$$

which defines a nonanticipative function $\pi^* : (t, x^*) \to u^*$.

Now we can use a similar argument to show that $\pi^* \in \Pi$, and therefore, according to Lemma 5.1, (6.8) is an optimal FB solution. However, note that the filter described by (6.7) might prove to be a more suitable implementation of π^*. (See Fig. 6.1.) In fact, (6.7) only requires storing u^* on the interval $(t - h, t)$, while in (6.8) we need x^* on the whole interval $(0, t)$. It should be clear from the dis-

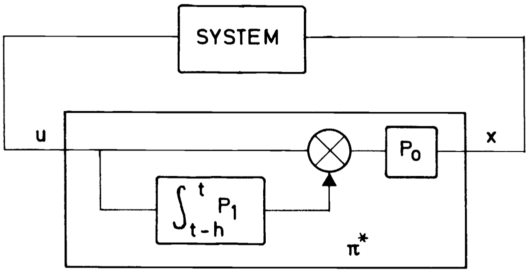


FIG. 3

cussion above that this implementation preserves the existence and uniqueness property of π^*.

Example 2. *Separation theorem*; *white measurement noise*. Consider the stochastic vector processes y and z defined by

$$dy(t) = [A(t)y(t) + B(t)u(t)]\,dt + C_1(t)\,dv_1(t), \tag{6.9}$$

$$dz(t) = D(t)y(t)\,dt + dv_2; \qquad z(0) = 0, \tag{6.10}$$

where $v = \begin{pmatrix} v_1 \\ v_2 \end{pmatrix}$ is a vector Wiener-process of type (3.9), $y(0)$ in Gaussian, $Ey(0) = a$, and v and $y(0)$ are independent. All matrix functions are bounded. The problem is to determine a nonanticipative function $\pi:(t, z) \to u(t) = \pi(t, z)$ so as to minimize

$$E\left\{\int_0^T (y'Q_3y + u'Q_2u)\,dt\right\}, \tag{6.11}$$

where Q_3 is nonnegative definite and bounded.

Now, this is clearly a problem of the type discussed in § 3. In fact, define x to be $\begin{pmatrix} y \\ z \end{pmatrix}$ and H to be $(0, I)$. Moreover, in $V_s(x, u)$, define Q_1 to be $\begin{pmatrix} Q_3 & 0 \\ 0 & 0 \end{pmatrix}$ and put $\alpha(t) = t$. Therefore, Π will be our class of admissible control laws from which a π minimizing (6.11) is to be selected.

For this problem we have an innovation process (5.11) given by

$$dw = dz_0 - D(t)\hat{y}_0(t)\,dt; \qquad w(0) = 0, \tag{6.12}$$

where $\hat{y}_0(t)$ is defined as in (5.12), and, according to Corollary 5.2 (condition (ii)), (5.6) is the optimal solution of our basic SOL problem (5.3) to determine a $u \in \mathscr{U}_w$ so as to minimize (6.11). Now, let $\bar{y}_0, \bar{y}^*, g_i$ and y_i^* be the subvectors in "y-position" of $\bar{x}_0, \bar{x}^*, m_i$ and x_i^* respectively, and let K_1 be the corresponding submatrix of K. Then, if Φ is defined as in Example 1, we have

$$\bar{y}_0(t) = \Phi(t, 0)a, \tag{6.13}$$

$$g_i(t, s) = \Phi(t, s)P(s)\,d_i(s) \quad \text{for } t \geqq s, \tag{6.14}$$

$$K_1(t, s) = \Phi(t, s)B(s), \tag{6.15}$$

where d_i is the ith column of D', and P is the conditional covariance

$$P(t) = E\{[y_0(t) - \hat{y}_0(t)][y_0(t) - \hat{y}_0(t)]'\}. \tag{6.16}$$

In fact, since y_0 and v_2 are independent, we have

$$\begin{aligned} E\{y_0(t)w'(s)\} &= \int_0^s E\{y_0(t)[y_0(\tau) - \hat{y}_0(\tau)]'\}D'(\tau)\,d\tau \\ &= \int_0^s \Phi(t,\tau)P(\tau)D'(\tau)\,d\tau, \end{aligned}$$

for the two last terms in

$$y_0(t) = \Phi(t,\tau)[y_0(\tau) - \hat{y}_0(\tau)] + \Phi(t,\tau)\hat{y}_0(\tau) + \int_\tau^t \Phi(t,\tau)C_1(\tau)\,dv_1$$

are orthogonal to $y_0(\tau) - \hat{y}_0(\tau)$, and therefore (6.14) follows from the definition of $m_i(t, s)$. Then, (5.7) and (5.8) belong to the family of problems:

$$\begin{aligned} &\min \int_s^T (y'Q_3y + u'Q_2u)\,dt \\ &\text{when} \\ &\frac{dy}{dt} = Ay + Bu \quad \text{for } t \geqq s, \end{aligned} \tag{6.17}$$

where $s = 0$ and $y(0) = a$ for (5.7), and $y(s) = P(s)\,d_i(s)$ for (5.8).

Hence, we have the following feedback solutions:

$$\bar{u}^*(t) = L(t)\bar{y}^*(t), \tag{6.18}$$

$$u_i^*(t, s) = L(t)y_i^*(t, s), \tag{6.19}$$

where L can be found in any textbook on the linear-quadratic regulator problem. This inserted into (5.6) yields

$$u^*(t) = L(t)\hat{y}^*(t), \tag{6.20}$$

where $\hat{y}^*(t) = E_t y^*(t)$ is given by (5.9) (see Corollary 5.2), and therefore satisfies the stochastic differential equation

$$\begin{aligned} d\hat{y}^*(t) &= [A(t)\hat{y}^*(t) + B(t)u^*(t)]\,dt + P(t)D'(t)\,dw(t), \\ \hat{y}^*(0) &= a. \end{aligned} \tag{6.21}$$

Now, the innovation process (6.12) can also be written

$$dw = dz^* - D\hat{y}^*\,dt, \tag{6.22}$$

for the control-dependent terms of z^* and $\hat{y}^*$ cancel out. Then, if Ψ is the transition matrix

$$\begin{aligned} \frac{\partial\Psi}{\partial t}(t, s) &= [A(t) + B(t)L(t) - P(t)D'(t)D(t)]\Psi(t, s), \\ \Psi(s, s) &= I, \end{aligned} \tag{6.23}$$

we have a nonanticipative function $\pi^*:(t, z^*) \to u^*$ which is defined by

$$u^*(t) = L(t)\Psi(t,0)a + \int_0^t L(t)\Psi(t,s)P(s)D'(s)\,dz^*(s) \tag{6.24}$$

and hence belongs to Π (Lemma 3.1). Therefore, according to Lemma 5.1, among all control laws in Π, the function π^* gives the smallest value to (6.11). However, it should be clear from the discussion at the end of § 5 that π^* can safely be implemented by the linear filter defined by (6.20), (6.21) and (6.22). Of course this is an advantage, for in this way there is no need to store old z^*.

Example 3. *Separation theorem*; *colored measurement noise*. We shall consider the preceding example (Example 2) modified in the following way: The observation process z is no longer defined by (6.10), but

$$z(t) = H_1 y(t) + n(t), \tag{6.25}$$

where H_1 is a constant matrix and n is a colored noise term generated by

$$dn(t) = D(t)n(t)\,dt + dv_2 \tag{6.26}$$

with $n(0) = 0$. Also $y(0) = a$ is assumed to be deterministic.

Again we have a problem of the type discussed in § 3, for define x to be $\begin{pmatrix} y \\ n \end{pmatrix}$, put $H = (H_1, I)$, and let Q_1 and α be defined as in the previous example. Therefore, the problem is to determine $\pi \in \Pi$ so as to minimize (6.11) when $u(t) = \pi(t, z)$.

Since $HC = I + H_1C_1$, we shall further assume that $(I + H_1C_1)^{-1}$ exists and is bounded. Then, we have an innovation process (5.11) given by

$$dw = (I + H_1C_1)^{-1}[dz_0 - H_1A\hat{y}_0\,dt - D\hat{n}\,dt], \tag{6.27}$$

where $w(0) = 0$ and $\hat{y}_0(t)$ and $\hat{n}(t)$ are defined as in (5.12).

Now, apart from the definition of the innovation process, we have the same problem as in the preceding example, and the optimal solution u^* of our basic SOL problem is given by (6.20) and (6.21). The innovation process can now be expressed in terms of z^* and $\hat{y}^*$:

$$dw = (I + H_1C_1)^{-1}[dz^* - Dz^*\,dt - (H_1A + H_1BL - DH_1)\hat{y}^*\,dt], \tag{6.28}$$

for due to (6.20) and $\hat{n}(t) = z^*(t) - H_1\hat{y}^*(t)$ which is an immediate consequence of (6.25), we have only added terms which cancel out. Then, because of (6.20), (6.21) and (6.28), we have

$$\begin{aligned} u^*(t) = {} & L(t)\Psi(t,0)a + \int_0^t \Gamma(t,s)D(s)\,ds H_1 a \\ & + \int_0^t \left[\Gamma(t,s) + \int_s^t \Gamma(t,\tau)D(\tau)\,d\tau\right] dz^*(s), \end{aligned} \tag{6.29}$$

where Γ and the transition matrix Ψ are defined by

$$\frac{\partial\Psi}{\partial t}(t,s) = [A + BL - PD'(I + H_1C_1)^{-1}(H_1A + H_1BL - DH_1)]\Psi(t,s),$$

$$\Psi(s,s) = I,$$

$$\Gamma(t,s) = L(t)\Psi(t,s)P(s)D'(s)(I + H_1C(s))^{-1}.$$

To obtain (6.29) we have used the fact that

$$z^*(t) = H_1 a + \int_0^t dz^*$$

and applied the stochastic Fubini theorem. Now, (6.29) clearly defines a function $\pi^*:(t, z^*) \to u^*$ which belongs to $\mathscr{L} \subset \Pi$. Therefore, according to Lemma 5.1, π^* is an optimal control law in the class Π. However, as usual we will find it more convenient to implement π^* by the linear filter defined by (6.20), (6.21) and (6.28).

Appendix A.

LEMMA A.1. *The SOL problem posed in the beginning of § 3 has an optimal solution which is unique up to a (t, ω)-equivalence.*

Proof. We introduce the following notation:

$$J(u) = EV_0(x_0 + Ku, u).$$

Now, there is a sequence $u_n \in \mathscr{U}$, $n = 1, 2, 3, \cdots$, such that

$$\lim_{n\to\infty} J(u_n) = \inf_{u\in\mathscr{U}} J(u) = \rho \geqq 0.$$

Then, for each $\varepsilon > 0$ the parallelogram identity yields

$$EV_0(K(u_m - u_n), u_m - u_n) = 2J(u_m) + 2J(u_n) - 4J\left(\frac{u_n + u_m}{2}\right)$$

$$< 2\left(\rho + \frac{\varepsilon}{4}\right) + 2\left(\rho + \frac{\varepsilon}{4}\right) - 4\rho = \varepsilon$$

for sufficiently large m and n. Therefore $\{u_n\}$ is a Cauchy sequence in $L_2([0, T] \times \Omega, \mathscr{B} \times \mathfrak{S}, \lambda \times P)$ (with norm $\|\cdot\| = (\int E|\cdot|^2\,dt)^{1/2}$; λ is the Lebesgue measure) defining a limit point u^* which, due to completeness, clearly satisfies conditions (i) and (ii) in the definition of $\mathscr{U}$. Moreover, since

$$\lim_{n\to\infty} \|u_n - u^*\| = 0, \qquad \lim_{n\to\infty} E|u_n(t) - u^*(t)|^2 = 0$$

for almost all t, and hence u^* satisfies condition (iii), too. Therefore, $u^* \in \mathscr{U}$. It remains to show that u^* is optimal. However this is the case, for it is not hard to see that $|J(u_n) - J(u^*)| \leqq \gamma\|u_n - u^*\|$, where γ is a constant, and hence $J(u^*) = \rho$. Moreover, if $u^0 \in \mathscr{U}$ and $J(u^0) = \rho$, the parallelogram identity implies that $\|u^* - u^0\| = 0$, for

$$J\left(\frac{u^* + u^0}{2}\right) \geqq \rho.$$

Therefore, $u^0 = u^*$ $\lambda \times P$-a.e., and hence the asserted uniqueness property is true.

Appendix B. Let $y(t)$ be a stochastic vector process with finite second order moments and mean $E\{y(t)\} = \bar{y}(t)$, and let $w(t)$ be a vector process with zero mean and orthogonal increments described by (4.1). The inverse functions $t \to \beta_i^{-1}(t)$ are uniquely defined except for at most enumerably many t. Then define

$$n_i(t, s) = \left[\frac{\partial}{\partial\sigma} E\{y(t) w_i(\beta_i^{-1}(\sigma))\}\right]_{\sigma = \beta_i(s)} \tag{B.1}$$

for which we shall use the shorthand notation

$$n_i(t, s) = \frac{\partial}{\partial\beta_i}E\{y(t)w_i(s)\}. \tag{B.2}$$

Now, we can formulate a lemma which slightly generalizes results which may be found in [13], [9] and [17], for example.

LEMMA B.1. *The wide sense conditional expectation of $y(t)$ with respect to $\{w(s); 0 \leqq s \leqq \tau\}$ is given by*

$$\hat{E}_\tau y(t) = \bar{y}(t) + \sum_i \int_0^\tau n_i(t, s)\, dw_i(s). \tag{B.3}$$

Proof (cf. [17, Lemma 2.1]). Since, by definition, $\hat{E}_\tau y(t)$ must have a representation of type (B.3), it only remains to show that n_i is given by (B.1). For $s \leqq \tau$, $w_i(s)$ and $y(t) - \hat{E}_\tau y(t)$ are orthogonal, and therefore

$$E\{y(t)w_i(s)\} = \int_0^s n_i(t, \tau)\, d\beta_i(\tau),$$

or, with $\sigma = \beta_i(s)$,

$$E\{y(t)w(\beta_i^{-1}(\sigma))\} = \int_{\beta_i(0)}^{\sigma} n_i(t, \beta_i^{-1}(\theta))\, d\theta,$$

which yields (B.1).

Appendix C. Problem (6.6) has an optimal feedback solution

$$u^*(t) = P_0(t)x^*(t) + \int_{t-h}^{t} P_1(t, s)u^*(s)\, ds, \tag{C.1}$$

where

$$P_1(t, s) = \Lambda(t, s + h)B_2(s + h) - \int_t^T R(t, \sigma, t)\Lambda(\sigma, s + h)\, d\sigma B_2(s + h),$$

$$P_0(t) = \Lambda(t, t) - \int_t^T R(t, \tau, t)\Lambda(\tau, t)\, d\tau.$$

Here,

$$\Lambda(t, s) = -Q_2^{-1}(t)\left[\int_t^T K'(\tau, t)Q_1(\tau)\Phi(\tau, s)\, d\tau + K'(T, t)Q_1(T)\Phi(T, s)\right],$$

where K is defined by (3.12) and R is the resolvent kernel given by

$$\begin{aligned} R(t, \tau, s) - P(t, \tau) &= -\int_s^T P(t, \sigma)R(\sigma, \tau, s)\, d\sigma \\ &= -\int_s^T R(t, \sigma, s)P(\sigma, \tau)\, d\sigma \end{aligned}$$

with

$$P(t, s) = \Lambda(t, s)B_1(s) + \Lambda(t, s + h)B_2(s + h).$$

In fact, by the method used in [17, § 4], we have

$$u^*(t) + \int_s^T P(t, \tau)u^*(\tau)\,d\tau = \Lambda(t, s)x^*(s) + \int_{s-h}^s \Lambda(t, \tau + h)B_2(\tau + h)u^*(\tau)\,d\tau$$

from which (C.1) follows by the same argument as in [17]. (Also see [1] and [15] where versions of this problem are discussed in detail.)

REFERENCES

[1] Y. ALEKAL, P. BRUNOVSKÝ, D. H. CHYUNG AND E. B. LEE, *The quadratic problem for systems with time delays*, IEEE Trans. Automatic Control, AC-16 (1971), pp. 473–687.

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

[3] A. BENSOUSSAN, *On the separation principle for distributed parameter systems*, IFAC Symp. Control of Distributed Parameter Systems, Banff, Canada, 1971.

[4] H. CRAMÉR, *On the structure of purely non-deterministic stochastic processes*, Ark. Mat., 4 (1961), pp. 249–266.

[5] M. H. A. DAVIS AND P. P. VARAIYA, *Information states in linear stochastic systems*, J. Math. Anal. Appl., 37 (1972), pp. 384–402.

[6] ———, *Dynamic programming conditions for partially observable stochastic systems*, Memo. ERL-M304, Electronics Research Laboratory, University of California, Berkeley, 1971.

[7] J. L. DOOB, *Stochastic Processes*, John Wiley, New York, 1953.

[8] W. H. FLEMING, *Optimal continuous-parameter stochastic control*, SIAM Rev., 11 (1969), pp. 470–509.

[9] P. FROST, *Nonlinear estimation in continuous time systems*, Rep. 6304-4, Center for Systems Research, Stanford University, Stanford, Calif., 1968.

[10] I. I. GIKHMAN AND A. V. SKOROKHOD, *Introduction to the Theory of Random Processes*, W. B. Saunders, Philadelphia, 1969.

[11] 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.

[12] T. HIDA, *Canonical representations of Gaussian processes and their applications*, Mem. Coll. Sci. Univ. Kyoto Ser. A. Math., 33 (1960/61), pp. 109–155.

[13] T. KAILATH, *An innovations approach to least squares estimation, Part I: Linear filtering in additive white noise*, IEEE Trans. Automatic Control, AC-13 (1968), pp. 646-655.

[14] ———, *The innovations approach to detection and estimation theory*, Proc. IEEE, 58 (1970), pp. 680–695.

[15] H. N. KOIVO AND E. B. LEE, *Controller syntheses for linear systems with retarded state and control variables and quadratic cost*, Automatica, 8 (1972), pp. 203–208.

[16] H. J. KUSHNER, *Stochastic Stability and Control*, Academic Press, New York, 1967.

[17] A. LINDQUIST, *Optimal control of linear stochastic systems with applications to time lag systems*, Information Sci., 4 (1972).

[18] ———, *A theorem on duality between estimation and control for linear stochastic systems with time delay*, J. Math. Anal. Appl., 37 (1972), pp. 516–536.

[19] F. SMITHIES, *Integral Equations*, Cambridge University Press, New York, 1958.

[20] H. S. WITSENHAUSEN, *Separation of estimation and control for discrete time systems*, Proc. IEEE, 59 (1971), pp. 1557–1566.

[21] W. M. WONHAM, *On the separation theorem of stochastic control*, this Journal, 6 (1968), pp. 312–326.

[22] ———, *Random differential equations in control theory*, Probabilistic Methods in Applied Mathematics, vol. 2, A. T. Bharucha-Reid, ed., Academic Press, New York, 1970, pp. 131–212.

[23] L. E. ZACHRISSON, *A proof of the separation theorem in control theory*, IOS Rep. R 23, Royal Institute of Technology, Stockholm, Sweden, 1968.

[24] ———, *On a separation theorem for almost separated systems*, to appear.

SIAM J. Control
Vol. 11, No. 2, May 1973

SOLUTION OF A DUAL PROBLEM IN OPTIMAL CONTROL THEORY[†]

EARL R. BARNES[‡]

Abstract. We consider a fixed endpoint problem in optimal control theory. The problem is given a dual formulation and an iterative procedure for solving the dual problem is described. The results of two numerical experiments are discussed to illustrate how the method works in particular instances.

1. Introduction. Consider a control system governed by the linear differential equation

$$\dot{x} = A(t)x + B(t)u(t), \qquad x(0) = x_0. \tag{1.1}$$

$x(t) \in E_n$ is an n-vector, $u(t) \in E_m$ is an m-vector, and $A(t)$ and $B(t)$ are continuous matrices of dimensions $n \times n$ and $n \times m$ respectively.

Let $f(t, x)$ and $h(t, u)$ be continuous functions, defined for all $x \in E_n$, $u \in E_m$, and for all t in an interval $[0, T]$, $T > 0$. We assume further that f and h are convex and twice continuously differentiable in the variables x and u respectively. Consider the problem of minimizing the cost functional.

$$c(u) = \int_0^T [f(t, x(t)) + h(t, u(t))]\, dt, \tag{1.2}$$

subject to the differential equation constraint (1.1), together with a terminal constraint $x(T) = x_1$ imposed on x, and subject to the requirement that the functions u allowed in the problem be measurable and assume values in a given compact convex constraint set $\Omega \subset E_m$. This class of u's will be referred to as the class of admissible controls, and will be denoted by $\mathscr{C}$. This problem is called a fixed endpoint optimal control problem. We shall refer to an admissible control which solves it as an optimal control. The existence of an optimal control is assumed throughout our discussion.

According to Pontryagin's maximum principle, in order that the control u^* be optimal for the system (1.1)–(1.2), it is necessary that there exist a function ψ and a nonpositive constant η_0 satisfying

$$\dot{\psi} = -A^{\mathrm{T}}(t)\psi - \eta_0 \frac{\partial f}{\partial x}(t, x^*(t)), \qquad 0 \leqq t \leqq T, \tag{1.3}$$

such that

$$\max_{u \in \Omega} \{B^{\mathrm{T}}(t)\psi(t) \cdot u + \eta_0 h(t, u)\} = B^{\mathrm{T}}(t)\psi(t) \cdot u^*(t) + \eta_0 h(t, u^*(t)), \tag{1.4}$$

for almost all $t \in [0, T]$. The symbol $^{\mathrm{T}}$ denotes transpose.

[†] Received by the editors November 9, 1971, and in revised form July 2, 1972. Contributed at the SIAM National Meeting held at the University of Washington, Seattle, June 28–30, 1971.

[‡] Visiting member of Mathematics Research Center, University of Wisconsin, Madison, Wisconsin 53706, from IBM Thomas J. Watson Research Center, Yorktown Heights, New York 10598. This work was supported by the United States Army under Contract DA-31-124-ARO-D-462.

In what follows we shall make a normality assumption on our problem. This is the assumption that the constant η_0 appearing in the maximum principle is <0. Then without loss of generality, we shall assume $\eta_0 = -1$.

With this assumption in force, let ψ be as in the maximum principle and let $\eta = \psi(T)$. Consider the minimization problem

$$\min \eta \cdot (x_1 - x(T)) + \int_0^T [f(t, x(t)) + h(t, u(t))]\, dt,$$

where the minimization is taken subject to (1.1) and $u \in \mathscr{C}$. If we apply the maximum principle to this problem and use the transversality condition, we see that its solution is u^*, the optimal control for the system (1.1)–(1.2). This discussion shows that we can solve the optimal control problem for the system (1.1)–(1.2) as a free endpoint problem, if we can determine the proper multiplier η. An algorithm for solving free endpoint problems has been described by the author in [1]. We therefore view the hard part of our control problem as that of determining the multiplier η. It is known (cf. [2], [3]) that η is characterized by the property that it maximizes the continuously differentiable function g, defined for each $\lambda \in E_n$ by

$$g(\lambda) = \min_{u \in \mathscr{C}} \lambda \cdot (x_1 - x(T)) + \int_0^T [f(t, x(t)) + h(t, u(t))]\, dt. \tag{1.5}$$

Let $u(t, \lambda)$ denote the admissible control which accomplishes the minimization in (1.5), and let $x(t, \lambda)$ denote the corresponding solution of (1.1). Then

$$\nabla g(\lambda) = x_1 - x(T, \lambda). \tag{1.6}$$

This result is proved in [2, Theorem 2.1]. It is also shown there that the function g is concave. Therefore, assuming that the function $u(t, \lambda)$ (and hence the gradient $\nabla g(\lambda)$) can be determined for any given λ, it is conceivable that some of the gradient methods from nonlinear programming theory can be used to maximize g, thus giving the desired multiplier η. The optimal control is of course given by $u^*(t) = u(t, \eta)$.

The problem

$$\max_{\lambda \in E_n} g(\lambda) \tag{1.7}$$

is said to be dual to the optimal control problem formulated above for the system (1.1)–(1.2).

The first attempt to solve optimal control problems by solving their duals was made by Neustadt in [4] for a time optimal problem. This work was later extended to fixed time problems similar to the one we consider here by Neustadt and Paiwonsky in [3] and by Meditch and Neustadt in [5]. Other papers dealing with computing optimal controls by dual methods include [6], [7], [8], and [9]. We shall remark on the advantages of the dual method when we come to the numerical results in § 3.

Our purpose in this paper is to describe an iterative procedure for solving the problem (1.7) for the case where h is a strictly convex quadratic form in u and f is a strongly convex function of x. Without loss of generality we shall assume that $h(t, u) = u \cdot D(t)u$,[1] where $D(t)$ is a positive definite $m \times m$ matrix for each $t \in [0, T]$.

[1] This assumption is merely a matter of convenience. It will be clear from our discussion that we need only assume that $h(t, u)$ is convex in u and satisfies $\frac{1}{2}\nabla^2 h(t, u) \geqq D(t), t \in [0, T]$.

The requirement that f be strongly convex in x means that there exist positive definite $n \times n$ matrices $C(t)$ and $C_1(t)$ such that $C(t) \leqq \frac{1}{2}\nabla^2 f(t, x) \leqq C_1(t)$ for all $t \in [0, T]$ and for all $x \in E_n$. The inequality $\frac{1}{2}\nabla^2 f(t, x) \geqq C(t)$ means that $\frac{1}{2}\nabla^2 f(t, x) - C(t)$ is positive semidefinite.

A typical step in an iterative procedure for solving a problem of the form (1.7) involves selecting an estimate λ_1 of the solution, and a direction s satisfying $s \cdot \nabla g(\lambda_1) > 0$. An improved estimate λ_2 is then determined by the formula $\lambda_2 = \lambda_1 + \sigma_1 s$, where σ_1 is defined by the requirement that

$$g(\lambda_1 + \sigma_1 s) = \max_{\sigma \geqq 0} g(\lambda_1 + \sigma s). \tag{1.8}$$

In general, computing σ_1 satisfying (1.8) is a very costly procedure. This is especially true when the function g is hard to evaluate for a given value of the argument, as is the case for our dual control problem. The special feature of our algorithm is that it does not require the solution of a problem of the form (1.8) at each step. However, in order to implement the method the matrix $C(t)$ must be known.

2. The iterative procedure. Before presenting our iterative procedure we describe how the function g can be evaluated for a given $\lambda \in E_n$. Let J be the functional defined on $\mathscr{C}$ by

$$J(u) = -\lambda \cdot x(T) + \int_0^T [f(t, x(t)) + u(t) \cdot D(t)u(t)]\, dt.$$

The problem of evaluating $g(\lambda)$ and $\nabla g(\lambda)$ is equivalent to minimizing J on $\mathscr{C}$. Construct a sequence $\{u_k\} \subset \mathscr{C}$ as follows:

Step 2.1(a). Let $u_1 \in \mathscr{C}$ be chosen arbitrarily.

Step 2.1(b). If $u_1, \cdots, u_k$ have been chosen, choose $\tilde{u}_k$ to be the solution of the problem

$$\min_{u \in \mathscr{C}} -\lambda \cdot x(T) + \int_0^T [\nabla f(t, x_k) \cdot x + u \cdot Du]\, dt.$$

Pontryagin's maximum principle determines $\tilde{u}_k$ explicitly.

Step 2.1(c). Let $u_{k+1} = u_k + \alpha_k(\tilde{u}_k - u_k)$ where α_k is chosen so that

$$J(u_k + \alpha_k(\tilde{u}_k - u_k)) = \min_{0 \leqq \alpha \leqq 1} J(u_k + \alpha(\tilde{u}_k - u_k)).$$

Remark. It is not necessary to compute α_k satisfying (2.1c) with a high degree of accuracy. In fact, we can take

$$\alpha_k = \frac{\|\tilde{u}_k - u_k\|_D^2}{\|\tilde{x}_k - x_k\|_{C_1}^2 + \|\tilde{u}_k - u_k\|_D^2},$$

where x_k and $\tilde{x}_k$ are the solutions of (1.1) corresponding to u_k and $\tilde{u}_k$ respectively. The norm $\|\cdot\|_D$ is defined by

$$\|u\|_D^2 = \int_0^T u(t) \cdot D(t)u(t)\, dt,$$

and $\|\cdot\|_{C_1}$ is defined similarly.

THEOREM 2.1. *The sequence $\{u_k\}$ constructed in* (2.1) *converges pointwise to $u(t, \lambda)$. Moreover, the values $J(u_k)$ converge to $J(u(t, \lambda))$ at the rate of a geometric progression.*

This theorem is proved in [1, Theorem 2.1]. It follows that the sequence $\{u_k\}$ converges to $u(t, \lambda)$ at the rate of a geometric progression, with respect to the norm $\|\cdot\|_D$. And from this it follows that the values $x_1 - x_k(T)$ converge to $\nabla g(\lambda)$ at the rate of a geometric progression.

We turn now to the description of our iterative procedure for solving (1.7).

Step 2.2(a). Select two constants θ and β satisfying $0 < \theta < 4$ and $\beta > 1$.

Step 2.2(b). Let $\lambda_1 \in E_n$ be chosen arbitrarily. If $\lambda_1, \cdots, \lambda_k \in E_n$ have been chosen, choose $\lambda_{k+1} = \lambda_k + \theta\sigma_k s_k$ where s_k is any vector in E_n satisfying

$$s_k \cdot \nabla g(\lambda_k) \geqq |\nabla g(\lambda_k)|^2, \qquad |s_k|^2 \leqq \beta|\nabla g(\lambda_k)|^2,$$

and σ_k is defined by

$$\sigma_k = \frac{|\nabla g(\lambda_k)|^2}{\int_0^T (|C^{-1/2}(t)A^{\mathrm{T}}(t)s_k|^2 + |D^{-1/2}(t)B^{\mathrm{T}}(t)s_k|^2)\,dt}.$$

THEOREM 2.2. *Let u^* denote the optimal control for the system* (1.1)–(1.2). *The sequence of admissible controls $\{u(t, \lambda_k)\}$, corresponding to the sequence $\{\lambda_k\}$ generated in* (2.2), *converges to u^* with respect to the norm $\|\cdot\|_D$.*

Before proving this theorem we must prove a lemma.

LEMMA 2.1. *Let γ and λ be any two vectors in E_n and let $\Delta x(t) = x(t, \gamma) - x(t, \lambda)$, and $\Delta u(t) = u(t, \gamma) - u(t, \lambda)$. Then*

$$\begin{aligned} &\int_0^T (|C^{1/2}\Delta x|^2 + |D^{1/2}\Delta u|^2)\,dt \\ &\leqq \int_0^T \left(\left|\frac{1}{2}C^{-1/2}A^{\mathrm{T}}(\gamma - \lambda)\right|^2 + \left|\frac{1}{2}D^{-1/2}B^{\mathrm{T}}(\gamma - \lambda)\right|^2\right) dt. \end{aligned} \tag{2.3}$$

Proof. Since the function $u(t, \lambda)$ solves a convex free endpoint problem we can characterize it by Pontryagin's maximum principle. According to this principle there exists a function $\psi(t, \lambda)$ satisfying

$$\dot{\psi} = -A^{\mathrm{T}}\psi + \frac{\partial f}{\partial x}(t, x(t, \lambda)), \qquad 0 \leqq t \leqq T, \quad \psi(T, \lambda) = \lambda, \tag{2.4}$$

such that

$$\begin{aligned} &\max_{u\in\Omega} \{B^{\mathrm{T}}(t)\psi(t, \lambda)\cdot u - u\cdot D(t)u\} \\ &\qquad = B^{\mathrm{T}}(t)\psi(t, \lambda)\cdot u(t, \lambda) - u(t, \lambda)\cdot D(t)u(t, \lambda), \end{aligned} \tag{2.5}$$

for almost all $t \in [0, T]$. The same statement holds if we replace λ by γ.

Let $\Delta\psi(t) = \psi(t,\gamma) - \psi(t,\lambda), \Delta x(t) = x(t,\gamma) - x(t,\lambda)$ and $\Delta u(t) = u(t,\gamma) - u(t,\lambda)$. Then

$$\begin{aligned} \Delta\dot{\psi}\cdot\Delta x &= -A^{\mathrm{T}}\Delta\psi\cdot\Delta x + \Delta x\cdot\left(\frac{\partial f}{\partial x}(t, x(t,\gamma)) - \frac{\partial f}{\partial x}(t, x(t,\lambda))\right) \\ &\geqq -A^{\mathrm{T}}\Delta\psi\cdot\Delta x + 2\Delta x\cdot C(t)\Delta x \\ &= -\Delta\psi\cdot(\Delta\dot{x} - B\Delta u) + 2\Delta x\cdot C(t)\Delta x. \end{aligned}$$

We rewrite this as

$$\frac{d}{dt}(\Delta\psi \cdot \Delta x) \geqq B^{\mathrm{T}}\Delta\psi \cdot \Delta u + 2\Delta x \cdot C\Delta x$$

and integrate to obtain

$$(\gamma - \lambda) \cdot (x(T, \gamma) - x(T, \lambda)) \geqq \int_0^T (B^{\mathrm{T}}\Delta\psi \cdot \Delta u + 2\Delta x \cdot C\Delta x)\, dt. \tag{2.6}$$

According to (2.5) the expression

$$\begin{aligned} &u(t) \cdot Du(t) - B^{\mathrm{T}}\psi(t, \lambda) \cdot u(t) - \{u(t, \lambda) \cdot Du(t, \lambda) - B^{\mathrm{T}}\psi(t, \lambda) \cdot u(t, \lambda)\} \\ &= (2Du(t, \lambda) - B^{\mathrm{T}}\psi(t, \lambda)) \cdot (u(t) - u(t, \lambda)) + (u(t) - u(t, \lambda)) \cdot D(u(t) - u(t, \lambda)) \end{aligned} \tag{2.7}$$

is $\geqq 0$ for almost all $t \in [0, T]$, for any $u \in \mathscr{C}$. In particular if we take $u(t) = u(t, \lambda) + \alpha(u(t, \gamma) - u(t, \lambda))$ where $0 < \alpha < 1$, we obtain, by taking α sufficiently small, that

$$(2Du(t, \lambda) - B^{\mathrm{T}}\psi(t, \lambda)) \cdot (u(t, \gamma) - u(t, \lambda)) \geqq 0.$$

This inequality clearly remains valid if we interchange the roles of λ and γ. This result gives us that

$$\begin{aligned} B^{\mathrm{T}}\Delta\psi \cdot \Delta u &= u(t, \lambda) \cdot Du(t, \lambda) - B^{\mathrm{T}}\psi(t, \gamma) \cdot u(t, \lambda) \\ &\quad - (u(t, \gamma) \cdot Du(t, \gamma) - B^{\mathrm{T}}\psi(t, \gamma) \cdot u(t, \gamma)) \\ &\quad + u(t, \gamma) \cdot Du(t, \gamma) - B^{\mathrm{T}}\psi(t, \lambda) \cdot u(t, \gamma) \\ &\quad - (u(t, \lambda) \cdot Du(t, \lambda) - B^{\mathrm{T}}\psi(t, \lambda) \cdot u(t, \lambda)) \\ &= (2Du(t, \gamma) - B^{\mathrm{T}}\psi(t, \gamma)) \cdot (u(t, \lambda) - u(t, \gamma)) \\ &\quad + (2Du(t, \lambda) - B^{\mathrm{T}}\psi(t, \lambda)) \cdot (u(t, \gamma) - u(t, \lambda)) \\ &\quad + 2\Delta u \cdot D\Delta u \\ &\geqq 2\Delta u \cdot D\Delta u. \end{aligned}$$

Going back to (2.6) we now have

$$(\gamma - \lambda) \cdot (x(T, \gamma) - x(T, \lambda)) \geqq 2\int_0^T (\Delta x \cdot C\Delta x + \Delta u \cdot D\Delta u)\, dt. \tag{2.8}$$

Furthermore,

$$\begin{aligned} (\gamma - \lambda) \cdot (x(T, \gamma) - x(T, \lambda)) &= (\gamma - \lambda) \cdot \Delta x(T) \\ &= \int_0^T [(\gamma - \lambda) \cdot AC^{-1/2}C^{1/2}\Delta x + (\gamma - \lambda) \cdot BD^{-1/2}D^{1/2}\Delta u]\, dt \\ &= \int_0^T [C^{-1/2}A^{\mathrm{T}}(\gamma - \lambda) \cdot C^{1/2}\Delta x + D^{-1/2}B^{\mathrm{T}}(\gamma - \lambda) \cdot D^{1/2}\Delta u]\, dt \\ &\leqq \left(\int_0^T [|C^{-1/2}A^{\mathrm{T}}(\gamma - \lambda)|^2 + |D^{-1/2}B^{\mathrm{T}}(\gamma - \lambda)|^2]\, dt\right)^{1/2} \\ &\quad \cdot \left(\int_0^T [|C^{1/2}\Delta x|^2 + |D^{1/2}\Delta u|^2]\, dt\right)^{1/2}. \end{aligned} \tag{2.9}$$

This inequality combines with (2.8) to give the result (2.3).

Proof of Theorem 2.2. For $k \geqq 1$ we have

$$\begin{aligned} g(\lambda_{k+1}) - g(\lambda_k) &= \int_0^1 \frac{d}{d\tau} g(\lambda_k + \tau\theta\sigma_k s_k)\, d\tau \\ &= \int_0^1 [\nabla g(\lambda_k + \tau\theta s_k) - \nabla g(\lambda_k)] \cdot \theta\sigma_k s_k\, d\tau \\ &\qquad + \nabla g(\lambda_k) \cdot \theta\sigma_k s_k \\ &= -\int_0^1 [x(T, \lambda_k + \tau\theta\sigma_k s_k) - x(T, \lambda_k)] \cdot \theta\sigma_k s_k\, d\tau \\ &\qquad + \theta\sigma_k s_k \cdot \nabla g(\lambda_k). \end{aligned} \tag{2.10}$$

Let $\Delta x(t, \tau) = x(t, \lambda_k + \tau\theta\sigma_k s_k) - x(t, \lambda_k)$ and $\Delta u(t, \tau) = u(t, \lambda_k + \tau\theta\sigma_k s_k) - u(t, \lambda_k)$. Then, as in (2.9), we have

$$\begin{aligned} \Delta x(T, \tau) \cdot s_k &\leqq \left(\int_0^T [|C^{-1/2}A^{\mathrm{T}} s_k|^2 + |D^{-1/2}B^{\mathrm{T}} s_k|^2]\, dt\right)^{1/2} \\ &\quad \cdot \left(\int_0^T [|C^{1/2}\Delta x|^2 + |D^{1/2}\Delta u|^2]\, dt\right)^{1/2}. \end{aligned}$$

And by Lemma 2.1, it follows that

$$\begin{aligned} \Delta x(T, \tau) \cdot s_k &\leqq \frac{\theta\sigma_k\tau}{2} \int_0^T (|C^{-1/2}A^{\mathrm{T}} s_k|^2 + |D^{-1/2}B^{\mathrm{T}} s_k|^2)\, dt \\ &= \frac{\theta\tau}{2} |\nabla g(\lambda_k)|^2. \end{aligned}$$

Putting this into (2.10) we have

$$\begin{aligned} g(\lambda_{k+1}) - g(\lambda_k) &= -\theta\sigma_k \int_0^1 \Delta x(T, \tau) \cdot s_k\, d\tau + \theta\sigma_k s_k \cdot \nabla g(\lambda_k) \\ &\geqq \int_0^1 -\frac{\theta^2\sigma_k\tau}{2} |\nabla g(\lambda_k)|^2\, d\tau + \theta\sigma_k |\nabla g(\lambda_k)|^2 \\ &= \theta\sigma_k\left(1 - \frac{\theta}{4}\right) |\nabla g(\lambda_k)|^2. \end{aligned} \tag{2.11}$$

Let $g(\eta) = \max g(\lambda)$, $\lambda \in E_n$. We then have

$$\begin{aligned} g(\eta) - g(\lambda_1) &\geqq \sum_{k=1}^{\infty} [g(\lambda_{k+1}) - g(\lambda_k)] \\ &\geqq \sum_{k=1}^{\infty} \theta\sigma_k\left(1 - \frac{\theta}{4}\right) |\nabla g(\lambda_k)|^2. \end{aligned} \tag{2.12}$$

From (2.2b) it follows that

$$\sigma_k \geqq \frac{1}{\beta \int_0^T (|C^{-1/2}A^{\mathrm{T}}|^2 + |D^{-1/2}B^{\mathrm{T}}|^2)\, dt} > 0.$$

Equation (2.12) now shows that

$$(2.13)\qquad \lim_{k\to\infty} \nabla g(\lambda_k) = \lim_{k\to\infty} (x_1 - x(T, \lambda_k)) = 0.$$

Because of our normality assumption, the sequence $\{\lambda_k\}$ is bounded. We shall not prove this result here. The proof of an analogous result will be given in [10] along with a geometric interpretation of the maximum principle and the dual problem. The discussion in [10] adapts readily to the present situation. It now follows from (2.13) and (2.8), with $\gamma = \eta$ and $\lambda = \lambda_k$, that

$$\int_0^T (u(t, \lambda_k) - u^*(t)) \cdot D(t)(u(t, \lambda_k) - u^*(t))\, dt \to 0,$$

as $k \to \infty$. This completes the proof of Theorem 2.2.

Remark. If for some k the s_k in (2.2b) satisfies $C^{-1/2}(t)A^{\mathrm{T}}(t)s_k = 0$ and $D^{-1/2}(t)B^{\mathrm{T}}(t)s_k = 0$ for all $t \in [0, T]$, then the scalar σ_k is undefined. But if this happens it is clear from (2.9) that $\Delta x(T, \tau) \cdot s_k = 0$ for each $\tau \in [0, 1]$. In this case (2.11) implies that

$$g(\lambda_k + \theta\sigma_k s_k) - g(\lambda_k) \geqq \theta\sigma_k |\nabla g(\lambda_k)|^2$$

for any pair of members $\theta, \sigma_k > 0$. It therefore follows that either $\nabla g(\lambda_k) = 0$ (in which case $u(t, \lambda_k)$ is an optimal control) or g has no finite maximum value, which is impossible if the original control problem has a solution, as we assume. This shows that the iterative scheme (2.3) is well-defined if our problem has a solution.

Note. If the $n \times n$ matrix

$$H = \frac{1}{2}\int_0^T [A(t)C^{-1}(t)A^{\mathrm{T}}(t) + B(t)D^{-1}(t)B^{\mathrm{T}}(t)]\, dt$$

is nonsingular, we can obtain an algorithm for solving (1.7) which resembles Newton's method. This is obtained from (2.2) by replacing (2.2b) by

$$\lambda_{k+1} = \lambda_k + H^{-1}\nabla g(\lambda_k).$$

Using this definition of λ_{k+1} and proceeding as in (2.10)–(2.11) we obtain

$$\begin{aligned} g(\lambda_{k+1}) - g(\lambda_k) &\geqq \int_0^1 [\nabla g(\lambda_k + \tau H^{-1}\nabla g(\lambda_k)) - \nabla g(\lambda_k)] \cdot H^{-1}\nabla g(\lambda_k)\, d\tau \\ &\qquad + \nabla g(\lambda_k) \cdot H^{-1}\nabla g(\lambda_k) \\ &\geqq -\int_0^1 \tau\nabla g(\lambda_k) \cdot H^{-1}\nabla g(\lambda_k)\, d\tau + \nabla g(\lambda_k) \cdot H^{-1}\nabla g(\lambda_k) \\ &= \tfrac{1}{2}\nabla g(\lambda_k) \cdot H^{-1}\nabla g(\lambda_k). \end{aligned}$$

From this it follows that $\lim_{k\to\infty} \nabla g(\lambda_k) = 0$ and the convergence of $u(t, \lambda_k)$ to $u^*(t)$ with respect to the norm $\|\cdot\|_D$ follows as in the proof of Theorem 2.2.

In the remainder of this paper we discuss the implementation of our algorithm on a computer. There are two points to be examined. In the first place, it is clear that the algorithm (2.1) for computing the gradients must be terminated after a finite number of steps. We must give a criterion for terminating these iterations. In the second place, we must show that the continuous problem (1.1)–(1.2) can be approximated as closely as we like by a discrete problem since a computer can only solve discrete problems.

To resolve the first point, let δ_j denote the approximation to $\nabla g(\lambda_k)$ obtained on the jth iterate of (2.1). It has been shown in [1] that

$$|\delta_{j+1} - \nabla g(\lambda_k)|^2 \leqq \alpha|\delta_j - \nabla g(\lambda_k)|^2,$$

where α is a computable constant between 0 and 1. It follows that

$$\begin{aligned}|\delta_{j+1} - \delta_j| &\geqq |\delta_j - \nabla g(\lambda_k)| - |\nabla g(\lambda_k) - \delta_{j+1}| \\ &\geqq (1 - \sqrt{\alpha})|\delta_j - \nabla g(\lambda_k)|.\end{aligned} \tag{2.14}$$

Thus by comparing two successive iterates of the sequence $\{\delta_j\}$, we can determine how accurately we have approximated $\nabla g(\lambda_k)$.

THEOREM 2.3. *For each $k = 1, 2, \cdots$, let $j = j(k)$ correspond to the iterate $\delta_{j(k)}$ of* (2.1) *used to approximate $\nabla g(\lambda_k)$. If*

$$\sum_{k=1}^{\infty} |\delta_{j(k)} - \nabla g(\lambda_k)| < \infty,$$

then the algorithm (2.2) *converges when $\delta_{j(k)}$ is used in place of $\nabla g(\lambda_k)$. Equation* (2.14) *provides a method for determining when the condition of the theorem is satisfied.*

Proof. If $\delta_{j(k)}$ is used in place of $\nabla g(\lambda_k)$ in (2.2), s_k will be computed from the requirements

$$s_k \cdot \delta_{j(k)} \geqq |\delta_{j(k)}|^2, \qquad |s_k|^2 \leqq \beta|\delta_{j(k)}|^2. \tag{2.15}$$

The other formulas in (2.2) remain as they are with $\delta_{j(k)}$ replacing $\nabla g(\lambda_k)$. Equation (2.10) and the first equation in (2.11) remain valid. Let $e_k = \delta_{j(k)} - \nabla g(\lambda_k)$. Then from (2.11) we have

$$\begin{aligned}g(\lambda_{k+1}) - g(\lambda_k) &= -\theta\sigma_k \int_0^1 \Delta x(T, \tau) \cdot s_k \, d\tau + \theta\sigma_k s_k \cdot \nabla g(\lambda_k) \\ &\geqq -\frac{\theta^2\sigma_k}{4}|\delta_{j(k)}|^2 + \theta\sigma_k s_k \cdot (\delta_{j(k)} - e_k) \\ &\geqq \theta\sigma_k\left(1 - \frac{\theta}{4}\right)|\delta_{j(k)}|^2 - \theta\sigma_k s_k \cdot e_k.\end{aligned} \tag{2.16}$$

The vectors $\theta\sigma_k s_k = \lambda_{k+1} - \lambda_k$ are clearly uniformly bounded. It therefore follows from the hypothesis of our theorem that the series

$$\sum_{k=1}^{\infty} \theta\sigma_k s_k \cdot e_k$$

converges. By summing the left-hand side of (2.16) we see that $\lim_{k\to\infty} \delta_{j(k)} = 0$. Finally, since

$$\nabla g(\lambda_k) = \delta_{j(k)} - e_k,$$

we have $\lim_{k\to\infty} \nabla g(\lambda_k) = 0$ and it follows as in the proof of Theorem 2.2 that $\lim_{k\to\infty} \|u(\cdot, \lambda_k) - u^*\|_D = 0$. This completes the proof of Theorem 2.3.

Remark. Theorem 2.3 does not indicate that the gradients $\nabla g(\lambda_k)$ need to be approximated with a high degree of accuracy for small k. However, we have found that for best results, all gradients should be computed at least as accurately as it is desired to have the constraint $x(T) = x_1$ satisfied.

The problem of approximating the continuous problem by a discrete problem is discussed in the next section.

3. Some computational results. We can compute the sequence $\{\lambda_k\}$ generated in (2.2), and the corresponding sequence $\{u(t, \lambda_k)\}$, at least approximately, with the aid of a computer. In this section we give computer results for two examples. For simplicity we take f to be a quadratic function $x \cdot Cx$ in each case. For purposes of the computer, the problem (1.1)–(1.2) must be discretized. We accomplish this in our examples by replacing (1.1) by the difference equation

$$\begin{aligned} &x(1) = x_0, \\ &x(i+1) = x(i) + hA(i)x(i) + hB(i)u(i), \qquad i = 1, \cdots, N-1, \end{aligned} \tag{3.1}$$

and the functional (1.2) by the finite sum

$$h \sum_{i=1}^{N-1} \{x(i) \cdot C(i)x(i) + u(i) \cdot D(i)u(i)\},^{2} \tag{3.2}$$

where $N > 1$ is a positive integer and $h = T/(N - 1)$. The optimization problem is now to minimize (3.2) subject to (3.1), the endpoint constraint $x(N) = x_1$, and $u(i) \in \Omega$ for $i = 1, \cdots, N - 1$.

The results of §§ 1 and 2 adapt readily to the system (3.1)–(3.2), but here a word of caution is in order. In computing $\tilde{u}_k(i)$, $i = 1, \cdots, N - 1$, according to (2.1b), we must use the discrete maximum principle. This principle says that for each i, $\tilde{u}_k(i)$ is determined by the requirement that

$$\max_{u \in \Omega} B^{\mathrm{T}}(i)\psi(i+1) \cdot u - u \cdot Du = B^{\mathrm{T}}(i)\psi(i+1) \cdot \tilde{u}_k(i) - \tilde{u}_k(i) \cdot D\tilde{u}_k(i), \tag{3.3}$$

where

$$\psi(N) = \lambda_k$$

and

$$\psi(j) = \psi(j+1) + hA^{\mathrm{T}}(j)\psi(j+1) - 2hC(j)x_k(j),$$

$j = 1, \cdots, N - 1$. The point is that one must be careful to use the advanced argument $i + 1$ of ψ in (3.3).

Remark. The optimization problem for the system (3.1)–(3.2) is a convex quadratic programming problem with convex constraints on the variables $u(i)$. In the case of linear constraints, there are already quadratic programming algorithms which are particularly tailored for solving such problems (cf. [11], [12]). However, in the case of nonlinear constraints, there is no universally accepted way to proceed. One tries to select a method that is well suited to the particular form of his constraints. But in any case, it has been observed that in general the difficulty in solving a nonlinear programming problem grows with the size of the problem. For a discussion of this see [13]. In some cases, a given convex programming problem can be given an equivalent dual formulation which reduces the

[2] In (3.1) and (3.3), and in the sequel, we use the argument i where, strictly speaking, we mean $(i - 1)h$.

problem to a sequence of smaller problems. This provides the motivation for studying computational methods for dual problems in nonlinear programming and in control theory. This decomposition of a large problem into a sequence of small problems is exactly the benefit we gain here. In place of the problem (3.1)–(3.2) we have substituted a sequence of significantly smaller problems of the form (3.3). We use our method for the solution of two example problems below. The advantages of the decomposition technique are more apparent in the second example where we have nonlinear constraints on the control variables $u(i)$.

In our initial experiments with the algorithm (2.2), the convergence turned out to be discouragingly slow. We took this to mean that it is probably not wise to ignore the problem of optimizing the step size, as in (1.8), at each iteration. We found that by sufficiently increasing the step size $\theta\sigma_k$ predicted in (2.2b) we could significantly reduce the number of iterations required for convergence. This suggested a method of modifying our algorithm to possibly increase efficiency.

This method involves selecting two constants r_1 and r_2 between 0 and 1, and a large value of θ, and iterating according to (2.2). If at some stage $|\nabla g(\lambda_{k+1})| \geqq r_1|\nabla g(\lambda_k)|$, replace θ by $\max\{2, r_2\theta\}$ and continue the iterative procedure. r_1 should always be chosen close to 1, since even when the step size is computed according to (1.8), the ratios $|\lambda_{k+1} - \eta|/|\lambda_k - \eta|$ can be very close to 1. There is no reason to suspect that the ratios $|\nabla g(\lambda_{k+1})|/|\nabla g(\lambda_k)|$ will behave differently. Also, r_2 should not be chosen too small, since then, our search technique may fail to find a value of θ close to the optimum. This modified version of (2.2) clearly converges since it reduces to (2.2) if the values $|\nabla g(\lambda_k)|$ fail to converge to zero at the rate of a geometric progression.

In the discrete versions of our examples we took $h = .1$. Also, in computing $u(i, \lambda_k)$ according to (2.2) we used $u(i, \lambda_{k-1})$ as our initial approximation of $u(i, \lambda_k)$ for $k \geqq 2$. This provides a good initial approximation for large k as the numerical results show. In Example 1 we took $N = 21$ corresponding to $T = 2$, and in Example 2 we took $N = 61$ corresponding to $T = 6$. In both cases convergence was defined by the criterion that the L_∞-norm of the gradient vector $\nabla g(\lambda_k)$ be less than 10^{-3}. Calculations were done on a UNIVAC 1108. The computing time for the first example was 1.62 seconds. The second example is a much larger problem because of the greater number of x variables involved, and because of the larger number of discrete variables that appear because of the longer time interval. The additional difficulty presented by the nonlinear constraints was easily resolved by our decomposition technique. This problem required 4.61 seconds of computer time. In both cases the ascent directions were taken as $s_k = \nabla g(\lambda_k)$, and the algorithm (2.1) was terminated when two successive iterates satisfied $|\delta_{j+1} - \delta_j| < 10^{-4}$. Faster convergence could undoubtedly be obtained by a more clever choice of ascent directions. On this point the reader should see [14].

Example 1.

$$\text{Minimize} \quad \int_0^2 [.5(x_1^2 + x_2^2) + 10(u_1^2 + u_2^2)]\,dt,$$

subject to

$$\frac{d}{dt}\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 1 & 2 \\ -1 & .5 \end{pmatrix}\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} + \begin{pmatrix} 5 & -1 \\ 2 & 4 \end{pmatrix}\begin{pmatrix} u_1 \\ u_2 \end{pmatrix},$$

and subject to the initial and terminal conditions

$$x_1(0) = x_2(0) = .5, \qquad x_1(2) = x_2(2) = .5$$

on the x variables. The u variables are required to satisfy $|u_i(t)| \leqq .2$, $i = 1, 2$.

The results of our numerical experiment for this example are reported in Table 1. We used the modified version of our algorithm starting with $\theta = 40$. This is significantly larger than the value of θ predicted in (2.3). We also used $r_1 = .9$ and $r_2 = .5$.

TABLE 1(a)

k	$\max_i \left\lvert \frac{\partial g}{\partial \lambda_i}(\lambda) \right\rvert$	θ	Step size $\theta\sigma_k$	λ_k	No. iterations of (2.2) required to compute $\nabla g(\lambda_k)$
1	1.6533	40	3·48	(.000, .000)	7
2	.5419	40	2.24	(1.990, 5.769)	11
3	.5680	20	1.09	(.844, 4.552)	8
4	.2841	20	.83	(1.464, 4.094)	5
5	.1341	20	2.01	(1.227, 4.005)	7
6	.0479	20	1.24	(1.179, 3.735)	7
7	.0360	20	2.24	(1.136, 3.677)	6
8	.0242	20	.79	(1.141, 3.596)	5
9	.0081	20	1.85	(1.121, 3.592)	4
10	.0038	20	1.89	(1.117, 3.577)	3
11	.0018	20	1.67	(1.115, 3.570)	3
12	.0009			(1.114, 3.567)	2

TABLE 1(b)

i	$x_1(i)$	$x_2(i)$	$u_1(i)$	$u_2(i)$
1	.50	.50	−.20	−.15
2	.56	.37	−.20	−.09
3	.60	.25	−.20	−.05
4	.62	.14	−.20	−.12
6	.58	−.04	−.20	.02
7	.53	−.04	−.20	.05
8	.46	−.12	−.16	.07
9	.40	−.18	−.13	.09
10	.33	−.23	−.09	.10
11	.27	−.26	−.06	.11
12	.20	−.28	−.04	.11
13	.15	−.27	.00	.11
14	.10	−.25	.01	.11
15	.05	−.22	.03	.10
16	.02	−.19	.04	.10
17	.00	−.15	.05	.09
18	−.01	−.11	.05	.08
19	−.02	−.07	.06	.07
20	−.01	−.03	.06	.06
21	.00018	.0009		

Example 2.

$$\text{Minimize} \quad 2\int_0^6 \left[\sum_{i=1}^{4} x_i^2(t) + \sum_{i=1}^{4} u_i^2(t)\right] dt,$$

subject to

$$\frac{d}{dt}\begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix} = \begin{pmatrix} 0 & 1 & 0 & 0 \\ -.5 & 0 & 0 & .5 \\ 0 & 0 & 0 & 1 \\ 0 & -.5 & -2 & 0 \end{pmatrix}\begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix} + \begin{pmatrix} 0 & 0 \\ 1 & 0 \\ 0 & 0 \\ 0 & 1 \end{pmatrix}\begin{pmatrix} u_1 \\ u_2 \end{pmatrix},$$

and subject to the initial and terminal conditions

$$x_1(0) = .2, \quad x_2(0) = -.25, \quad x_3(0) = -.5, \quad x_4(0) = .25,$$
$$x_i(6) = 0, \qquad i = 1, \cdots, 4,$$

on the x variables. The u variables are required to satisfy the constraints

$$u_1^2(t) + u_2^2(t) \leqq .1.$$

The results of our numerical experiment for this example are reported in Table 2. Again we started the modified version of our algorithm with $\theta = 40$, $r_1 = .9, r_2 = .5$.

We conclude this section with a proof that the solution of the discrete problem (3.1)–(3.2) converges to the solution of the continuous problem (1.1)–(1.2), in an appropriate sense, as $N \to \infty$. In (3.2) the function $f(t, x)$ is assumed to be a quadratic form. However, our proof uses only the convexity property of f. Therefore our result holds for a wider class of problems, and in particular, for the problem (1.1)–(1.2).

THEOREM 3.1. *Let* $u = u(1), \cdots, u(N-1)$ *denote the solution of the discrete problem* (3.1)–(3.2). *Let u also denote the admissible control for the continuous*

TABLE 2(a)

k	$\max_i \left\lvert \frac{\partial g}{\partial \lambda_i}(\lambda_k) \right\rvert$	θ	Step size $\theta\sigma_k$	λ_k	No. iterations of (2.2) required to compute $\nabla g(\lambda_k)$
1	.3503	40	7.49	(.000, .000, .000, .000)	7
2	.2291	40	8.21	(.883, −2.138, −2.623, −.573)	5
3	.1495	40	8.71	(1.472, −3.461, −4.506, −.710)	5
4	.0968	40	9.23	(1.843, −4.222, −5.809, −.708)	5
5	.0618	40	9.93	(2.067, −4.636, −6.703, −.684)	5
6	.0387	40	10.90	(2.191, −4.845, −7.317, −.660)	4
7	.0236	40	12.11	(2.254, −4.936, −7.740, −.649)	4
8	.0140	40	13.33	(2.282, −4.964, −8.027, −.644)	4
9	.0077	40	14.78	(2.290, −4.964, −8.214, −.644)	3
10	.0040	40	16.10	(2.290, −4.952, −8.329, −.645)	3
11	.0019	40	17.24	(2.287, −4.939, −8.394, −.645)	2
12	.0009	40		(2.284, −4.930, −8.427, −.646)	2

TABLE 2(b)

i	$x_i(i)$	$x_2(i)$	$x_3(i)$	$x_4(i)$	$u_1(i)$	$u_2(i)$
1	.20	−.25	−.50	.25	.17	−.23
6	.09	−.12	−.29	.59	−.05	−.31
11	.06	.02	.02	.62	−.01	−.31
16	.09	.11	.28	.31	−.07	−.22
21	.15	.08	.35	−.08	−.05	−.07
26	.17	−.02	.25	−.34	−.02	.27
31	.14	−.12	.07	−.35	.11	.29
36	.06	−.15	−.07	−.19	.16	.17
41	.00	−.11	−.13	.00	.15	−.01
46	−.04	−.02	−.10	.11	.08	−.15
51	−.03	.03	−.04	.10	.00	−.19
56	−.01	.04	.00	.03	−.08	−.11
61	.0001	−.0002	.0009	−.00002		

problem defined by

$$u(t) = u(i) \quad \textit{for} \quad (i-1)h \leqq t < ih, \quad i = 1, 2, \cdots, N-1.$$

Let u^* denote the optimal control for the problem (1.1)–(1.2). *Then u converges to u^* with respect to the norm $\|\cdot\|_D$.*

Proof. By Lee's lemma [15, p. 242],

$$c(u) - c(u^*) \geqq \int_0^T B^{\mathrm{T}}\psi \cdot u^* - u^* \cdot Du^* - \{B^{\mathrm{T}}\psi \cdot u - u \cdot Du\}\, dt, \tag{3.4}$$

where ψ is the function defined by (1.3). If in (2.7) we take $\lambda = \eta$, the solution of (1.7), we have $\psi(t, \lambda) = \psi(t)$, $u(t, \lambda) = u^*(t)$ so that the expression (2.7) is precisely the integrand on the right in (3.4). Since

$$(2D(t)u^*(t) - B^{\mathrm{T}}(t)\psi(t)) \cdot D(t)(u(t) - u^*(t)) \geqq 0,$$

we have, by (2.7) and (3.4),

$$c(u) - c(u^*) \geqq \int_0^T (u(t) - u^*(t)) \cdot D(t)(u(t) - u^*(t))\, dt. \tag{3.5}$$

Cullum [16, Cor. 5.1] has shown that the cost $c(u)$ approaches $c(u^*)$ as $N \to \infty$. It therefore follows from (3.5) that

$$\lim_{N \to \infty} \|u - u^*\|_D = 0.$$

This completes the proof of Theorem 3.1.

REFERENCES

[1] E. R. BARNES, *A geometrically convergent algorithm for solving optimal control problems*, this Journal, 10 (1972), pp. 434–443.

[2] B. N. PSHENICHNIY, *Linear optimal control problems*, this Journal, 4 (1966), pp. 577–593.

[3] L. W. NEUSTADT AND B. H. PAIWONSKY, *On synthesizing optimal controls*, Procceedings of Second IFAC Congress, Butterworths, London, 1965.

[4] L. W. NEUSTADT, *Synthesizing time optimal control systems*, J. Math. Anal. Appl., 1 (1960), pp. 484–493.

[5] J. S. MEDITCH AND L. W. NEUSTADT, *An application of optimal control to midcourse guidance*, Proceedings of Second IFAC Congress, Butterworths, London, 1965.

[6] J. H. EATON, *An iterative solution to time optimal control*, J. Math. Anal. Appl., 5 (1962), pp. 329–344.

[7] D. G. LUENBERGER, *A primal-dual algorithm for the computation of optimal control*, Computing Methods in Optimization Problems. II, L. A. Zadeh, L. W. Neustadt and A. V. Balakrishnan, eds., Academic Press, New York, 1969, pp. 223–234.

[8] E. POLAK, *On primal and dual methods for solving discrete optimal control problems*, Ibid., pp. 317–330.

[9] E. J. FADDEN AND E. G. GILBERT, *Computational aspects of the time optimal control problem*, Computing Methods in Optimization Problems, A. V. Balakrishnan and L. W. Neustadt, eds., Academic Press, New York, 1964, pp. 167–189.

[10] E. R. BARNES, *A procedure for solving certain dual optimal control problems*, this Journal, 10 (1972), pp. 377–392.

[11] P. WOLFE, *The simplex method for quadratic programming*, Econometrica, 27 (1959), pp. 382–398.

[12] E. M. L. BEALE, *On quadratic programming*, Naval Res. Logist. Quart., 6 (1959), pp. 227–244.

[13] P. WOLFE, *Convergence theory in nonlinear programming*, Integer and Nonlinear Programming, J. Abadie, ed., North Holland, Amsterdam, 1970.

[14] B. N. PSHENICHNII, *On the acceleration of convergence of algorithms for solving optimal control*, Computing Methods in Optimization Problems. II, L. A. Zadeh, L. W. Neustadt and A. V. Balakrishnan, eds., Academic Press, New York, 1969, pp. 331–352.

[15] E. B. LEE, *A sufficient condition in the theory of optimal control*, this Journal, 1 (1963), pp. 241–245.

[16] J. CULLUM, *An explicit procedure for discretizing continuous, optimal control problems*, J. Optimization Theory Appl., 8 (1971), pp. 15–33.

SIAM J. Control
Vol. 11, No. 2, May 1973

EXTENSIONS OF THE CONCEPT OF REGULAR SYNTHESIS AS A SUFFICIENT CONDITION FOR OPTIMALITY*

THOMAS G. HACK†

Abstract. A generalized regular synthesis is defined and examined using the approach of Boltyanskii [2]. Special attention is given to the action of the marked trajectories on the lower order manifolds. The yield of this analysis is twofold. First, sufficient conditions for the general autonomous control problem are obtained. In addition, the notion of a modified regular synthesis is exhibited as a second sufficient condition for optimality. The subsequent discussions show how the latter result can be applied to differential games.

1. Introduction.

In [2] Boltyanskii introduces the notion of a "regular synthesis" as a sufficient condition for the time optimal control problem. Mirica, in [5], applies the methods of Berkovitz [1] to obtain a more general sufficiency theorem. His underlying assumptions differ slightly from those of Boltyanskii but the fundamental hypothesis required is the regular synthesis.

Our purpose in this paper is to define a regular synthesis for the general autonomous control problem and to examine its "marked trajectories" employing a broader approach than that used by Mirica and Boltyanskii. In this way, we not only generalize Boltyanskii's results (to the general autonomous control problem), but also establish a basic set of properties of the synthesis which can be used in the solution of other problems. The investigation of regular synthesis is carried out in § 3.

The method of attack is the same as in [2], but the argument is executed in a generalized setting—a situation which gives rise to some new insights. The most interesting of these is the equation which relates the derivatives of the value function with the dual variables on the lower order manifolds.

Unfortunately the concept of regular synthesis in its original form cannot be applied to differential games. In § 4 we introduce a slightly weaker hypothesis which we call a "modified regular synthesis" and obtain sufficient conditions for optimality in a restricted class of admissible controls. The modified regular synthesis can also be used to obtain sufficient conditions for saddle points in differential games (see [4]).

Section 5 contains a comparison of regular synthesis and modified regular synthesis as they apply to differential games. The reason for our adoption of the notion of modified regular synthesis for differential games is made clear by the explanation of the failure of regular synthesis in that area.

2. Statement of the problem and preliminary notions.

In this paper, we shall be concerned with control processes which can be described by a system of ordinary differential equations:

$$\frac{dx}{dt} = f(x, u), \tag{2.1}$$

* Received by the editors August 30, 1971.

† Bell Telephone Laboratories, Incorporated, Naperville, Illinois 60540.

where $x(t)$ in R^n is a vector defining the state of the system at each instant of time t. The control variable u occurring in (2.1) may take any value in a given control set $U \subset R^r$.

In addition, we are given a region V contained in R^n and a k-dimensional smooth manifold T lying in V or on the boundary of V. The set T is called the terminal manifold. Here $k < n$ and we allow the possibility of T being a single point. For each x in V, the set $A(x)$ is defined to be the set of all piecewise continuous controls $u(t)$ taking values in U which transfer the phase point moving in accordance with (2.1) from the position x to the terminal manifold T; the phase point is required to remain in V. If $u \in A(x_0)$, u is said to be an admissible control transferring x_0 to T and the solution x of (2.1) is called the admissible trajectory corresponding to u. For $u \in A(x_0)$, we define

$$J(u) = g(x(t_1)) + \int_{t_0}^{t_1} f^0(x(t), u(t))\, dt, \tag{2.2}$$

where g is a real-valued function continuously differentiable in a neighborhood of T, f^0 is a real-valued function satisfying certain smoothness requirements which we shall state explicitly in §§ 3 and 4, and $[t_0, t_1]$ is the domain of definition of u. The basic problem we shall consider is the following: given $x_0 \in V$, find an admissible control u which transfers x_0 to T and minimizes J over all $u \in A(x_0)$.

The control u which solves this problem and its corresponding trajectory x will be called optimal in V. Clearly, if the region V is changed, this control may cease to be optimal.

The well-known necessary condition (see [7]) for our basic problem is the maximum principle which may be stated as follows:

In order that a control u and the corresponding trajectory x be optimal, it is necessary that there exist a nonzero absolutely continuous vector function

$$\hat{\lambda}(t) = (\lambda^0(t), \lambda(t)) = (\lambda^0(t), \lambda^1(t), \cdots, \lambda^n(t))$$

satisfying (2.3) through (2.6):

$$\frac{d\lambda}{dt} = -H_x(\hat{\lambda}, x(t), u(t)), \qquad \frac{d\lambda^0}{dt} = 0, \tag{2.3}$$

where

$$H(\hat{\lambda}, x, u) = \hat{\lambda}\hat{f}(x, u), \qquad \hat{f} = (f^0, f).$$

$$H(\hat{\lambda}(t), x(t), u(t)) = \max_{u \in U} H(\hat{\lambda}(t), x(t), u), \qquad t_0 \leqq t \leqq t_1. \tag{2.4}$$

$$H(\hat{\lambda}(t), x(t), u(t)) = 0; \qquad \lambda^0(t) = \text{const.} \leqq 0. \tag{2.5}$$

(2.6) The vector $\lambda^0(t_1)g_x(x(t_1)) - \lambda(t_1)$ is orthogonal to T.

In this paper we shall present additional relations which when satisfied along with the maximum principle guarantee optimality. These relations will take the form of a regular synthesis or a modified regular synthesis, concepts which are defined in §§ 3 and 4 respectively.

It is convenient at this time to present the following lemma, due to Boltyanskii [2], which plays a fundamental role in the proof of our results. We use the notion of a piecewise-smooth set as defined by Boltyanskii in [2].

LEMMA 2.1. *In the region V let there be given a piecewise-smooth set M of dimension $\leqq n - 1$ and the function $S(x)$, continuous in V, continuously differentiable in $V - M$ which satisfies the condition*

$$S_x(x)f(x,u) \leqq f^0(x,u). \tag{2.7}$$

Then if $u(t)$, $t_0 \leqq t \leqq t_1$, is a piecewise continuous control taking values in U and transferring the phase point from position x_0 to x_1 in accordance with (2.1), *and if $x(t)$ is the corresponding trajectory,*

$$\int_{t_0}^{t_1} f^0(x(t), u(t))\, dt \geqq S(x_1) - S(x_0). \tag{2.8}$$

3. Regular synthesis. In [2], the notion of regular synthesis is presented in relation to the fixed endpoint problem, that is, T is a single point in V. We now generalize that concept to fit our basic problem.

DEFINITION 3.1. Let N be a piecewise-smooth set of dimension $\leqq n - 1$ and $P^k, \cdots, P^n$ be piecewise-smooth sets, such that

$$P^k \subset P^{k+1} \subset \cdots \subset P^n = V. \tag{3.1}$$

Let T be a k-dimensional smooth manifold lying in V or on the boundary of V. In addition, let $v(x)$ be a function defined in V and taking values in U. The sets (3.1) and the function $v(x)$ effect a *regular synthesis* for (2.1) in the region V if the following conditions are satisfied:

(A) If $T \subset V$, we set $P^{k-1} = T$; if $T \subset \partial V$, we set $P^{k-1} = \varnothing$. Note that in the former case, $k - 1$ does not indicate the dimension of T. Every component of the set $P^i - (P^{i-1} \cup N)$, $i = k, \cdots, n$, is an i-dimensional smooth manifold. We call these components i-dimensional cells. The function v is continuous and continuously differentiable on each cell and has a continuously differentiable extension defined on a neighborhood of the closure of the cell.

(B) Each cell is designated either as a cell of the first kind or a cell of the second kind. All n-dimensional cells are cells of the first kind and all k-dimensional cells are cells of the second kind. If σ is any i-dimensional cell of the first kind, then there exists a unique $(i - 1)$-dimensional cell $\pi(\sigma)$ such that the following condition is satisfied: if $x \in \sigma$, there exists a unique trajectory of the equation

$$\frac{dx}{dt} = f(x, v(x)) \tag{3.2}$$

passing through x, which leaves the cell σ after a finite time. At this time of departure, it strikes $\pi(\sigma)$ at a nonzero angle and approaches it with a nonzero velocity. If σ is any i-dimensional cell of the second kind, there exists a unique $(i + 1)$-dimensional cell of the first kind $\Sigma(\sigma)$ such that from each point $x \in \sigma$, there issues a unique trajectory of (3.2) which moves in the cell $\Sigma(\sigma)$ and intersects σ only at the initial point x. We also assume that v is continuously differentiable in $\sigma \cup \Sigma(\sigma)$

and can be extended to a continuously differentiable function in a neighborhood of $\sigma \cup \Sigma(\sigma)$.

(C) The functions f^0 and f are continuously differentiable in some open set E containing $V \times U$.

(D) From each point of the set N, there issues a trajectory of (3.2) (not necessarily unique) which remains in V and reaches T in a finite time. This trajectory is said to be marked.

(E) For $x \in V - N$, condition (B) ensures that the trajectory of (3.2) passing through x may be extended from cell to cell. We further require that it remain in V, reach T in a finite time after traveling through at most a finite number of cells and not be tangent to T. This trajectory is also said to be marked.

(F) All marked trajectories satisfy the maximum principle.

(G) The functional J defined in (2.2) computed along these trajectories is assumed to be a continuous function of the initial point x_0 in V. In particular, if $x_0 \in N$, the value of J computed along the various trajectories transferring x_0 to T must be the same.

DEFINITION 3.2. For each $\gamma \in V - N$, the unique trajectory transferring γ to T in a regular synthesis will be called the *marked trajectory originating at* γ and denoted by $\phi(t, \gamma)$. For $\gamma \in N$, each of the trajectories transferring γ to T will be called a marked trajectory originating at γ. For convenience in notation, we let $\phi(t, \gamma)$ denote one of the trajectories.

We consider each marked trajectory as starting at time t_0, an arbitrarily chosen constant. This is permissible because we are dealing with an autonomous system. The time at which the marked trajectory originating at γ strikes T is designated by $t_f(\gamma)$. Thus we have

$$\phi(t_0, \gamma) = \gamma, \quad \gamma \in V, \tag{3.3}$$

$$\phi(t_f(\gamma), \gamma) \in T, \quad \gamma \in V. \tag{3.4}$$

DEFINITION 3.3. For $\gamma \in V$, we define

$$w(\gamma) = g(\phi(t_f(\gamma), \gamma)) + \int_{t_0}^{t_f(\gamma)} f^0(\phi(t, \gamma), v(\phi(t, \gamma)))\, dt.$$

The function w is called the *value function*.

Note that condition G of Definition 3.1 states that w is continuous in V.

THEOREM 1. *If a regular synthesis for* (2.1) *is effected in the region* V, *then the marked trajectories are optimal.*

Since the argument is rather long, the proof of Theorem 1 will be broken down into a series of lemmas. To further simplify the exposition we make the following assumption concerning the cells of a regular synthesis: each m-dimensional cell σ is given parametrically by equations

$$x = X_\sigma(\delta),$$

where $\delta = (\delta^1, \cdots, \delta^m)$ ranges over some open set N_σ in m-dimensional space, that is, $\sigma = X_\sigma(N_\sigma)$. This type of representation is assumed solely for the purpose of making the argument less cumbersome; the existence of a regular synthesis alone is sufficient to establish the validity of Theorem 1 (see [4]).

Let σ denote an arbitrary m-dimensional cell, $k \leqq m \leqq n$. Then by conditions (B) and (E) of regular synthesis, the marked trajectories originating in σ pass through the same sequence of cells of the first kind, $\sigma_1, \sigma_2, \cdots, \sigma_q$. Here we have set $\sigma_1 = \sigma$ if σ is a cell of the first kind and $\sigma_1 = \Sigma(\sigma)$ if σ is a cell of the second kind.

DEFINITION 3.4. For each $\delta \in N_\sigma$, we define $t_i(\delta)$ to be the instant of time at which $\phi(t, X_\sigma(\delta))$ strikes $\pi(\sigma_i)$. Note that $t_q(\delta) = t_f(X_\sigma(\delta))$. In addition, $x_i(\delta)$ is defined to be the point at which $\phi(t, X_\sigma(\delta))$ strikes $\pi(\sigma_i)$; that is,

$$x_i(\delta) = \phi(t_i(\delta), X_\sigma(\delta)), \qquad \delta \in N_\sigma, \qquad i = 1, \cdots, q. \tag{3.5}$$

LEMMA 3.1. *The functions t_i, x_i, $i = 1, \cdots, q$, are continuously differentiable on N_σ.*

Proof. For $i = 1, \cdots, q$, condition (A) of regular synthesis asserts that $v|_{\sigma_i}$ has an extension (denoted by v_i) which is continuously differentiable in a neighborhood G_i of $\bar{\sigma}_i$. In G_i, we let $y_i(t; s, \gamma)$ denote the solution of the system

$$\frac{dx}{dt} = f(x, v_i(x)), \qquad x(s) = \gamma. \tag{3.6}$$

For each $\delta \in N_\sigma$, it follows that

$$\begin{aligned} \phi(t, X_\sigma(\delta)) &= y_i(t; t_{i-1}(\delta), x_{i-1}(\delta)), \\ t_{i-1}(\delta) &\leqq t \leqq t_i(\delta), \qquad i = 1, \cdots, q. \end{aligned} \tag{3.7}$$

Let δ_0 be any point of N_σ. We shall first show that the functions $t_1(\delta)$, $x_1(\delta)$ are continuously differentiable in a neighborhood of δ_0 and thus are continuously differentiable in N_σ. Setting $i = 1$, we have for $\delta \in N_\sigma$,

$$\phi(t, X_\sigma(\delta)) = y_1(t; t_0, X_\sigma(\delta)), \qquad t_0 \leqq t \leqq t_1(\delta). \tag{3.8}$$

There exists an $\alpha_1 > 0$ such that the solution $y_1(t; t_0, X_\sigma(\delta_0))$ of

$$\frac{dx}{dt} = f(x, v_1(x)), \qquad x(t_0) = X_\sigma(\delta_0)$$

can be extended in G_1 to the interval $t_0 - \alpha_1 \leqq t \leqq t_1(\delta_0) + \alpha_1$. Now from standard theorems on the continuity and differentiability of solutions of differential equations with respect to initial conditions (see [6]), we obtain the existence of a positive number r such that for

$$t_0 - \alpha_1 < t < t_1(\delta_0) + \alpha_1, \quad |s - t_0| < r, \quad |\delta - \delta_0| < r$$

the solution $y_1(t; s, X_\sigma(\delta))$ of

$$\frac{dx}{dt} = f(x, v_1(x)), \qquad x(s) = X_\sigma(\delta)$$

is a continuously differentiable function of each of its arguments.

Let σ_1 have dimension m. Then $\pi(\sigma_1)$ is an $(m - 1)$-dimensional smooth manifold and $x_1(\delta_0) \in \pi(\sigma_1)$. Since $\pi(\sigma_1)$ has the representation

$$x = X_1(\beta^1, \cdots, \beta^{m-1}), \qquad X_1 \equiv X_{\pi(\sigma_1)},$$

there exists a point $\beta_0 = (\beta_0^1, \cdots, \beta_0^{m-1})$ such that

$$x_1(\delta_0) = X_1(\beta_0).$$

From (3.5) and (3.7), we have the equality

$$x_1(\delta_0) = \phi(t_1(\delta_0), X_\sigma(\delta_0)) = y_1(t_1(\delta_0); t_0, X_\sigma(\delta_0)),$$

and thus

$$X_1(\beta_0) = y_1(t_1(\delta_0); t_0, X_\sigma(\delta_0)). \tag{3.9}$$

For δ in a neighborhood of δ_0, we consider the equation

$$X_1(\beta^1, \cdots, \beta^{m-1}) = y_1(t; t_0, X_\sigma(\delta)) \tag{3.10}$$

which defines $t, \beta^1, \cdots, \beta^{m-1}$ as implicit functions of δ. The solutions $t(\delta)$, $\beta(\delta)$ of (3.10) will give the time and place at which $\phi(t, X_\sigma(\delta))$ strikes $\pi(\sigma_1)$.

The functional matrix of the equations (3.10) is

$$\left(\frac{\partial X_1}{\partial \beta}, \frac{-dy_1}{dt}\right). \tag{3.11}$$

At $t = t_1(\delta_0)$, $\beta = \beta_0$, $\delta = \delta_0$, the matrix in (3.11) has rank m since the trajectory $\phi(t, X_\sigma(\delta_0))$ is not tangent to $\pi(\sigma_1)$ and does not approach $\pi(\sigma_1)$ with a zero velocity. Now by the implicit function theorem, we conclude that

$$t_1(\delta), \beta^1(\delta), \cdots, \beta^{m-1}(\delta) \in C^{(1)}(N_1(\delta_0)), \tag{3.12}$$

where $N_1(\delta_0)$ denotes some neighborhood of δ_0. Also, clearly

$$x_1(\delta) = X_1(\beta^1(\delta), \cdots, \beta^{m-1}(\delta)) \in C^{(1)}(N_1(\delta_0)). \tag{3.13}$$

It follows from (3.12) and (3.13) that

$$x_1, t_1 \in C^{(1)}(N_\sigma). \tag{3.14}$$

Now $\pi(\sigma_1)$ is either a cell of the first kind ($\sigma_2 = \pi(\sigma_1)$) or a cell of the second kind ($\sigma_2 = \Sigma(\pi(\sigma_1))$). In any event, we direct our attention to the flow of trajectories in σ_2 and let $y_2(t; s, \gamma)$ be the solution of the system

$$\frac{dx}{dt} = f(x, v_2(x)), \qquad x(s) = \gamma; \tag{3.15}$$

then for each $\delta \in N_\sigma$,

$$\phi(t, X_\sigma(\delta)) = y_2(t; t_1(\delta), x_1(\delta)), \qquad t_1(\delta) \leqq t \leqq t_2(\delta).$$

From (3.14) and standard theorems on the continuity and differentiability of solutions of differential equations with respect to initial conditions, we deduce the existence of an $\alpha_2 > 0$ such that $y_2(t; t_1(\delta), x_1(\delta))$ is a continuously differentiable function of t and δ for

$$t_1(\delta_0) - \alpha_2 < t < t_2(\delta_0) + \alpha_2, \qquad \delta \in N_2(\delta_0),$$

where $N_2(\delta_0)$ is a sufficiently small neighborhood of δ_0. First, let us assume that $\sigma_2 = \pi(\sigma_1)$. Then we must have that $\pi(\sigma_2)$ is an $(m - 2)$-dimensional smooth manifold. Since $\pi(\sigma_2)$ has the representation

$$x = X_2(\theta^1, \cdots, \theta^{m-2}), \qquad X_2 \equiv X_{\pi(\sigma_2)},$$

there exists a point $\theta_0 = (\theta_0^1, \cdots, \theta_0^{m-2})$ such that

$$x_2(\delta_0) = X_2(\theta_0).$$

In addition we have that

$$X_2(\theta_0) = y_2(t_2(\delta_0); t_1(\delta_0), x_1(\delta_0)).$$

For $\delta \in N_2(\delta_0)$, we consider the equation

$$X_2(\theta^1, \cdots, \theta^{m-2}) = y_2(t; t_1(\delta), x_1(\delta)) \tag{3.16}$$

which defines $t, \theta^1, \cdots, \theta^{m-2}$ as implicit functions of δ. The solutions $t(\delta), \theta(\delta)$ of (3.16) will give the time and place at which $\phi(t, X_\sigma(\delta))$ strikes $\pi(\sigma_2)$.

The $n \times (m - 1)$ functional matrix of the equations (3.16) is

$$\left(\frac{\partial X_2}{\partial \theta}, \frac{-dy_2}{dt}\right). \tag{3.17}$$

At $t = t_2(\delta_0), \theta = \theta_0, \delta = \delta_0$, the matrix in (3.17) has rank $m - 1$ since $\phi(t, X_\sigma(\delta))$ is not tangent to $\pi(\sigma_2)$ and does not approach $\pi(\sigma_2)$ with a zero velocity. Thus if we apply the implicit function theorem to $m - 1$ of the equations in (3.16), we can conclude that

$$t_2(\delta), \theta^1(\delta), \cdots, \theta^{m-2}(\delta) \in C^{(1)}(N_3(\delta_0)), \tag{3.18}$$

where $N_3(\delta_0)$ is a neighborhood of δ_0 contained in $N_2(\delta_0)$. We also have

$$x_2(\delta) = X_2(\theta^1(\delta), \cdots, \theta^{m-2}(\delta)) \in C^{(1)}(N_3(\delta_0)). \tag{3.19}$$

By (3.18) and (3.19), we arrive at the desired result,

$$x_2, t_2 \in C^{(1)}(N_\sigma). \tag{3.20}$$

The second case occurs when $\pi(\sigma_1)$ is a cell of the second kind. In this situation, $\sigma_2 = \Sigma(\pi(\sigma_1))$ is a m-dimensional cell and $\pi(\sigma_2)$ is an $(m - 1)$-dimensional cell. The proof of (3.20) is the same as the above, except that (3.17) is an $n \times m$ matrix of rank m.

We continue in this fashion treating each of the cells σ_i similarly. The final result is

$$x_i, t_i \in C^{(1)}(N_\sigma), \qquad i = 1, \cdots, q, \tag{3.21}$$

and thus the proof of Lemma 3.1 is complete.

In what follows, we denote the gradients of x_i, t_i by $x_{i\delta}$, $t_{i\delta}$, $i = 1, \cdots, q$.

LEMMA 3.2. *For any cell σ, $w^*(\delta) \in C^{(1)}(N_\sigma)$, where $w^*(\delta) \equiv w(X_\sigma(\delta))$, $\delta \in N_\sigma$.*

Proof. Let σ be any m-dimensional cell and let $\sigma_i, x_i, t_i, v_i, G_i, i = 1, \cdots, q$, be defined as before. Now, by definition, we have for $\delta \in N_\sigma$,

$$t_f(\delta) = t_q(\delta). \tag{3.22}$$

This fact and (3.5) yield

$$x_q(\delta) = \phi(t_f(\delta), X_\sigma(\delta)), \qquad \delta \in N_\sigma. \tag{3.23}$$

By assumption, g is continuously differentiable in a neighborhood of T. Hence Lemma 3.1 and (3.23) imply that

$$g(\phi(t_f(\delta), X_\sigma(\delta))) \in C^{(1)}(N_\sigma). \tag{3.24}$$

Recalling that

$$w^*(\delta) = g(\phi(t_f(\delta), X_\sigma(\delta))) + \int_{t_0}^{t_f(\delta)} f^0(\phi(t, X_\sigma(\delta)), v(\phi(t, X_\sigma(\delta))))\,dt, \tag{3.25}$$

we need only show that

$$\int_{t_0}^{t_f(\delta)} f^0(\phi(t, X_\sigma(\delta)), v(\phi(t, X_\sigma(\delta))))\,dt \in C^{(1)}(N_\sigma). \tag{3.26}$$

But we can rewrite (3.26) as follows:

$$\begin{aligned} &\int_{t_0}^{t_f(\delta)} f^0(\phi(t, X_\sigma(\delta)), v(\phi(t, X_\sigma(\delta))))\,dt \\ &\quad = \sum_{i=1}^{q} \int_{t_{i-1}(\delta)}^{t_i(\delta)} f^0(y_i(t; t_{i-1}(\delta), x_{i-1}(\delta)), v_i(y_i(t; t_{i-1}(\delta), x_{i-1}(\delta))))\,dt \end{aligned} \tag{3.27}$$

by expressing $\phi(t, X_\sigma(\delta))$ as in (3.7). Now if we make use of the fact that v_i, y_i, $i = 1, \cdots, q$, are smooth functions of their arguments, Lemma 3.1, and standard theorems (for example, [3, Chap. VII, Thm. 11]), we may conclude that each term of the sum in (3.27) is continuously differentiable in N_σ. This implies (3.26), which completes the proof.

It is convenient at this time to introduce additional notation. Condition (F) of regular synthesis requires that each marked trajectory satisfy the maximum principle. For each $\gamma \in V - N$, $\phi(t, \gamma)$ defined in $t_0 \leqq t \leqq t_f(\gamma)$ denotes a marked trajectory. We let $\hat{\lambda}(t, \gamma)$ on $t_0 \leqq t \leqq t_f(\gamma)$ represent the absolutely continuous function associated with $\phi(t, \gamma)$ by the maximum principle. For reference, we list the properties of $\hat{\lambda}(t, \gamma)$:

$$\hat{\lambda}(t, \gamma) \neq 0, \quad \lambda^0(t, \gamma) \leqq 0, \quad t_0 \leqq t \leqq t_f(\gamma). \tag{3.28}$$

$$\frac{d\lambda^0}{dt} = 0,$$

$$\begin{aligned} \frac{d\lambda}{dt}(t, \gamma) &= -H_x(\hat{\lambda}(t, \gamma), \phi(t, \gamma), v(\phi(t, \gamma))) \\ &= -\sum_{i=0}^{n} \lambda^i(t, \gamma) f_x^i(\phi(t, \gamma), v(\phi(t, \gamma))). \end{aligned} \tag{3.29}$$

$$\max_{u \in U} H(\hat{\lambda}(t, \gamma), \phi(t, \gamma), u) = H(\hat{\lambda}(t, \gamma), \phi(t, \gamma), v(\phi(t, \gamma))) = 0, \quad t_0 \leqq t \leqq t_f(\gamma). \tag{3.30}$$

$$\begin{aligned} &\text{The vector } \lambda^0(t_f(\gamma), \gamma) g_x(\phi(t_f(\gamma), \gamma)) - \lambda(t_f(\gamma), \gamma) \\ &\qquad\qquad \text{is orthogonal to } T \text{ at } \phi(t_f(\gamma), \gamma). \end{aligned} \tag{3.31}$$

The next lemma establishes the relationship between the partial derivatives of the value function w and the multiplier functions $\lambda(t, y)$. The reader is invited to compare this result with those of Berkovitz [1] derived under the assumption of a "regular decomposition."

LEMMA 3.3. *Let σ be any cell of the regular synthesis. Then for any $\delta \in N_\sigma$, the vector*

$$(\lambda^0(t_0, X_\sigma(\delta)), \lambda(t_0, X_\sigma(\delta))\frac{\partial X_\sigma}{\partial \delta}(\delta))$$

is a constant multiple of $(1, w_\delta^*(\delta))$.

Proof. Let δ_0 be an arbitrary but fixed point of N_σ. We shall show that there exists a constant c such that

$$(\lambda^0(t_0, X_\sigma(\delta)), \lambda(t_0, X_\sigma(\delta))\frac{\partial X_\sigma}{\partial \delta}(\delta)) = c(1, w_\delta^*(\delta)). \tag{3.32}$$

To verify (3.32), it is sufficient to show that if $\hat{p} = (p_0, p_1, \cdots, p_m)$ is any vector orthogonal to $(1, w_\delta^*(\delta_0))$, then $\hat{p}$ is also orthogonal to

$$\left(\lambda^0(t_0, X_\sigma(\delta_0)), \lambda(t_0, X_\sigma(\delta_0))\frac{\partial X_\sigma}{\partial \delta^1}(\delta_0), \cdots, \lambda(t_0, X_\sigma(\delta_0))\frac{\partial X_\sigma}{\partial \delta^m}(\delta_0)\right).$$

Here m is the dimension of the cell σ. This is the task we undertake at this time. Let $\hat{p} = (p_0, p_1, \cdots, p_m)$ be a vector satisfying

$$p_0 + pw_\delta^*(\delta_0) = 0, \tag{3.33}$$

where $p \equiv (p_1, \cdots, p_m)$. Let $\phi(t) = \phi(t, y_0)$, that is, $\phi(t)$ is the marked trajectory starting from $y_0 = X_\sigma(\delta_0)$. Further, let $\phi_\alpha(t) = \phi(t, X_\sigma(\delta_0 + \alpha p))$. Using the notation introduced above, these trajectories pass through q cells of the first kind $\sigma_1, \cdots, \sigma_q$, striking $\pi(\sigma_j)$ at points $x_j(\delta_0)$ and $x_j(\delta_0 + \alpha p)$ at times $t_j \equiv t_j(\delta_0)$ and $t_{j\alpha} \equiv t_j(\delta_0 + \alpha p)$.

We are going to calculate the limit

$$\lim_{\alpha \to 0} \frac{1}{\alpha}[w^*(\delta_0 + \alpha p) - w^*(\delta_0)]\lambda^0(t_0, y_0).$$

Since we know that the limit exists by Lemma 3.2, we may choose any sequence $\alpha_n \to 0$ to calculate it. Thus for each $j = 1, \cdots, q$, we may assume that either $t_{j\alpha_n} \geqq t_j$, $n = 1, 2, \cdots$, or $t_{j\alpha_n} \leqq t_j$, $n = 1, 2, \cdots$. We shall give the proof for the case $t_{j\alpha_n} \geqq t_j$, $n = 1, 2, \cdots$; $j = 1, \cdots, q$. The proofs for the other possibilities are similar. To simplify the notation we shall drop the subscript n, assume $t_{j\alpha} \geqq t_j$. and take the limit as $\alpha \to 0$. Setting $t_{0\alpha} = t_0$ and recalling that $\lambda^0(t, y_0)$ is constant on $t_0 \leqq t \leqq t_q$, we have

$$\frac{1}{\alpha}[w^*(\delta_0 + \alpha p) - w^*(\delta_0)]\lambda^0(t_0, y_0)$$

$$= \frac{1}{\alpha}\left[\int_{t_0}^{t_{q\alpha}} f^0(\phi_\alpha(t), v(\phi_\alpha(t)))\,dt - \int_{t_0}^{t_q} f^0(\phi(t), v(\phi(t)))\,dt\right.$$

$$\left. + g(x_q(\delta_0 + \alpha p)) - g(x_q(\delta_0))\right]\lambda^0(t_0, y_0) \qquad \text{(cont.)}$$

$$
\begin{aligned}
&= \frac{1}{\alpha}\sum_{j=1}^{q}\int_{t_{j-1,\alpha}}^{t_j}[f^0(\phi_\alpha(t), v(\phi_\alpha(t))) - f^0(\phi(t), v(\phi(t)))]\lambda^0(t, \gamma_0)\,dt \\
&\quad + \frac{1}{\alpha}\sum_{j=1}^{q-1}\int_{t_j}^{t_{j\alpha}}[f^0(\phi_\alpha(t), v(\phi_\alpha(t))) - f^0(\phi(t), v(\phi(t)))]\lambda^0(t, \gamma_0)\,dt \\
&\quad + \frac{1}{\alpha}[g(x_q(\delta_0 + \alpha p)) - g(x_q(\delta_0))]\lambda^0(t_q, \gamma_0) \\
&\quad + \frac{1}{\alpha}\int_{t_q}^{t_{q\alpha}} f^0(\phi_\alpha(t), v(\phi_\alpha(t)))\,dt\lambda^0(t_q, \gamma_0).
\end{aligned}
$$

Evaluating the limits of the terms in the second sum, we obtain, for $j = 1, \cdots, q-1$,

$$
\begin{aligned}
(3.34)\qquad &\lim_{\alpha\to 0}\frac{1}{\alpha}\int_{t_j}^{t_{j\alpha}}[f^0(\phi_\alpha(t), v(\phi_\alpha(t))) - f^0(\phi(t), v(\phi(t)))]\lambda^0(t, \gamma_0)\,dt \\
&= t_{j\delta}(\delta_0)p[f^0(\phi(t_j), v_j(\phi(t_j))) - f^0(\phi(t_j), v_{j+1}(\phi(t_j)))]\lambda^0(t_j, \gamma_0),
\end{aligned}
$$

where $t_{j\delta} \equiv \partial t_j/\partial\delta$. Here we have made use of the fact that for $t_j < t < t_{j\alpha}$, $\phi(t) \in \sigma_{j+1}$ and $\phi_\alpha(t) \in \sigma_j$. Similarly, we have

$$
(3.35)\qquad \lim_{\alpha\to 0}\frac{1}{\alpha}\int_{t_q}^{t_{q\alpha}} f^0(\phi_\alpha(t), v(\phi_\alpha(t)))\,dt = t_{q\delta}(\delta_0)pf^0(\phi(t_q), v_q(\phi(t_q)))\lambda^0(t_q, \gamma_0).
$$

Since $F(x) = f^0(x, v_j(x))$ is continuously differentiable on σ_j and can be extended to a continuously differentiable function in a neighborhood of $\bar{\sigma}_j$, we can write

$$
\begin{aligned}
&\frac{1}{\alpha}\int_{t_{j-1,\alpha}}^{t_j}[f^0(\phi_\alpha(t), v(\phi_\alpha(t))) - f^0(\phi(t), v(\phi(t)))]\lambda^0(t, \gamma_0)\,dt \\
&\quad = \frac{1}{\alpha}\int_{t_{j-1,\alpha}}^{t_j}[f_x^0(\bar{\phi}(t), v(\bar{\phi}(t))) + f_u^0(\bar{\phi}(t), v(\bar{\phi}(t)))v_x(\bar{\phi}(t))][\phi_\alpha(t) - \phi(t)]\lambda^0(t, \gamma_0)\,dt \\
&\quad = \int_{t_{j-1,\alpha}}^{t_j}\frac{1}{\alpha}G_\alpha(t)[\phi_\alpha(t) - \phi(t)]\,dt,
\end{aligned}
$$

where $\bar{\phi}(t)$ lies on the line segment joining $\phi_\alpha(t)$ and $\phi(t)$ and

$$
\begin{aligned}
G_\alpha(t) = {}&[H_x(\hat{\lambda}(t, \gamma_0), \bar{\phi}(t), v(\bar{\phi}(t))) - \lambda(t, \gamma_0)f_x(\bar{\phi}(t), v(\bar{\phi}(t)))] \\
&+ [H_u(\hat{\lambda}(t, \gamma_0), \bar{\phi}(t), v(\bar{\phi}(t))) - \lambda(t, \gamma_0)f_u(\bar{\phi}(t), v(\bar{\phi}(t)))]v_x(\bar{\phi}(t)).
\end{aligned}
$$

Now it easily follows from Lemma 3.1 that there exists a constant C such that

$$
(3.36)\qquad |\phi_\alpha(t) - \phi(t)| \leqq C\alpha, \qquad t_0 \leqq t \leqq t_q,
$$

for α sufficiently small.

Moreover, the function $G_\alpha(t)$ is bounded on $t_{j-1,\alpha} \leqq t \leqq t_j$ uniformly in α for α sufficiently small. As $\alpha \to 0$, the function $(1/\alpha)G_\alpha(t)[\phi_\alpha(t) - \phi(t)]$ converges almost everywhere to

$$
\begin{aligned}
&\frac{\partial\phi}{\partial\delta}(t, X_\sigma(\delta_0))p\{[H_x(\hat{\lambda}(t, \gamma_0), \phi(t), v(\phi(t))) - \lambda(t, \gamma_0)f_x(\phi(t), v(\phi(t)))] \\
(3.37)\qquad &\quad + [H_u(\hat{\lambda}(t, \gamma_0), \phi(t), v(\phi(t))) - \lambda(t, \gamma_0)f_u(\phi(t), v(\phi(t)))]v_x(\phi(t))\}.
\end{aligned}
$$

We note that

$$H_u(\hat{\lambda}(t, \gamma_0), \phi(t), v(\phi(t)))v_x(\phi(t))\frac{\partial \phi}{\partial \delta}(t, X_\sigma(\delta_0))p = 0,$$

since by the maximum principle the function (considered as a function of δ alone, that is, with t fixed)

$$H(\hat{\lambda}(t, \gamma_0), \phi(t), v(\phi(t, X_\sigma(\delta))))$$

has a maximum at $\delta = \delta_0$.

Taking this into account and using the dominated convergence theorem, we obtain

$$\begin{aligned}
&\lim_{\alpha \to 0} \frac{1}{\alpha} \int_{t_{j-1,\alpha}}^{t_j} [f^0(\phi_\alpha(t), v(\phi_\alpha(t))) - f^0(\phi(t), v(\phi(t)))]\lambda^0(t, \gamma_0)\, dt \\
&\quad = \int_{t_{j-1}}^{t_j} [H_x(\lambda(t, \gamma_0), \phi(t), v(\phi(t))) - \lambda(t, \gamma_0) f_x(\phi(t), v(\phi(t))) \\
&\qquad - \lambda(t, \gamma_0) f_u(\phi(t), v(\phi(t)))v_x(\phi(t))]\frac{\partial \phi}{\partial \delta}(t, X_\sigma(\delta_0))p\, dt.
\end{aligned} \tag{3.38}$$

Rewriting (3.38) with the help of (3.29) we find that

$$\begin{aligned}
&\lim_{\alpha \to 0} \frac{1}{\alpha} \int_{t_{j-1,\alpha}}^{t_j} [f^0(\phi_\alpha(t), v(\phi_\alpha(t))) - f^0(\phi(t), v(\phi(t)))]\lambda^0(t, \gamma_0)\, dt \\
&\quad = \int_{t_{j-1}}^{t_j} \left\{ -\frac{d\lambda}{dt}(t, \gamma_0)\frac{\partial \phi}{\partial \delta}(t, X_\sigma(\delta_0))p \right. \\
&\qquad \left. - \lambda(t, \gamma_0)\frac{\partial}{\partial \delta}[f(\phi(t, X_\sigma(\delta_0)), v(\phi(t, X_\sigma(\delta_0))))]p \right\} dt \\
&\quad = -\int_{t_{j-1}}^{t_j} \frac{d}{dt}\left\{ \lambda(t, \gamma_0)\frac{\partial \phi}{\partial \delta}(t, X_\sigma(\delta_0))p \right\} dt \\
&\quad = \left[\lambda(t_{j-1}^+, \gamma_0)\frac{\partial \phi}{\partial \delta}(t_{j-1}^+, X_\sigma(\delta_0)) - \lambda(t_j^-, \gamma_0)\frac{\partial \phi}{\partial \delta}(t_j^-, X_\sigma(\delta_0)) \right] p, \\
&\hspace{20em} j = 1, \cdots, q.
\end{aligned} \tag{3.39}$$

Set $x_0(\delta) = X_\sigma(\delta)$. The following relations can be obtained by straightforward calculation:

$$\begin{aligned}
&\frac{\partial \phi}{\partial \delta}(t_j^+, X_\sigma(\delta_0)) = -t_{j\delta}(\delta_0) f(\phi(t_j, X_\sigma(\delta_0)), v_{j+1}(\phi(t_j, X_\sigma(\delta_0)))) + x_{j\delta}(\delta_0), \\
&\hspace{20em} j = 0, \cdots, q - 1, \\
&\frac{\partial \phi}{\partial \delta}(t_j^-, X_\sigma(\delta_0)) = -t_{j\delta}(\delta_0) f(\phi(t_j, X_\sigma(\delta_0)), v_j(\phi(t_j, X_\sigma(\delta_0)))) + x_{j\delta}(\delta_0), \\
&\hspace{20em} j = 1, \cdots, q.
\end{aligned} \tag{3.40}$$

Thus from (3.34), (3.35), (3.39) and (3.40) we have

$$
\begin{aligned}
&\lim_{\alpha\to 0}\frac{1}{\alpha}[w^*(\delta_0+\alpha p)-w^*(\delta_0)]\lambda^0(t_0,\gamma_0)\\
&\quad=\sum_{j=1}^{q}\{\lambda(t_{j-1}^+,\gamma_0)[-t_{j-1,\delta}(\delta_0)f(\phi(t_{j-1},X_\sigma(\delta_0)),v_j(\phi(t_{j-1},X_\sigma(\delta_0))))+x_{j-1,\delta}(\delta_0)]p\\
&\qquad\quad-\lambda(t_j^-,\gamma_0)[-t_{j\delta}(\delta_0)f(\phi(t_j,X_\sigma(\delta_0)),v_j(\phi(t_j,X_\sigma(\delta_0))))+x_{j\delta}(\delta_0)]p\}\\
&\qquad\quad+\sum_{j=1}^{q-1}\lambda^0(t_j,\gamma_0)[f^0(\phi(t_j),v_j(\phi(t_j)))-f^0(\phi(t_j),v_{j+1}(\phi(t_j)))]t_{j\delta}(\delta_0)p\\
&\qquad\quad+\lambda^0(t_q,\gamma_0)f^0(\phi(t_q),v_q(t_q))t_{q\delta}(\delta_0)p+\lambda^0(t_q,\gamma_0)g_x(x_q(\delta_0))x_{q\delta}(\delta_0)p\\
&\quad=-\sum_{j=1}^{q-1}[\lambda^0(t_j,\gamma_0)f^0(\phi(t_j),v_{j+1}(\phi(t_j)))+\lambda(t_j^+,\gamma_0)f(\phi(t_j),v_{j+1}(\phi(t_j)))]t_{j\delta}(\delta_0)p\\
&\qquad\quad+\sum_{j=1}^{q-1}[\lambda^0(t_j,\gamma_0)f^0(\phi(t_j),v_j(\phi(t_j)))+\lambda(t_j^-,\gamma_0)f(\phi(t_j),v_j(\phi(t_j)))]t_{j\delta}(\delta_0)p\\
&\qquad\quad+\lambda^0(t_q,\gamma_0)f^0(\phi(t_q),v_q(\phi(t_q)))t_{q\delta}(\delta_0)p+\lambda(t_q^-,\gamma_0)f(\phi(t_q),v_q(\phi(t_q)))t_{q\delta}(\delta_0)p\\
&\qquad\quad+\sum_{j=1}^{q}[\lambda(t_{j-1}^+,\gamma_0)x_{j-1,\delta}(\delta_0)-\lambda(t_j^-,\gamma_0)x_{j\delta}(\delta_0)]p\\
&\qquad\quad-\lambda(t_0^+,\gamma_0)f(\phi(t_0),v_1(\phi(t_0)))t_{0\delta}(\delta_0)p+\lambda^0(t_q,\gamma_0)g_x(x_q(\delta_0))x_{q\delta}(\delta_0)p.
\end{aligned}
$$

By (3.30),

$$
\begin{aligned}
H(\hat{\lambda}(t_j^+,\gamma_0),\phi(t_j),v_{j+1}(\phi(t_j)))&=0, &\qquad j&=0,\cdots,q-1,\\
H(\hat{\lambda}(t_j^-,\gamma_0),\phi(t_j),v_j(\phi(t_j)))&=0, & j&=1,\cdots,q,\\
t_{0\delta}(\delta_0)&=0,
\end{aligned}
$$

so that

$$
\begin{aligned}
&\lim_{\alpha\to 0}\frac{1}{\alpha}[w^*(\delta_0+\alpha p)-w^*(\delta_0)]\lambda^0(t_0,\gamma_0)\\
&\quad=\sum_{j=1}^{q}[\lambda(t_{j-1}^+,\gamma_0)x_{j-1,\delta}(\delta_0)-\lambda(t_j^-,\gamma_0)x_{j\delta}(\delta_0)]p+\lambda^0(t_q,\gamma_0)g_x(x_q(\delta_0))x_{q\delta}(\delta_0)p\\
&\quad=\lambda(t_0^+,\gamma_0)\frac{\partial X_\sigma(\delta_0)}{\partial\delta}p.
\end{aligned}
$$

The last equality follows from (3.31) and the fact that

$$
\lambda(t_j^+,\gamma_0)x_{j\delta}(\delta_0)=\lambda(t_j^-,\gamma_0)x_{j\delta}(\delta_0),\qquad j=1,\cdots,q. \tag{3.41}
$$

Now since $(p_0,p_1,\cdots,p_m)$ was assumed to be orthogonal to $(1,w_\delta^*(\delta_0))$, it follows that

$$
p_0\lambda^0(t_0,\gamma_0)+p\lambda(t_0,\gamma_0)\frac{\partial X_\sigma(\delta_0)}{\partial\delta}=0. \tag{3.42}
$$

Thus $\hat{p} = (p_0, p_1, \cdots, p_m)$ is orthogonal to the vector

$$\left(\lambda^0(t_0, \gamma_0), \lambda(t_0, \gamma_0)\frac{\partial X_\sigma(\delta_0)}{\partial \delta}\right)$$

and this completes the proof of Lemma 3.3.

LEMMA 3.4. *Let σ be a cell of dimension m, $k \leqq m \leqq n$. If $m < n$, we make the following assumption: for each $\gamma \in \sigma$, $\lambda^0(t, \gamma) < 0$ for $t_0 \leqq t \leqq t_f(\gamma)$. Then if (x, u) is an admissible pair transferring x_0 to x_1 in σ,*

$$w(x_0) - w(x_1) \leqq \int_{t_0}^{t_1} f^0(x(t), u(t))\, dt,$$

where $x(t_0) = x_0, x(t_1) = x_1$.

Proof. Note that the assumption made concerning $\lambda^0(t, \gamma)$ holds automatically if σ is an n-dimensional cell, since, by Lemma 3.3, $\lambda^0(t, \gamma) = 0$ for some t would imply $\hat{\lambda}(t, \gamma) = 0$, which is impossible. In σ we define

$$\lambda^*(\gamma) = -\frac{\lambda(t_0, \gamma)}{\lambda^0(t_0, \gamma)}.$$

Consider the closed interval $[t_0 + \alpha, t_1 - \alpha]$. Since $x(t) \in \sigma$ for these values of t, there exists a partition $t_0 + \alpha = s_0 < s_1 < \cdots < s_r = t_1 - \alpha$ of $[t_0 + \alpha, t_1 - \alpha]$ and functions δ_i, $i = 1, \cdots, r$, absolutely continuous on $[s_{i-1}, s_i]$, such that

$$x(t) = X_\sigma(\delta_i(t)), \qquad s_{i-1} \leqq t \leqq s_i. \tag{3.43}$$

Since $w(X_\sigma(\delta)) \in C^{(1)}(N_\sigma)$ by Lemma 3.2, $w(x(t)) = w(X_\sigma(\delta_i(t)))$ is absolutely continuous on $s_{i-1} \leqq t \leqq s_i$, $i = 1, \cdots, m$. Setting $w^*(\delta) = w(X_\sigma(\delta))$, we obtain

$$\begin{aligned} w(x(s_i)) - w(x(s_{i-1})) &= \int_{s_{i-1}}^{s_i} \frac{d}{dt}(w(x(t)))\, dt \\ &= \int_{s_{i-1}}^{s_i} \frac{\partial w^*(\delta_i(t))}{\partial \delta} \frac{d\delta_i}{dt} dt. \end{aligned} \tag{3.44}$$

From Lemma 3.3,

$$w_\delta^*(\delta_i(t)) = \frac{\lambda(t_0, X_\sigma(\delta_i(t)))}{\lambda^0(t_0, X_\sigma(\delta_i(t)))} \frac{\partial X_\sigma(\delta_i(t))}{\partial \delta}, \qquad s_{i-1} \leqq t \leqq s_i, \qquad i = 1, \cdots, m. \tag{3.45}$$

Also

$$\frac{dx}{dt} = \frac{\partial X_\sigma(\delta_i(t))}{\partial \delta} \frac{d\delta_i(t)}{dt}, \qquad s_{i-1} \leqq t \leqq s_i, \qquad i = 1, \cdots, m. \tag{3.46}$$

Thus from (3.44) through (3.46),

$$\begin{aligned} w(x(s_{i-1})) - w(x(s_i)) &= \int_{s_{i-1}}^{s_i} -\frac{\lambda(t_0, X_\sigma(\delta_i(t)))}{\lambda^0(t_0, X_\sigma(\delta_i(t)))}\frac{dx}{dt}\,dt \\ &= \int_{s_{i-1}}^{s_i} -\frac{\lambda(t_0, X_\sigma(\delta_i(t)))}{\lambda^0(t_0, X_\sigma(\delta_i(t)))} f(X_\sigma(\delta_i(t)), u(t))\,dt \\ &\leqq \int_{s_{i-1}}^{s_i} f^0(X_\sigma(\delta_i(t)), u(t))\,dt. \end{aligned}$$

Therefore

$$w(x(t_0 + \alpha)) - w(x(t_1 - \alpha)) \leqq \int_{t_0+\alpha}^{t_1-\alpha} f^0(x(t), u(t))\,dt.$$

Letting $\alpha \to 0$, we get

$$w(x(t_0)) - w(x(t_1)) \leqq \int_{t_0}^{t_1} f^0(x(t), u(t))\,dt.$$

We are now in a position to complete the proof of Theorem 1. Put $M = P^{n-1} \cup N$. Then M is a piecewise-smooth set of dimension $n - 1$. The set $V - M$ consists precisely of the n-dimensional cells. From Lemma 3.3, we have for any n-dimensional cell σ:

$$w_x(\gamma) = \frac{\lambda(t_0, \gamma)}{\lambda^0(t_0, \gamma)}, \qquad w \in C^{(1)}(V - M).$$

From the maximum principle the following relation obtains:

$$\begin{aligned} w_x(\gamma) f(\gamma, u) &= \frac{\lambda(t_0, \gamma)}{\lambda^0(t_0, \gamma)} \cdot f(\gamma, u) \\ &\leqq f^0(\gamma, u), \qquad \gamma \in \sigma, \quad u \in U. \end{aligned}$$

Now, from Lemma 2.1, it follows that if (x, u) is an admissible pair in V, transferring x_0 to $x_1 \in T$,

$$w(x_0) - w(x_1) \leqq \int_{t_0}^{t_1} f^0(x(t), u(t))\,dt;$$

that is,

$$w(x_0) \leqq g(x_1) + \int_{t_0}^{t_1} f^0(x(t), u(t))\,dt.$$

This proves Theorem 1.

4. Modified regular synthesis. In § 3 we showed that the existence of a regular synthesis insured the optimality of the marked trajectories. Unfortunately the concept of regular synthesis cannot be directly applied to differential games because the continuity assumptions on f^0 and f are too strong. Even the simplest examples fail to satisfy these requirements. This phenomenon will be explained more fully in § 5.

Forced to relax our continuity assumptions if we are to deal with this situation, we introduce the notion of a modified regular synthesis.

DEFINITION 4.1. Let N be a piecewise-smooth set of dimension $\leqq n - 1$ and $P^k, \cdots, P^n$ be piecewise-smooth sets, such that

$$P^k \subset P^{k+1} \subset \cdots \subset P^n = V. \tag{4.1}$$

Let T be a k-dimensional smooth manifold lying in V or on the boundary of V. In addition, let $v(x)$ be a function defined in V and taking values in U. The sets (4.1) and the function $v(x)$ effect a modified regular synthesis for (2.1) in the region V if the following conditions are satisfied:

(A) If $T \subset V$, we set $P^{k-1} = T$; if $T \subset \partial V$, we set $P^{k-1} = \varnothing$. Every component of the set $P^i - (P^{i-1} \cup N)$, $i = k, \cdots, n$, is an i-dimensional smooth manifold. Every component of $N - (N \cap T)$ is a smooth manifold in V of dimension $\leqq n - 1$. As before, we call these components i-dimensional cells, where i refers to the dimension of the manifold. The function v is continuous and continuously differentiable on each cell σ and has a continuously differentiable extension defined on a neighborhood of $\bar{\sigma}$.

(B) Same as condition (B) of regular synthesis.

(C) Corresponding to each cell σ, there exist an open set E_σ containing $\bar{\sigma} \times U$ and functions $f^0_\sigma(x, u)$, $f_\sigma(x, u)$ in $C^{(1)}(E_\sigma)$ such that

$$f^0(x, u) = f^0_\sigma(x, u), \qquad f(x, u) = f_\sigma(x, u), \qquad x \in \sigma, \qquad u \in U.$$

(D) We do not require that the trajectories issuing from points in N be unique, but for each point $x \in N$ we must be able to select one of the trajectories issuing from x so that the family of trajectories thus chosen satisfies condition (B).

(E) Same as condition (E) of regular synthesis.

(F) As in § 3 we denote the marked trajectory originating at γ by $\phi(t, \gamma)$, $t_0 \leqq t \leqq t_f(\gamma)$. Let σ be any cell of the synthesis. For each $\gamma \in \sigma$, there exists a function

$$\hat{\lambda}(t, \gamma) = (\lambda^0(t, \gamma), \lambda^1(t, \gamma), \cdots, \lambda^n(t, \gamma)),$$

absolutely continuous on

$$t_{i-1}(\delta) \leqq t \leqq t_i(\delta), \qquad i = 1, \cdots, q,$$

where $\gamma = X_\sigma(\delta)$ and t_i, $i = 1, \cdots, q$, are defined as in § 3. In addition, the following relations are satisfied:

$$\frac{d\lambda}{dt} = -H_x(\hat{\lambda}(t, \gamma), \phi(t, \gamma) \cdot v(\phi(t, \gamma))), \tag{4.2}$$

$$t_{i-1}(\delta) \leqq t \leqq t_i(\delta), \qquad i = 1, \cdots, q,$$

where $\lambda = (\lambda^1, \cdots, \lambda^n)$.

$$\lambda^0(t, \gamma) = \text{const.} < 0, \qquad t_{i-1}(\delta) \leqq t \leqq t_i(\delta) \tag{4.3}$$

for $i = 1, \cdots, q$.

$$\lambda(t_i^-(\delta), \gamma)x_{i\delta}(\delta) = \lambda(t_i^+(\delta), \gamma)x_{i\delta}(\delta) \tag{4.4}$$

for $i = 1, \cdots, q-1$.

$$H(\lambda(t,\gamma), \phi(t,\gamma), u) \leqq H(\lambda(t,\gamma), \phi(t,\gamma), v(\phi(t,\gamma))) = 0 \tag{4.5}$$

for $u \in U$, $t_0 \leqq t \leqq t_q(\delta)$. The vector

$$\lambda^0(t_q(\delta), \gamma) g_x(x_q(\delta)) - \lambda(t_q(\delta), \gamma)$$

is orthogonal to T at $x_q(\delta)$.

(G) Same as condition (G) of regular synthesis.

The existence of a modified regular synthesis implies the optimality of the marked trajectories but only in a restricted class of admissible trajectories. The inability to use Lemma 2.1 under the assumption of modified regular synthesis accounts for the weaker result. This will be explained more fully below, but first we shall present a few lemmas that parallel those of § 3.

In what follows we use the notation introduced in § 3. The next three lemmas are reformulations of Lemmas 3.1 through 3.3 for the case of a modified regular synthesis. The proofs are the same.

Lemma 4.1. *In a modified regular synthesis, the functions $t_i, x_i, i = 1, \cdots, q$, are continuously differentiable on N_σ.*

Lemma 4.2. *For any cell σ of a modified regular synthesis, $w^*(\delta) \in C^{(1)}(N_\sigma)$, where $w^*(\delta) = w(X_\sigma(\delta))$, $\delta \in N_\sigma$.*

Lemma 4.3. *For any cell σ of a modified regular synthesis and any $\delta \in N_\sigma$, the vector*

$$\left(\lambda^0(t_0, X(\delta)), \lambda(t_0, X_\sigma(\delta)) \frac{\partial X_\sigma(\delta)}{\partial \delta}\right)$$

is a constant multiple of $(1, w_\sigma^*(\delta))$.

Lemma 4.4. *Let σ be an m-dimensional cell in a modified regular synthesis, $k \leqq m \leqq n$. Then if (x, u) is an admissible pair transferring x_0 to x_1 in σ,*

$$w(x_0) - w(x_1) \leqq \int_{t_0}^{t_1} f^0(x(t), u(t))\, dt,$$

where $x(t_0) = x_0$, $x(t_1) = x_1$.

Proof. First we note that for each $\gamma \in \sigma$, $\lambda^0(t, \gamma) < 0$ for $t_0 \leqq t \leqq t_f(\gamma)$ by (4.3). The proof of Lemma 3.4 now suffices.

Definition 4.2. Let C_1 and C_2 be any pair of curves such that the final point of C_1 coincides with the initial point of C_2; then the curve C made up of the two adjacent arcs C_1 and C_2, in that order, is called the *fusion* of C_1, C_2.

Lemma 4.4 deals with curves lying entirely in a single cell. By repeated application of Lemma 4.4, we can handle curves which are finite fusions of curves, each of which lies entirely in one cell. Thus we have the following sufficiency condition.

Theorem 2. *Under the assumption of a modified regular synthesis, the marked trajectories are optimal in the class of curves obtained by the operation of finite fusion on curves each of which lies entirely in one cell.*

5. Further remarks. Earlier it was mentioned that the concept of regular synthesis was not suitable as a sufficient condition for differential games. The reason for this is that the continuity assumptions on $f^0(x, u)$ and $f(x, u)$ are too strong in a regular synthesis. In differential games, there are two control variables, say u and v. One is to be chosen to maximize J, the other to minimize. Thus in this situation f^0 and f are replaced by $h^0(x, u, v)$ and $h(x, u, v)$ and we look for a saddle point (u^*, v^*) of

$$J(u, v) = g(x(t_1)) + \int_{t_0}^{t_1} h^0(x(t), u(t), v(t))\, dt$$

subject to

$$\frac{dx}{dt} = h(x, u(t), v(t)).$$

Assuming that $v^*(x)$ is an optimal strategy, the differential game reduces to a control problem in u with

$$f^0(x, u) = h^0(x, u, v^*(x)),$$

$$f(x, u) = h(x, u, v^*(x)).$$

Even in the simplest examples v^* is discontinuous across cell boundaries. The same then is true of f^0 and f and we must relax the continuity requirements if we are to solve the differential game problem by a reduction to a control problem. This is what we have done in § 4 and this theory is applicable to differential games. See [4, Chap. 3] for a more detailed treatment of differential games.

As an alternative, one may use the repairable decomposition approach to attack this problem. The concept of a repairable decomposition was introduced by L. C. Young in [8] in relation to the time-optimal control problem. The application of this theory to differential games, which depends heavily on Theorem 2, can be found in [4].

REFERENCES

[1] L. D. Berkovitz, *Necessary conditions for optimal strategies in a class of differential games and control problems*, this Journal, 5 (1967), pp. 1–24.

[2] V. G. Boltyanskii, *Sufficient conditions for optimality and the justification of the dynamic programming method*, this Journal, 4 (1966), pp. 326–361.

[3] L. M. Graves, *Theory of Functions of Real Variables*, 2nd ed., McGraw-Hill, New York, 1956.

[4] T. G. Hack, *Sufficient conditions in the theory of optimal control and differential games*, Doctoral thesis, Applied Mathematics Dept., Purdue University, West Lafayette, Ind., 1970.

[5] S. Mirica, *On the admissible synthesis in optimal control theory and differential games*, this Journal, 7 (1969), pp. 292–316.

[6] L. S. Pontryagin, *Ordinary Differential Equations*, Addison-Wesley, Reading, Mass., 1952.

[7] L. S. Pontryagin, V. G. Boltyanskii, R. V. Gamkrelidze, and E. F. Mishchenko, *The Mathematical Theory of Optimal Processes*, Interscience, New York, 1962.

[8] L. C. Young, *Lectures on the Calculus of Variations and Optimal Control Theory*, W. B. Saunders, Philadelphia, 1969.

SIAM J. CONTROL
Vol. 11, No. 2, May 1973

KANTOROVICH ESTIMATES FOR SOME CONVEX PROGRAMS*

A. A. GOLDSTEIN† AND R. A. TAPIA‡

Abstract. If C is a closed convex subset of a Hilbert space H, and f a real-valued function defined and twice Fréchet differentiable on C, we consider whether, given a point $x_0 \in H$, it is possible to determine if a local minimizer of f in C lies nearby, and if so, to estimate its distance.

Let C be a closed convex subset of a Hilbert space H, and let f be a real-valued function which is defined and twice Fréchet differentiable on H. The problem we consider is the following:

Given a point $x_0 \in H$ is it possible to check whether a local minimizer of f in C lies nearby, and if so, to estimate its distance?

We denote by P the projection operator for C, viz. $P(x) \in C$ and $\|P(x) - x\| \leqq \|y - x\|$ for all $x \in H$ and $y \in C$. It follows that $\|P(x) - P(y)\| \leqq \|x - y\|$ for all $x, y \in H$ and $\|P'(x)\| \leqq 1$ whenever $P'(x)$ exists. In this paper, differentiable means Fréchet differentiable.

THEOREM. *Given* $x_0 \in H$ *and* $r > 0$, *suppose that* f *is twice differentiable in* $S = \{x : \|x - x_0\| \leqq r\}$ *and* P *is differentiable in* $S_\delta = \{x - \delta \nabla f(x) : x \in S]$, *where* δ *satisfies*:

(i) $0 < \delta \leqq 1/\|f''(x_0)\|$.

Suppose further that $f''(x) = f''(x, \cdot, \cdot)$ *satisfies*

(ii) $f''(x_0, h, h) \geqq \mu\|h\|^2$ *for some* $\mu > 0$, *all* $h \in H$, *and*

(iii) $\|f''(x) - f''(y)\| \leqq L_1\|x - y\|$ *for* $x, y \in S$.

Finally, assume that

(iv) $\|P'(x) - P'(y)\| \leqq L_2\|x - y\|$ *for* x, y *in* S_δ.

Let

$$K = \delta L_1 + L_2(3 + \delta r L_1)^2$$

and

$$\eta = \|x_0 - P(x_0 - \delta \nabla f(x_0))\|.$$

Then if $h = (K\eta/(\delta\mu)^2) < \frac{1}{2}$ *and* $r_0 = (\delta\mu(1 - \sqrt{1 - 2h)}/K) \leqq \eta$, *the sphere* $S_0 = \{x : \|x - x_0\| \leqq r_0\}$ *contains a local minimizer for* f *in* C *which is unique in the larger sphere* $\{x : \|x - x_0\| \leqq \min(r, \delta\mu(1 + \sqrt{1 - 2h)}/K)\}$.

Proof. By the formula $f''(x, l, m) = \langle Q(x)l, m\rangle$ we identify $f''(x)$ with a bounded linear self-adjoint operator $Q(x)$.

* Received by the editors January 27, 1971, and in revised form October 26, 1971. This research was supported by the United States Army under Contract DA-31-124-ARO-D-462 and by the Air Force Office of Scientific Research.

† Department of Mathematics, Michigan State University, East Lansing, Michigan 48823.

‡ Department of Mathematical Sciences, Rice University, Houston, Texas 77001.

By (ii) and (iii), we have for $x \in S_0$ that $(\langle Q(x)l, l\rangle/\|l\|^2) \geqq \mu - L_1\|x - x_0\| \geqq \mu - L_1 r_0 \geqq \mu\sqrt{1 - 2h} > 0$, since $h < \frac{1}{2}$ and $L_1\delta/K \leqq 1$. Let $\alpha = \mu\sqrt{1 - 2h}/2$. By continuity the set $D = \{x| \in H : (\langle Q(x)l, l\rangle/\|l\|^2) \geqq \alpha\}$ contains S_0 in its interior. If $x \in S_0$ and $x = P(x - \nabla f(x))$, then there exists $\varepsilon > 0$ such that $B = \{y \in H: \|x - y\| \leqq \varepsilon\} \subset D$. From [3, p. 122] f is strictly convex on any convex subset of D; hence on $B \cap C$. Since $x = P(x - \nabla f(x))$ and $x \in B \cap C$, the projection of $x - \nabla f(x)$ onto $B \cap C$ is also x. It follows from [3, p. 123] that x is a global minimizer of f on $B \cap C$. This means that x is a local minimizer of f on C.

Now, $x = P(x - \nabla f(x))$ if and only if $x = P(x - v\nabla f(x))$ for any $v > 0$. Let $T(x) = x - P(x - \delta\nabla f(x))$. The point x will be a local minimizer of f on C if $T(x) = 0$.

Suppose A is a linear operator from H to itself and $\|A\| < 1$. Then $(I - A)^{-1}$ exists and $\|(I - A)^{-1}\| \leqq (1 - \|A\|)^{-1}$. It follows that $T'(x_0)$ has an inverse if $\|I - \delta Q(x_0)\| \leqq \max\{|1 - \delta\mu|, |1 - \delta\|Q(x_0)\||\} < 1$, and this is so because $0 < \delta \leqq 1/\|f''(x_0)\|$. Moreover, $\|T'(x_0)^{-1}\| \leqq (\delta\mu)^{-1}$. A further computation shows that $\|T'(x) - T'(y)\| \leqq K\|x - y\|$ if $x, y \in S$, and finally $\|T(x_0)\| = \eta$. The theorem follows now by application of the Kantorovich estimate (see, for example, [3, p. 143] or [4, p. 3]).

Remark 1. If $S_\delta \subset C$ we can set $L_2 = 0$, since P restricted to C is the identity operator.

Remark 2. Newton's method can be used to find the local minimizer.

Remark 3. In the case of unconstrained minimization, i.e., $C = H$, we have $P = I$ and $L_2 = 0$; hence the conditions of the theorem reduce to

$$h = L_1\|\nabla f(x_0)\|/\mu^2 < \tfrac{1}{2} \quad \text{and} \quad r_0 = \mu(1 - \sqrt{1 - 2h})/L_2 \leqq r.$$

It is interesting, but not surprising, that due to cancellation we do not have to specify a value for δ. This special case of the theorem has been obtained independently by J. E. Dennis Jr. (see [2, p. 450)].

Remark 4. The reader familiar with the statement of Kantorovich's theorem will obviously question the strict inequality $(h < \frac{1}{2})$ in the above theorem. The following example shows that the theorem is not true if $h = \frac{1}{2}$. Let $C = H = R^1$ and $f(x) = x^3$. Choose $x_0 = r = \frac{1}{2}$, $\mu = 3$ and $L_1 = 6$. It follows that $\eta = \frac{3}{4}$; hence $h = \frac{1}{2}$ and $r_0 = r$. However, f has no local minima. Clearly the theorem fails because we can no longer guarantee, when x is on the boundary of S_0, that f is convex on an open set containing x.

Remark 5. Let x_0 be a local minimizer of f in C and assume the hypotheses of the theorem above at x_0. Then if $\hat{x}$ is sufficiently close to x_0, the conditions of the theorem will hold at $\hat{x}$.

Proof. Observe that for some neighborhood containing x_0, K will be bounded above and $\delta\mu$ will be bounded below. Also η can be made arbitrarily small by choosing $\hat{x}$ in this neighborhood sufficiently close to x_0.

Remark 6. Assume that x_0 is a local minimizer of f on the unit sphere of H. Assume that f satisfies the same conditions as the theorem, namely f'' exists on S and (ii) and (iii) hold. Assume that $\|x_0\| = 1$ and $\nabla f(x_0) = 0$ do not hold simultaneously. Then if $\hat{x}$ is sufficiently close to x_0 the hypotheses of the theorem can be fulfilled at $\hat{x}$.

Proof. Clearly $P(x) = x/\max(1, \|x\|)$. The operator P is not differentiable at x if $\|x\| = 1$. However, if $\|x\| < 1$, then $P'(x, h) = h$ and if $\|x\| > 1$, then

$$P'(x, h) = \frac{-1}{\|x\|^3}(x\langle x, h\rangle - h\langle x, x\rangle)$$

and

$$P''(x, h, k) = \frac{-1}{\|x\|^5}[\langle x, x\rangle(x\langle h, k\rangle + h\langle x, k\rangle + k\langle x, h\rangle) - 3x\langle x, h\rangle\langle x, k\rangle].$$

If follows that $\|P''(x)\| \leqq 6$ if $\|x\| \neq 1$.

Let $Q(\delta) = \|x_0 - \delta\nabla f(x_0)\|^2 - 1 = \delta^2\|\nabla f(x_0)\|^2 - 2\delta\langle x_0, \nabla f(x_0)\rangle + (\|x_0\|^2 - 1)$. Clearly $Q(\delta) = 0$ for at most two values of δ, unless of course both $\|x_0\| = 1$ and $\nabla f(x_0) = 0$, but this case has been excluded. We may therefore pick $r, \delta > 0$, such that $\|x - \delta\nabla f(x)\| \neq 1$ and $\delta \leqq 1/\|f''(x)\|$ whenever $\|x - x_0\| < r$. The remark now follows from an argument similar to the one used in the previous remark.

Remark 7. In constrained minimization it is usual for a local minimizer to lie on the boundary of the convex set C and for the gradient not to vanish there. As the previous remark implies, projection operators are seldom differentiable on the boundary. Hence the important rôle δ plays in the theorem is to move the set S_δ away from the boundary of C. We now illustrate these points and the theorem with an example.

Example. As before let $H = R^1$ and $f(x) = x^3$. We are interested in minimizing f on the closed convex set $C = [1, +\infty)$. Let $r = +\infty$, $\mu = |f''(x_0)|$, $\delta = \mu^{-1}$ and $L_1 = 6$. The projection operator is differentiable everywhere except $x = 1$; also we may choose $L_2 = 0$. Now, since $\delta^{-1} = 6x_0$ we have that $S_\delta = \{y : y = x - x^2/(2x_0), -\infty < x < +\infty\}$ and S_δ is strictly outside of $[1, +\infty)$ if $x_0 < 2$. It is also not difficult to show that $h = |1 - x_0|/|x_0|$, $r_0 = |x_0|(1 - \sqrt{1 - 2h})$ and $h < \frac{1}{2}$ if $\frac{2}{3} < x_0 < 2$. Hence the conditions of the theorem are satisfied for any x_0 in $(\frac{2}{3}, 2)$ and we can guarantee the existence of a local minimizer x^* of f on C, satisfying

$$x_0\sqrt{1 - 2h} \leqq x^* \leqq x_0(2 - \sqrt{1 - 2h}).$$

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

REFERENCES

[1] A. V. BALAKRISHNAN, *An operator theoretic formulation of a class of control problems and a steepest descent method of solution*, this Journal, 1 (1962), pp. 109–127.

[2] J. E. DENNIS, JR., *Toward a unified convergence theory for Newton-like methods*, Nonlinear Functional Analysis and Applications, L. B. Rall, ed., Academic Press, New York, 1971, pp. 425–472.

[3] A. A. GOLDSTEIN, *Constructive Real Analysis*, Harper and Row, New York, 1967.

[4] L. B. RALL AND R. A. TAPIA, *The Kantorovich theorem and error estimates for Newton's method*, MRC Rep. 1043, Mathematics Research Center, University of Wisconsin, Madison, 1970.

SIAM J. CONTROL
Vol. 11, No. 2, May 1973

NEW CONDITIONS FOR EXACTNESS OF A SIMPLE PENALTY FUNCTION*

STEPHEN HOWE†

Abstract. We consider penalty function methods for finding the maximum of a function f over the set

$$S = \{x \in R^n : g_i(x) \leqq 0 \text{ for } i = 1, \cdots, m \text{ and } h_j(x) = 0 \text{ for } j = 1, \cdots, p\}.$$

New conditions, extending earlier work done by Pietrzykowski, are presented under which the penalty function

$$P(x, \mu) = \mu f(x) - \sum_{i=1}^{m} g_i^+(x) - \sum_{j=1}^{p} |h_j(x)|$$

is locally exact. The relationships among the new conditions, Pietrzykowski's conditions, and Kuhn–Tucker constraint qualifications are explored.

1. Introduction. Consider the constrained maximization problem with objective function f and constraint set

$$S = \{x \in R^n : g_i(x) \leqq 0, i = 1, \cdots, m \text{ and } h_j(x) = 0, j = 1, \cdots, p\},$$

where $f, g_1, \cdots, g_m, h_1, \cdots, h_p$ are continuously differentiable real-valued functions on R^n. Much recent work has been devoted to the reduction of this problem to a single unconstrained maximization. Several advances in this direction have been made using the penalty function defined, for $\mu > 0$ and $x \in R^n$, by

$$P(x, \mu) = \mu f(x) - \sum_{i=1}^{m} g_i^+(x) - \sum_{j=1}^{p} |h_j(x)|,$$

where

$$g_i^+(x) = \begin{cases} g_i(x) & \text{if } g_i(x) > 0, \\ 0 & \text{otherwise.} \end{cases}$$

In 1967, Zangwill [4] presented an important result for the constrained problem having inequality constraints only: if this problem is convex (i.e., f is concave and each g_i is convex), if it satisfies Slater's condition (for some $x \in R^n$, $g_i(x) < 0$ each i), and if the problem has a solution x^*, then for any μ sufficiently small $P(\cdot, \mu)$ has a global unconstrained maximum at x^*. In this case we say that the function P is *globally exact*. Note in Zangwill's result the use of Slater's condition, which is a constraint qualification validating the Kuhn–Tucker necessary optimality criterion for convex, inequality-constrained problems.

In the general case, where the problem also has equality constraints and is no longer assumed to be convex, one is interested in the possible existence of local constrained maxima. In 1969, Pietrzykowski [2], [3] proved a theorem, similar

* Received by the editors December 21, 1971, and in revised form May 23, 1972.

† Department of Statistics, University of North Carolina, Chapel Hill, North Carolina. Now at Department of Mathematics, Rensselaer Polytechnic Institute, Troy, New York 12181. This work was supported in part by the Office of Naval Research under Grant N000-14-67-A-0321-0003 (NR 047-096).

to Zangwill's result, for such points: If f has an isolated (strong) local maximum x_0 on S such that the set of "active gradients"

$$\{\nabla h_j(x_0), j = 1, \cdots, p \text{ and } \nabla g_i(x_0) \text{ for each } i, \text{ where } g_i(x_0) = 0\}$$

is linearly independent, then for μ sufficiently small the function $P(\cdot, \mu)$ has an unconstrained isolated local maximum at x_0. In this context we say that P is *locally exact*. Again, note in this theorem the presence of a constraint qualification (linearly independent active gradients) validating the Kuhn–Tucker criterion.

In this paper a new condition guaranteeing local exactness of the function P is given. The new condition, unlike that of Pietrzykowski, depends upon the objective function f. The differences between these conditions are further illustrated in the first two examples of § 3. This paper also deals with the relationship, suggested by the results of Zangwill and Pietrzykowski, between constraint qualifications and exactness of the penalty function. The last example of § 3 discounts this suggested relationship.

2. Result. Let the general constrained maximization problem, and the penalty function $P(x, \mu)$, be defined as in the Introduction.

THEOREM. *For some point* $x_0 \in S$, *let* $I = \{i : g_i(x_0) = 0\}$, *and suppose the following condition is satisfied*:

G: *For every nonzero* $y \in R^n$, *such that* $\nabla g_i(x_0)^{\mathrm{T}} y \leqq 0$ *for each* $i \in I$ *and* $\nabla h_j(x_0)^{\mathrm{T}} y = 0$ *for each* $j = 1, \cdots, p$, *it follows that* $\nabla f(x_0)^{\mathrm{T}} y < 0$.

Then there is a $\mu_0 > 0$ *such that if* $0 < \mu \leqq \mu_0$ *then* $P(x, \mu)$ *has an unconstrained isolated local maximum at* x_0.

Remark. It can be shown that, under these conditions, x_0 is an isolated local maximum of f on S.

Proof. Since $g_1, \cdots, g_m$ are continuous, there is a $\delta > 0$ such that $g_i(x) < 0$ for every $x \in B(x_0, \delta)$ and $i \notin I$. Here $B(x_0, \delta)$ is the open ball of radius δ about x_0. Now suppose, to get a contradiction, that there is a sequence $\{\mu_n\}$ of positive scalars such that $\mu_n \to 0$ as $n \to \infty$ and where $P(x, \mu_n)$ does not have an unconstrained isolated local maximum at x_0 for each n. Then for each n there is a point x_n such that $0 < \|x_n - x_0\| < \min\{\delta, \mu_n\}$ but for which $P(x_n, \mu_n) \geqq P(x_0, \mu_n)$. Denote $y_n = x_n - x_0$ for each n, and note that $y_n \to 0$ as $n \to \infty$. Consider the sequence $\{y_n/\|y_n\|\}$ of vectors in the compact set $\{z \in R^n : \|z\| = 1\}$. This sequence must have a subsequential limit $y_0 \in R^n$ such that $\|y_0\| = 1$. Assume without loss of generality that $y_n/\|y_n\| \to y_0$ as $n \to \infty$.

We shall now show that $\nabla g_i(x_0)^{\mathrm{T}} y_0 \leqq 0$ for each $i \in I$. Suppose, conversely, that $\nabla g_r(x_0)^{\mathrm{T}} y_0 > 0$ for some $r \in I$. Then for each n,

$$\begin{aligned}\frac{1}{\|y_n\|} g_r(x_n) &= \frac{1}{\|y_n\|}[g_r(x_n) - g_r(x_0)] \\ &= \nabla g_r(x_0)^{\mathrm{T}} \frac{y_n}{\|y_n\|} + \frac{o(\|y_n\|)}{\|y_n\|}.\end{aligned}$$

This last expression converges to $\nabla g_r(x_0)^{\mathrm{T}} y_0$ as $n \to \infty$, thus for n sufficiently large $g_r(x_n) > 0$ so that $g_r^+(x_n) = g_r(x_n)$. But then it follows that, for n sufficiently large,

$$P(x_n, \mu_n) - P(x_0, \mu_n) \leqq \mu_n[f(x_n) - f(x_0)] - g_r(x_n),$$

or, equivalently,

$$\frac{1}{\|y_n\|}[P(x_n, \mu_n) - P(x_0, \mu_n)] \leqq \mu_n \nabla f(x_0)^{\mathrm{T}} \frac{y_n}{\|y_n\|} - \nabla g_r(x_0)^{\mathrm{T}} \frac{y_n}{\|y_n\|} + \frac{o(\|y_n\|)}{\|y_n\|}.$$

But the latter right-hand expression converges to $-\nabla g_r(x_0)^{\mathrm{T}} y_0 < 0$ as $n \to \infty$, implying that for large n, $P(x_n, \mu_n) < P(x_0, \mu_n)$. This contradiction establishes the claim that $\nabla g_i(x_0)^{\mathrm{T}} y_0 \leqq 0$ for each $i \in I$.

Similarly, it can be shown that $\nabla h_j(x_0)^{\mathrm{T}} y_0 = 0$ for $j = 1, \cdots, p$. Hence, from the assumption that G holds, it follows that $\nabla f(x_0)^{\mathrm{T}} y_0 < 0$. But then

$$\frac{1}{\|y_n\|}[f(x_n) - f(x_0)] = \nabla f(x_0)^{\mathrm{T}} \frac{y_n}{\|y_n\|} + \frac{o(\|y_n\|)}{\|y_n\|}$$

$$\to \nabla f(x_0)^{\mathrm{T}} y_0 < 0 \quad \text{as } n \to \infty,$$

implying that for n sufficiently large $f(x_n) < f(x_0)$ and hence $P(x_n, \mu_n) < P(x_0, \mu_n)$. This final contradiction establishes the theorem.

3. Examples.

Example 1. Let $f, g_1, g_2 : R \to R$ be defined by $f(x) = x, g_1(x) = x, g_2(x) = -x$. Both constraints are active at $x_0 = 0$, which is the only feasible point and hence the constrained maximum. Since there is no nonzero $y \in R$ such that $yg_i'(0) \leqq 0$ for $i = 1, 2$, G holds trivially, but the gradients $g_1'(0) = 1$ and $g_2'(0) = -1$ are not linearly independent. Thus the new condition G may hold when Pietrzykowski's condition does not.

Example 2. Let $f, g : R \to R$ be defined by $f(x) = x^3$, $g(x) = x$. The constraint is active at the constrained maximum $x_0 = 0$. Note that $g'(0) = 1$, and that for $y = -1$, $yg'(0) = -1 \leqq 0$ but $yf'(0) = 0$. Therefore, Pietrzykowski's condition does not imply G.

Example 3. Let $f, g_1, g_2 : R^2 \to R$ be defined by

$$f(x, y) = y + x^4,$$
$$g_1(x, y) = y,$$
$$g_2(x, y) = y^3 + x^6,$$

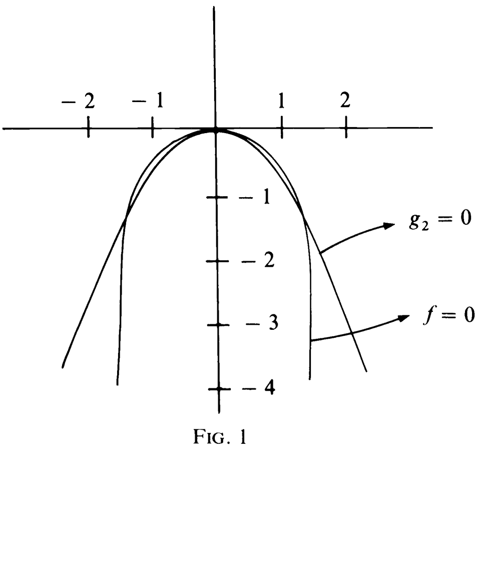


FIG. 1

(see Fig. 1). Consider the weak constraint qualification, shown by Gould and Tolle [4] to be both necessary and sufficient for validation of the Kuhn–Tucker criterion. It can be verified that this constraint qualification holds at the isolated local constrained maximum $(0, 0)$. It is not true, however, that for μ sufficiently small the point $(0, 0)$ is a local constrained maximum of $P((x, y), \mu)$. To see this, define $h: R \to R$ by $h(x) = P((x, 0), \mu) = \mu x^4 - x^6$ for any $\mu > 0$, noting that h has a local *minimum* at $x = 0$ for any $\mu > 0$. Hence, for an isolated local maximum x_0, it is not true that any Kuhn–Tucker constraint qualification will guarantee local exactness of the penalty function $P(x, \mu)$.

REFERENCES

[1] F. J. GOULD AND JON W. TOLLE, *A necessary and sufficient constraint qualification for constrained optimization*, SIAM J. Appl. Math., 20 (1971), pp. 164–172.

[2] T. PIETRZYKOWSKI, *The potential method for conditional maxima in the locally compact metric spaces*, Numer. Math., 14 (1970), pp. 325–329.

[3] ———, *An exact potential method for constrained maxima*, SIAM J. Numer. Anal., 6 (1969), pp. 299–304.

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

SIAM J. CONTROL
Vol. 11, No. 2, May 1973

STABILITY OF LINEAR SYSTEMS WITH CONTROL DEPENDENT NOISE*

U. G. HAUSSMANN†

Abstract. Consider a system described by the autonomous stochastic differential equation

$$dx = (Ax - Bu)\,dt + C(u)\,dw_1 + D\,dw_2,$$

where $C(\cdot)$ is linear in the control variable u and w_1, w_2 are two independent Wiener processes. It is known that if the pair (A, B) is stabilizable and if $C(\cdot)$ is sufficiently small, then there exists a control $u = \phi(x)$ such that the corresponding diffusion process $x(\cdot)$ possesses an invariant probability measure with finite second moment, and hence there exists a control which minimizes the expected value with respect to the invariant measure of a quadratic cost functional. In the present work necessary and sufficient conditions on the structure of the system are given such that, without requiring $C(\cdot)$ to be small, a control exists which induces an invariant probability measure with finite second moment.

1. Introduction.

Consider a system described by the stochastic differential equation

$$\dot{x} = Ax - Bu - C(u)\dot{w}_1 + D\dot{w}_2, \tag{1.1}$$

where x is the n-dimensional state vector, u the m-dimensional control vector, $\dot{w}_1, \dot{w}_2$ are independent Gaussian white noise processes and C is linear in u. Equation (1.1) could be rewritten as

$$\dot{x} = Ax - [B + \bar{C}(\dot{w}_1)]u + D\dot{w}_2 \tag{1.2}$$

so that C can represent linear random disturbances in the control matrix B. Wonham has shown [1] that if the pair (A, B) is stabilizable and if C is sufficiently small, then a linear feedback control u exists for which the corresponding process $x(\cdot)$ possesses an invariant probability distribution with finite second moment. We shall say that the control u stabilizes x. In this case a control can be found which minimizes the long-run average (i.e., the expectation with respect to the invariant measure) of a quadratic cost functional [1], [2], [9].

In some recent work [2], the author proved that under certain conditions on the structure of the system, C did not have to be small in order for a stabilizing control (hence an optimal control) to exist. Now the "best" conditions have been derived. Loosely speaking they are the following. Assume that the system (1.1) can be decomposed into a sequence of coupled subsystems, and that for each subsystem a stabilizing control which does not produce any noise in that subsystem exists if the interactions are neglected. Then a linear feedback control u can be found which stabilizes the whole system no matter how large C is. This is proved in § 3, after the problem has been precisely stated in § 2.

* Received by the editors February 10, 1972, and in revised form June 16, 1972.

† Department of Mathematics, University of British Columbia, Vancouver 8, British Columbia, Canada. This work was supported by the National Research Council of Canada under Grant A8051.

The converse is established in § 4. If in the above procedure one of the subsystems cannot be stabilized without producing noise in that subsystem, then C can be chosen so large that no control can stabilize the system, provided A has no purely imaginary eigenvalues and DD' is positive definite. D' denotes the transpose of D. The basic tool in the proof of the converse is an instability theorem due to Wonham [10]. His proof employs the partial differential equations approach. It is interesting to observe that a more direct proof using probabilistic arguments can also be given.

Two examples in §5 conclude the article. It should be mentioned that Krasovskii has considered a related, but simpler problem [3].

2. The stabilization problem.

Let us first state the problem precisely. Consider the stochastic differential equation

$$dx = Ax\,dt - Bu\,dt - C(u)\,dw_1 + D\,dw_2, \qquad t \geqq 0, \tag{2.1}$$

where x is a vector in R^n, Euclidean n-space with norm $\|x\| = \{\sum_{i=1}^n |x_i|^2\}^{1/2}$, $n \geqq 2$, u is in R^m, and w_1, w_2 are independent Wiener processes of dimension d and d_0 respectively. $C(u)$ is given by

$$C(u) = \sum_{i=1}^{m} C^i u_i, \tag{2.2}$$

and A, B, C^i, D are real constant matrices of corresponding dimensions.

If in (2.1) u has the form $u = \phi(x)$ with

$$\|\phi(x) - \phi(y)\| \leqq k\|x - y\|, \qquad x, y \in R^n, \tag{2.3}$$

then (2.1).is an equation of Itô's type, so that if the random variable $x(0)$ is independent of the increments of w_1, w_2, then (2.1) determines a diffusion process (see [4])

$$X_\phi = \{x(t): t \geqq 0\}. \tag{2.4}$$

We shall say that μ, a probability measure on the Borel sets of R^n, is *invariant* provided that whenever $x(0)$ has distribution μ then $x(t)$ has distribution μ for every $t > 0$. Moreover a diffusion process is *stable* if at least one invariant probability measure exists, and for any invariant measure μ,

$$\mathscr{E}_\mu\{\|x\|^2\} \equiv \int_{R^n} \|x\|^2 \mu(dx) < \infty. \tag{2.5}$$

Finally, the control $u = \phi(x)$ *stabilizes* the system (2.1) if ϕ satisfies (2.3) and the corresponding process (2.4) is stable.

The problem is to find conditions under which stabilizing controls exist. The significance of the problem lies in the fact that in [1] it is proved that a control which will minimize a quadratic cost functional in the steady state can be found, provided a stabilizing control exists.

Let $\mathscr{L}_u$ be the differential operator given by

$$\mathscr{L}_u V(x) = \tfrac{1}{2}\operatorname{tr}\{C(u)'V_{xx}C(u) + D'V_{xx}D\} + (Ax - Bu)'V_x, \tag{2.6}$$

where tr A is the trace of A, V_x denotes the vector $\partial V/\partial x$ and V_{xx} denotes the matrix $\partial^2 V/\partial x^2$. If we assume $u = \phi(x)$ where ϕ satisfies (2.3), then $\mathscr{L}_u \equiv \mathscr{L}_\phi$ is the differential generator of X_ϕ. The next theorem is basic.

THEOREM 2.1. *If there exist two real constant matrices K and P with $P \geqq 0$* (i.e., P is symmetric and $x'Px \geqq 0$ for all $x \in R^n$) *and positive scalars λ and ρ, such that*

$$(2.7) \qquad \mathscr{L}_\phi V(x) \leqq \lambda - \rho\|x\|^2, \qquad x \in R^n,$$

where $V(x) = x'Px$ and $\phi(x) = Kx$, then ϕ stabilizes (2.1).

Proof. The result follows directly from Theorems 1 and 2 of [8].

In [2] we proved that if the range of any matrix C^i is contained in the "stable" part of R^n, then (2.1) can be stabilized no matter how large C^i is. In the next section we determine the weakest conditions which C^i must satisfy in order that the same conclusions hold.

3. Stabilization with arbitrary noise intensity. For any $n \times n$ matrix F let $\alpha(\lambda; F)$ be the minimal polynomial of F. It factors into

$$\alpha(\lambda; F) = \alpha_+(\lambda; F)\alpha_-(\lambda; F),$$

where all the zeros of $\alpha_+(\lambda; F)$ lie in the closed right half-plane, and those of $\alpha_-(\lambda; F)$ in the open left half-plane. For any $m \times n$ matrix K, we set

$$E_\pm(K) = \{x \in R^n : \alpha_\pm(A - BK; A - BK)x = 0\}.$$

As α_+ and α_- are coprime it follows that $R^n = E_+(K) \oplus E_-(K)$, although the direct sum is in general not orthogonal. $E_-(0)$ and $E_+(0)$ correspond to the asymptotically stable and unstable modes of A respectively. Let $T_\pm$ be the projection of R^n onto $E_\pm(0)$ along $E_\mp(0)$. Then $T_\pm A = AT_\pm \equiv A_\pm$. It follows that A_- restricted to $E_-(0)$ is stable, i.e., has all its eigenvalues in the open left half-plane.

Suppose we set $y = T_+x$, $z = T_-x$, $T_\pm C(u) = C_\pm(u)$, $T_\pm B = B_\pm$, $T_\pm D = D_\pm$; then (2.1) splits into

$$(3.1) \qquad dy = (A_+y - B_+u)\,dt - C_+(u)\,dw_1 + D_+\,dw_2,$$

$$(3.2) \qquad dz = (A_-z - B_-u)\,dt - C_-(u)\,dw_1 + D_-\,dw_2.$$

This leads us to the next result. Let us write $\mathscr{L}_u^+$ and $\mathscr{L}_u^-$ for the differential operators associated with (3.1) and (3.2) respectively, cf. (2.6).

LEMMA 3.1. *If there exist matrices $P_+ \geqq 0$ and K_+ and positive constants λ_+, ρ_+ such that with $u = K_+y$,*

$$(3.3) \qquad \mathscr{L}_u^+(y'P_+y) \leqq \lambda_+ - \rho_+\|y\|^2, \qquad y \in E_+(0),$$

then there exist P, K, ρ, λ as required in Theorem 2.1, *so that* (2.1) *can be stabilized.*

Proof. The proof is very similar to that of [2, Theorem 4.1]. As A_- is stable on $E_-(0)$, it follows that there is a matrix $P_- \geqq 0$ and a positive constant λ_-, such that with $u = 0$,

$$(3.4) \qquad \mathscr{L}_u^-(z'P_-z) \leqq \lambda_- - \|z\|^2, \qquad z \in E_-(0);$$

see [2]. If we set $V(x) = (y'P_+y + \beta z'P_-z)$ with $\beta > 0$, then for β sufficiently small and $u = K_+T_+x$, it follows that

$$\mathscr{L}_uV(x) \leqq \lambda_+ + \beta\lambda_- - \rho\|x\|^2$$

for some $\rho > 0$; see [2, Theorem 4.1]. Now Theorem 2.1 gives the result.

We write span $\{A\}$ for the span of the columns of A. Let us define

$$R_B[A] = \operatorname{span}\{B, AB, A^2B, \cdots, A^{n-1}B\}.$$

We now have the following theorem.

THEOREM 3.2. *If A is an $n \times n$ matrix with no eigenvalues with negative real part, and if B is an $n \times m$ matrix, then:*

(i) *for any $m \times n$ matrix K, $E_-(K) \subseteq R_B[A]$;*

(ii) *there exists a matrix K_0 such that $E_-(K_0) = R_B[A]$.*

Proof. Since $R_B[A]$ is invariant under A and under $A - BK$, then there exist induced maps $\bar{A}$ and $\overline{A - BK}$ on $R^n/R_B[A]$. Moreover $\bar{A} = \overline{A - BK}$. Let $\overline{E_-(K)}$ be the set in $R^n/R_B[A]$ induced by $E_-(K)$. Then

$$\begin{aligned} \overline{E_-(K)} &= \{\bar{x} : \alpha_-(\overline{A - BK}; \overline{A - BK})\bar{x} = \bar{0}\} \\ &= \{\bar{x} : \alpha_-(\bar{A};\bar{A})\bar{x} = \bar{0}\} \qquad (3.5) \\ &= \overline{E_-(0)} \\ &= \bar{0}. \end{aligned}$$

Hence

$$E_-(K) \subseteq R_B[A].$$

To prove (ii) we observe that by controllability on $R_B[A]$, there exists K_0 such that $(A - BK_0)$ on $R_B[A]$ is stable. Hence $E_-(K_0) \supseteq R_B[A]$ and the result follows.

We now give a decomposition of R^n which will yield the stability result. At each stage of the decomposition we look for controls in the kernel of the noise matrix. We make the following change in notation:

$$y^1 = x, \quad u^1 = u, \quad A^1 = A, \quad B^1 = B, \quad C^1(u) = C(u),$$
$$D^1 = D, \quad E^1_\pm = E_\pm(0), \quad m_1 = m.$$

Our equation is then

$$dy^1 = A^1y^1\,dt - B^1u^1\,dt - C^1(u^1)\,dw_1 + D^1\,dw_2. \qquad (3.6)$$

With $T^1_\pm = T_\pm$, $y^1_\pm = T^1_\pm y^1$, equation (3.6) becomes

$$dy^1_- = A^1_-y^1_-\,dt - T^1_-B^1u^1\,dt - T^1_-C^1(u^1)\,dw_1 + T^1_-D^1\,dw_2, \qquad (3.7)$$

$$dy^1_+ = A^1_+y^1_+\,dt - T^1_+B^1u^1\,dt - T^1_+C^1(u^1)\,dw_1 + T^1_+D^1\,dw_2. \qquad (3.8)$$

Set $C^1_+(u^1) = T^1_+C^1(u^1)$ for $u^1 \in R^{m_1}$, and let $\mathcal{N}^1$, the kernel of C^1_+, have dimension q_1. With $m_2 = m_1 - q_1$ we define the map $\bar{S}^1$ as follows: $\bar{S}^1$ maps R^{m_1} onto R^{q_1} such that it is an isomorphism between $\mathcal{N}^1$ and R^{q_1}, with kernel $\mathcal{N}^1_\perp$, the orthogonal complement of $\mathcal{N}^1$. Let S^1 be the generalized inverse of $\bar{S}^1$. Then $B^1_+ \equiv T^1_+B^1S^1$ maps R^{q_1} into E^1_+, and we have the following result.

LEMMA 3.3. *The largest* (*over all matrices* K^1 *mapping* E^1_+ *into* $\mathcal{N}^1$) *subset of* E^1_+ *consisting only of stable modes of* $(A^1_+ - T^1_+B^1K^1)$ *is the subspace* $R_{B^1_+}[A^1_+]$.

Proof. S^1 maps R^{q_1} onto $\mathcal{N}^1$, so we may consider the matrix $A^1_+ - T^1_+B^1S^1K$ where K maps E^1_+ into R^{q_1}. But this matrix is $A^1_+ - B^1_+K$. Now Theorem 3.2 yields that for some K_0 the set of stable modes is $R_{B^1_+}[A^1_+]$, and no other K can do better. Hence the best K^1 is S^1K_0.

LEMMA 3.4. *If* $R_{B^1_+}[A^1_+] = E^1_+$, *then there exists a control* u *which stabilizes* (2.1) *no matter what* C *is.*

Proof. By Lemma 3.1, we need only consider (3.8) but not (3.7). If we restrict ourself to controls $u^1 \in \mathcal{N}^1$, (3.8) becomes

$$dy^1_+ = A^1_+y_1\,dt - B^1_+v\,dt + T^1_+D^1\,dw_2, \qquad v \in R^{q_1},$$

and (A^1_+, B^1_+) is controllable. It follows that $P_+, K_+, \lambda_+, \rho_+$ satisfying (3.3) with equality exist. Hence the result follows by Lemma 3.1.

Lemma 3.3 tells us that this is the best we can hope to do using the present decomposition. The other extreme is that $R_{B^1_+}[A^1_+]$ is $\{0\}$. In that case we shall show later that instability results for suitable C.

If $R_{B^1_+}[A^1_+]$ lies between these two extremes we can continue the decomposition. However, let us now proceed in general giving a definition by induction.

$$(3.9) \qquad dy^i = A^iy^i\,dt - B^iu^i\,dt - C^i(u^i)\,dw_1 + D^i\,dw_2, \qquad u^i \in R^{m_i}.$$

Set $E^i_\pm = \{x \in E^{i-1}_+ : \alpha_\pm(A^i; A^i)x = 0\}$ and let $T^i_\pm$ be the projection of E^{i-1}_+ onto $E^i_\pm$ along $E^i_\mp$. With $y^i_\pm = T^i_\pm y^i$, $T^i_\pm A^i = A^iT^i_\pm = A^i_\pm$ we have from (3.9),

$$(3.10) \qquad dy^i_- = A^i_-y^i_-\,dt - T^i_-B^iu^i\,dt - T^i_-C^i(u^i)\,dw_1 + T^i_-D^i\,dw_2,$$

$$(3.11) \qquad dy^i_+ = A^i_+y^i_+\,dt - T^i_+B^iu^i\,dt - T^i_+C^i(u^i)\,dw_1 + T^i_+D^i\,dw_2.$$

Set $C^i_+(u^i) = T^i_+C^i(u^i)$, and let $\mathcal{N}^i$, the kernel of C^i_+, have dimension q_i. Then $m_{i+1} \equiv m_i - q_i$. Define $\bar{S}^i, \bar{S}^i_\perp$ to map R^{m_i} onto $R^{q_i}, R^{m_{i+1}}$, respectively, such that $\bar{S}^i$ is an isomorphism between $\mathcal{N}^i$ and R^{q_i}, $\bar{S}^i_\perp$ is one between $\mathcal{N}^i_\perp$, the orthogonal complement of $\mathcal{N}^i$, and $R^{m_{i+1}}$, and they have kernel $\mathcal{N}^i_\perp, \mathcal{N}^i$ respectively. Let $S^i, S^i_\perp$ be the generalized inverse of $\bar{S}^i, \bar{S}^i_\perp$ respectively. Then $B^i_+ \equiv T^i_+B^iS^i$ maps R^{q_i} into E^i_+ and we have as in Lemma 3.3, that $R_{B^i_+}[A^i_+]$ is the largest set of modes in E^i_+ which can be stabilized without introducing any noise through C^i_+. By Theorem 3.2 there exists K^i mapping E^i_+ into R^{q_i} such that the stable modes of $(A^i_+ - B^i_+K^i)$ are $R_{B^i_+}[A^i_+]$. Then with $u^i = S^iK^iy^i_+ + S^i_\perp u^{i+1}$, $u^{i+1} \in R^{m_{i+1}}$, $A^{i+1} = A^i_+ - B^i_+K^i$, $B^{i+1} = T^i_+B^iS^i_\perp$, $C^{i+1}(u^{i+1}) = C^i_+(S^i_\perp u^{i+1})$, $D^{i+1} = T^i_+D^i$, $y^{i+1} = y^i_+$, (3.11) becomes

$$(3.12) \quad dy^{i+1} = A^{i+1}y^{i+1}\,dt - B^{i+1}u^{i+1}\,dt - C^{i+1}(u^{i+1})\,dw_1 + D^{i+1}\,dw_2.$$

Finally we define the following sets:

$$\mathcal{M}(\mathcal{V}) = \{Bu : u \in R^m, \text{span}\,\{C(u)\} \subseteq \mathcal{V}\}$$
$$= \{Bu : u \in R^m, \bar{C}_ju \in \mathcal{V}, j = 1, 2, \cdots, d\},$$

where $(\bar{C}_j)_{lm} = C^m_{lj}$. Then

$$\mathcal{M}(\mathcal{V}) = B \bigcap_{j=1}^{d} \bar{C}_j^{-1}(\mathcal{V});$$

i.e., the image under B of the intersection of the inverse images under $\bar{C}_j$ of $\mathscr{V}$. Let $\mathscr{R}_0 = E_-^1$, the stable modes of A, and let $\mathscr{R}_i = \operatorname{span}\{\mathscr{R}_{i-1}, R_x[A] : x \in \mathscr{M}(\mathscr{R}_{i-1}\}$, $i > 0$. It follows that the $\mathscr{R}_k$ are an increasing sequence of subspaces of $\mathscr{R}^n$.

LEMMA 3.5. *If $\mathscr{R}_k = \mathscr{R}_{k-1}$, then $\mathscr{R}_{k+1} = \mathscr{R}_{k-1}$. $\mathscr{R}_{k+1} = \mathscr{R}_k$ for $k \geqq$* rank B.

Proof. If $\mathscr{R}_k = \mathscr{R}_{k-1}$, then $\mathscr{R}_{k+1} = \operatorname{span}\{\mathscr{R}_{k-1}, R_x[A] : x \in \mathscr{M}(\mathscr{R}_{k-1})\} = \mathscr{R}_k$. Hence the first conclusion follows.

If $\mathscr{R}_{i+1} \neq \mathscr{R}_i$, then $\mathscr{R}_{l+1} \supset \mathscr{R}_i$ and $\dim \mathscr{R}_{i+1} \geqq 1 + \dim \mathscr{R}_i$. This increase in dimension is due to an increase in the dimension of $\mathscr{M}(\mathscr{R}_i)$, i.e., $\dim \mathscr{M}(\mathscr{R}_{i+1}) \geqq 1 + \dim \mathscr{M}(\mathscr{R}_i)$. As $\mathscr{M}(\mathscr{R}_i) \subseteq \operatorname{span}\{B\}$, then we can increase the dimension at most rank B times. The result follows.

LEMMA 3.6.

$$\mathscr{R}_i = \operatorname{span}\{E_-^j : j = 1, 2, \cdots, i+1\}, \qquad i \geqq 0.$$

Proof. Certainly $\mathscr{R}_0 = E_-^1$. Assume the result holds for $i = k-1$. As

$$\begin{aligned}\{u : \operatorname{span}\{C(u)\} \subseteq \mathscr{R}_{k-1}\} &= \{u : T_+^k T_+^{k-1} \cdots T_+^1 C(u) = 0\} \\ &= \operatorname{span}\{\mathscr{N}^1, S_\perp^1 \mathscr{N}^2, \quad \cdots, S_\perp^1 S_\perp^2 \cdots S_\perp^{k-1} \mathscr{N}^k\},\end{aligned}$$

then

$$\begin{aligned}\operatorname{span}\{R_x[A] : x \in \mathscr{M}(\mathscr{R}_{k-1})\} &= \operatorname{span}\{R_{Bu}[A] : \operatorname{span}\{C(u)\} \subseteq \mathscr{R}_{k-1}\} \\ &= \operatorname{span}\Big\{R_{Bu}[A] : u \in \prod_{l=1}^{r-1} S_\perp^l \mathscr{N}^r, r = 1, 2, \cdots, k\Big\} \\ &= \operatorname{span}\Big\{R_x[A], R_{Bu}[A] : x \in \mathscr{M}(\mathscr{R}_{k-2}), u \in \prod_{l=1}^{k-1} S_\perp^l \mathscr{N}^k\Big\}.\end{aligned}$$

Hence,

$$\begin{aligned}\mathscr{R}_k &= \operatorname{span}\{\mathscr{R}_{k-1}, R_x[A] : x \in \mathscr{M}(\mathscr{R}_{k-1})\} \\ &= \operatorname{span}\Big\{\mathscr{R}_{k-1}, R_{Bu}[A] : u \in \prod_{l=1}^{k-1} S_\perp^l \mathscr{N}^k\Big\} \\ &= \operatorname{span}\Big\{\mathscr{R}_{k-1}, T_+^k T_+^{k-1} \cdots T_+^1 R_{Bu}[A] : u \in \prod_{l=1}^{k-1} S_\perp^l \mathscr{N}^k\Big\}\end{aligned}$$

as $\mathscr{R}_{k-1} = \operatorname{span}\{E_-^1, E_-^2, \cdots, E_-^k\}$. But $T_+^1 R_{Bu}[A] = R_{T_+^1 Bu}[T_+^1 A] = R_{T_+^1 Bu}[A_+^1]$. As $E_-^{i+1} = R_{B_+^i}[A_+^i]$, then $T_+^{i+1} B_+^i = 0$, or $A_+^{i+1} = T_+^{i+1} A^{i+1} = T_+^{i+1}(A_+^i - B_+^i K^i) = T_+^{i+1} A_+^i$. Hence $R_{T_+^2 T_+^1 Bu}[A_+^2] = T_+^2 R_{T_+^1 Bu}[A_+^1]$. We conclude that

$$\begin{aligned}T_+^k T_+^{k-1} \cdots T_+^1 \operatorname{span}\Big\{R_{Bu}[A] : u \in \prod_{l=1}^{k-1} S_\perp^l \mathscr{N}^k\Big\} \\ &= \operatorname{span}\{R_x[A_+^k] : x \in T_+^k \cdots T_+^1 B S_\perp^1 \cdots S_\perp^{k-1} \mathscr{N}^k\} \\ &= \operatorname{span}\{R_{T_+^k B^k S^k v}[A_+^k] : v \in R^{q_k}\} \\ &= R_{B_+^k}[A_+^k] = E^{k+1}.\end{aligned}$$

Hence

$$\mathscr{R}_k = \operatorname{span}\{\mathscr{R}_{k-1}, E_-^{k+1}\}$$

and the lemma is proved.

THEOREM 3.7. *If $\mathscr{R}_m = R^n$, then there exists a linear control u which stabilizes* (2.1) *no matter how large C is.*

Proof. By assumption $R^n = \operatorname{span}\{E_-^1, E_-^2, \cdots, E_-^l\}$. By Lemma 3.1, if there exist $P_+^1, K_+^1, \lambda_+^1, \rho_+^1$ such that (3.3) is satisfied on E_+^1 then (2.1) can be stabilized. But $E_+^1 = E_-^2 \oplus E_+^2$, and so again by Lemma 3.1, P_+^1, K_+^1, λ_+^1, ρ_+^1 exist if $P_+^2, K_+^2, \lambda_+^2, \rho_+^2$ exist and satisfy (3.3) on E_+^2. This continues on up to the necessity of finding P_+^{l-1}, K_+^{l-1}, λ_+^{l-1}, ρ_+^{l-1} satisfying (3.3) on E_+^{l-1}. However $E_+^{l-1} = E_-^l = R_{B_+^{l-1}}[A_+^{l-1}]$, so that on E_+^{l-1} the system is controllable with controls that do not produce any control dependent noise. Hence P_+^{l-1}, K_+^{l-1}, λ_+^{l-1}, $\rho_+^{l-1} = 1$ can be constructed as in the case $C = 0$. This establishes the theorem.

We observe that if $\mathscr{R}_m \neq R^n$ then at some stage, $R_{B_+^i}[A_+^i] = \{0\}$ and so $E_-^{i+1} = \{0\}$. However, since $E_+^i \neq \{0\}$, $E_+^{i+1} \neq \{0\}$, and so some unstable modes remain. If C is sufficiently small on these modes, then the system can still be stabilized; however, we shall show that if C is sufficiently large on these unstable modes, then no control can stabilize.

If we restrict the decomposition of R^m into subspaces spanned by the coordinate axes, rather than the kernel of C and its orthogonal complement, we derive a useful corollary.

Set

$$S_0 = \{\varnothing\},$$

$$S_{k+1} = \left\{ i \notin \bigcup_{l=1}^{k} S_l : \operatorname{span}\{C^i\} \subseteq \operatorname{span}\{E_-^1, R_{b^j}[A] : j \in S_k\} \right\},$$

where b^j is the jth column of B and C^i is as in (2.2).

COROLLARY 3.8. *If $i \in \bigcup_{k=1}^m S_k$, then* (2.1) *can be stabilized no matter how large C^i is.*

4. Instability. We begin with a general instability theorem due to Wonham [10]. Wonham uses the partial differential equations approach to prove the result; however, this is not very intuitive, and so we shall give a probabilistic proof. We consider now

$$dx(t) = f(x(t))\,dt + \sigma(x(t))\,dw(t), \qquad t \geqq 0. \tag{4.1}$$

Assume:

(i) x, f are in R^n, σ is an $n \times n$ matrix.
(ii) $\{w(t) : t > 0\}$ is an n-dimensional Wiener process.
(iii) $x(0)$ is a random variable, ind.ependent of the increments of $w(t)$.
(iv) f and σ are Lipschitz continuous.
(v) There is a constant $c > 0$ such that

$$y'\sigma(x)\sigma(x)'y \geqq c\|y\|^2$$

for all x, y in R^n.

It follows that the unique solution of (4.1) is a continuous strong Feller process $X = \{x(t): t \geqq 0\}$. Observe that condition (v) was not required in the previous section; however we are going to use a result from [5] which does require this condition. Next we assume that the process X determined by (4.1) is *positive*; i.e. for any nonempty, connected, bounded, open set G with boundary Γ,

$$\mathscr{E}_x\{\tau_\Gamma\} \equiv \mathscr{E}\{\tau_\Gamma | x(0) = x\} < \infty$$

for any x in the complement of G, where τ_Γ is the first time Γ is reached. It is known [6] that under assumptions (i)–(v) positivity is equivalent to the existence of an invariant measure μ, and moreover this probability measure is unique.

Wonham proved the following result [5, Lemma 3.2].

LEMMA 4.1. *Assume $L(x) \geqq 0$ is locally Hölder continuous, as well as all the above hypotheses. Then*

$$\mathscr{E}_\mu\{L(x)\} < \infty$$

if and only if

$$\mathscr{E}_x\left\{\int_0^{\tau_\Gamma} L[x(t)]\,dt\right\} < \infty, \qquad x \in R^n.$$

Now we have an instability theorem established originally by Wonham.

THEOREM 4.2. *Let X be the diffusion process determined by* (4.1) *under the preceding assumptions* (i)–(v), *and let $\mathscr{L}$ be its differential generator. Let $L(x) \geqq 0$ be locally Hölder continuous, and assume there exist two real-valued functions V_1, V_2, such that:*

(a) *For some $r < \infty$, V_1 and V_2 are defined and are twice continuously differentiable for $\|x\| \geqq r$.*

(b) *There exists a sequence $\{x_n\}$ with $\|x_n\| \to \infty$ such that $V_1(x_n) \to \infty$ as $n \to \infty$.*

(c) $$V_2(x) \geqq 0 \quad \textit{if } \|x\| \geqq r.$$

(d) $$\limsup_{\rho\to\infty} \frac{\max\{V_1(x): \|x\| = \rho\}}{\min\{V_2(x): \|x\| = \rho\}} = 0,$$

(e) $$\mathscr{L}V_1(x) \geqq 0, \quad \mathscr{L}V_2(x) \leqq L(x) \quad \textit{for } \|x\| \geqq r.$$

If X is positive with stationary measure μ, then

$$\mathscr{E}_\mu\{L(x)\} = +\infty.$$

Proof. Let $\Gamma_n = \{x: \|x\| = n\}$, $\Gamma = \{x: \|x\| = r\}$, and write τ_n for τ_{Γ_n}, and τ for τ_Γ. Also let $\bar{\tau}_n = \min\{\tau, \tau_n\}$. By considering the process obtained by stopping the positive process X at $\bar{\tau}_n$, it can be shown for any function V satisfying (a), that

$$\mathscr{E}_x\{V[x(\bar{\tau}_n)]\} - V(x) = \mathscr{E}_x\left\{\int_0^{\bar{\tau}_n} \mathscr{L}V[x(s)]\,ds\right\} \tag{4.2}$$

if $r < \|x\| < n$. We call this domain D_n.

Let $\alpha_n = \Pr\{\bar{\tau}_n = \tau_n\}$, $\beta_n = \Pr\{\bar{\tau}_n = \tau\}$. By (e) and (4.2) we have

$$\begin{aligned} &\mathscr{E}_x\{V_1[x(\tau_n)] | \bar{\tau}_n = \tau_n\}\alpha_n + \mathscr{E}_x\{V_1[x(\tau)] | \bar{\tau}_n = \tau\}\beta_n - V_1(x) \\ &\qquad = \mathscr{E}_x\left\{\int_0^{\bar{\tau}_n} \mathscr{L}V_1[x(s)]\,ds\right\} \geqq 0 \end{aligned} \tag{4.3}$$

and

$$\begin{aligned}&\mathscr{E}_x\{V_2[x(\tau_n)]|\bar{\tau}_n = \tau_n\}\alpha_n + \mathscr{E}_x\{V_2[x(\tau)]|\bar{\tau}_n = \tau\}\beta_n - V_2(x)\\ &\quad \leqq \mathscr{E}_x\left\{\int_0^{\bar{\tau}_n} L[x(t)]\,dt\right\} \leqq \mathscr{E}_x\left\{\int_0^{\tau_n} L[x(t)]\,dt\right\}\end{aligned} \tag{4.4}$$

as $L(x) \geqq 0$. We call the last integral $I(x)$. By the lemma it suffices to show $I(x) = +\infty$ for some x (hence for all x).

From (d) it follows that given $\varepsilon > 0$, there exists N such that if $\|x\| = n > N$, then $V_1(x) \leqq \varepsilon V_2(x)$. Hence (4.3) and (4.4) yield

$$\begin{aligned}I(x) &\geqq \mathscr{E}_x\{V_2[x(\tau)]|\bar{\tau}_n = \tau\}\beta_n - V_2(x)\\ &\quad + \frac{1}{\varepsilon}[V_1(x) - \mathscr{E}_x\{V_1[x(\tau)]|\bar{\tau}_n = \tau\}\beta_n].\end{aligned}$$

Because of (b) we may conclude that $V_1(x) \leqq 0$ for $x \in \Gamma$, but $V_1(x) > 0$ for some x with $\|x\| > r$. Finally using (c), we observe

$$I(x) \geqq \frac{V_1(x)}{\varepsilon} - V_2(x).$$

Now let $\varepsilon \to 0$ to conclude $I(x) = +\infty$. This proves the theorem.

Remark. Observe that (b) could be weakened to:

(b′) $V_1(x) \leqq 0$ for $\|x\| = r$, and there exists x with $\|x\| > r$ such that $V_1(x) > 0$.

Let us now apply this result to an equation of the type of (2.1). If P is an $n \times n$ matrix, we let $\Gamma(P)$ be the matrix with *ij*th entry

$$\Gamma_{ij}(P) = \operatorname{tr}\{C^{i\prime}PC^j\}.$$

If we consider controls $u = \phi(x)$ satisfying (2.3) then conditions (i)–(iv) are satisfied. Moreover if $DD' > 0$ (i.e. DD' is real, symmetric and $x'DD'x > 0$ for all $x \neq 0$), then (v) is also taken care of. Hence if an invariant measure exists it is unique and the process is positive. We let $L(x) = x'x = \|x\|^2$.

THEOREM 4.3. *Assume all the eigenvalues of A lie in the open right half-plane, and assume that the kernel of B contains the kernel of C. If there exists* $P > 0$ *such that*

$$u'\Gamma(Q)u > u'B'QP^{-1}QBu \tag{4.5}$$

for any u not in the kernel of B, where

$$Q = \int_{-\infty}^{0} e^{tA'}P\,e^{tA}\,dt, \tag{4.6}$$

then no control u satisfying (2.3) *can stabilize* (2.1).

Proof. If no invariant measure μ exists, then the result is obvious, so assume that the process is positive.

Let $V_2(x) = \theta x'x$. If $u = \phi(x)$ satisfies (2.3), then $\mathscr{L}_u V_2(x) \leqq \theta k_1(1 + \|x\|^2) \leqq \|x\|^2$ for $\|x\| \geqq r$, some $r > 0$, provided $\theta > 0$ is sufficiently small. Let $V_1(x) = (x'Qx)^q$. If $0 < q < 1$ then (a) through (d) of Theorem 4.2 are satisfied.

If $\mathscr{L}_u V_1 \geqq 0$ for $\|x\| \geqq r$, then our result follows by Theorem 4.2.

From (4.6) it follows that

$$A'Q + QA = P.$$

Hence for $\|x\| > 0$,

$$\begin{aligned}\mathscr{L}_u V_1(x) = q(x'Qx)^{q-1}\{&u'\Gamma(Q)u + 2(q-1)u'\Gamma(Qxx'Q)u/(x'Qx)\\ &+ \operatorname{tr}(D'QD) + 2(q-1)\operatorname{tr}(D'Qxx'QD)/(x'Qx)\\ &+ x'Px - 2u'B'Qx\}.\end{aligned}$$

But from the positivity of Q and the Schwarz inequality, it follows that $\Gamma(Qxx'Q) \leqq (x'Qx)\Gamma(Q)$. Also $\operatorname{tr}(D'Qxx'QD) = \|D'Qx\|^2 \leqq (x'Qx)\operatorname{tr}(D'QD)$. Then

$$\mathscr{L}_u V_1(x) \geqq q(x'Qx)^{q-1}\{(2q-1)u'\Gamma(Q)u + (2q-1)\operatorname{tr}(D'QD) + x'Px - 2u'B'Qx\}.$$

If $q \geqq \frac{1}{2}$ we may drop the trace term, and then complete the square to find

$$\mathscr{L}_u V_1(x) \geqq q(x'Qx)^{q-1}\{y'y + u'[(2q-1)\Gamma(Q) - B'QP^{-1}QB]u\}, \tag{4.7}$$

where

$$y = \sqrt{P} - (\sqrt{P})^{-1}QBu.$$

From (4.7) we see $\mathscr{L}V_1(x) \geqq 0$ provided

$$(2q-1)\Gamma(Q) \geqq B'QP^{-1}QB. \tag{4.8}$$

We demand also of course that $1 > q \geqq \frac{1}{2}$.

Since $u'\Gamma(Q)u = \operatorname{tr}\{C(u)'QC(u)\}$ then the kernel of $\Gamma(Q)$ is the same as the kernel of C. Now by considering (4.5) on the orthogonal complement of the null space of $\Gamma(Q)$ we conclude that (4.8) holds for $1 > q \geqq \frac{1}{2}$. This completes the proof.

Finally we can deduce the converse of Theorem 3.7; however we must assume that A has no eigenvalues lying on the imaginary axis, and that $DD' > 0$. We use the notation of the previous section.

THEOREM 4.4. *Assume that $DD' > 0$ and that A has no eigenvalues lying on the imaginary axis. If $\mathscr{R}_m \neq R^n$, then for C sufficiently large, no control $u = \phi(x)$ satisfying* (2.3) *can stabilize* (2.1).

Proof. If $\mathscr{R}_m \neq R^n$, then the decomposition breaks down at some stage. Hence for some k,

$$R_{B^k_+}[A^k_+] = E^{k+1}_- = \{0\}$$

so that $B^k_+ = 0$, i.e. $T^k_+ B^k u = 0$ for any u in $\mathscr{N}^k$, the kernel of C^k_+. Moreover from (3.11) we have

$$dy^k_+ = A^k_+ y^k_+ dt - T^k_+ B^k u dt - C^k_+(u)\,dw_1 + D^k_+ dw_2. \tag{4.9}$$

Now (4.9) has the form (2.1) and the kernel of C^k_+ is contained in the kernel of $T^k_+ B^k$. Moreover A^k_+ has no eigenvalues in the open left half-plane, and by hypothesis none on the imaginary axis. We can almost apply Theorem 4.3.

Suppose u gives rise to an invariant measure μ. Then μ projected on E^k_+ is an invariant measure for (4.9). Call it $\tilde{\mu}$; i.e., as $\mathscr{R}_{k-1}$ is a complement of E^k_+, then for A in E^k_+, $\tilde{\mu}(A) \equiv \mu(A \times \mathscr{R}_{k-1})$. Hence if (4.5) holds, then

$$\int_{E^k_+} \|y\|^2 \tilde{\mu}(dy) = +\infty,$$

and if $x = y + z$, $y \in E^k_+$, $z \in \mathscr{R}_{k-1}$, then

$$\begin{aligned}\int_{R^n} \|x\|^2 \mu(dx) &= \int_{R^n} \|(y + z)\|^2 \mu(dx) \\ &\geqq \int_{R^n} \|y\|^2 \mu(dx) \\ &= \int_{E^k_+} \|y\|^2 \tilde{\mu}(dy) = +\infty.\end{aligned}$$

But with $P = I$, (4.5) becomes

$$\text{(4.10)} \qquad \operatorname{tr}\{C^k_+(u)' Q C^k_+(u)\} > \|Q T^k_+ B^k u\|^2$$

for u not in the kernel of $T^k_+ B^k$. It is evident that this inequality can be satisfied for all u if $C(u)$ is sufficiently large, i.e. each of the matrices C^i has a sufficiently large norm. This proves the theorem.

Suppose now that such a subspace E^k_+ exists on which (2.1) cannot be stabilized if C is large, i.e. if (4.10) is satisfied. But let us examine the case when $u = Kx$, i.e. u is assumed to be linear in x. Moreover we neglect controls in the kernel of B, so assume the range of K is disjoint (except for 0) from the kernel of B. Also assume (A, B) is stabilizable [7]. We set $\hat{A} = A - BK$ and put

$$T(P) = \int_0^\infty e^{t\hat{A}'} K' \Gamma(P) K e^{t\hat{A}}\, dt.$$

For convenience we identify (4.9) with (2.1). Then it is known [1], [2], that $u = Kx$ stabilizes (4.9) if

$$T(I) < I.$$

Using a proof similar to that of Theorem 4.3 except putting $\hat{A}$ for A and $K'\Gamma(I)K$ for P, we see that $u = Kx$ does not stabilize (4.9) if

$$K'\Gamma[T(I)]K > K'\Gamma(I)K,$$

or

$$\text{(4.11)} \qquad K'\Gamma[T(I) - I]K > 0.$$

As the null space of B contains that of Γ we conclude that (4.11) holds if and only if

$$T(I) > I.$$

Hence we see that for $n \geqq 2$ there is a possibility that we cannot decide the stability of (2.1).

5. Examples.

Example 5.1.

$$A = \begin{pmatrix} 1 & 1 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & -1 \end{pmatrix}, \qquad B = \begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 1 \end{pmatrix},$$

$$C^1 = \alpha \begin{pmatrix} 1 & 2 \\ 1 & 0 \\ 1 & -1 \end{pmatrix}, \qquad C^2 = \alpha \begin{pmatrix} 2 & 2 \\ 0 & 2 \\ -2 & 4 \end{pmatrix}, \qquad C^3 = \beta \begin{pmatrix} 3 & 4 \\ 1 & 2 \\ -1 & 0 \end{pmatrix}.$$

With $\gamma = -\beta/\alpha$ we have

$$\mathscr{R}_0 = \operatorname{span} \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}, \qquad \mathscr{M}(\mathscr{R}_0) = \operatorname{span} \begin{pmatrix} \gamma \\ \gamma \\ 1+\gamma \end{pmatrix},$$

$$\mathscr{R}_1 = \operatorname{span} \left\{ \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}, \begin{pmatrix} 1 \\ 1 \\ 0 \end{pmatrix} \right\}, \qquad \mathscr{M}(\mathscr{R}_1) = \operatorname{span} \begin{pmatrix} 4\gamma \\ \gamma \\ 1+\gamma \end{pmatrix},$$

$$\mathscr{R}_2 = R^3,$$

and so this system can be stabilized no matter how large α and β are (provided $\alpha \neq 0$).

Example 5.2.

$$A = \begin{pmatrix} 1 & 1 & 0 \\ 0 & 2 & 1 \\ 0 & 0 & -1 \end{pmatrix}, \qquad B = \begin{pmatrix} 0 & 1 \\ 1 & 0 \\ 1 & 0 \end{pmatrix},$$

$$C^1 = \alpha \begin{pmatrix} 1 & 2 \\ 1 & 2 \\ 1 & -1 \end{pmatrix}, \qquad C^2 = \beta \begin{pmatrix} 1 & 2 \\ 1 & 2 \\ -1 & 1 \end{pmatrix}.$$

If $\gamma = -\alpha/\beta$, $\beta \neq 0$,

$$\mathscr{R}_0 = \operatorname{span} \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}, \qquad \mathscr{M}(\mathscr{R}_0) = \operatorname{span} \begin{pmatrix} \gamma \\ 1 \\ 1 \end{pmatrix},$$

$$\mathscr{R}_1 = \operatorname{span} \left\{ \begin{pmatrix} \gamma \\ 1 \\ 1 \end{pmatrix}, \begin{pmatrix} 1+\gamma \\ 3 \\ -1 \end{pmatrix}, \begin{pmatrix} 4+\gamma \\ 5 \\ 1 \end{pmatrix}, \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix} \right\} = R^3$$

for all γ, and the system can be stabilized.

Observe that if $\beta = 0$, then

$$\mathscr{M}(\mathscr{R}_0) = \text{span}\begin{pmatrix}1\\0\\0\end{pmatrix}, \qquad \mathscr{R}_1 = \text{span}\left\{\begin{pmatrix}1\\0\\0\end{pmatrix}, \begin{pmatrix}0\\0\\1\end{pmatrix}\right\}.$$

But now $\mathscr{M}(\mathscr{R}_1) = \mathscr{M}(\mathscr{R}_0)$ and so $\mathscr{R}_2 = \mathscr{R}_1 \neq R^3$, and the system cannot be stabilized for α large. Hence the elimination of some noise leads to instability. The point of course is that it is no longer possible to obtain the same linear combinations.

Finally it should be noted that most of the results hold in the more general case when a term $F(x)\,dw_3$ is added to the right side of (2.1), provided F satisfies certain restrictions. Here $F(x) = \sum_{i=1}^{n} F^i x_i$. For example, Theorem 4.3 holds with (4.5) modified to

$$(4.5)' \qquad u'\Gamma(Q)u > u'B'Q[\Delta(Q) + P]^{-1}QBu,$$

where $\Delta(Q)$ is a matrix constructed from F in the same way $\Gamma(Q)$ is constructed from C. On the other hand the problem of giving conditions which guarantee that a stabilizing control exists no matter how large F is, is more difficult and probably less interesting since sufficient conditions will be very strong. For some results see [2].

Acknowledgment. I thank a referee for pointing out the present simpler proof of Theorem 3.2.

REFERENCES

[1] W. M. Wonham, *Random Differential Equations in Control Theory*, Probabilistic Methods in Applied Mathematics, vol. 2, A. T. Bharucha-Reid, ed., Academic Press, New York, 1969.

[2] U. G. Haussmann, *Optimal stationary control with state and control dependent noise*, this Journal, 9 (1971), pp. 184–198.

[3] N. N. Krasovskii, *Stabilization of systems in which noise is dependent on the value of the control signal*, Engrg. Cybernetics, 2 (1965), pp. 94–102.

[4] E. B. Dynkin, *Markov Processes*, vol. I, Academic Press, New York, 1965.

[5] W. M. Wonham, *A Liapunov method for the estimation of statistical averages*, J. Differential Equations, 2 (1966), pp. 195–207.

[6] R. Z. Khasminskii, *Ergodic properties of recurrent diffusion processes and stabilization of the solution to the Cauchy problem for parabolic equations*, Theor. Probability Appl., 5 (1960), pp. 179–196.

[7] W. M. Wonham, *On pole assignment in multi-input controllable linear systems*, IEEE Trans. Automatic Control, AC-12 (1967), pp. 660–665.

[8] M. Zakai, *A Lyapunov criterion for the existence of stationary probability distributions for systems perturbed by noise*, this Journal, 7 (1969), pp. 390–397.

[9] U. G. Haussmann, *Stabilization of linear systems with multiplicative noise*, Lecture Notes in Mathematics, vol. 291, Ruth F. Curtain, ed., Springer-Verlag, Berlin, 1972, pp. 125–131.

[10] W. M. Wonham, *Optimal stationary control of a linear system with state-dependent noise*, this Journal, 5 (1967), pp. 486–500.

SIAM J. Control
Vol. 11, No. 2, May 1973

ERRATA: PERTURBATIONS OF LINEAR CONTROL SYSTEMS*

JERALD P. DAUER†

A portion of Theorem 6 and its proof are not valid. The sentence "If B is constant we can choose D constant." should be deleted from Theorem 6. This can be seen from the following example, due to G. Franklin (Stanford University):

$$\dot{x}_1 = u,$$

$$\dot{x}_2 = u.$$

Constant 2×1 perturbations of this system are not completely controllable in L^∞.

Proof of Theorem 6. Under the transformation $z(t) = X^{-1}(t)x(t)$, system (1) becomes

$$\dot{z} = C(t)u,$$

where $C(t) = X^{-1}(t)B(t)$. Let D_1 be a measurable $n \times m$ matrix function whose rows are linearly independent over I and such that

$$|C(t) - D_1(t)| < \varepsilon/\|X\|_\infty$$

for all $t \in I$. Let f be the function defined continuously on the compact set $S = \{y^* : |y^*| = 1\}$ by

$$f(y^*) = \int_I |y^* D_1(s)|\, ds.$$

We have $f(y^*) > 0$ for all y^* in S, hence

$$\inf_{y^* \in S} f(y^*) > 0.$$

Theorem 1 then shows that the system

$$\dot{z} = D_1(t)u$$

is completely controllable in L^∞. This is equivalent to the system

$$\dot{y} = A(t)y + X(t)D_1(t)u$$

being completely controllable in L^∞. Hence define $D(t) = X(t)D_1(t)$ on I.

* This Journal, 9 (1971), pp. 393–400. Received by the editors September 28, 1971.

† Department of Mathematics, University of Nebraska, Lincoln, Nebraska 68508.

SIAM J. Control
Vol. 11, No. 3, August 1973

LION AND MAN: THE BOUNDARY CONSTRAINT*

JAMES O. FLYNN†

Abstract. We consider a generalization of Rado's pursuit problem. A pursuer, moving in a circular arena with speed bounded by 1, tries to get as close as possible to an evader, moving along the circumference with speed bounded by $w \geqq 1$. Using the approach of Ryll-Nardzewski and Varaiya and Lin, we formulate this problem as a differential game. Then we determine the value and construct optimal strategies for both players. Our results are compared with Isaacs' on pursuit in the half-plane.

Introduction. A lion L and a man M, confined to a circular arena, move with speeds bounded by 1 and $w \geqq 1$, respectively. L and M, continually aware of each other's position, have contrary objectives. L wants to get as close as possible to M who wants to keep as far away as possible from L. One can view this conflict, quite naturally, as a differential game with a payoff from the lion to the man equal to the "smallest" distance between the two.

In this paper we adopt that viewpoint and find a solution for the *special case* where *the man is forced to stay on the boundary of the arena.* An informal description of that solution takes up the first section. In the second, we formulate the problem as a game using the approach developed by Ryll-Nardzewski [6, pp. 113–126] and Varaiya and Lin [8]. Then in the third, fourth and fifth, we construct optimal pursuit and evasion strategies. Finally, in the last section, we compare our problem with *pursuit in the half-plane.* Our method of solution involves separating the game into a min-max pursuit part and a max-min evasion part. This technique, tracing back to the control theory approach of Pontryagin [5], is motivated largely by Halpern [2] and Varaiya and Lin [8] and justified, for the most part, by results in Varaiya and Lin [8]. That this method has limited applicability is indicated by the examples in § 6.

We wish to thank L. Dubins and D. Blackwell for introducing us to the *general problem* (described in the first paragraph) and for stressing the importance of a rigorous approach. We wish to thank L. Dubins and Ronald Stern for their helpful comments and criticisms of an earlier version of this paper. We have benefited from Isaacs' [3] solution to *pursuit in the half-plane* and from Gerald J. Smith's [7] unpublished results on the problem of *pursuit in the circle.* In closing, we note that the *general problem* is a generalization of R. Rado's *Lion and Man* problem [4] and Isaacs' game of *pursuit in the half-plane* [3, pp. 261–265]. Apparently Isaacs was the first one to formulate it [3, pp. 265, 270]. Our solution to the *general problem* is in the process of being submitted for publication.

1. Results.

In this section we describe the trajectories and the payoff which result when both players use optimal strategies. For the present we assume an intuitive grasp of the problem leaving the formulation for the next section.

* Received by the editors October 7, 1971, and in revised form July 15, 1972.

† Graduate School of Business, University of Chicago, Chicago, Illinois 60637. This research was supported in part by the Army Research Office, Office of Naval Research, and Air Force Office of Scientific Research under Contract Nonr-225(52)(NR-342-022) at the Statistics Department, Stanford University.

We begin by replacing the arena by the unit circle with center O. Some definitions follow. Let (x, y) denote L's position in a rectangular coordinate system with origin O and M's position fixed at $(1, 0)$ (see Fig. 1). Define $\alpha = \arctan(y/x)$ as the angle between OL and the X-axis, $\theta = \arctan[y/(1-x)]$ as the angle between the line LM and the X-axis and $\rho = [(1-x)^2 + y^2]^{1/2}$ as the distance $|LM|$.

The term *speed* always refers to the norm of a player's right-hand velocity vector. Such a vector, of course, need not exist. However, when L's right-hand velocity vector does exist, we denote by ϕ the angle which it makes away from the direction LM. For ϕ positivity corresponds to measuring in a clockwise arc, e.g., ϕ is positive in Fig. 1. It follows from a result of Smith [7], than whenever L and M travel at maximum speeds,

$$\begin{aligned} dx/dt &= \cos(\theta + \phi) - \psi w y, \\ dy/dt &= -\sin(\theta + \phi) + \psi w x, \end{aligned} \tag{1}$$

where $\psi = +1$ if M moves clockwise and $\psi = -1$ if M moves counterclockwise.

The best way to describe M's optimal strategy is to draw a picture. Examine the unit circle depicted in Fig. 2. Let A be an arbitrary point on the circle, B the point on the line segment OA whose distance from O is $1/w$, and C_1 and C_2 the points of intersection of the circle with the perpendicular to OA at B. M, starting at A, plays optimally by obeying the following rule. If the distance $\rho \leqq (w-1)/w$, move directly to C_1 or C_2 at maximum speed, selecting C_1 if $\theta < 0$, C_2 if $\theta \geqq 0$. (The selection for $\theta = 0$ is, of course, arbitrary.) if $\rho > (w-1)/w$, wait until $\rho = (w-1)/w$ and then act as directed above. Upon reaching C_1 or C_2 continue with the rule which applies to A.

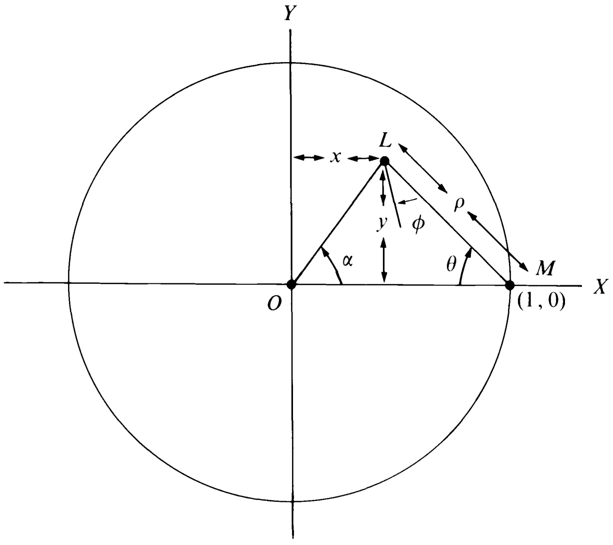


FIG. 1

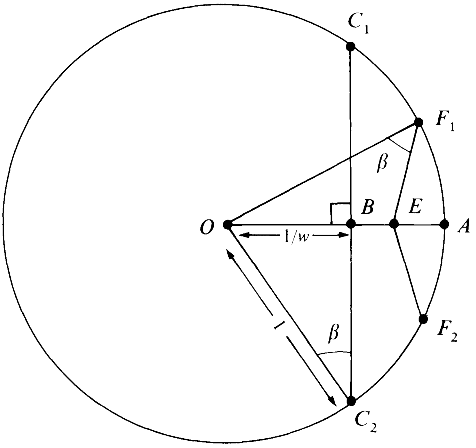


FIG. 2

L, starting at O, plays optimally by staying on the radius OM until his distance from O is $1/w$. As indicated by (13), L can achieve such a position in a finite time regardless of M's motion. What L should do afterwards, however, depends on M.

Some historical remarks are appropriate. Radial pursuit strategies first appear in J. Littlewood's [4, pp. 135–136] exposition on the case of equal speeds. More recently they appear in Gerald Smith's [7] attack on the problem of *pursuit in the circle.* Smith establishes that, for any $\varepsilon > 0$, L is a radial strategy which brings him to a point Q on OM which satisfies $|OQ| \geqq (1/w)|OM| - \varepsilon$. He attributes that result to a suggestion given by Dubins. In this paper we obtain a stronger result (see (13)) by exploiting the fact that M is restricted to the boundary.

Turn to Fig. 2. Suppose B and A denote the respective starting positions of L and M. Then, if M moves at full speed toward C_1, optimal play directs L at full speed along BC_1, while if M moves at full speed toward C_2, optimal play directs L at full speed along BC_2. The minimum value of $|LM|$ occurs when M reaches C_1 or C_2. (This fact is verified in the proof of Lemma 3.)

Using elementary arguments we can show that if L and M play as described in the above paragraph, then the minimum distance is achieved when $t = t_0 \equiv (1/w)\arccos(1/w)$. At that time $\rho = v^*$ and $|\theta| = \beta$, where

$$v^* = (1/w)[(w^2 - 1)^{1/2} - \arccos(1/w)] \tag{2}$$

and

$$\beta = \arcsin(1/w). \tag{3}$$

The resulting (x, y)-path satisfies

$$\begin{aligned} x(t) &= (1/w)(\cos(wt) + wt\sin(wt)),\\ |y(t)| &= (1/w)(\sin(wt) - wt\cos(wt)) \quad \text{for } 0 \leqq t \leqq t_0. \end{aligned} \tag{4}$$

The sign of $y(t)$ depends, of course, on the direction M chooses. Note that (4) is

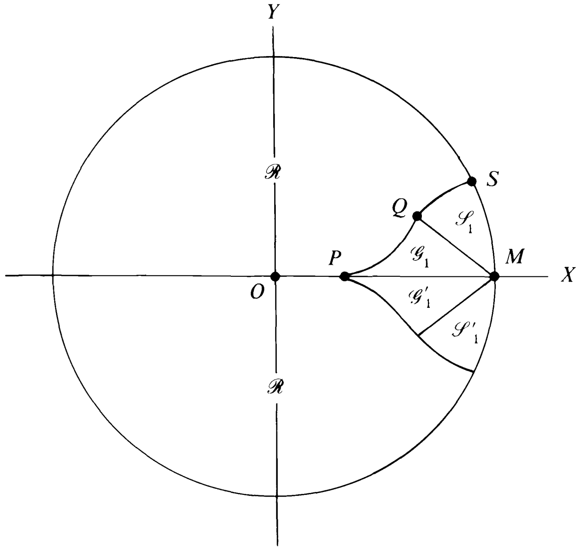


FIG. 3

the parametric representation of a section of the involute of a circle with radius $1/w$ and center O.

In Fig. 3, we divide the (x, y) representations into the sets $\mathscr{R}, \mathscr{S} = \mathscr{S}_1 \cup \mathscr{S}'_1$ and $\mathscr{G} = \mathscr{G}_1 \cup \mathscr{G}'_1$. We will obtain different results for each set. Let P correspond to $\rho = (w - 1)/w$ and $\theta = 0$, Q to $\rho = v^*$ and $\theta = \beta$, PQ to the curve traced by $(x(t), |y(t)|)$, and QS to an arc of the circle having radius v^* and center M. $\mathscr{S}_1$ is the closed set included within the arcs QS, SM and the line QM. $\mathscr{G}_1$ is the closed set included within the lines QM, PM and the curve PQ. $\mathscr{S}'_1$ and $\mathscr{G}'_1$ are the reflections of $\mathscr{S}_1$ and $\mathscr{G}_1$. And $\mathscr{R}$ is the set of points in the circle not included in $\mathscr{S}_1, \mathscr{G}_1, \mathscr{S}'_1$ or $\mathscr{G}'_1$.

We are going to describe optimal trajectories which start from positions in $\mathscr{G}$. We begin by considering points on the line segment PM (see Fig. 3). Turn to Fig. 2. Let E be a point on the radius OA where distance from O is at least $1/w$. Let F_1 and F_2 be the two points on the circle satisfying $\sphericalangle OF_1E = \sphericalangle DF_2E = \beta$ included within the minor arc C_1C_2. Suppose L starts at E and M starts at A. Then, if M moves at full speed toward C_1, optimal play directs L at full speed along EF_1, while if M moves at full speed toward C_2, optimal play directs L at full speed along EF_2. This time the minimum value of $|LM|$ occurs when M reaches F_1 or F_2. This fact will be used later; its verification is left to the reader.

Now we consider points in $\mathscr{G}_1$ which do not lie on PM. The situation in $\mathscr{G}'_1$ is, of course, analogous. In Fig. 4, E and A denote the respective starting positions of L and M. B, C_1 and C_2 are defined as before, and F is the unique point on the minor arc AC_2 satisfying $\sphericalangle OFE = \beta$. If M moves at full speed toward C_2, then optimal play directs L at full speed along EF. As before, the minimum value of $|LM|$ occurs when M reaches F. Again, the verification of this fact is left to the reader.

We can easily show that if L and M, starting from a position in $\mathscr{G}$, play as described above, then the minimum value of $|LM|$ is $W(x, y)$, where (x, y) denotes

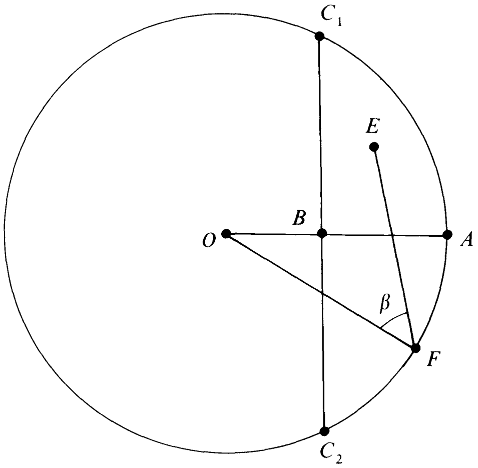


FIG. 4

the starting position and $W(x, y)$ satisfies (6). Also, while L and M remain in $\mathscr{G}$, their directions $\phi(t)$ and $\psi(t)$ (see (1)) satisfy

$$(5) \qquad \psi(t) = \begin{cases} +1 & \text{for } \theta(t) > 0, \\ \pm 1 & \text{for } \theta(t) = 0, \\ -1 & \text{for } \theta(t) < 0, \end{cases}$$

$$|\phi(t)| \\ = \arccos \left[\frac{(\cos\theta(t) - \rho(t))(\rho^2(t) - 2\rho(t)\cos\theta(t) + \cos^2\beta)^{1/2} + \sin|\theta(t)| \sin\beta}{\rho^2(t) - 2\rho(t)\cos\theta(t) + 1} \right],$$

and

$$\operatorname{sgn}\phi(t) = \operatorname{sgn}\psi(t).$$

In § 5 we will describe optimal play by L versus arbitrary play by M. When both L and M play optimally, a position in $\mathscr{G}$ leads to a position in $\mathscr{S}$. As shown in Lemma 2, L is unable to stop $|LM|$ from increasing in $\mathscr{S}$. Eventually the game drifts into $\mathscr{R}$. Given a position in $\mathscr{R}$, L's optimal strategy takes him to the center and then along the radius line to the point P (see Fig. 3) which is in $\mathscr{G}$.

The optimal payoff or value V is a function of the initial (x, y)-position. Specifically $V = W(x, y)$, where

$$(6) \qquad W(x, y) = \begin{cases} v^* \quad \text{for } (x, y) \in \mathscr{R}, \\ ((1 - x)^2 + y^2)^{1/2} \quad \text{for } (x, y) \in \mathscr{S}, \\ v^* - (1/w)\{[w^2(x^2 + y^2) - 1]^{1/2} \\ \qquad\qquad - \arccos[x/(x^2 + y^2)^{1/2}] \\ \qquad\qquad - \arccos[1/w(x^2 + y^2)^{1/2}]\} \quad \text{for } (x, y) \in \mathscr{G}. \end{cases}$$

With a little patience one can verify that Isaacs' main equations [3, pp. 67, 69] hold on the interior of $\mathscr{G}$ for the value function $W(x, y)$ and the directions ϕ and ψ satisfying (5). However, we do not use this fact to support any of our results.

One closing remark is appropriate. The idea behind L's strategy in $\mathscr{G}$—that L heads directly towards a point F, where $|\sphericalangle OFL| = \beta$—generalizes Isaacs' results on *pursuit in the half-plane* [3, pp. 260–265]. We will discuss the relationship between *pursuit in the circle* and *pursuit in the half-plane* in the last section.

2. Formulation. Let $\underset{\sim}{R}$ represent the real numbers and let $\underset{\sim}{R}^2$ represent the space $\underset{\sim}{R} \times \underset{\sim}{R}$ with the metric determined by the norm $\|\cdot\|$, where $\|(r_1, r_2)\| = (r_1^2 + r_2^2)^{1/2}$. Denote the unit circle in $\underset{\sim}{R}^2$ by $\underset{\sim}{C}$, the closed unit disc in $\underset{\sim}{R}^2$ by $\underset{\sim}{D}$ and the time axis $[0, \infty)$ by $\underset{\sim}{T}$. Define:

$$\underset{\sim}{L}(l) = \{\underset{\sim}{l} \mid \underset{\sim}{l} : \underset{\sim}{T} \to \underset{\sim}{D}, \underset{\sim}{l}(0) = l \text{ and } \|\underset{\sim}{l}(t') - \underset{\sim}{l}(t'')\| \leqq \|t' - t''\|$$
$$\text{for all } t', t'' \in \underset{\sim}{T}\} \quad \text{for } l \in \underset{\sim}{D},$$
$$\underset{\sim}{M}(\mathscr{M}) = \{\underset{\sim}{\mathscr{M}} \mid \underset{\sim}{\mathscr{M}} : \underset{\sim}{T} \to \underset{\sim}{C}, \underset{\sim}{\mathscr{M}}(0) = \mathscr{M} \text{ and } \|\underset{\sim}{\mathscr{M}}(t') - \underset{\sim}{\mathscr{M}}(t'')\| \leqq w\|t' - t''\|$$
$$\text{for all } t', t'' \in \underset{\sim}{T}\} \quad \text{for } \mathscr{M} \in \underset{\sim}{C},$$
$$\underset{\sim}{L} = \bigcup_{l \in D} \underset{\sim}{L}(l), \underset{\sim}{M} = \bigcup_{\mathscr{M} \in C} \underset{\sim}{M}(\mathscr{M}) \text{ and } P(\underset{\sim}{l}, \underset{\sim}{\mathscr{M}}) = \inf_{t \in \underset{\sim}{T}} \|\underset{\sim}{l}(t) - \underset{\sim}{\mathscr{M}}(t)\|$$
$$\text{for } (\underset{\sim}{l}, \underset{\sim}{\mathscr{M}}) \in \underset{\sim}{L} \times \underset{\sim}{M}.$$

$\underset{\sim}{L}(l)$ and $\underset{\sim}{M}(\mathscr{M})$ are trajectories for L and M originating from the respective positions l and $\mathscr{M}$, while $P(\underset{\sim}{l}, \underset{\sim}{\mathscr{M}})$ is the payoff from L to M when L uses $\underset{\sim}{l}$ and M uses $\underset{\sim}{\mathscr{M}}$. Using Ascoli's theorem, one can show that $\underset{\sim}{L}$ and $\underset{\sim}{M}$ are compact metric spaces with respect to the topology of uniform convergence on each compact subset of T. Clearly, P is a continuous function on the compact space $\underset{\sim}{L} \times \underset{\sim}{M}$.

We adopt the approach of Ryll-Nardzewski [6, pp. 113–126] and Varaiya and Lin [8] in defining strategies. Let $(l, m) \in \underset{\sim}{D} \times \underset{\sim}{C}$. A mapping $\pi : \underset{\sim}{M}(\mathscr{M}) \to \underset{\sim}{L}(l)$ is a pursuit strategy at $(l, \mathscr{M})$ if it satisfies the information constraint: for any $\underset{\sim}{\mathscr{M}}'$, $\underset{\sim}{\mathscr{M}}'' \in \underset{\sim}{M}(\mathscr{M})$, $\underset{\sim}{\mathscr{M}}'(t) = \underset{\sim}{\mathscr{M}}''(t)$ for $0 \leqq t \leqq t'$ implies that $\pi(\underset{\sim}{\mathscr{M}}')(t) = \pi(\underset{\sim}{\mathscr{M}}'')(t)$ for $0 \leqq t \leqq t'$. In much the same manner we define evasion strategies. Denote by $\Pi(l, \mathscr{M})$ and $H(l, \mathscr{M})$, respectively, the sets of pursuit and evasion strategies at $(l, \mathscr{M})$.

Remark 1. We would like to define the outcome resulting from L choosing $\pi \in \Pi(l, \mathscr{M})$ and M choosing $\eta \in H(l, \mathscr{M})$ as any pair $(\underset{\sim}{l}, \underset{\sim}{\mathscr{M}}) \in \underset{\sim}{L}(l) \times \underset{\sim}{M}(\mathscr{M})$ satisfying $\pi(\underset{\sim}{\mathscr{M}}) = \underset{\sim}{l}$ and $\eta(\underset{\sim}{l}) = \underset{\sim}{\mathscr{M}}$. Unfortunately, as in the first example in § 6, that system might not have a solution. We can guarantee existence (but not uniqueness) if we require that π and η are continuous. However, such a restriction is much too stringent. Varaiya and Lin [8] propose the following definition to resolve this difficulty.

DEFINITION. $(\underset{\sim}{l}, \underset{\sim}{\mathscr{M}}) \in \underset{\sim}{L}(l) \times \underset{\sim}{M}(\mathscr{M})$ is an *outcome* of $(\pi, \eta) \in \Pi(l, \mathscr{M}) \times H(l, \mathscr{M})$ if there exist sequences $\langle \underset{\sim}{l}_n \rangle_{n=1}^{\infty} \subset \underset{\sim}{L}(l)$ and $\langle \underset{\sim}{\mathscr{M}}_n \rangle_{n=1}^{\infty} \subset \underset{\sim}{M}(\mathscr{M})$ such that $\lim_n \underset{\sim}{l}_n = \lim_n \pi(\underset{\sim}{\mathscr{M}}_n) = \underset{\sim}{l}$ and $\lim_n \underset{\sim}{\mathscr{M}}_n = \lim_n \eta(\underset{\sim}{l}_n) = \underset{\sim}{\mathscr{M}}$.

We denote the set of outcomes of (π, η) by $O(\pi, \eta)$. Because of the compactness of $\underset{\sim}{L}(l) \times \underset{\sim}{M}(\mathscr{M})$, we can use the arguments of Varaiya and Lin [8] to show that

$O(\pi, \eta)$ is a nonempty compact subset of $\underset{\sim}{L}(l) \times \underset{\sim}{M}(\mathcal{M})$. Then, using the continuity of P, we can show that

$$\sup_{\underset{\sim}{\mathcal{M}} \in \underset{\sim}{M}(\mathcal{M})} P(\pi(\underset{\sim}{\mathcal{M}}), \underset{\sim}{\mathcal{M}}) \geqq P(\underset{\sim}{l}, \underset{\sim}{\mathcal{M}}) \geqq \inf_{\underset{\sim}{l} \in \underset{\sim}{L}(l)} P(\underset{\sim}{l}, \eta(\underset{\sim}{l})) \tag{7}$$

holds for every $\pi \in \Pi(l, \mathcal{M})$, $\eta \in H(l, \mathcal{M})$, and $(\underset{\sim}{l}, \underset{\sim}{\mathcal{M}}) \in O(\pi, \eta)$.

We define the game $\Gamma(l, \mathcal{M})$ as in Varaiya and Lin [8, p. 150]. L selects a $\pi \in \Pi(l, \mathcal{M})$ while M independently selects an $\eta \in H(l, \mathcal{M})$. The payoff from L to M is $P(\underset{\sim}{l}, \underset{\sim}{\mathcal{M}})$, where $(\underset{\sim}{l}, \underset{\sim}{\mathcal{M}})$ is an arbitrary point in $O(\pi, \eta)$. If $\pi^* \in \Pi(l, \mathcal{M})$ and $\eta^* \in H(l, \mathcal{M})$ satisfy

$$\max_{(\underset{\sim}{l}, \underset{\sim}{\mathcal{M}}) \in O(\pi^*, \eta)} P(\underset{\sim}{l}, \underset{\sim}{\mathcal{M}}) \leqq \max_{(\underset{\sim}{l}, \underset{\sim}{\mathcal{M}}) \in O(\pi^*, \eta^*)} P(\underset{\sim}{l}, \underset{\sim}{\mathcal{M}}) = \min_{(\underset{\sim}{l}, \underset{\sim}{\mathcal{M}}) \in O(\pi^*, \eta^*)} P(\underset{\sim}{l}, \underset{\sim}{\mathcal{M}}) \leqq \min_{(\underset{\sim}{l}, \underset{\sim}{\mathcal{M}}) \in O(\pi, \eta^*)} P(\underset{\sim}{l}, \underset{\sim}{\mathcal{M}}) \tag{8}$$

for all $\pi \in \Pi(l, \mathcal{M})$ and $\eta \in H(l, \mathcal{M})$, then π^* and η^* are optimal pursuit and evasion strategies, and the value of the game, which we denote by $V(l, \mathcal{M})$, is the value of $P(\underset{\sim}{l}, \underset{\sim}{\mathcal{M}})$ common to $(\underset{\sim}{l}, \underset{\sim}{\mathcal{M}}) \in O(\pi^*, \eta^*)$.

Because of (7), we can replace our game with two simpler problems. In the first, a min-max pursuit problem (see Halpern [2]), L announces a strategy $\pi \in \Pi(l, \mathcal{M})$ to M who selects $\underset{\sim}{\mathcal{M}} \in \underset{\sim}{M}(\mathcal{M})$ and receives $P(\pi(\underset{\sim}{\mathcal{M}}), \underset{\sim}{\mathcal{M}})$ from L. Similarly, in the second, a max-min evasion problem, M announces a strategy $\eta \in H(l, \mathcal{M})$ to L who selects $\underset{\sim}{l} \in \underset{\sim}{L}(l)$ and pays $P(\underset{\sim}{l}, \eta(\underset{\sim}{l}))$ to M. If the optimal payoffs in these two problems are equal, then their solutions can be combined to form a solution to the game. We have the following theorem.

THEOREM 1. *If $\pi^* \in \Pi(l, \mathcal{M})$ and $\eta^* \in H(l, \mathcal{M})$ satisfy*

$$\sup_{\underset{\sim}{\mathcal{M}} \in \underset{\sim}{M}(\mathcal{M})} P(\pi^*(\underset{\sim}{\mathcal{M}}), \underset{\sim}{\mathcal{M}}) = \inf_{\underset{\sim}{l} \in L(l)} P(\underset{\sim}{l}, \eta^*(\underset{\sim}{l})), \tag{9}$$

then they are optimal pursuit and evasion strategies and

$$V(l, \mathcal{M}) = \sup_{\underset{\sim}{\mathcal{M}} \in \underset{\sim}{M}(\mathcal{M})} P(\pi^*(\underset{\sim}{\mathcal{M}}), \underset{\sim}{\mathcal{M}}).$$

Proof. By (7), (9) implies (8). (For additional details, see Theorem 7 of Varaiya and Lin [8].)

Remark 2. The continuity of P is essential. In § 6 we give an example of a discontinuous P for which (7) does not hold.

3. The program. Eventually we will show that the rule described in paragraph 4 of § 1 defines an optimal evasion strategy for every $(l, \mathcal{M}) \in \underset{\sim}{D} \times \underset{\sim}{C}$. We denote that strategy by η^* (or more properly $\eta^*[l, \mathcal{M}]$). In § 4 we prove the following theorem (see (6)).

THEOREM 2. $\inf_{\underset{\sim}{l} \in \underset{\sim}{L}(l)} P(\underset{\sim}{l}, \eta^*[l, \mathcal{M}](\underset{\sim}{\mathcal{M}})) = W(x, y)$, *where* (x, y) *represents* $(l, \mathcal{M})$.

Then, in the next section, we define a mapping $\pi^*[l, \mathcal{M}] \in \Pi(l, \mathcal{M})$ and prove Theorem 3.

THEOREM 3. $\sup_{\underset{\sim}{\mathcal{M}} \in \underset{\sim}{M}(\mathcal{M})} P(\pi^*[l, \mathcal{M}]\ (\underset{\sim}{\mathcal{M}}), \underset{\sim}{\mathcal{M}}) = W(x, y)$, *where* (x, y) *represents* $(l, \mathcal{M})$.

Our principal result follows easily from Theorems 1, 2 and 3.

THEOREM 4. *The game corresponding to $(l, \mathcal{M}) \in \underset{\sim}{D} \times \underset{\sim}{C}$ has an optimal pursuit strategy, $\pi^*[l, \mathcal{M}]$, an optimal evasion strategy, $\eta^*[l, \mathcal{M}]$, and a value equal to $W(x, y)$, where (x, y) represents $(l, \mathcal{M})$.*

4. Evasion. Assume that M uses η^*. Let $(l, \mathscr{M}) \in \underset{\sim}{D} \times \underset{\sim}{C}$. L's objective is to minimize $P(\underset{\sim}{l}, \eta^*(\underset{\sim}{l}))$. We call $(l, \mathscr{M})$ an *attack position* (AP) if $\rho \leqq (w - 1)/w$ and a *nonattack position* (NAP) if $\rho > (w - 1)/w$. Starting from any NAP, L can always reach an AP by going to the center and walking along the radius OM. Hence we call any position where $\rho = (w - 1)/w$ and $\theta = 0$ a *reachable attack position* (RAP).

Let $(l, \mathscr{M})$ be any AP and let $\underset{\sim}{l} \in \underset{\sim}{L}(l)$. Define

$$\tau_0(\underset{\sim}{l}, \mathscr{M}) = \inf\{t | (\underset{\sim}{l}(t), \eta^*(\underset{\sim}{l})(t)) \text{ is a NAP}\}.$$

L stays in an AP up to time $\tau_0(\underset{\sim}{l}, \mathscr{M})$. We have the following lemma.

LEMMA 1. *Let* $(l, \mathscr{M})$ *be an AP and let* $\underset{\sim}{l} \in \underset{\sim}{L}(l)$. *If* $\theta(0) \geqq 0, \theta(0) < 0$ *then* $\theta(t) \geqq 0, \theta(t) < 0$, *respectively, for all* $0 \leqq t \leqq \tau_0(\underset{\sim}{l}, \mathscr{M})$.

Proof. The line OL cannot rotate as rapidly as the line OM unless the distance $|OL| \leqq 1/w$.

Again let $(l, \mathscr{M})$ be an AP. Define

$$Q(\underset{\sim}{l}, \mathscr{M}) = \inf\{\|\underset{\sim}{l}(t) - \eta^*(\underset{\sim}{l})(t)\| \,|0 \leqq t \leqq \tau_0(\underset{\sim}{l}, \mathscr{M})\}.$$

Let $(l^*, \mathscr{M}^*)$ be any RAP and let

$$Q^* = \inf_{\underset{\sim}{l} \in \underset{\sim}{L}(l^*)} Q(\underset{\sim}{l}, \mathscr{M}^*).$$

One can easily show that

$$\inf_{\underset{\sim}{l} \in \underset{\sim}{L}(l)} P(\underset{\sim}{l}, \eta^*[l, \mathscr{M}](\underset{\sim}{l})) = \min\{Q^*, \inf_{\underset{\sim}{l} \in \underset{\sim}{L}(l)} Q(\underset{\sim}{l}, \mathscr{M})\}. \tag{10}$$

We call $(l, \mathscr{M})$ a *bad attack position* (BAP) if

$$\inf_{\underset{\sim}{l} \in \underset{\sim}{L}(l)} Q(\underset{\sim}{l}, \mathscr{M}) \geqq Q^*.$$

Clearly the optimal payoffs at a NAP, a BAP and a RAP are equal to Q^*.

One important consequence of the next lemma is that $|\theta(0)| \geqq \beta$ implies $Q(\underset{\sim}{l}, \mathscr{M}) = \rho(0)$ for every $\underset{\sim}{l} \in \underset{\sim}{L}(l)$. Another is that L can do nothing to prevent M from increasing in $\mathscr{S}$.

LEMMA 2. *If* $(l, \mathscr{M})$ *represents an AP*, $\underset{\sim}{l} \in \underset{\sim}{L}(l)$ *and* $0 \leqq t' \leqq t'' \leqq \tau_0(\underset{\sim}{l}, \mathscr{M})$, *then either* (a) $|\theta(t'')| \leqq |\theta(t')| \leqq \beta$ *or* (b) $|\theta(t')| \geqq \beta$ *implies* (c) $\rho(t'') \geqq \rho(t')$.

Proof. Assume $\theta(0) \geqq 0$. It follows from the arguments of Lemma 1 that $\alpha(t)$ is nondecreasing, $\alpha(t) \geqq 0$ and $\theta(t) \geqq 0$ for $0 \leqq t \leqq \tau_0(\underset{\sim}{l}, \mathscr{M})$. Also, a simple geometric argument establishes that $\theta(t') \leqq \beta$ and $\|\underset{\sim}{l}(t') - \underset{\sim}{\mathscr{M}}(t')\| \leqq (w - 1)/w$ imply $\alpha(t') + \theta(t') < \pi/2$. But the latter, $\alpha(t'') \geqq \alpha(t')$ and $\theta(t'') \leqq \theta(t')$ imply $\rho'(t'') \geqq \rho(t')$. Hence, (a) implies (c). Finally, since M's velocity component along the line $\underset{\sim}{l}(t)\underset{\sim}{\mathscr{M}}(t)$ is $w \sin \theta(t)$, $\rho(t)$ can decrease only when $\theta(t) \leqq \beta$. The latter and the fact that (a) implies (c) tells us that (b) implies (c). Similar arguments work when $\theta(0) < 0$.

As before let $(l^*, \mathscr{M}^*)$ be a RAP. In the proof of the next lemma we will show that the trajectory which directs L at full speed towards F minimizes $Q(\underset{\sim}{l}, \mathscr{M}^*)$ (see Fig. 5). We will exploit this fact later. For the present, note that the trajectory in question is identical to the one described in paragraph 7 of § 1 and that the resulting (x, y)-coordinates trace the curve PQ depicted in Fig. 3.

LEMMA 3. $Q^* = v^*$ (see (3) and (10)).

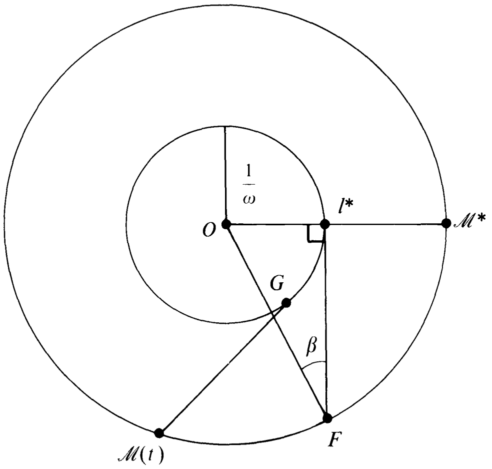


FIG. 5

Proof. Take any $\underset{\sim}{l} \in \underset{\sim}{L}(l^*)$ and $t \leqq \tau_0(\underset{\sim}{l}, \mathscr{M}^*)$. If $t < t_0 = (1/w)(\pi/2 - \beta)$, then

$$\|\underset{\sim}{l}(t) - \underset{\sim}{\mathscr{M}}(t)\| \geqq (1/w)(w^2 - 2w\cos wt + 1)^{1/2} - t,$$

the distance between L and M which results when L moves at full speed along the line joining l^* and $\underset{\sim}{\mathscr{M}}(t)$. One can show that the latter $> v^*$ when $t < t_0$. If $\underset{\sim}{l}$ directs L at full speed towards F, then $\underset{\sim}{\mathscr{M}}(t_0) = F$ and $\|\underset{\sim}{l}(t_0) - \underset{\sim}{\mathscr{M}}(t_0)\| = v^*$. Hence, $\|\underset{\sim}{l}(t) - \underset{\sim}{\mathscr{M}}(t)\| \geqq v^*$ for $t \leqq t_0$ with possible equality at $t = t_0$.

Suppose $t > t_0$. Denote by $\underset{\sim}{C}'$ the circle having radius $1/w$ and center O. Draw a line through $\underset{\sim}{\mathscr{M}}(t)$ tangent to $\underset{\sim}{C}'$ such that the point of tangency G lies to the right of M as M standing at $\underset{\sim}{\mathscr{M}}(t)$ faces O (see Fig. 5). Clearly, L cannot head directly towards $\underset{\sim}{\mathscr{M}}(t)$ without passing through the interior of $\underset{\sim}{C}'$. But L cannot pass through the interior of $\underset{\sim}{C}'$ without violating $t \leqq \tau_0(\underset{\sim}{l}, \mathscr{M}^*)$. It is not difficult to show that the trajectory where L moves along the arc l^*G and then heads toward $\underset{\sim}{\mathscr{M}}(t)$ on the line $G\underset{\sim}{\mathscr{M}}(t)$ minimizes $\|\underset{\sim}{l}(t) - \underset{\sim}{\mathscr{M}}(t)\|$ among all $\underset{\sim}{l} \in \underset{\sim}{L}(l^*)$ satisfying $t \leqq \tau_0(\underset{\sim}{l}, \mathscr{M}^*)$. But under that trajectory, $\rho = (w - 1)/w$ and $\theta = 0$ when L is at G. Hence, $\|\underset{\sim}{l}(t) - \underset{\sim}{\mathscr{M}}(t)\| \geqq v^*$ for $t \geqq t_0$.

Let $(l, \mathscr{M})$ be an AP, where $|\theta| \leqq \beta$. It follows from the remarks preceding Lemma 3 that $(l, \mathscr{M})$ is a BAP unless its (x, y)-representation lies in region $\mathscr{G}$ (see Fig. 3). The next lemma applies to that region.

LEMMA 4. *If* $(l, \mathscr{M})$ *is a position where* $(x, y) \in \mathscr{G}$, *then* $\min_{\underset{\sim}{l} \in \underset{\sim}{L}(l)} Q(\underset{\sim}{l}, \mathscr{M}) = W(x, y)$ (see (6)).

Proof. By the arguments of Lemma 3, we need only consider straight-line trajectories. By Lemma 2, we only have to consider paths along which $d\theta/dt > 0$. From (1)

$$d\rho/dt = w\sin\theta - \cos\phi,$$

$$d\theta/dt = (1/\rho)(w(\cos\theta - \rho) - \sin\phi)$$

and

$$\frac{d\rho}{d\theta} = F(\phi) = \left(\frac{d\rho}{dt}\right)\Big/\left(\frac{d\theta}{dt}\right).$$

The last expression is, of course, defined since $d\theta/dt > 0$. We want a function $\phi(t)$ which minimizes $F(\phi)$ for all t in $[0, \tau_0(\underset{\sim}{l}, \mathscr{M})]$. By Lemma 3 we can assume that $(\rho, \theta) \neq ((w-1)/w, 0)$.

Using routine but lengthy calculus arguments we can show that $\phi(t)$ satisfying (5) attains the minimum.

Now we can prove Theorem 2.

Proof. Let $(l, \mathscr{M})$ be any position with representation (x, y). For $(x, y) \in \mathscr{G}$, the result follows from (10), Lemmas 3 and 4, and the fact that $v^* \geqq W(x, y)$. For $(x, y) \in \mathscr{S}$, the result follows from (10) and Lemmas 2 and 3. Finally, Lemma 2 and the remarks preceding Lemma 4 imply that

$$\mathscr{R} = \{\text{NAP}\} \cup \{\text{BAP}\},$$

taking care of $(x, y) \in \mathscr{R}$.

5. Pursuit. Let $(l, \mathscr{M}) \in D \times C$ have the representation (x, y). We will construct a mapping π^* (or more precisely $\pi^*[l, \mathscr{M}]$) that satisfies

$$\pi^* \in \Pi(l, \mathscr{M}) \tag{11}$$

and

$$\sup_{\underset{\sim}{\mathscr{M}} \in \underset{\sim}{M}(\mathscr{M})} P(\pi^*(\underset{\sim}{\mathscr{M}}), \underset{\sim}{\mathscr{M}}) = W(x, y) \tag{12}$$

(see Theorem 3).

The construction presents no problems when $(x, y) \notin \mathscr{G}$. Certainly any strategy will suffice when $(x, y) \in \mathscr{S}$. When $(x, y) \in \mathscr{R}$ let L move to the center O. Once at O let him employ the strategy π^*, where

$$\pi^*(\underset{\sim}{\mathscr{M}})(t) = \begin{cases} (1/w)(\sin wt)(\underset{\sim}{\mathscr{M}}(t)), & 0 \leqq t \leqq \pi/2w, \\ (1/w)\underset{\sim}{\mathscr{M}}(t), & t \geqq \pi/2w, \end{cases} \tag{13}$$

for $\underset{\sim}{\mathscr{M}} \in \underset{\sim}{M}$. Under π^*, L arrives at a RAP at time $t = \pi/2w$. Note that (13) is written in standard polar coordinate notation. The reader should verify that $\pi^* \in \Pi$. We have reduced $(x, y) \in \mathscr{R}$ to $\rho = (w-1)/w$ and $\theta = 0$. We break $(x, y) \in \mathscr{G}$ into three cases.

Case 1.

$$\rho < (w-1)/w \quad \text{and} \quad \theta = 0. \tag{14}$$

We return to the usual rectangular coordinate system. Rotate coordinates so that $l = (l_1, l_2) = (1-\rho, 0)$ and $\mathscr{M} = (\mathscr{M}_1, \mathscr{M}_2) = (1, 0)$. Define $\tau_1(\underset{\sim}{l}) = \inf\{t \mid \|\underset{\sim}{l}(t)\| = 1\}$ for $\underset{\sim}{l} \in \underset{\sim}{L}(l)$. $\tau_1(\underset{\sim}{l})$ is the first time that $\underset{\sim}{l}$ hits the boundary. Let $\underset{\sim}{\mathscr{M}} = (\underset{\sim}{\mathscr{M}}_1, \underset{\sim}{\mathscr{M}}_2) \in \underset{\sim}{M}(\mathscr{M})$.

For Case 1 we define π^* by

$$\pi^*(\underset{\sim}{\mathscr{M}}) = \underset{\sim}{l}^{\rho} = (\underset{\sim}{l}_1^{\rho}, \underset{\sim}{l}_2^{\rho}), \tag{15}$$

where

$$\underset{\sim}{l}_1^\rho(t) = \begin{cases} 1 - \rho + (1/w(1 - \rho))(w^2(1 - \rho)^2 - 1)^{1/2}t, & 0 \leqq t \leqq \tau_1(\underset{\sim}{l}^\rho), \\ \underset{\sim}{l}_1^\rho(\tau_1(\underset{\sim}{l}^\rho)), \quad t \geqq \tau_1(\underset{\sim}{l}^\rho), & \end{cases}$$

and

$$\underset{\sim}{l}_2^\rho(t) = \begin{cases} (1/w^2(1 - \rho)) \arcsin(\underset{\sim}{\mathscr{M}}_2(t)), & 0 \leqq t \leqq \tau_1(\underset{\sim}{l}^\rho), \\ \underset{\sim}{l}_2^\rho(\tau_1(\underset{\sim}{l}^\rho)), \quad t \geqq \tau_1(\underset{\sim}{l}^\rho). & \end{cases}$$

Turn to Fig. 2. Represent L's starting position by E and M's by A. If L uses π^* and M moves at full speed towards C_1 or C_2, then L moves at full speed towards F_1 or F_2, respectively. The distance $|LM|$ is minimized when M reaches F_1 or F_2 (see § 1, paragraph 10).

LEMMA 5. π^* (Case 1) *satisfies* (11) *and* (12).

Proof. By using $\rho < (w - 1)/w$, one can show that

$$\left(\frac{d\underset{\sim}{l}_1^\rho(t)}{dt}\right)^2 + \left(\frac{d\underset{\sim}{l}_2^\rho(t)}{dt}\right)^2 \leqq 1,$$

where $\underset{\sim}{l}_1^\rho$ and $\underset{\sim}{l}_2^\rho$ satisfy (15). Hence (11) holds. We assert that the best that M can do against π^* when (14) holds is to go to F_1 or F_2 at full speed. This follows from a direct geometric argument which uses the fact that $d\underset{\sim}{l}_1^\rho(t)/dt > 0$ when $\rho < (w - 1)/w$. The rest follows from the remarks preceding this lemma and paragraphs 10 through 12 of § 1.

Case 2.

$$(16) \qquad\qquad (x, y) \in \mathscr{G} \quad \text{and} \quad \theta \neq 0.$$

We assume $(x, y) \in \mathscr{G}_1$ (see Fig. 3); the case $(x, y) \in \mathscr{G}'_1$ is similar. Turn to Fig. 4. Let E denote L's starting position and A, M's. Denote by $\underset{\sim}{l}^{\rho,\theta}$ the trajectory, where L moves at full speed along the line segment EF. Let $\underset{\sim}{\mathscr{M}} \in \underset{\sim}{M}(\mathscr{M})$. Define

$$\tau_2(\underset{\sim}{\mathscr{M}}) = \inf\{t | \underset{\sim}{l}^{\rho,\theta}(t) \text{ lies on the radius } O\underset{\sim}{\mathscr{M}}(t)\}.$$

Clearly, $(\underset{\sim}{l}^{\rho,\theta}(\tau_2(\underset{\sim}{\mathscr{M}})), \underset{\sim}{\mathscr{M}}(\tau_2(\underset{\sim}{\mathscr{M}})))$ satisfies (14) whenever $\tau_2(\underset{\sim}{\mathscr{M}}) < \infty$ (see Case 1).

Now we define π^* for Case 2. Let $\pi^*(\underset{\sim}{\mathscr{M}}) = \underset{\sim}{l}^{\rho,\theta}$ if $\tau_2(\underset{\sim}{\mathscr{M}}) = \infty$. Otherwise, let $\pi^*(\underset{\sim}{\mathscr{M}})(t) = \underset{\sim}{l}^{\rho,\theta}(t)$ for $0 \leqq t \leqq \tau_2(\underset{\sim}{\mathscr{M}})$. Then, at time $\tau_2(\underset{\sim}{\mathscr{M}})$, let π^* select a trajectory according to the procedure specified by (15) for Case 1.

If L uses π^* and M uses η^*, then M moves at full speed along the minor arc AC while L moves at full speed along the line segment EF (see Fig. 4). By using π^*, L benefits from any deviations on M's part.

LEMMA 6. π^* (Case 2) *satisfies* (11) *and* (12).

Proof. (11) is immediate. For (12), observe that L does better when $\tau_2(\underset{\sim}{\mathscr{M}}) < \infty$. The remarks preceding this lemma and paragraphs 11 and 12 of § 1 take care of (12) when $\tau_2(\underset{\sim}{\mathscr{M}}) = \infty$.

Case 3.

$$(17) \qquad\qquad \rho = (w - 1)/w \quad \text{and} \quad \theta = 0 \quad \text{(RAP's)}.$$

RAP's present special problems. Our first impulse is to define a strategy by (15). Unfortunately, $\underset{\sim}{l}_1^\rho(t) = 1/w$ for all t when $\rho = (w - 1)/w$, so there is nothing to prevent M from staying within a small arc about $\mathcal{M}$. We get around this problem by defining an appropriate sequence of strategies.

Because of (13) we can assume that $w > 1$, since if $w = 1$ then (13) implies that $v^* = 0$. Let k be the smallest integer $\geqq 1/(w-1)$. For each integer $n \geqq k$ we will find a strategy π^n and a real number t_n that satisfy

$$\min_{0 \leqq t \leqq t_n} \|\pi^n(\underset{\sim}{\mathcal{M}})(t) - \underset{\sim}{\mathcal{M}}(t)\| \leqq v^* + 1/n \tag{18}$$

for every $\underset{\sim}{\mathcal{M}} \in \underset{\sim}{M}(\mathcal{M})$. Given such objects we can construct an optimal strategy π^* as follows.

L begins with π^k, then, at time t_k, he heads toward the center O. Once at O, he returns to a RAP by employing (13). Observe that the first return to a RAP takes less than $t_k + 1 + \pi/2w$ time units. After L's nth return to a RAP, he employs π^{n+k} for t_{n+k} units of time, heads back toward O and then, using (13), returns to a RAP for the $(n+1)$st time. As before, the $(n+1)$st return to a RAP takes less than $t_{n+k} + 1 + \pi/2w$ time units. Evidently, π^* satisfies (11). That π^* also satisfies (12) follows easily from

$$\min_{0 \leqq t \leqq A_n} \|\pi^*(\underset{\sim}{\mathcal{M}})(t) - \underset{\sim}{\mathcal{M}}(t)\| \leqq v^* + 1/n$$

for every $\underset{\sim}{\mathcal{M}} \in \underset{\sim}{M}(\mathcal{M})$, where

$$A_n = \sum_{j=k}^{n} (t_j + 1 + \pi/2w) < \infty .$$

Our construction of π^n involves allowing a delay of time $1/nw$ in L's information about M's position. Suppose M chooses $\underset{\sim}{\mathcal{M}} \in \underset{\sim}{M}(\mathcal{M})$. Define $\underset{\sim}{\mathcal{M}}^n$ by

$$\underset{\sim}{\mathcal{M}}^n(t) = \begin{cases} \mathcal{M} & \text{for } 0 \leqq t \leqq 1/nw, \\ \underset{\sim}{\mathcal{M}}(t - 1/nw) & \text{for } t \geqq 1/nw. \end{cases}$$

Under the strategy π^n, L starts by moving at full speed towards the point $\mathcal{M}$. He continues on this path until time $1/nw$ when he reaches a point on the radius $O\mathcal{M}$ whose distance from $\mathcal{M}$ is $(w-1)/w - 1/nw$. Thereafter, he follows the trajectory which pursues $\underset{\sim}{\mathcal{M}}^n$ according to the rule (15).

LEMMA 7. π^* (Case 3) *satisfies* (11) *and* (12).

Proof. We have to show that π^n satisfies (18) for some finite t_n. Using (15) and Lemma 5 one can show that

$$\min_{0 \leqq t \leqq t_n} \|\underset{\sim}{L}(t) - \underset{\sim}{\mathcal{M}}^n(t)\| < v^*,$$

where

$$t_n = 1/n + (n+1)/(1+2n)^{1/2}.$$

The result follows from

$$\|\underset{\sim}{\mathcal{M}}(t) - \underset{\sim}{\mathcal{M}}^n(t)\| \leqq 1/n$$

and the triangle inequality.

Remark 3. For large n, the paths that result when M and L start at a RAP and use π^n and η^*, respectively, converge uniformly to the paths described in paragraphs 7 and 8 of § 1.

Now we can prove Theorem 3.

Proof. The theorem follows from the remarks of the second paragraph of this section and Lemmas 5, 6 and 7.

6. Examples. Does L have a strategy $\hat{\pi}$, allowing him to stay on the radius OM and keep his distance from M equal to $(w - 1)/w$? Does M have a strategy $\hat{\eta}$, which prevents L from maintaining a position on OM whose distance from M is $(w - 1)/w$? As we shall see in Example 1, the answer to both questions is "Yes" (see Remark 1). In Example 2 we define a payoff function $\hat{P}$ for which

$$\sup_{\underset{\sim}{\mathscr{M}} \in \underset{\sim}{M}(\mathscr{M})} P(\hat{\pi}(\underset{\sim}{\mathscr{M}}), \underset{\sim}{\mathscr{M}}) < \inf_{\underset{\sim}{l} \in \underset{\sim}{L}(l)} P(\underset{\sim}{l}, \hat{\eta}(\underset{\sim}{l}))$$

(see Remark 2).

Example 1. Let $l = (l_1, l_2) = (1/w, 0)$ and $\mathscr{M} = (\mathscr{M}_1, \mathscr{M}_2) = (1, 0)$. Define $\hat{\pi} \in \Pi(l, \mathscr{M})$ by $\hat{\pi}(\underset{\sim}{\mathscr{M}})(t) = (1/w)\underset{\sim}{\mathscr{M}}(t)$ for $\underset{\sim}{\mathscr{M}} \in \underset{\sim}{M}(\mathscr{M})$. The definition of $\hat{\eta} \in H(l, \mathscr{M})$ is more involved. For $\underset{\sim}{l} \in \underset{\sim}{L}(l)$, let

$$\sigma'(\underset{\sim}{l}) = \sup\{t \mid \|\underset{\sim}{l}(s)\| = 1/w \text{ for all } 0 \leqq s \leqq t\}$$

and

$$\sigma''(\underset{\sim}{l}) = \sup\{t \mid \underset{\sim}{l}(s) = l \text{ for all } 0 \leqq s \leqq t \leqq \sigma'(\underset{\sim}{l})\}.$$

Define $\hat{\eta}$ by

$$\hat{\eta}(\underset{\sim}{l})(t) = \begin{cases} (\cos wt, \sin wt) & \text{for } \sigma''(\underset{\sim}{l}) > 0 \\ \text{and } t \geqq 0, & \\ (w\underset{\sim}{l}_1(t), -w\underset{\sim}{l}_2(t)) & \text{for } \sigma''(\underset{\sim}{l}) = 0 \\ \text{and } 0 \leqq t \leqq \sigma'(\underset{\sim}{l}), & \\ (w\underset{\sim}{l}_1(\sigma'(\underset{\sim}{l})), -w\underset{\sim}{l}_2(\sigma'(\underset{\sim}{l}))) & \\ \text{for } \sigma''(\underset{\sim}{l}) = 0 \quad \text{and} \quad t \geqq \sigma'(\underset{\sim}{l}). & \end{cases}$$

We have the following theorem (see Remark 1).

THEOREM 5. *There do not exist $\underset{\sim}{l} \in \underset{\sim}{L}(l)$ and $\underset{\sim}{\mathscr{M}} \in \underset{\sim}{M}(\mathscr{M})$ satisfying $\hat{\eta}(\underset{\sim}{l}) = \underset{\sim}{\mathscr{M}}$ and $\hat{\pi}(\underset{\sim}{\mathscr{M}}) = \underset{\sim}{l}$.*

Proof. Suppose that there exists such an $\underset{\sim}{l}$. Then $\sigma''(\underset{\sim}{l}) = 0$ since $\sigma''(\underset{\sim}{l}) > 0$ implies that $\underset{\sim}{l}(t) = l$ *and* $\underset{\sim}{l}(t) = (1/w)(\cos wt, \sin wt)$ for $0 \leqq t \leqq \sigma''(\underset{\sim}{l})$. Also $\sigma'(\underset{\sim}{l}) = \infty$ since $\underset{\sim}{l}(t) = (1/w)\underset{\sim}{\mathscr{M}}(t)$. But $\sigma''(\underset{\sim}{l}) = 0$ and $\sigma'(\underset{\sim}{l}) = \infty$ imply that $\underset{\sim}{l}(t) = (\underset{\sim}{l}_1(t), -\underset{\sim}{l}_2(t))$ or, equivalently, $\underset{\sim}{l}_2(t) = 0$ for all $t \geqq 0$. But the latter and $\sigma'(\underset{\sim}{l}) = \infty$ imply that $\sigma''(\underset{\sim}{l}) = \infty$: a contradiction.

Compare Example 1 with Isaacs' "perpetuated dilemma" [3, pp. 137, 149].

Example 2. Take $l, \mathscr{M}, \hat{\pi}$ and $\hat{\eta}$ as in Example 1. For $\underset{\sim}{l} \in \underset{\sim}{L}(l)$ and $\underset{\sim}{\mathscr{M}} \in M(\underset{\sim}{\mathscr{M}})$, let

$$\hat{P}(\underset{\sim}{l}, \underset{\sim}{\mathscr{M}}) = \begin{cases} 0, & \text{if } \underset{\sim}{l}(t) = (1/w)\underset{\sim}{\mathscr{M}}(t) \text{ for all } t \geqq 0, \\ 1, & \text{otherwise.} \end{cases}$$

Clearly, $\hat{P}(\hat{\pi}(\underset{\sim}{\mathscr{M}}), \underset{\sim}{\mathscr{M}}) = 0$ for all $\underset{\sim}{\mathscr{M}} \in \underset{\sim}{M}(\mathscr{M})$ while $\hat{P}(\underset{\sim}{l}, \hat{\eta}(\underset{\sim}{l})) = 1$ for all $\underset{\sim}{l} \in \underset{\sim}{L}(l)$.

7. Pursuit in the half-plane. Replacing the circular arena with a half-plane leads to a second problem, a version of which appears in Isaacs [3, pp. 260–265]. We are going to discuss the relationships between these two problems.

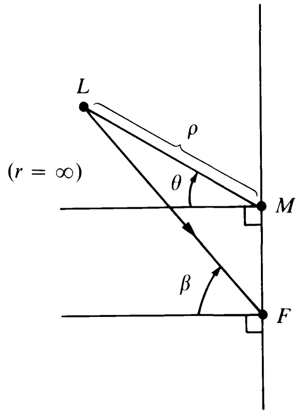


FIG. 6

The most striking difference is illustrated by the following fact, which, according to Dubins, was discovered by Blackwell. In *pursuit in the half-plane* the players can ignore trajectories which lead to an increase in their distance from the boundary. Hence, if play starts with the evader on the boundary he remains there. This is not true for *pursuit in the circle.* One can show that in the latter game the evader can do better by traveling along small chords than by staying on the circumference.

Now let us compare the game where the man is *restricted* to the boundary of a circular arena with the game where the man is restricted to a half-plane and *starts on the boundary.* Denote the radius of the arena by r where $r = \infty$ corresponds to the case where the arena is a half-plane. Represent the starting positions by (ρ, θ) where ρ is the distance $|LM|$ and θ is the angle that the line LM makes with the normal to the boundary at M (see Fig. 6). Denote by $V_r(\rho, \theta)$ the value corresponding to the point (ρ, θ) and the radius r.

The game in the half-plane is closely related to the game in region $\mathcal{G}$ of the circle. Except for the case $w = 1$, the strategies which are optimal for the half-plane are direct generalizations of the strategies which are optimal in region $\mathcal{G}$. For example, given the initial position in Fig. 6, M's optimal strategy would direct him downward at full speed while L's corresponding trajectory would direct him at full speed toward F. In the rather odd case $w = 1$, L has no optimal strategy in the half-plane, and region $\mathcal{G}$ of the circle consists of the point M. By improving on the heuristic arguments of Flynn [1], one can prove the following theorem.

THEOREM 6. *If* $0 < r < \infty$, *then* $V_r(\rho, \theta) = rV_1(\rho/r, \theta)$. *Moreover,*

$$\lim_{r \to \infty} V_r(\rho, \theta) = V_\infty(\rho, \theta) = \begin{cases} \rho \cos(\beta - |\theta|) & \textit{for } |\theta| \leqq \beta, \\ \rho \quad \textit{for } |\theta| \geqq \beta. \end{cases}$$

REFERENCES

[1] J. FLYNN, *Some bounded pursuit games with simple motion*, Tech. Rep. 153, Statistics Department, Stanford University, Stanford, Calif., 1970.

[2] B. HALPERN, *The robot and the rabbit—a pursuit problem*, Amer. Math. Monthly, 76 (1969), pp. 140–144.

[3] R. ISAACS, *Differential Games*, John Wiley, New York, 1965.

[4] J. LITTLEWOOD, *Lion and Man. A Mathematician's Miscellany*, Methuen, London, 1963.

[5] L. PONTRYAGIN, *On the theory of differential games*, Russian Math. Surveys, 21 (1966), pp. 193–246.

[6] C. RYLL-NARDZEWSKI, *A Theory of Pursuit and Evasion*: *Advances in Game Theory*, Princeton University Press, Princeton, 1964.

[7] GERALD J. SMITH, *A pursuit evasion game*, unpublished manuscript, 1968.

[8] P. VARAIYA AND J. LIN, *Existence of saddle points in differential games*, this Journal, 7 (1969), pp. 141–157.

SIAM J. Control
Vol. 11, No. 3, August 1973

A UNIQUENESS THEOREM FOR LINEAR CONTROL SYSTEMS WITH COINCIDING REACHABLE SETS*

M. L. J. HAUTUS† and G. J. OLSDER‡

Abstract. Two multivariable linear control systems are considered with control $u(t)$ satisfying the inequality $\|u(t)\|_p \leqq 1, 1 \leqq p \leqq \infty$, and with coinciding reachable sets. Under certain conditions (of which $p \neq 2$ seems to be the most remarkable), it is shown that the control systems have equal system matrices and equal control matrices up to the signs and the ordering of the columns. The proof depends on a theorem of Banach on rotations.

1. Introduction and the results.

Consider the control system described by the vector differential equation

$$\dot{x}(t) = Ax(t) + Bu(t), \tag{1}$$

where x is an n-vector, u is an r-vector and A and B are constant matrices of appropriate sizes. The control $u(t)$ is constrained to lie in a set Ω_p. The set Ω_p is defined as the set of all measurable vector functions $v(t)$ with $\|v\|_p \leqq 1, 1 \leqq p \leqq \infty$, where $\|\cdot\|_p$ denotes the l_p-norm, i.e.,

$$\|v\|_p = \left\{ \sum_{i=1}^{r} |v_i|^p \right\}^{1/p} \quad \text{for } 1 \leqq p < \infty \tag{2a}$$

and

$$\|v\|_\infty = \max_{1 \leqq i \leqq r} |v_i|, \tag{2b}$$

and where v_i are the components of vector v.

In order to define the reachable set, a coordinate transformation $x \to y$ will be made from $R^n \to R^n$ by

$$y = e^{-tA}x.$$

In y-space, differential equation (1) becomes

$$\dot{y} = e^{-tA}Bu(t).$$

The reachable set $R_p(t)$ is defined as the set of all points in y-space that can be reached at time t if $y(0) = 0$ by using all possible control functions:

$$R_p(t) = \left\{ y \middle| y = \int_0^t e^{-\tau A}Bu(\tau)\,d\tau,\ u \in \Omega_p \right\}, \qquad t \geqq 0.$$

It can be shown [4, p. 107] that for each $t \geqq 0$ the reachable set is convex and compact.

We will consider another linear control system,

$$\dot{\bar{x}}(t) = \bar{A}\bar{x}(t) + \bar{B}\bar{u}(t), \tag{3}$$

* Received by the editors April 25, 1972.

† Department of Mathematics, Eindhoven University of Technology, Eindhoven, The Netherlands.

‡ Department of Applied Mathematics, Twente University of Technology, Enschede, The Netherlands.

where $\bar{x}$ is an $\bar{n}$-vector and $\bar{u}$ is an $\bar{r}$-vector. Matrices $\bar{A}$ and $\bar{B}$ have constant elements. The control $\bar{u}(t)$ will belong to $\bar{\Omega}_p$. The set $\bar{\Omega}_p$ is defined in the same way as Ω_p; the only difference is that in this case r must be replaced by $\bar{r}$ in (2a) and (2b). The reachable set $\bar{R}_p(t)$ for system (3) is defined in an analogous way as $R_p(t)$ for system (1).

The results of this paper are given in the following theorem.

THEOREM 1. *Consider systems* (1) *and* (3) *with* $u \in \Omega_p, \bar{u} \in \bar{\Omega}_p$ *respectively and* $1 \leqq p \leqq \infty$. *If*

(i) $p \neq 2$,

(ii) $\operatorname{rank}(B) = r$, $\operatorname{rank}(\bar{B}) = \bar{r}$,

(iii) $n = \bar{n}$,

(iv) *systems* (1) *and* (3) *are controllable* [4],

(v) $R_p(t) = \bar{R}_p(t)$ *for* $0 \leqq t < \delta$, δ *being some positive number*,

then $r = \bar{r}$, $A = \bar{A}$ *and there is a one-to-one correspondence between the columns* b_i *of* B *and the columns* $\bar{b}_i$ *of* $\bar{B}$, *viz.*, $b_i = \pm \bar{b}_{j_i}$, $i = 1, \cdots, r; \{j_1, \cdots, j_r\} = \{1, \cdots, r\}$.

Remark 1. If in x-space the minimum time needed to steer system (1) from a point x_0 to the origin $x = 0$ by means of a control $u \in \Omega_p$ is given by $T(x_0)$, and if $\bar{T}(x_0)$ corresponds in the same way to system (3), then condition (v) of Theorem 1 can be read as $T(x_0) = \bar{T}(x_0)$ for all x_0 in some neighborhood of $x = 0$. By a slight modification of the proof of Theorem 1, one can show that condition (v) can be weakened to

(v′) $R_p(t) = \bar{R}_p(t)$ for $t \in S$, where S is a set of real numbers with an accumulation point.

The proof of Theorem 1, which will be given in § 3, rests on a theorem of Banach on rotations [1].

Theorem 1 is a generalization of a result due to Hájek [3]. Hájek only considers systems of the kind (1) and (3) with one control component (i.e., $r = \bar{r} = 1$) and with $p = \infty$. His proof is quite different from ours. Hájek considers the particular shape of the reachable sets. In fact he only uses those boundary points of these sets where the outward normals of all supporting hyperplanes at such a point span an $(n - 1)$-dimensional space.

2. A theorem of Banach and a preliminary result.

DEFINITION 1. An $m \times m$-matrix Z is called a *q-isometry*, with $1 \leqq q \leqq \infty$, if for all m-vectors ζ,

$$\|\zeta' Z\|_q = \|\zeta'\|_q.$$

($'$ denotes the transpose and for row vectors $\|\cdot\|_q$ also denotes the l_q-norm.)

THEOREM 2. *A q-isometry is regular. If* $q \neq 2$, *then in every column and every row of such a q-isometry only one element* $\neq 0$ *appears. These nonzero elements equal* ± 1.

The proof will be omitted because Theorem 2 with $1 \leqq q < \infty$ is a special case of a theorem of Banach on rotations [1, pp. 178, 179]. The proof for the case $q = \infty$ is immediate. Note that a 2-isometry is an orthogonal matrix.

LEMMA 1. *Suppose we are given two matrices* G *and* $\bar{G}$ *of sizes* $n \times r$ *and* $n \times \bar{r}$ *respectively with* $\operatorname{rank}(G) = r$ *and* $\operatorname{rank}(\bar{G}) = \bar{r}$. *If for all n-vectors* η *and some*

q with $1 \leqq q \leqq \infty$,

$$\|\eta' G\|_q = \|\eta' \bar{G}\|_q, \tag{4}$$

then $r = \bar{r}$ and $\bar{G} = GZ$, where Z is a q-isometry.

Proof. It easily follows from (4) that the columns of G and $\bar{G}$ span the same linear space, which will be denoted by S. Because the column vectors of $\bar{G}$ are linearly independent, as are the column vectors of G, we have $r = \bar{r}$.

G and $\bar{G}$ will be considered in another coordinate system. Let $\{e_1, \cdots, e_n\}$ be a new basis in such a way that $E = [e_1, \cdots, e_n]$ is an orthogonal matrix and in addition that $e_1, \cdots, e_r$ span S. In the new basis G and $\bar{G}$ are transformed to H and $\bar{H}$:

$$G = EH, \quad \bar{G} = E\bar{H}.$$

Note that the last $n - r$ rows of H and $\bar{H}$ have only zero elements. If the original vector is denoted by v in the new basis (i.e., $\eta = Ev$), then

$$\|\eta' G\|_q = \|v' H\|_q = \|\tilde{v}' \tilde{H}\|_q, \qquad \|\eta' \bar{G}\|_q = \|v' \bar{H}\|_q = \|\tilde{v}' \tilde{\bar{H}}\|_q, \tag{5}$$

where $\tilde{H}$ and $\tilde{\bar{H}}$ denote the square nonsingular matrices consisting of the first r rows of H and $\bar{H}$ respectively, and where $\tilde{v}$ is the r-vector consisting of the first r elements of v. Equations (4) and (5) yield

$$\|\tilde{v}' \tilde{H}\|_q = \|\tilde{v}' \tilde{\bar{H}}\|_q$$

for arbitrary $\tilde{v}$, or, if $\tilde{v}' \tilde{H} = \xi'$,

$$\|\xi'\|_q = \|\xi' (\tilde{H})^{-1} \tilde{\bar{H}}\|_q$$

for arbitrary r-vectors ξ. According to Definition 1, the matrix $(H)^{-1}\tilde{\bar{H}}$ is a q-isometry, to be denoted by Z. Hence

$$\tilde{\bar{H}} = HZ,$$

and

$$\bar{G} = E\bar{H} = \begin{bmatrix} E \end{bmatrix} \begin{bmatrix} \tilde{\bar{H}} \\ 0 \end{bmatrix} = \begin{bmatrix} E \end{bmatrix} \begin{bmatrix} \tilde{H}Z \\ 0 \end{bmatrix} = EHZ = GZ,$$

which proves the lemma.

3. Proof of Theorem 1. It follows from the equality of the reachable sets and the fact that these sets are compact that

$$\max_{u \in \Omega_p} \int_0^t \eta' e^{-\tau A} B u(\tau)\, d\tau = \max_{\bar{u} \in \bar{\Omega}_p} \int_0^t \eta' e^{-\tau \bar{A}} \bar{B} \bar{u}(\tau)\, d\tau \tag{6}$$

for an arbitrary n-vector η and $0 \leqq t < \delta$. Equation (6) can be rewritten as

$$\int_0^t \max_{\|v\|_p \leqq 1} \eta' e^{-\tau A} B v\, d\tau = \int_0^t \max_{\|\bar{v}\|_p \leqq 1} \eta' e^{-\tau \bar{A}} \bar{B} \bar{v}\, d\tau. \tag{7}$$

It is well known that the maximum of $\eta' e^{-\tau A} B v$ with respect to v and subject to $\|v\|_p \leqq 1$ equals [5, pp. 29, 30] $\|\eta' e^{-\tau A} B\|_q$, where $q = p/(p - 1)$, i.e., $p^{-1} + q^{-1} = 1$. Hence (7) reads

$$\int_0^t \|\eta' e^{-\tau A} B\|_q\, d\tau = \int_0^t \|\eta' e^{-\tau \bar{A}} \bar{B}\|_q\, d\tau,$$

from which

$$\|\eta' e^{-\tau A}B\|_q = \|\eta' e^{-\tau \bar{A}}\bar{B}\|_q, \qquad 0 \leqq \tau < \delta,$$

and this formula is valid for all n-vectors η. Because it was assumed that rank $(B) = r$, rank $(\bar{B}) = \bar{r}$, it follows that rank $(e^{-\tau A}B) = r$, rank $(e^{-\tau \bar{A}}\bar{B}) = \bar{r}$ and hence Lemma 1 can be applied resulting in $r = \bar{r}$ and

$$e^{-\tau A}B = e^{-\tau \bar{A}}\bar{B}Z(\tau), \qquad 0 \leqq \tau < \delta,$$

where $Z(\tau)$ is a q-isometry for each $\tau \in [0, \delta)$.

From Theorem 2 it follows that for $p \neq 2$ and hence $q \neq 2$ only a finite number of different q-isometries of the same size exists. A sequence of distinct numbers $\tau_i \in [0, \delta)$ can be found in such a way that

$$e^{-\tau A}B = e^{-\tau \bar{A}}\bar{B}Z_0 \quad \text{for } \tau = \tau_1, \tau_2, \cdots,$$

for some q-isometry Z_0. The functions $e^{-\tau A}B$ and $e^{-\tau \bar{A}}\bar{B}Z_0$ are analytic in τ and are equal for $\tau = \tau_1, \tau_2, \cdots$. The series $\{\tau_i\}$ has at least one accumulation point and hence,

$$e^{-tA}B = e^{-t\bar{A}}\bar{B}Z_0 \quad \text{for all } t.$$

Equivalently,

$$e^{tA}B = e^{t\bar{A}}\bar{B}Z_0 \quad \text{for all } t.$$

The function $e^{tA}B$ is the weighting pattern [2] of the following control system:

$$\begin{aligned} \dot{x} &= Ax + Bu, \\ z &= Cx, \quad C = I, \end{aligned} \tag{8}$$

where I denotes the identity matrix. In the same way, $e^{t\bar{A}}\bar{B}Z_0$ corresponds to

$$\begin{aligned} \dot{x} &= \bar{A}x + \bar{B}Z_0u, \\ z &= \bar{C}x, \quad \bar{C} = I. \end{aligned} \tag{9}$$

Systems (8) and (9) are controllable and observable and hence are minimal realizations of their weighting patterns. Because these patterns are equal, a constant nonsingular matrix P exists in such a way that [2, p. 113]

$$PAP^{-1} = \bar{A}, \qquad PB = \bar{B}Z_0, \qquad CP^{-1} = \bar{C}. \tag{10}$$

The last equation (10) reads $P = I$ and hence, $A = \bar{A}, B = \bar{B}Z_0$. Application of Theorem 2 yields a one-to-one correspondence between the columns of B and the columns of $\bar{B}$ as mentioned in the statement of Theorem 1. This completes the proof.

Remark 2. Theorem 1 is not true for $p = 2$. An example of two systems for which the conditions of Theorem 1 are satisfied (except $p \neq 2$) in such a way that the results are not true, is:

$n = \bar{n} = r = \bar{r} = p = 2;$

$$A = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}, \quad B = \begin{bmatrix} 1 & 1 \\ -1 & 1 \end{bmatrix}, \quad \bar{A} = \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix}, \quad \bar{B} = \begin{bmatrix} \sqrt{2} & 0 \\ 0 & \sqrt{2} \end{bmatrix}.$$

One can easily convince oneself that conditions (ii), (iii), (iv) and (v) of Theorem 1 are satisfied for these systems.

Acknowledgment. The authors are indebted to J. J. A. M. Brands for valuable suggestions for the proof of Theorem 1.

REFERENCES

[1] S. BANACH, *Théorie des opérations linéaires*, Chelsea, New York, 1963.

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

[3] O. HÁJEK, *Identification of control systems by performance*, Math. Systems Theory, 5 (1971), pp. 349–352.

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

[5] D. G. LUENBERGER, *Optimization by Vector Space Methods*, John Wiley, New York, 1969.

SIAM J. Control
Vol. 11, No. 3, August 1973

GLOBAL CONTROLLABILITY OF LINEAR SYSTEMS WITH POSITIVE CONTROLS*

STEPHEN H. SAPERSTONE†

Abstract. A linear autonomous differential control system is considered, where the zero control is an extreme point of the restraint set. Necessary and sufficient conditions are given for global controllability in the case of bounded or unbounded scalar control.

1. Introduction. In a previous paper by the author and Yorke [5], necessary and sufficient conditions were given for local controllability of differential systems of the form

$$\dot{x} = Ax + bu, \tag{1.1}$$

where $x, b \in R^d$, A is a real constant $d \times d$ matrix, and $u:R^+ \rightarrow \Omega = [0, 1]$ is measurable. R^+ denotes the nonnegative real numbers.

The standard results on controllability (cf. [3]) depend on the assumption that the zero control is interior to Ω. The elimination of this assumption leads to, in the case of scalar controls, the consideration of only positive-valued (or only negative-valued) controllers. This was first considered in [5]. It was found that such systems were locally controllable if and only if they were of an "oscillatory" nature, and satisfied the usual algebraic controllability condition.

The present paper extends those results to obtain conditions for global controllability with both bounded and unbounded positive controls. That is, we characterize those systems of the form (1.1) with bounded positive control ($\Omega = [0, 1]$) so that the origin can be steered to any point in R^d. We also obtain (Lemma 5.6) sufficient conditions for this in the event Ω does not even contain the zero controller as a boundary point.

2. Definitions and background. Let Ω be an interval in R^+ and U_Ω denote the set of all measurable functions $u:R^+ \rightarrow \Omega$. Unless otherwise specified, $x(t) = x(t; u(\cdot))$ denotes the unique solution of (1.1) with initial condition $x(0) = 0$. Let $K_\Omega^+(t) \overset{\text{def}}{=} \{x(t; u(\cdot)) : u \in U_\Omega\}$ be the *reachable set at time* $t \geqq 0$, and $C(A, b) \overset{\text{def}}{=} [b, Ab, \cdots, A^{d-1}b]$ be the *controllability matrix* of (1.1).

DEFINITION 2.1. The system (1.1) is *locally controllable* at $x = 0$ if there exists $T < \infty$ such that $K_\Omega^+(T)$ contains a neighborhood of $x = 0$.

DEFINITION 2.2. The system (1.1) is *globally controllable* from $x = 0$ if $K_\Omega^+ \overset{\text{def}}{=} \bigcup_{t \geqq 0} K_\Omega(t) = R^d$.

From [5], we have the following.

THEOREM 2.3. *The system* (1.1) *with* $\Omega = [0, 1]$ *is locally controllable at* $x = 0$ *if and only if* rank $C(A, b) = d$ *and no eigenvalue of* A *is real.*

Henceforth we can assume that d is even in the event (1.1) is locally controllable with restraint set $\Omega = [0, 1]$.

* Received by the editors December 1, 1970, and in revised form August 17, 1972.

† Department of Mathematics, George Mason University, Fairfax, Virginia 22030.

3. Global controllability with unbounded positive controls. Theorem 2.3 has an immediate corollary if we admit unbounded controls.

COROLLARY 3.1. *Consider the system* (1.1) *with* $\Omega = [0, \infty)$. *There exists* $T < \infty$ *such that* $K^+_\Omega(T) = R^d$ *if and only if* (i) rank $C(A, b) = d$, *and* (ii) *no eigenvalue of* A *is real.*

Proof. If (i) and (ii) are satisfied, $K^+_{[0,1]}(T)$ contains a neighborhood of the origin. Replacing $[0, 1]$ by $[0, j], j = 1, 2, 3, \cdots$, we obtain by the linearity of (1.1), $K^+_{[0,j]}(T) = jK^+_{[0,1]}(T)$, where $jK^+_{[0,1]}(T)$ denotes the product of every response in $K^+_{[0,1]}(T)$ by j. Hence, $K^+_{[0,\infty)}(T) = \bigcup_{j=1}^{\infty} jK^+_{[0,1]}(T) = R^{2n}$. Conversely, the necessity of conditions (i) and (ii) follows easily from an argument similar to that used to prove Theorem 2.3. Q.E.D.

Thus for $\Omega = [0, \infty)$, the conditions in Corollary 3.1 characterize *uniform* global controllability at the origin. That is, there is a fixed time $T < \infty$ by which every point in R^d can be reached from the origin. Later on we will show if Ω is restricted to $[0, 1]$, we can only say that any point in R^d can be reached in finite time under suitable conditions. There is no uniform time T as in the case of $\Omega = [0, \infty)$. In fact, $K^+_{[0,1]}(t)$ is a proper subset of R^d for all $t \geqq 0$.

We now show that the minimum time to reach a neighborhood of the origin in the case when $\Omega = [0, 1]$ is the same minimum time to reach all of R^d in the case when $\Omega = [0, \infty)$.

COROLLARY 3.2. *Define*

$$T_1 = \inf\{t \in R^+ : 0 \in \operatorname{int} K^+_{[0,1]}(t)\},$$

$$T_\infty = \inf\{t \in R^+ : K^+_{[0,\infty)}(t) = R^d\}.$$

Then $T_1 = T_\infty$.

Proof. It is clear from the proof of Corollary 3.1 that $T_\infty \leqq T_1$. Now suppose $T_\infty < T_1$. Choose t so that $T_\infty < t < T_1$. Then $0 \in \partial K^+_{[0,1]}(t)$. We will show that $0 \in \partial K^+_{[0,\infty)}(t)$. This would imply that $K^+_{[0,\infty)}(t)$ is a proper subset of R^d, so $t \leqq T_\infty$, a contradiction.

For convenience, set $K_j = K^+_{[0,j]}(t)$ and $K_\infty = K^+_{[0,\infty)}(t)$. Since $0 \in \partial K_1, K_1$ lies on one side of a hyperplane with outer normal η through $x = 0$. Thus $\eta x \leqq 0$ for all $x \in K_1$. Now choose any $x_0 \in K_\infty$. By construction in the proof of the last corollary, $x_0 \in K_j$ for some positive integer j. Since $K_j = jK_1$, let $x_0 = jx, x \in K_1$. Then $\eta x_0 = \eta j x = j(\eta x) \leqq 0$. Hence, x_0 lies on the same side of η as does K_1. Thus $0 \in \partial K_\infty$.

Remark 3.3. Let $T_c = \inf\{t \in R^+ : 0 \in \operatorname{int} K^+_{[0,c]}(t)\}$. Then $T_c = T_\infty$ for all $c > 0$.

It is significant that the time to reach a neighborhood of the origin T_1 (or T_∞) need not be small. For example, $T_1 = \pi/\beta$ for the motion of a simple pendulum given by $\ddot{\theta} + \beta^2\theta = u$, for $u \in U_{[0,1]}$. Intuitively this makes sense, for if we desire to steer the pendulum from the origin $(\theta, \dot{\theta}) = (0, 0)$ to a point where θ is negative and $\dot{\theta}$ is positive, we would have to apply the controller u for some arbitrarily short period of time t_0 to set the pendulum in motion, and then wait a half period π/β before the natural return of the pendulum carries it to the desired point. In general, for a system of the form (1.1), the value of T_1 is difficult to compute, or even obtain good bounds for. (See [5, § 7] for an example.)

4. A lemma on almost periodic functions. Let f be a continuous real-valued almost periodic (a.p.) function defined on R^+ such that $f(t) \not\equiv 0$. The integral $M\{f\} \overset{\text{def}}{=} \lim_{T\to\infty} (1/T)\int_0^T f(t)\,dt$ exists and is called the *mean value of* f (cf. [1, p. 39]). The following lemma is needed to establish part of Theorem 5.1. For any interval $J \subset R^+$, let $l(J)$ denote the length of J.

LEMMA 4.1. *Let f be a real valued a.p. function defined on R^+ such that $f(t) \not\equiv 0$ and $M\{f\} = 0$. There exists a denumerable family of disjoint open intervals $\{J_k\}_{k=1}^{\infty}$ and positive numbers r and δ such that*

(i) $\lim_{k\to\infty} (\sup J_k) = \infty$,

(ii) $l(J_k) = \delta, k = 1, 2, 3, \cdots$,

(iii) $f(t) > r/2$ *for all* $t \in \bigcup_{k=1}^{\infty} J_k$.

Proof. Let $P = \{t \in R^+ : f(t) > 0\}$. P is the union of a disjoint family $\{P_j\}$ of open intervals such that $f(t) = 0$ when t is an endpoint of P_j. Moreover, $\lim_{j\to\infty} (\sup P_j) = \infty$. Otherwise, there exists a $T > 0$ such that $f(t) \leqq 0$ for all $t \geqq T$. Hence by Lemma 4.1 of [5], $f \equiv 0$ on R^+, a contradiction. Likewise each P_j is of finite length.

Since the endpoints of each P_j are zeros of f, there exists for each j some $\tau_j \in P_j$ such that $f(\tau_j) = \sup_{t\in P_j} f(t) \overset{\text{def}}{=} r_j$. Consequently, $\limsup_{j\to\infty} r_j = r > 0$. For if $\limsup_{j\to\infty} r_j = 0$, then $f(t) \to 0$ as $t \to \infty$ on J. Again an application of Lemma 4.1 of [5] shows $f \equiv 0$ on R^+.

There exists a subsequence $\{r_{j_k}\}$ of $\{r_j\}$ with $\lim_{k\to\infty} r_{j_k} = r$. All but a finite number of the r_{j_k} are greater than $3r/4$. Relabel and denote them by $\{r_k\}$. Similarly, relabel the points τ_{j_k} and the intervals P_{j_k}. By uniform continuity of f on R^+ (cf. [1, p. 35]), there exists $\delta = \delta(r)$ such that $|f(t) - f(\tau_k)| < r/4$ whenever $|t - \tau_k| < \delta$ for each τ_k. Hence, $f(t) > f(\tau_k) - r/4 \geqq r/2$ for $|t - \tau_k| < \delta$.

Note that $\lim_{k\to\infty} \tau_k = \infty$. Otherwise $\lim_{k\to\infty} \tau_k = \tau < \infty$ and therefore $f(\tau) = r$. This is impossible since there are points arbitrarily close to τ, where f is nonpositive. Set $J_k = (\tau_k - \delta/2, \tau_k + \delta/2)$ for each k. The conditions of the lemma are now satisfied.

5. Global controllability with bounded positive controls. We now state and prove a theorem on global controllability for the system (1.1). The theorem is analogous to one for global controllability, where $\Omega = [-\varepsilon, \varepsilon]$ for some $\varepsilon > 0$ (cf. [4, p. 92]). The proof of the following theorem, though, is unlike that in [4].

THEOREM 5.1. *Consider the control process* (1.1) *in* R^d, *where* $\Omega = [0, 1]$. *Then* (1.1) *is globally controllable from* $x = 0$ *if and only if*

(i) *for each eigenvalue* λ *of* A, $\operatorname{Im} \lambda \neq 0, \operatorname{Re} \lambda \geqq 0$ *and*

(ii) $\operatorname{rank} C(A, b) = d$.

Proof. Necessity. Global controllability from $x = 0$ implies $0 \in \operatorname{int} K_\Omega^+$. Then $\operatorname{Im} \lambda \neq 0$ for each eigenvalue of A (Lemma 3.1 of [5]), and $\operatorname{rank} C(A, b) = d$ (Theorem 2.4 of [5]). Thus we may assume $d = 2n$. Observe that we have shown the system (1.1) is locally controllable at $x = 0$.

Transforming A to TAT^{-1} if necessary (where T is an invertible matrix), we may assume, without loss of generality, that A (or TAT^{-1}) has the canonical form (cf. [2, p. 358])

$$A = \begin{bmatrix} A_1 & & & & \\ & A_2 & & & \\ & & \cdot & & \\ & & & \cdot & \\ & & & & A_s \end{bmatrix}.$$

The elements not shown are zeros. Each A_k is a square matrix (of even dimension) of the form

$$A_k = \begin{bmatrix} S_k & I & 0 \cdots 0 \\ 0 & S_k & I \cdots 0 \\ \cdot & \cdot & \cdot \quad \cdot \\ \cdot & \cdot & \cdot \quad \cdot \\ \cdot & \cdot & \cdot \quad I \\ 0 & 0 & 0 \cdots S_k \end{bmatrix},$$

where 0 is the 2 × 2 zero matrix, I is the 2 × 2 identity matrix and

$$S_k = \begin{bmatrix} \alpha_k & \beta_k \\ -\beta_k & \alpha_k \end{bmatrix}.$$

The number of 2×2 S_k blocks in A_k is equal to the multiplicity m_k of the eigenvalue λ_k.

Now suppose there exists an eigenvalue $\lambda_k = \alpha_k + i\beta_k$, $\alpha_k < 0$. We will show that $K^+_{[0,1]} \neq R^{2n}$. Now consider only those components of the transformed system associated with the 2×2 S_k block in the lower right-hand corner of A_k. More precisely, consider the vector subspace

$$V_k = \{v \in R^{2n} : P_k(A)v = 0\},$$

where $P_k(y) = y^2 - 2\alpha_k y + (\alpha_k^2 + \beta_k^2)$, the minimal polynomial for S_k. In this two-dimensional subspace we obtain the system

$$\dot{x}_k = S_k x_k + b_k u, \tag{5.1}$$

where $x_k, b_k \in R^2$. Moreover, $b_k \neq 0$; otherwise the solution would remain at the origin in V_k for all $t \geqq 0$. Consequently, the solution to (1.1) could never be steered to a neighborhood of the origin, contradicting the local controllability at $x = 0$.

For ease of notation, we drop the subscripts in (5.1). Now define the function

$$V(x) = \tfrac{1}{2}\|x\|^2.$$

Then

$$\dot{V}(x) = x'\dot{x} = \alpha\|x\|^2 + (b'x)u,$$

where the prime indicates transpose. So $\dot{V}(x) > 0$ implies $\alpha\|x\|^2 + \|b\|\,\|x\| > 0$. Since $\alpha < 0$, this yields

$$\|x\| < |\alpha|^{-1}\|b\|.$$

Hence, for $\|x\| \geqq |\alpha|^{-1}\|b\|$ (letting $u = 1$) we have $\dot{V}(x) \leqq 0$. Going back to the $2n$-dimensional system (1.1), we find if $\alpha_k < 0$, then for any point $x(t) \in K^+_{[0,1]}(t)$ the components of $x(t)$ described by (5.1) are bounded in norm by $|\alpha_k|^{-1}\|b_k\|$. Thus, $K^+_{[0,1]}$ is a proper subset of R^{2n}.

Sufficiency. Conversely, suppose conditions (i) and (ii) are satisfied. Suppose $K^+_{[0,1]} \neq R^{2n}$. Choose any $x_0 \in R^{2n} \backslash K^+_{[0,1]}$. Then $x_0 \notin K^+_{[0,1]}(t)$ for all $t \geqq 0$. Since $K^+_{[0,1]}(t)$ is convex, $K^+_{[0,1]}(t)$ lies on one side of a hyperplane through x_0. Denote the hyperplane by its unit outward normal (row) vector, η. Henceforth, we say that such a hyperplane separates x_0 and $K^+_{[0,1]}(t)$ at time t. As shown in the proof of Theorem 2.1 of [5], there exists a fixed hyperplane η_0 which separates x_0 and $K^+_{[0,1]}(t)$ for any $t \geqq 0$. It follows for any $t \geqq 0$, and any $x(t) \in K^+_{[0,1]}(t)$, that

$$\eta_0[x(t) - x_0] \leqq 0. \tag{5.2}$$

We claim though, that for large enough t, this inequality does not hold.

After a change of variables in the variation of parameters formula, we can write $\eta_0 x(t) = \int_0^t \eta_0 e^{A\tau} bu(t - \tau)\, d\tau$. Let $\varphi(\tau) = \eta_0 e^{A\tau} b$. From Lemma 3.2 of [5] we see that if $\alpha = \max\{\alpha_1, \alpha_2, \cdots, \alpha_s\}$, then $\varphi(\tau) = e^{\alpha\tau}\tau^m\{v(\tau) + \mu(\tau)\}$, where $v(\tau) = \sum_{k=1}^s h_k \sin(\beta_k \tau + \zeta_k)$ and h_k, β_k, ζ_k are constants.

We obtain from Lemma 3.2 of [5] that $v(\tau) \not\equiv 0$ on R^+ and is almost periodic with zero mean. Furthermore, $\lim_{\tau\to\infty} \mu(\tau) = 0$. From Lemma 4.1, we establish the existence of a denumerable set of disjoint open intervals $\{I_k\}_{k=1}^\infty$, each lying in R^+ with $\lim_{k\to\infty}(\sup I_k) = \infty$ such that $l(I_k) = L$, and $v(\tau) > r/2$ for all $\tau \in \bigcup_{k=1}^\infty I_k$ for some positive numbers L and r. Now choose $T > 0$ so that $|\mu(\tau)| < r/4$ for all $\tau \geqq T$. Let k_0 be the smallest positive integer such that $T < \inf I_{k_0}$. Then for any integer $k \geqq k_0$, $v(\tau) + \mu(\tau) > r/4$ whenever $\tau \in I_k$. Hence, $\varphi(\tau) > r/4$ for all $\tau \in \bigcup_{k=k_0}^\infty I_k$. (Without loss of generality, we take $e^{\alpha\tau}\tau^m > 1$.) For each integer $k \geqq k_0$, define $t_k = \sup I_k$. Choose an integer $N \geqq k_0$ sufficiently large so that $(N + k_0)(N - k_0 + 1)(rL/8) > \|x_0\|$. Define the function $u \in U_{[0,1]}$ by

$$u(t_N - \tau) = \begin{cases} 1, & \tau \in \bigcup_{k=k_0}^{N} I_k, \\ 0, & \text{otherwise.} \end{cases}$$

Set $u_N(\tau) = u(t_N - \tau)$. Then

$$\begin{aligned} \eta_0 x(t_N) &= \int_0^{t_N} \varphi(\tau)u_N(\tau)\, d\tau = \int_T^{t_N} \varphi(\tau)u_N(\tau)\, d\tau \\ &\geqq \sum_{k=k_0}^{N} \int_{I_k} \varphi(\tau)u_N(\tau)\, d\tau > \sum_{k=k_0}^{N} \frac{r}{4}L = (N + k_0)(N - k_0 + 1)\frac{rL}{8}. \end{aligned}$$

Thus for $t = t_N$,

$$\eta_0 x(t) > \eta_0 x_0.$$

Hence, the inequality (5.2) is violated. Consequently, $K^+_{[0,1]} = R^{2n}$. This concludes the proof of the theorem.

We may reverse the flow in (1.1) to obtain the following.

COROLLARY 5.2. *Let $N_{[0,1]}$ denote the set of points in R^d which can be steered to the origin in finite time along solutions of* (1.1) *with* $\Omega = [0, 1]$. *Then* $N_{[0,1]} = R^d$ *if and only if*

(i) *for each eigenvalue* λ *of* A, $\operatorname{Im}\lambda \neq 0, \operatorname{Re}\lambda \leqq 0$, *and*

(ii) $\operatorname{rank} C(A, b) = d$.

Proof. The proof follows easily from [4, p. 84] and the preceding theorem.

DEFINITION 5.3. The control process given by (1.1) in R^d with $u(t) \in [a, c]$ is called *completely controllable* if for any two points $x_0, x_1 \in R^d$ there exists an admissible controller $u \in U_{[a,c]}$ which steers x_0 to x_1 in finite time.

Combining Theorem 5.1 with Corollary 5.2, we obtain the following.

THEOREM 5.4. *Suppose* $[a, c] = [0, 1]$. *Then the control process* (1.1) *is completely controllable if and only if*

(i) *all eigenvalues of* A *are pure imaginary* (*nonzero*), *and*

(ii) $\operatorname{rank} C(A, B) = d$.

If we admit unbounded positive controls, namely let $\Omega = [0, \infty)$, we can relax the condition required for the eigenvalues of A as stated in Theorem 5.4. We remark that the conditions given in the following corollary characterize complete controllability for the restraint set $\Omega = [0, \infty)$ (or equivalently, local controllability at $x = 0$ with restraint set $\Omega = [0, 1]$). The result follows directly from Corollary 3.1.

COROLLARY 5.5. *The control process* (1.1) *is completely controllable with* $\Omega = [0, \infty)$ *if and only if*

(i) *for each eigenvalue* λ *of* A, $\operatorname{Im}\lambda \neq 0$, *and*

(ii) $\operatorname{rank} C(A, b) = d$.

Up to now we have considered only those processes where the null control was a boundary point of the restraint set $\Omega = [0, 1]$ or $\Omega = [0, \infty)$. The following theorem characterizes those processes which allow us to take for the restraint set Ω any nonempty interval in R^1.

LEMMA 5.6. *The control process* (1.1) *is locally controllable at the origin for any restraint interval* $[a_1, a_2]$ *provided*

(i) *the eigenvalues of* A *are distinct and pure imaginary* (*nonzero*) *and*

(ii) $\operatorname{rank} C(A, b) = d$.

Proof. Observe that any controller v for which $\Omega = [a_1, a_2]$ can be written as $a_1 + (a_2 - a_1)u(\tau)$, where $u(\tau) \in [0, 1]$. Therefore the solution of (1.1) satisfying $x(0; v(\cdot)) = 0$ is

$$\begin{aligned} x(t) &= \int_0^t e^{A\tau} b[a_1 + (a_2 - a_1)u(t - \tau)]\, d\tau \\ &= \int_0^t e^{A\tau} A x_1 \, d\tau + (a_2 - a_1) \int_0^t e^{A\tau} b u(t - \tau)\, d\tau, \end{aligned}$$

where x_1 is chosen to be $a_1 A^{-1} b$. Since the first integral reduces to $e^{At}x_1 - x_1$, we see that the reachable set from the origin at time t is

$$K^+_{[a_1,a_2]}(t) = e^{At}x_1 - x_1 + (a_2 - a_1) \cdot K^+_{[0,1]}(t).$$

Conditions (i) and (ii) imply that there exists $T_0 > 0$ (finite) such that $x = 0$ is interior to $K^+_{[0,1]}(t)$ for all $t \geqq T_0$. Furthermore, there exists $\varepsilon = \varepsilon(T_0)$ so that the ball of radius ε with center at $x = 0$, $B_\varepsilon(0)$, is interior to $(a_2 - a_1) \cdot K^+_{[0,1]}(t)$. Since the eigenvalues of A are pure imaginary and are distinct, the free motion through x_1 is recurrent. In particular, there exists $t_1 \geqq T_0$ such that $\|e^{At_1}x_1 - x_1\| < \varepsilon/2$. It follows that the open ball

$$e^{At_1}x_1 - x_1 + B_\varepsilon(0) = B_\varepsilon(e^{At_1}x_1 - x_1)$$

contains $x = 0$ and lies in $K^+_{[a_1,a_2]}(t_1)$.

An immediate consequence of the previous lemma is the following.

COROLLARY 5.7. *Consider the control process in* R^d:

$$\dot{x} = Ax + bu + f,$$

where $f \in R^d$, $\Omega = [0, 1]$. *If conditions* (i) *and* (ii) *of Lemma* 5.6 *are satisfied, then for any* $x_0 \in R^d$ *there exists* $T > 0$ *such that the reachable set at time* T *from* x_0 *contains a neighborhood of* x_0.

Note: the reachable set at time t from x_0 is $\{x(t; u(\cdot)) : u \in U_\Omega, x(0; u(\cdot)) = x_0\}$.

REFERENCES

[1] H. BOHR, *Almost Periodic Functions*, Chelsea, New York, 1947.

[2] E. CODDINGTON AND N. LEVINSON, *Theory of Ordinary Differential Equations*, McGraw-Hill, New York, 1955.

[3] R. KALMAN, Y. HO AND K. NARENDA, *Controllability of linear dynamical systems*, Contribution to Differential Equations, 1 (1963), pp. 189–213.

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

[5] S. SAPERSTONE AND J. YORKE, *Controllability of linear oscillatory systems using positive controls*, this Journal, 9 (1971), pp. 253–262.

SIAM J. CONTROL
Vol. 11, No. 3, August 1973

TRACKING AND REGULATION IN LINEAR MULTIVARIABLE SYSTEMS*

W. M. WONHAM†

Abstract. For the multivariable control system described by

$$\dot{x} = Ax + Bu, \qquad y = Cx, \qquad z = Dx,$$

necessary and sufficient conditions are found for the existence of state feedback $u = Fx$ such that $\ker F \supset \ker C$ and $D \exp [t(A + BF)] \rightarrow 0$ as $t \rightarrow \infty$. It is assumed that $\ker C$ is A-invariant, or equivalently that a dynamic observer is utilized. A constructive version of the existence conditions is obtained under the further assumption that auxiliary integrating elements can be introduced by way of dynamic compensation, and a bound is given for the number of such elements required.

Introduction. A classical multivariable control problem requires the design of dynamic compensation to ensure the following behavior of the closed loop system: (a) Each of an assigned set of output variables converges to, or "tracks", a corresponding observed reference input from a specified function class. This is the servo problem. (b) Each of an assigned set of output variables converges to zero from arbitrary initial values, when the system is perturbed by disturbances, possibly not directly observable, but again in a specified function class. This is the regulator problem; the convergent variables are said to be "regulated", and zero represents the target or "set-point" value. Very often both features are present together: the system is required to track in the presence of disturbances. Indeed the formal distinction between the servo and regulator problems disappears if each tracking error (the difference between a reference input and the corresponding output) is regarded as a variable to be regulated. For this reason we restrict attention to the regulator problem alone.

As is usual in classical control theory (e.g., [1]), we assume that the disturbance variables, including any reference input variables, satisfy known time-invariant linear differential equations of finite order, which we lump with the equations, of similar type, representing the plant. In a suitable state space (cf. [2]) the combined system equations then take the standard form

$$\dot{x}(t) = Ax(t) + Bu(t), \tag{1}$$

$$y(t) = Cx(t), \tag{2a}$$

$$z(t) = Dx(t) \tag{2b}$$

defined, say, for $t > 0$. Here $y(\cdot)$ is the vector of directly observed outputs and $z(\cdot)$ is the output vector to be regulated. The control $u(\cdot)$ is required to be such that for every initial state $x(0+)$ there results

$$z(t) \rightarrow 0 \quad \text{as } t \rightarrow \infty. \tag{3}$$

* Received by the editors April 7, 1972, and in revised form August 25, 1972.

† Department of Electrical Engineering, University of Toronto, Toronto, Ontario, Canada. This research was supported in part by the National Research Council of Canada under Grant A-7399.

Typically the matrix triple (C, A, B) is not controllable or observable, or even stabilizable [3] or detectable [4]. If it were, the problem could be solved by standard methods. Stabilizability may fail because the exogenous variables (disturbance and reference inputs) are (say) polynomials in t and cannot be controlled; and detectability may fail if, for instance, certain disturbance variables happen to be decoupled from the observed outputs.

In this paper we discuss when and how a linear feedback control $u(\cdot)$ can be determined which regulates the output in the sense of (3), while respecting the constraint that only $y(\cdot)$ in (2a) can be observed directly. The problem is stated precisely in § 1, and necessary and sufficient conditions for solvability are obtained in § 2 and § 3. The main result, Theorem 2, represents an extension and completion of the results in [2] and [5]. The proof depends on some of the geometric ideas developed in [6] and [7].

Notation. $\mathbb{R}$ stands for the real numbers, $\mathbb{C}$ for the complex plane, $\mathbb{C}^+$ (resp. $\mathbb{C}^-$) for the closed right-half (resp. open left-half) complex plane. Script capitals denote linear spaces over $\mathbb{R}$, and Roman capitals denote linear transformations (*maps*). The image of a map (e.g., Im B) may be written as the corresponding script capital ($\mathscr{B}$). Invariably $A\colon \mathscr{X} \to \mathscr{X}$, $B\colon \mathscr{U} \to \mathscr{X}$, $C\colon \mathscr{X} \to \mathscr{Y}$ and $D\colon \mathscr{X} \to \mathscr{Z}$ are the maps associated with the standard system (1), (2). The symbol $\approx$ means vector space isomorphism (equality of dimension). Writing $d(\cdot)$ for dimension, we shall fix $d(\mathscr{X}) = n$. $\langle A|\mathscr{B}\rangle$ will denote the controllable subspace generated by the pair (A, B):

$$\langle A|\mathscr{B}\rangle = \mathscr{B} + A\mathscr{B} + \cdots + A^{n-1}\mathscr{B}.$$

If $\mathscr{R} \subset \mathscr{X}$ and $A\mathscr{R} \subset \mathscr{R}$, $A|\mathscr{R}$ is the restriction of A to $\mathscr{R}$. The *spectrum* of A, written $\sigma(A)$, is the set of n roots in $\mathbb{C}$ of the characteristic polynomial of A. A^{-1} is the functional inverse of A, i.e., $A^{-1}\mathscr{R} = \{x : Ax \in \mathscr{R}\}$. If $\alpha_F(\lambda)$ is the minimal polynomial (m.p.) of $A + BF$, we denote by $\alpha_F^+(\lambda)$ (resp. $\alpha_F^-(\lambda)$) the factor of $\alpha_F(\lambda)$ with roots in $\mathbb{C}^+$ (resp. $\mathbb{C}^-$), and define $\mathscr{X}^{\pm}(A + BF) = \ker \alpha_F^{\pm}(A + BF)$. When $F = 0$ the subscript (0) is dropped. If $x_1, x_2, \cdots$ are vectors in $\mathscr{X}$, $\{x_1, x_2, \cdots\}$ denotes their span.

1. Restricted regulator problem: algebraic formulation. Let $\mathscr{N}$ denote the unobservable subspace of (C, A):

$$\mathscr{N} = \bigcap_{i=1}^{n} \ker(CA^{i-1}).$$

Write $\bar{\mathscr{X}} = \mathscr{X}/\mathscr{N}$ and let $P : \mathscr{X} \to \bar{\mathscr{X}}$ be the canonical projection. The observable "reduced" system may be identified with the triple $(\bar{C}, \bar{A}, \bar{B})$ in the commutative diagram below:

$$\begin{tikzcd}
& \mathscr{X} \arrow[r, "A"] \arrow[dd, "P"'] & \mathscr{X} \arrow[dd, "P"] \arrow[dr, "C"] & \\
\mathscr{U} \arrow[ur, "B"] \arrow[dr, "\bar{B}"'] & & & \mathscr{Y}. \\
& \bar{\mathscr{X}} \arrow[r, "\bar{A}"'] & \bar{\mathscr{X}} \arrow[ur, "\bar{C}"'] &
\end{tikzcd}$$

Since $A\mathscr{N} \subset \mathscr{N}$, the induced map $\bar{A} : \bar{\mathscr{X}} \to \bar{\mathscr{X}}$ is well-defined, and as $\ker C \supset \mathscr{N}$ there exists a unique map $\bar{C} : \bar{\mathscr{X}} \to \mathscr{Y}$ with $\bar{C}P = C$; finally we set $\bar{B} = PB$.

Since the pair $(\bar{C}, \bar{A})$ is observable it is easily seen (cf. [3]) that maps $J : \bar{\mathscr{X}} \to \bar{\mathscr{X}}$ and $K : \mathscr{Y} \to \bar{\mathscr{X}}$ exist such that the auxiliary system

$$\dot{\bar{w}}(t) = J\bar{w}(t) + Ky(t) + \bar{B}u(t), \qquad t > 0, \tag{4}$$

is a "dynamic observer" for the system (1), (2) reduced mod $\mathscr{N}$. Namely we select K such that

$$\sigma(\bar{A} - K\bar{C}) \subset \mathbb{C}^-$$

and set $J = \bar{A} - K\bar{C}$, $\bar{e}(t) = \bar{x}(t) - \bar{w}(t)$. Since

$$\dot{\bar{x}}(t) = \bar{A}\bar{x}(t) + \bar{B}u(t), \qquad t > 0, \tag{5}$$

there follows $\dot{\bar{e}}(t) = J\bar{e}(t)\,(t > 0)$, so that $\bar{e}(t) \to 0\,(t \to \infty)$. Since $\sigma(J)$ can be assigned arbitrarily by suitable choice of K (see [3]), convergence can in principle be made arbitrarily exponentially fast. Now suppose $u(t) = \bar{F}\bar{w}(t)$ for some $\bar{F}$: $\bar{\mathscr{X}} \to \mathscr{U}$. From (5) and $\bar{e} = \bar{x} - \bar{w}$, one gets

$$\dot{\bar{x}} = (\bar{A} + \bar{B}\bar{F})\bar{x} - \bar{B}\bar{F}\bar{e}, \qquad t > 0,$$

or equivalently

$$\dot{x} = (A + BF)x - B\bar{F}\bar{e}, \qquad t > 0, \tag{6}$$

where $F = \bar{F}P$.

From (2b), (3) and (6) we conclude that regulation is achieved with the control law $u = \bar{F}\bar{w}$ if and only if

$$D\, e^{t(A+BF)}\left[x(0+) - \int_0^t e^{-s(A+BF)}B\bar{F}\bar{e}(s)\, ds\right] \to 0$$

as $t \to \infty$, for all $x(0+)$, $\bar{e}(0+)$. Equivalently,

$$D\, e^{t(A+BF)} \to 0, \qquad t \to \infty, \tag{7}$$

and

$$D\, e^{t(A+BF)} \int_0^t e^{-s(A+BF)}B\bar{F}\, e^{sJ}\, ds \to 0, \qquad t \to \infty. \tag{8}$$

If (7) is true then (8) will follow in view of our previous condition that J be stable. On this basis we shall ignore the term in $\bar{e}$ in (6), assume that the state $\bar{x}$ of the observable reduced system is directly observable at the start, and admit a priori all controls of the form $u = \bar{F}\bar{x}$. A control $u = Fx$ can be written in this form if and only if $F = \bar{F}P$ for some $\bar{F}$, that is, $\ker F \supset \mathscr{N}$, and this version of the observability constraint will be used in the sequel. We remark that exactly the same reasoning applies if the observer is chosen to be of minimal dynamic order given by

$$d(\ker \bar{C}) = d(\ker C) - d(\mathscr{N}),$$

along the lines of [8], [9].

Finally, since

$$\mathscr{X} = \mathscr{X}^+(A + BF) \oplus \mathscr{X}^-(A + BF)$$

and each summand is $(A + BF)$-invariant, a condition equivalent to (7) is $\mathscr{X}^+(A + BF) \subset \ker D$.

In this way we are led to formulate the following algebraic problem.

RESTRICTED REGULATOR PROBLEM (RRP). *Given the maps* $A:\mathscr{X} \to \mathscr{X}$, $B:\mathscr{U} \to \mathscr{X}$, $D:\mathscr{X} \to \mathscr{Z}$, *and a subspace* $\mathscr{N} \subset \mathscr{X}$ *with* $A\mathscr{N} \subset \mathscr{N}$, *find* $F:\mathscr{X} \to \mathscr{U}$ *such that*

$$\ker F \supset \mathscr{N}$$

and

$$\mathscr{X}^+(A + BF) \subset \ker D. \tag{9}$$

RRP is "restricted" in the sense that no provision is made for dynamic compensation other than that tacitly introduced by the observer. Actually, we shall exploit dynamic compensation later, using the technique of state space extension, much as in the "extended" decoupling problem studied in [7].

2. Solution of RRP.

In this section we obtain necessary and sufficient conditions for the solvability of RRP. As they stand, these conditions are not constructive in the sense of providing an algorithmic solution of the problem when a solution exists; nevertheless they can be made so in combination with state space extension, as will be shown in § 3.

THEOREM 1. *RRP is solvable if and only if there exists a subspace* $\mathscr{V} \subset \mathscr{X}$ *such that*

$$\mathscr{V} \subset \ker D \cap A^{-1}(\mathscr{V} + \mathscr{B}), \tag{10a}$$

$$\mathscr{X}^+(A) \cap \mathscr{N} + A(\mathscr{V} \cap \mathscr{N}) \subset \mathscr{V}, \tag{10b}$$

$$\mathscr{X}^+(A) \subset \langle A|\mathscr{B}\rangle + \mathscr{V}.$$

Before proving Theorem 1 we note various structural features of conditions (10). Introduce the family of subspaces

$$\mathbf{\mathscr{V}} = \{\mathscr{V} : \mathscr{V} \subset \ker D \cap A^{-1}(\mathscr{V} + \mathscr{B}) \quad \text{and} \quad A(\mathscr{V} \cap \mathscr{N}) \subset \mathscr{V}\}. \tag{11}$$

In general $\mathbf{\mathscr{V}}$ is not closed under addition and it is not true that $\mathbf{\mathscr{V}}$ contains a supremal element (in the semilattice of all subspaces of $\mathscr{X}$, partly ordered by ($\subset$) and with join operation ($+$)). However, as $\mathbf{\mathscr{V}}$ is nonempty ($0 \in \mathbf{\mathscr{V}}$) it always has, possibly many, maximal elements: by definition $\mathscr{V}^M \in \mathbf{\mathscr{V}}$ is *maximal* if $\mathscr{V} \in \mathbf{\mathscr{V}}$ and $\mathscr{V} \supset \mathscr{V}^M$ implies $\mathscr{V} = \mathscr{V}^M$. We have the following.

COROLLARY 1.1. *RRP is solvable if and only if*

$$\mathscr{X}^+(A) \cap \mathscr{N} \subset \ker D, \tag{12}$$

and for some maximal element $\mathscr{V}^M \in \mathbf{\mathscr{V}}$,

$$\mathscr{X}^+(A) \subset \langle A|\mathscr{B}\rangle + \mathscr{V}^M. \tag{13}$$

The difficulty in verifying (13) is that of effectively parametrizing the subfamily of all $\mathscr{V}^M$. Actually, in many cases which arise in practice it happens that $\mathbf{\mathscr{V}}$ does contain a (unique) supremal element, namely

$$\mathscr{V}^* \equiv \sup\{\mathscr{V} : \mathscr{V} \subset \ker D \cap A^{-1}(\mathscr{V} + \mathscr{B})\}; \tag{14}$$

that is, $\mathscr{V}^*$ satisfies the second defining condition in (11):

$$A(\mathscr{V}^* \cap \mathscr{N}) \subset \mathscr{V}^*. \tag{15}$$

Then, of course, the $\mathscr{V}^M$ all coincide with $\mathscr{V}^*$ and RRP is constructively solvable by the algorithm (cf. [6, Thm. 3.1]):

$$\begin{aligned} &\mathscr{V}^0 = \ker D, \\ &\mathscr{V}^j = \ker D \cap A^{-1}(\mathscr{V}^{j-1} + \mathscr{B}), \qquad j = 1, 2, \cdots, \\ &\mathscr{V}^* = \mathscr{V}^n. \end{aligned} \tag{16}$$

A sufficient condition for (15) to be true is included in the following.

COROLLARY 1.2. *Suppose*

$$A(\mathscr{N} \cap \ker D) \subset \ker D. \tag{17}$$

Then RRP is solvable if and only if

$$\mathscr{X}^+(A) \cap \mathscr{N} \subset \ker D \tag{18}$$

and

$$\mathscr{X}^+(A) \subset \langle A|\mathscr{B}\rangle + \mathscr{V}^*. \tag{19}$$

The proof of these results depends on four lemmas, of which the first was proved in [5].

LEMMA 1. *Let $\mathscr{K} \subset \mathscr{X}$. There exists a map $F: \mathscr{X} \to \mathscr{U}$ such that*

$$\mathscr{X}^+(A + BF) \subset \mathscr{K}$$

if and only if

$$\mathscr{X}^+(A) \subset \langle A|\mathscr{B}\rangle + \mathscr{T}^*,$$

where

$$\mathscr{T}^* = \sup\{\mathscr{T} : \mathscr{T} \subset \mathscr{K} \cap A^{-1}(\mathscr{T} + \mathscr{B})\}.$$

LEMMA 2. *Let $A: \mathscr{X} \to \mathscr{X}, A_1 : \mathscr{X} \to \mathscr{X}$ and $\mathscr{N} \subset \mathscr{X}$, with $A\mathscr{N} \subset \mathscr{N}$ and $A_1|\mathscr{N} = A|\mathscr{N}$. Then*

$$\mathscr{X}^+(A_1) \cap \mathscr{N} = \mathscr{X}^+(A) \cap \mathscr{N}.$$

Proof. Denote by α_1^+ (resp. α^+) the unstable factor of the m.p. of A_1 (resp. A). Let $x \in \mathscr{X}^+(A_1) \cap \mathscr{N}$. Then $x \in \ker \alpha_1^+(A_1)$. Since A coincides with A_1 on $\mathscr{N}$, $A^j x = A_1^j x$ ($j = 1, 2, \cdots$) and therefore

$$\alpha_1^+(A)x = \alpha_1^+(A_1)x = 0.$$

Let $\alpha_x(\lambda)$ be the A-m.p. of x. Then $\alpha_x|\alpha_1^+$, that is, the complex zeros of α_x belong to $\mathbb{C}^+$. Let $\alpha = \alpha^+\alpha^-$ be the m.p. of A. Then also $\alpha_x|\alpha$, and therefore $\alpha_x|\alpha^+$. Thus $x \in \ker \alpha^+(A)$; that is, $x \in \mathscr{X}^+(A) \cap \mathscr{N}$. We have shown that

$$\mathscr{X}^+(A_1) \cap \mathscr{N} \subset \mathscr{X}^+(A) \cap \mathscr{N},$$

and the reverse inclusion follows by symmetry.

LEMMA 3. *Let $A\mathscr{V} \subset \mathscr{V} + \mathscr{B}$* and *$A_1 = A + BF_1$. Then the relation*

$$\mathscr{X}^+(A) \subset \langle A|\mathscr{B}\rangle + \mathscr{V} \tag{20}$$

implies

$$\mathscr{X}^+(A_1) \subset \langle A_1|\mathscr{B}\rangle + \mathscr{V}.$$

Proof. By (20) and Lemma 1 (with $\mathscr{K} = \mathscr{T}^* = \mathscr{V}$) there exists $F:\mathscr{X} \to \mathscr{U}$ such that $\mathscr{X}^+(A + BF) \subset \mathscr{V}$. Setting $F_0 = F - F_1$, we have $\mathscr{X}^+(A_1 + BF_0) \subset \mathscr{V}$ and so, again by Lemma 1,

$$\mathscr{X}^+(A_1) \subset \langle A_1|\mathscr{B}\rangle + \mathscr{V}.$$

LEMMA 4. *For arbitrary* $F: \mathscr{X} \to \mathscr{U}$,

$$\mathscr{X}^+(A + BF) + \langle A|\mathscr{B}\rangle = \mathscr{X}^+(A) + \langle A|\mathscr{B}\rangle. \tag{21}$$

Proof. If $P: \mathscr{X} \to \mathscr{X}/\langle A|\mathscr{B}\rangle$ is the canonical projection and a bar denotes the induced map in $\mathscr{X}/\langle A|\mathscr{B}\rangle$, then $\overline{A + BF} = \bar{A}$ and

$$\begin{aligned} P\mathscr{X}^+(A + BF) &= \bar{\mathscr{X}}^+(\overline{A + BF}) \\ &= \bar{\mathscr{X}}^+(\bar{A}) \\ &= P\mathscr{X}^+(A). \end{aligned} \tag{22}$$

From (22), (21) follows at once.

Proof of Theorem 1. (*If*) Let $\mathscr{V}$ have the properties (10). Then

$$A\mathscr{V} \subset \mathscr{V} + \mathscr{B}, \qquad A(\mathscr{V} \cap \mathscr{N}) \subset \mathscr{V}. \tag{23}$$

By (23) there exists $F_0:\mathscr{X} \to \mathscr{U}$ such that

$$(A + BF_0)\mathscr{V} \subset \mathscr{V}, \qquad F_0(\mathscr{V} \cap \mathscr{N}) = 0.$$

Let

$$\mathscr{V} + \mathscr{N} = \hat{\mathscr{V}} \oplus \mathscr{V} \cap \mathscr{N} \oplus \hat{\mathscr{N}},$$

where $\hat{\mathscr{V}} \subset \mathscr{V}$ and $\hat{\mathscr{N}} \subset \mathscr{N}$. Define $F_1:\mathscr{X} \to \mathscr{U}$ such that $F_1|\mathscr{V} = F_0|\mathscr{V}$ and $F_1|\hat{\mathscr{N}} = 0$. Then $F_1\mathscr{N} = 0$ and $(A + BF_1)\mathscr{V} \subset \mathscr{V}$. Write $A_1 = A + BF_1$ and consider the commutative diagram

$$\begin{tikzcd}
 & \mathscr{X} \arrow[r, "A_1"] \arrow[d, "P"'] & \mathscr{X} \arrow[d, "P"] \arrow[dr, "D"] & \\
\mathscr{U} \arrow[ur, "B"] \arrow[dr, "\bar{B}"'] & & & \mathscr{Y}. \\
 & \bar{\mathscr{X}} \arrow[r, "\bar{A}_1"] & \bar{\mathscr{X}} \arrow[ur, "\bar{D}"'] &
\end{tikzcd}$$

Here $\bar{A}_1$ is the induced map on $\bar{\mathscr{X}} = \mathscr{X}/\mathscr{V}$, and $\bar{D}$ exists since $\ker D \supset \mathscr{V}$. Now $A_1|\mathscr{N} = A|\mathscr{N}$, so by Lemma 2 and (10b),

$$\mathscr{X}^+(A_1) \cap \mathscr{N} = \mathscr{X}^+(A) \cap \mathscr{N} \subset \mathscr{V}.$$

Thus

$$\begin{aligned} \mathscr{N} &= \mathscr{N} \cap \mathscr{X}^+(A_1) \oplus \mathscr{N} \cap \mathscr{X}^-(A_1) \\ &\subset \mathscr{X}^-(A_1) + \mathscr{V} \end{aligned}$$

so that

$$\bar{\mathscr{N}} \equiv P\mathscr{N} \subset \bar{\mathscr{X}}^-(\bar{A}_1). \tag{24}$$

Also, Lemma 3 and (10c) yield

$$\mathscr{X}^+(A_1) \subset \langle A_1|\mathscr{B}\rangle + \mathscr{V},$$

so

$$\bar{\mathscr{X}}^+(\bar{A}_1) \subset \langle \bar{A}_1 | \bar{\mathscr{B}} \rangle. \tag{25}$$

By (24) and (25) there exists $\bar{F}_2 : \bar{\mathscr{X}} \to \mathscr{U}$ such that $\ker \bar{F}_2 \supset \bar{\mathscr{N}}$ and $\bar{A}_1 + \bar{B}\bar{F}_2$ is stable. Define $F_2 = \bar{F}_2 P$. Then $F_2\mathscr{N} = \bar{F}_2\bar{\mathscr{N}} = 0$. Let $F = F_1 + F_2$. Then $F\mathscr{N} = 0$. Also, $F_2\mathscr{V} = 0$ implies

$$(A + BF)|\mathscr{V} = A_1|\mathscr{V},$$

so $(A + BF)\mathscr{V} \subset \mathscr{V}$. For the induced map $\overline{A + BF}$ on $\bar{\mathscr{X}}$ we have

$$\bar{\mathscr{X}}^+(\overline{A + BF}) = \bar{\mathscr{X}}^+(\bar{A}_1 + \bar{B}\bar{F}_2) = \bar{0},$$

so that

$$\mathscr{X}^+(A + BF) \subset \mathscr{V} \subset \ker D$$

as required.

(*Only if*) Let $\mathscr{V} = \mathscr{X}^+(A + BF)$. Then (10a) is clear from (9). Since $\ker F \supset \mathscr{N}$, we have

$$(A + BF)|\mathscr{N} = A|\mathscr{N}$$

and by Lemma 2,

$$\begin{aligned} \mathscr{V} \cap \mathscr{N} &= \mathscr{X}^+(A + BF) \cap \mathscr{N} \\ &= \mathscr{X}^+(A) \cap \mathscr{N}, \end{aligned}$$

so that $A(\mathscr{V} \cap \mathscr{N}) \subset \mathscr{V} \cap \mathscr{N}$, proving (10b). Finally

$$\begin{aligned} \mathscr{V} + \langle A|\mathscr{B}\rangle &= \mathscr{X}^+(A + BF) + \langle A|\mathscr{B}\rangle \\ &= \mathscr{X}^+(A) + \langle A|\mathscr{B}\rangle, \end{aligned}$$

by Lemma 4, and this verifies (10c).

Proof of Corollary 1.1. If (10) holds for some $\mathscr{V}$, then $\mathscr{V} \in \boldsymbol{\mathscr{V}}$. There is some $\mathscr{V}^M \in \boldsymbol{\mathscr{V}}$ with $\mathscr{V}^M \supset \mathscr{V}$, and (10) clearly holds for such $\mathscr{V}^M$. Furthermore

$$\mathscr{X}^+(A) \cap \mathscr{N} \subset \mathscr{V} \subset \ker D.$$

For the converse, set $\mathscr{V} = \mathscr{V}^M$. By (12), $\mathscr{X}^+(A) \cap \mathscr{N}$ is an A-invariant subspace of $\mathscr{N} \cap \ker D$, hence certainly belongs to each $\mathscr{V}^M$: indeed if $\mathscr{V} \in \boldsymbol{\mathscr{V}}$ then

$$[\mathscr{V} + \mathscr{X}^+(A) \cap \mathscr{N}] \cap \mathscr{N} = \mathscr{V} \cap \mathscr{N} + \mathscr{X}^+(A) \cap \mathscr{N}$$

and therefore, $\mathscr{V} + \mathscr{X}^+(A) \cap \mathscr{N} \in \boldsymbol{\mathscr{V}}$.

Proof of Corollary 1.2. By (17), $\mathscr{N} \cap \ker D \subset \mathscr{V}^*$, so that

$$\mathscr{V}^* \cap \mathscr{N} \subset \ker D \cap \mathscr{N} \subset \mathscr{V}^* \cap \mathscr{N}.$$

Thus $\mathscr{V}^* \cap \mathscr{N} = \ker D \cap \mathscr{N}$ is A-invariant, so $\mathscr{V}^* \in \boldsymbol{\mathscr{V}}$. Therefore $\mathscr{V}^M = \mathscr{V}^*$ for all $\mathscr{V}^M$, and the result follows by Corollary 1.1.

To conclude this section we give a description of the subspaces $\mathscr{V}^M$ which will be useful later.

PROPOSITION 1. *Let*

$$\mathscr{V}_0 = \bigcap_{i=1}^{n} A^{-i+1}(\ker D \cap \mathscr{N}).$$

Each subspace $\mathscr{V}^M$ *is of the form*

$$\mathscr{V}^M = \mathscr{V}_0 \oplus \mathscr{V}_1,$$

where

$$\mathscr{V}_1 = \sup \{\mathscr{V}: \ \mathscr{V} \subset \mathscr{W} \cap A^{-1}(\mathscr{B} + \mathscr{V}_0 + \mathscr{V})\}$$

and $\mathscr{W}$ *is a suitable complement of* $\mathscr{N} \cap \ker D$ *in* $\ker D$.

Proof. Suppose $\mathscr{V}^M$ is maximal, and write

$$\mathscr{V}^M = \mathscr{V}^M \cap \mathscr{N} \oplus \mathscr{V}_1^M,$$
$$\ker D = \ker D \cap \mathscr{N} \oplus \mathscr{V}_1^M \oplus \mathscr{W}_1,$$
$$\mathscr{W} = \mathscr{V}_1^M \oplus \mathscr{W}_1.$$

Next observe that

$$\mathscr{V}_0 = \sup \{\mathscr{V}: \mathscr{V} \subset \mathscr{N} \cap \ker D, A\mathscr{V} \subset \mathscr{V}\}.$$

Since $\mathscr{V}^M$ is maximal, $\mathscr{V}^M \supset \mathscr{V}_0$: indeed, the subspace

$$(\mathscr{V}^M + \mathscr{V}_0) \cap \mathscr{N} = \mathscr{V}^M \cap \mathscr{N} + \mathscr{V}_0$$

is A-invariant, hence, $\mathscr{V}^M \subset \mathscr{V}^M + \mathscr{V}_0 \in \mathbf{V}$. Since $\mathscr{V}^M \cap \mathscr{N}$ is A-invariant and $\mathscr{V}_0$ is supremal, it follows that $\mathscr{V}^M \cap \mathscr{N} = \mathscr{V}_0$. Next

$$A\mathscr{V}_1^M \subset A\mathscr{V}^M \subset \mathscr{V}^M \cap \mathscr{N} + \mathscr{V}_1^M + \mathscr{B}$$
$$= \mathscr{V}_1^M + \mathscr{B} + \mathscr{V}_0$$

and therefore, $\mathscr{V}_1^M \subset \mathscr{V}_1$. Since $\mathscr{V}_1 \cap \mathscr{N} = 0$, the subspace

$$(\mathscr{V}_0 + \mathscr{V}_1) \cap \mathscr{N} = \mathscr{V}_0$$

is A-invariant, hence, $\mathscr{V}^M \subset \mathscr{V}_0 + \mathscr{V}_1 \in \mathbf{V}$, so that $\mathscr{V}^M = \mathscr{V}_0 + \mathscr{V}_1$ and therefore, $\mathscr{V}_1^M = \mathscr{V}_1$.

3. Extended regulator problem. To exploit the advantages of dynamic compensation, we introduce an auxiliary dynamic element with equation

$$\dot{x}_a = B_a u_a, \qquad y_a = C_a x_a,$$

where $u_a \in \mathscr{U}_a$, $x_a \in \mathscr{X}_a$, $y_a \in \mathscr{Y}_a$; and $B_a : \mathscr{U}_a \approx \mathscr{X}_a$, $C_a : \mathscr{X}_a \approx \mathscr{Y}_a$. Write

$$\mathscr{U}_e = \mathscr{U} \oplus \mathscr{U}_a, \qquad \mathscr{X}_e = \mathscr{X} \oplus \mathscr{X}_a, \qquad \mathscr{Y}_e = \mathscr{Y} \oplus \mathscr{Y}_a$$

for the extended control, state and (observed) output spaces, and introduce the extended maps

$$\begin{aligned}
&A_e : \mathscr{X}_e \to \mathscr{X}_e; \qquad && A_e|\mathscr{X} = A, \qquad && A_e|\mathscr{X}_a = 0,\\
&B_e : \mathscr{U}_e \to \mathscr{X}_e; && B_e|\mathscr{U} = B, && B_e|\mathscr{U}_a = B_a,\\
&C_e : \mathscr{X}_e \to \mathscr{Y}_e; && C_e|\mathscr{X} = C, && C_e|\mathscr{X}_a = C_a,\\
&D_e : \mathscr{X}_e \to \mathscr{Z}; && D_e|\mathscr{X} = D, && D_e|\mathscr{X}_a = 0.
\end{aligned}$$

Observe that

$$\ker C_e = \ker C = \mathscr{N},$$

by use of a dynamic observer, and

$$\ker D_e = \ker D \oplus \mathscr{X}_a.$$

We define the *extended regulator problem* (ERP) as that of finding suitable $\mathscr{X}_a$ (that is, $d(\mathscr{X}_a)$) and then $F_e : \mathscr{X}_e \to \mathscr{U}_e$, such that

$$\ker F_e \supset \mathscr{N}$$

and

$$\mathscr{X}_e^+(A_e + B_eF_e) \subset \ker D \oplus \mathscr{X}_a.$$

The main result of this paper is the following.

THEOREM 2. *Let RRP be defined as in* § 1. *Then ERP is solvable if and only if, for RRP,*

$$\mathscr{X}^+(A) \cap \mathscr{N} \subset \ker D \tag{26}$$

and

$$\mathscr{X}^+(A) \subset \langle A|\mathscr{B}\rangle + \mathscr{V}^*, \tag{27}$$

where

$$\mathscr{V}^* = \sup\{\mathscr{V} : \mathscr{V} \subset \ker D \cap A^{-1}(\mathscr{V} + \mathscr{B})\}. \tag{28}$$

Furthermore, if ERP is solvable, it is possible to take

$$d(\mathscr{X}_a) \leqq d\left[(\mathscr{N} \cap \mathscr{V}^*)\Big/\left(\mathscr{N} \cap \bigcap_{i=1}^{n} A^{-i+1}\mathscr{V}^*\right)\right]. \tag{29}$$

For the proof we shall need the following.

LEMMA 5. *Suppose* $\mathscr{V} \subset \mathscr{X}$ *and* $\mathscr{N} \subset \mathscr{X}$ *with* $A\mathscr{N} \subset \mathscr{N}$. *Define extended spaces and maps with*

$$\mathscr{X}_a \approx (\mathscr{N} \cap \mathscr{V})\Big/\left(\mathscr{N} \cap \bigcap_{i=1}^{n} A^{-i+1}\mathscr{V}\right). \tag{30}$$

There exists a map $E : \mathscr{X}_e \to \mathscr{X}_e$ *with* $\operatorname{Im} E = \mathscr{X}_a$, *such that the subspace* $\mathscr{V}_e = (1 + E)\mathscr{V}$ *has the property*

$$\mathscr{N} \cap \bigcap_{i=1}^{n} A^{-i+1}\mathscr{V} + A_e(\mathscr{V}_e \cap \mathscr{N}) \subset \mathscr{V}_e.$$

Proof. Write $\mathscr{V}_0 = \mathscr{N} \cap \bigcap_{i=1}^{n} A^{-i+1}\mathscr{V}$ and let

$$\mathscr{V} \cap \mathscr{N} = \mathscr{V}_0 \oplus \mathscr{V}_1,$$

$$\mathscr{V} = \mathscr{V} \cap \mathscr{N} \oplus \mathscr{V}_2.$$

Let $E : \mathscr{X}_e \to \mathscr{X}_e$ be any map such that

$$\ker E \supset \mathscr{V}_0 \oplus \mathscr{V}_2,$$

$$\ker E \cap \mathscr{V}_1 = 0,$$

$$E\mathscr{V}_1 = \mathscr{X}_a.$$

Such a map exists by (30). Now,

$$\mathscr{V}_e \cap \mathscr{N} = [\mathscr{V}_0 \oplus (1 + E)\mathscr{V}_1 \oplus \mathscr{V}_2] \cap \mathscr{N} = \mathscr{V}_0,$$

so that

$$A_e(\mathscr{V}_e \cap \mathscr{N}) = A\mathscr{V}_0 \subset \mathscr{V}_0 = \mathscr{V}_e \cap \mathscr{N},$$

as required.

Proof of Theorem 2. (*If*) Choose $\mathscr{X}_a$ according to (30) (with $\mathscr{V}^*$ in place of $\mathscr{V}$). Construct $\mathscr{V}_e^* = (1 + E)\mathscr{V}^*$ as in Lemma 5. Thus

$$\mathscr{N} \cap \bigcap_{i=1}^{n} A^{-i+1}\mathscr{V}^* + A_e(\mathscr{V}_e^* \cap \mathscr{N}) \subset \mathscr{V}_e^*. \tag{31}$$

We shall verify that the conditions of Theorem 1 hold for the extended problem. Now

$$\mathscr{V}_e^* \subset \mathscr{V}^* \oplus \mathscr{X}_a \subset \ker D \oplus \mathscr{X}_a = \ker D_e$$

and

$$\begin{aligned} A_e\mathscr{V}_e^* &= A\mathscr{V}^* \\ &\subset \mathscr{V}^* + \mathscr{B} \\ &\subset (1 + E)\mathscr{V}^* + \mathscr{B} + \mathscr{X}_a \\ &= \mathscr{V}_e^* + \mathscr{B}_e. \end{aligned}$$

Next,

$$\begin{aligned} \mathscr{X}_e^+(A_e) \cap \mathscr{N} &= [\mathscr{X}^+(A) \oplus \mathscr{X}_a] \cap \mathscr{N} \\ &= \mathscr{X}^+(A) \cap \mathscr{N} \\ &\subset \mathscr{N} \cap \bigcap_{i=1}^{n} A^{-i+1}\mathscr{V}^* \quad \text{(by (26) and (28))} \\ &\subset \mathscr{V}_e^* \quad \text{(by (31))}, \end{aligned}$$

and this verifies the extended version of (10b). Finally,

$$\begin{aligned} \mathscr{X}_e^+(A_e) &= \mathscr{X}^+(A) \oplus \mathscr{X}_a \\ &\subset \langle A|\mathscr{B}\rangle + \mathscr{V}^* + \mathscr{X}_a \\ &= \langle A_e|\mathscr{B}_e\rangle + \mathscr{V}_e^*, \end{aligned}$$

which verifies the extension of (10c).

(*Only if*) Let Q be the projection on $\mathscr{X}$ along $\mathscr{X}_a$. Applied to ERP, Theorem 1 provides a subspace $\mathscr{V}_e \subset \mathscr{X}_e$ which satisfies the extended version of (10). In particular (10b) implies

$$\begin{aligned} \ker D_e &\supset \mathscr{X}_e^+(A_e) \cap \mathscr{N} \\ &= [\mathscr{X}^+(A) \oplus \mathscr{X}_a] \cap \mathscr{N} \\ &= \mathscr{X}^+(A) \cap \mathscr{N} \end{aligned}$$

so that

$$\begin{aligned} \mathscr{X}^+(A) \cap \mathscr{N} &\subset Q \ker D_e \\ &= \ker D, \end{aligned}$$

proving (26). Next, (10c) applied to ERP yields

$$\mathcal{X}_e^+(A_e) \subset \langle A_e|\mathcal{B}_e\rangle + \mathcal{V}_e$$

and so, with $\mathcal{V} = Q\mathcal{V}_e$,

$$(32)\qquad \begin{aligned}\mathcal{X}^+(A) &= Q\mathcal{X}_e^+(A_e)\\ &\subset \langle A|\mathcal{B}\rangle + \mathcal{V}.\end{aligned}$$

Finally we note that $\mathcal{V}_e \subset \ker D_e$ implies $\mathcal{V} \subset \ker D$, and $A_e\mathcal{V}_e \subset \mathcal{V}_e + \mathcal{B}_e$ implies

$$\begin{aligned}A\mathcal{V} &= AQ\mathcal{V}_e\\ &= QA_e\mathcal{V}_e\\ &\subset Q(\mathcal{V}_e + \mathcal{B}_e)\\ &= \mathcal{V} + \mathcal{B},\end{aligned}$$

so that $\mathcal{V} \subset \mathcal{V}^*$ and (27) follows from (32).

Remark. If $\mathcal{V}_e$ is a solution of ERP, then by (10b),

$$(33)\qquad A_e(\mathcal{V}_e \cap \mathcal{N}) \subset \mathcal{V}_e;$$

but it is not true in general that $A(\mathcal{V} \cap \mathcal{N}) \subset \mathcal{V}$ with $\mathcal{V} = Q\mathcal{V}_e$. Deduction of the last-written inclusion from (33) would be immediate if

$$Q(\mathcal{V}_e \cap \mathcal{N}) = Q\mathcal{V}_e \cap Q\mathcal{N};$$

and, as $\ker Q = \mathcal{X}_a$ and $\mathcal{N} \cap \mathcal{X}_a = 0$, this would be true if and only if

$$(34)\qquad (\mathcal{V}_e + \mathcal{N}) \cap \mathcal{X}_a = \mathcal{V}_e \cap \mathcal{X}_a.$$

But, in general, (34) fails; indeed the construction of Lemma 5 has

$$(\mathcal{V}_e + \mathcal{N}) \cap \mathcal{X}_a = \mathcal{X}_a, \qquad \mathcal{V}_e \cap \mathcal{X}_a = 0.$$

This heuristic reasoning suggests that in some cases ERP is solvable when RRP is not, a conjecture borne out by the following example.

Example. Let

$$A = \begin{bmatrix} 0 & 1 & 0\\ -1 & -2 & 0\\ 0 & 0 & 1\end{bmatrix}, \qquad B = \begin{bmatrix}0\\1\\0\end{bmatrix},$$

$$C = [0 \quad 0 \quad 1], \qquad D = [\alpha \quad -\alpha \quad -1].$$

Assume $\alpha \neq 0$. We have

$$\mathcal{N} = \left\{\begin{matrix}1 & 0\\ 0 & 1\\ 0 & 0\end{matrix}\right\}, \qquad \ker D = \left\{\begin{matrix}1 & 1\\ 1 & 0\\ 0 & \alpha\end{matrix}\right\},$$

$$\langle A|\mathcal{B}\rangle = \left\{\begin{matrix}1 & 0\\ 0 & 1\\ 0 & 0\end{matrix}\right\}, \qquad \mathcal{X}^+(A) = \left\{\begin{matrix}0\\ 0\\ 1\end{matrix}\right\}.$$

The algorithm (16) yields $\mathscr{V}^* = \ker D$, so that

$$\mathscr{X}^+(A) \subset \mathscr{X} = \langle A|\mathscr{B}\rangle + \mathscr{V}^*.$$

Since in addition $\mathscr{X}^+(A) \cap \mathscr{N} = 0$, Theorem 2 asserts that ERP is solvable. We find

$$\mathscr{N} \cap \bigcap_{i=1}^{3} A^{-i+1}\mathscr{V}^* = 0$$

so we can take

$$d(\mathscr{X}_a) = d(\mathscr{N} \cap \mathscr{V}^*) = d\left\{\begin{matrix}1\\1\\0\end{matrix}\right\} = 1.$$

Write explicitly

$$E(1 \quad 1 \quad 0)' = (1 \quad 1 \quad 0 \quad 1)';$$

then

$$\mathscr{V}_e^* = (1 + E)\mathscr{V}^* = \left\{\begin{matrix}1 & 1\\1 & 0\\0 & \alpha\\1 & 0\end{matrix}\right\}.$$

ERP is now essentially solved: it only remains to compute a feedback map F_e such that $\ker F_e \supset \mathscr{N}, (A_e + B_eF_e)\mathscr{V}_e^* \subset \mathscr{V}_e^*$, and the induced map $\overline{A_e + B_eF_e}$ on $\mathscr{X}_e/\mathscr{V}_e^*$ is stable. In this example these requirements lead to the unique result,

$$F_e = \begin{bmatrix}0 & 0 & 0 & 4\\0 & 0 & -\alpha^{-1} & 1\end{bmatrix}.$$

The reader is invited to draw the signal flow diagram. The given system has two stable, unobservable "modes" (with state variables x_1, x_2) and one unstable, observable mode (x_3). Only x_1, x_2 are controllable. The regulated variable is $z = \alpha x_1 - \alpha x_2 - x_3$. To interpret the role of dynamic compensation, suppose first that all states had been observable and consider the corresponding RRP, with $\mathscr{N} = 0$. A simple computation shows that output regulation is achieved, with state feedback $u = f_1x_1 + f_2x_2 + f_3x_3$, only if $f_1 + f_2 = 4$. But then $(A + BF)|\langle A|\mathscr{B}\rangle$ has the spectrum $\{1, 1 - f_1\}$, i.e., the controllable subsystem is destabilized. Returning to the example, we have that $\langle A|\mathscr{B}\rangle$ is unobservable. The compensator can be thought of as simulating the unobservable variable $x_1 + x_2$ according to the equation $\dot{x}_a = -x_a + \tilde{u}_a$. Then the feedback law

$$u = 4x_a, \qquad \tilde{u}_a = -\alpha^{-1}x_3 + 2x_a$$

produces the required internal instability. The resulting transfer function from the exponential "disturbance" $x_3(\cdot)$ to the regulated output is

$$\hat{z}(s)/\hat{x}_3(s) = -(s + 3)(s - 1)/(s + 1)^2.$$

Of course the example is artificial, and from a sensitivity viewpoint ill-posed, but it exhibits decoupling action as a distinct, fundamental role of linear state feedback and compensation. A second role, observer action, is inoperative by virtue of our initial assumption $\ker C = \mathscr{N}$; a third, pole-shifting, is not possible here because the controllable observable state subspace is zero.

To verify that RRP is not solvable we apply Proposition 1. Clearly,

$$\mathscr{N} \cap \ker D = \{(1 \quad 1 \quad 0)'\}$$

is not A-invariant, so $\mathscr{V}_0 = 0$. Every subspace $\mathscr{W}$, such that $\mathscr{W} \oplus N \cap \ker D = \ker D$, can be written

$$(35) \qquad \mathscr{W}_\mu = \left\{ \mu \begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix} + \begin{bmatrix} 1 \\ 0 \\ \alpha \end{bmatrix} \right\} = \left\{ \begin{matrix} \mu + 1 \\ \mu \\ \alpha \end{matrix} \right\}.$$

Application of (16), with $\mathscr{W}_\mu$ in place of $\ker D$, yields

$$(36) \qquad \mathscr{V}_\mu^M = 0.$$

To satisfy the condition (13) of Corollary 1.1 would require

$$(37) \qquad \begin{Bmatrix} 0 \\ 0 \\ 1 \end{Bmatrix} \subset \begin{Bmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 0 \end{Bmatrix} + \mathscr{V}_\mu^M$$

and this is clearly incompatible with (36).

To conclude the general discussion, we note that if $A(\mathscr{V}^* \cap \mathscr{N}) \subset \mathscr{V}^*$, the right side of (29) reduces to 0, so that RRP is solvable whenever ERP is solvable, and no dynamic compensation is required. Examination of the proof of Lemma 5 together with condition (12) of Corollary 1.1 reveals that dynamic compensation is necessary only if the stable, unobservable state subspace $\mathscr{X}^-(A) \cap \mathscr{N}$ is nonzero. This fact and the foregoing example indicate that the decoupling action of the adjoined dynamics is realized via the simulation only of stable unobservable modes, but we shall not develop this assertion in detail here.

Finally, there was nothing special about the partition $\mathbb{C} = \mathbb{C}^+ \cup \mathbb{C}^-$. In practice one might well require that convergence in (3) take place with exponents in some "good" subset $\mathbb{C}_g \subset \mathbb{C}^-$, where $\mathbb{C}_g$ is symmetric with respect to the real axis. Writing $\mathbb{C}_b$ for the "bad" complement $\mathbb{C} - \mathbb{C}_g$, and replacing $\mathscr{X}^+$ (resp. $\mathscr{X}^-$) by the corresponding subspaces $\mathscr{X}_b$ (resp. $\mathscr{X}_g$), one obtains the corresponding results by exactly the same procedure.

4. Concluding remark. Our main result (Theorem 2) can be paraphrased by saying that regulation is achievable, at least with dynamic compensation, if and only if (i) the unstable unobservable modes of the system are nulled at the regulated output, and (ii) output stabilization is possible without regard to observability constraints. The theory is constructive, reasonably complete, and offers significant structural insight. Its main shortcoming is that no condition is imposed to ensure internal stability, that is, stability of all controllable and observable modes. This more practical problem is completely solved in a sequel article [10].

REFERENCES

[1] O. I. ELGERD, *Control Systems Theory*, McGraw-Hill, New York, 1967.

[2] S. P. BHATTACHARYYA AND J. B. PEARSON, *Error systems and the servo-mechanism problem*, Princeton Conference on Information and Systems Sciences, Princeton University, Princeton, 1971.

[3] W. M. WONHAM, *On pole assignment in multi-input controllable linear systems*, IEEE Trans. Automatic Control, AC-12 (1967), pp. 660–665.

[4] ———, *On a matrix Riccati equation of stochastic control*, this Journal, 6 (1968), pp. 681–697.

[5] S. P. BHATTACHARYYA, J. B. PEARSON AND W. M. WONHAM, *On zeroing the output of a linear system*, Information and Control, 20 (1972), pp. 135–142.

[6] W. M. WONHAM AND A. S. MORSE, *Decoupling and pole assignment in linear multivariable systems: a geometric approach*, this Journal, 8 (1970), pp. 1–18.

[7] A. S. MORSE AND W. M. WONHAM, *Decoupling and pole assignment by dynamic compensation*, this Journal, 8 (1970), pp. 317–337.

[8] D. G. LUENBERGER, *Observers for multivariable systems*, IEEE Trans. Automatic Control, AC-11 (1966), pp. 190–197.

[9] W. M. WONHAM, *Dynamic observers: geometric theory*, Ibid., AC-15 (1970), pp. 258–259.

[10] W. M. WONHAM AND J. B. PEARSON, *Regulation and internal stabilization in linear multivariable systems*, Control Systems Rep. 7212, Dept. of Electrical Engineering, University of Toronto, 1972.

SIAM J. CONTROL
Vol. 11, No. 3, August 1973

A KUHN–TUCKER ALGORITHM*

HILBERT K. SCHULTZ†

Abstract. An algorithm for the solution of the nonlinear programming problem with linear constraints is given based on an attempt to satisfy the Kuhn–Tucker conditions. The algorithm utilizes "nearly" active constraints at a feasible point to avoid zigzagging and the convergence proof is based on general conditions developed by Zangwill.

1. Introduction. This paper presents an algorithm and convergence proof for solving the nonlinear programming problem of finding $\bar{x}$ such that

$$f(\bar{x}) = \min f(x), \qquad x \in X = \{x \in R^n | g_i(x) \leqq 0, i = 1, \cdots, m\}, \tag{1.1}$$

where

$$g_i(x) = n_i x - b_i \leqq 0, \qquad i = 1, \cdots, m, \tag{1.2}$$

i.e., linear constraints. The algorithm is based on attempting to satisfy the Kuhn–Tucker conditions at a point $x \in X$ and was suggested by Rosen (1965). Preliminary computational results by Teorey (1971) (§ 3) were very good.

Section 2 considers the nonlinear programming problem in a more general setting and presents some minor modifications of the work of Zangwill (1969). The conditions given there are sufficient for algorithmic convergence and are used in § 3 where the Kuhn–Tucker algorithm is stated and convergence is verified.

2. The minimization problem and conditions for algorithmic convergence. The problem is to find $\bar{x}$ such that

$$f(\bar{x}) = \min_{x \in X} f(x), \qquad \bar{x} \in X \subset R^n. \tag{2.1}$$

Let

$$U = \{x \in X | f(x) = f(\bar{x})\}, \tag{2.2}$$

i.e., U is the set of solution points. Assume further that there exists a set

$$S \subset X, \qquad S \neq \varnothing, \tag{2.3}$$

called the set of stationary points. We make no other assumptions on X, U, or S at this time.

DEFINITION. An algorithm to solve (2.1) generating a sequence of points $x_1, x_2, \cdots$ is said to be *convergent* if the limit of every convergent subsequence is a stationary point, i.e., it is in S.

Remark 1. This definition is the one commonly used in mathematical programming, but it only makes sense in the case where X is compact, since, for example, it considers an algorithm generating the points $1, 2, 3, \cdots$ to be convergent otherwise.

* Received by the editors October 26, 1971, and in revised form September 20, 1972.

† Department of Mathematics, Wisconsin State University, Oshkosh, Wisconsin 54091. This research was supported by the National Science Foundation under Grant GJ-362 and an NSF Institutional Grant.

Remark 2. We adopt the convention that for a finite sequence the last element is assumed to be repeated so that an infinite sequence is generated (for theoretical purposes). Thus for an algorithm generating a finite sequence to be convergent the terminal point must be a stationary point.

Remark 3. We shall use the notation of Zangwill (1969) for subsequences. That is, if $x_1, x_2, \cdots$ is a sequence, we shall denote subsequences by $\{x_j\}, j \in K$, where K is an *infinite* subset of the positive integers. We shall also adopt the convention of using primes to denote the transpose of matrices only, while vectors will be either a row or column vector depending on the context.

Remark 4. We note that this development could be applied to more general spaces.

Let $x_1, x_2, \cdots$ be a sequence generated by an algorithm and let Z be a continuous function from X into R. Suppose that the algorithm satisfies the following conditions (Zangwill (1969, p. 244)):

(i) if the algorithm terminates at x_j, then $x_j \in S$. If the algorithm generates an infinite sequence of points; then

(ii) for all k, there exists L_k such that for all $l \geqq L_k$, $Z(x_l) \leqq Z(x_k)$; and

(iii) if $x_j \to \bar{x}$, $j \in K \subset \{1, 2, \cdots\}$, and $\bar{x} \notin S$, there exists x_k such that $Z(x_k) < Z(\bar{x})$.

Condition (ii) says that for each k, from some point on the terms are less than or equal to $Z(x_k)$. Thus an increase in Z over its value at x_k is only allowed a finite number of times for all k. The next condition, (iii), is the condition which requires using ε active constraints in gradient projection and feasible direction methods.

The definition we have taken for convergence is the one most commonly used (implicitly) in mathematical programming. The main difference between the approach taken here and that of Zangwill (1969, p. 235) is that his definition of convergence and sufficient conditions for convergence place requirements on the algorithm for determining when there is no solution. However, that sort of approach detracts from the generality of the theory since most authors treat the existence of solutions separately from the algorithm.

THEOREM 1 (Sufficiency). *If an algorithm satisfies* (i)–(iii) *for a continuous* Z, *then the algorithm is convergent, i.e., it satisfies Definition* 1.

The proof of this theorem requires only a minor modification of that of Zangwill (1969, p. 242), and is given in Schultz (1971). We note that for this sufficiency theorem we only need Z to be lower semicontinuous but that in the necessity theorem below we need Z to be upper semicontinuous. Hence for simplicity and comparison we have assumed Z to be continuous in the statement of the theorems.

THEOREM 2 (Necessity). *If an algorithm is convergent, with a continuous function* f, *if the* x_j, *for all* j, *lie in a compact set, if* $S = U$, *and if the algorithm terminates whenever* $x_j \in U$, *then the algorithm satisfies* (i)–(iii) *with* $Z = f$.

The proof is essentially the same as that of Zangwill (1969, p. 242) and is given in Schultz (1971).

We note that the hypotheses of Theorem 2 are common assumptions of many algorithms. We only require the continuity of f, the equivalence of stationary points and solutions to the minimization problem, the x_j remaining in a compact set, and that the algorithm recognizes a solution when it finds one.

We also note that convergence of the algorithm could be proven directly without the use of Zangwill's condition. Also other general theories such as Cea (1971), Daniel (1970), Elkin (1968) and Ortega and Rheinbolt (1970) could be used in conjunction with the theory presented in § 3 to prove convergence. However, the author felt that Zangwill's ideas would be more appropriate since they are not only sufficient but also necessary with minor additional assumptions.

3. A Kuhn–Tucker algorithm and convergence proof. In the sequel we shall require some additional notation. If $x = (x_1, x_2, \cdots, x_n)$ and $y = (y_1, y_2, \cdots, y_n)$ are vectors in R^n, then

$$x \leqq y \Leftrightarrow x_i \leqq y_i \quad \text{for } i = 1, \cdots, n,$$

$$x \leq y \Leftrightarrow x \leqq y \quad \text{and} \quad x \neq y,$$

$$x < y \Leftrightarrow x_i < y_i \quad \text{for } i = 1, \cdots, n.$$

In examining algorithms for solving the problem of finding $\bar{x}$ such that

$$f(\bar{x}) = \min f(x), \qquad x \in X = \{x | g_i(x) \leqq 0, i = 1, \cdots, m\},$$

it becomes clear that two very broad and important classes of algorithms can be distinguished. The first class of algorithms is based on consideration of the entire feasible set X at each point. The versions of conditional gradient, gradient projection and Newton's method as discussed in Levitin and Polyak (1966) are examples of such techniques. Generally speaking, the convergence theory of such methods is easier to establish but the subproblem at each iteration is of a much higher degree of complexity than methods of class two. The second class of methods consists of those based only on the active constraints (those constraints for which $g_i(x) = 0$). Techniques such as those of Abadie (1970), Goldfarb (1969), McCormick (1970a), (1970b), Ritter (1971), Rosen (1960), Zangwill (1967), and Zoutendijk (1960), (1970) all fall into this category. In proving convergence of such algorithms, some sort of anti-zigzagging procedure is usually required. One way of accomplishing this is to use ε tolerances (Zoutendijk (1960), Demjanov (1967), Zangwill (1969), Polak (1971)) and this technique will be used in proving convergence of the Kuhn–Tucker algorithm given below.

In the Kuhn–Tucker algorithm we consider the linearly constrained nonlinear programming problem of finding $\bar{x}$ such that

$$f(\bar{x}) = \min f(x), \quad x \in X = \{x | n_i x - b_i \leqq 0, i = 1, \cdots, m\}, \tag{3.1}$$

$$n_i n_i = 1, \quad i \in J = \{1, 2, \cdots, m\}. \tag{3.2}$$

We use the notation:

$$I(x_j, \varepsilon_j) = \{i \in J | -\varepsilon_j \leqq n_i x_j - b_i \leqq 0\}, \quad \varepsilon_j > 0; \tag{3.3}$$

$$I(x_j) = I(x_j, 0); \tag{3.4}$$

(3.5) N_{ε_j} = matrix of unit normals containing all n_i (as column vectors) with $i \in I(x_j, \varepsilon_j)$;

(3.6) N_I = matrix of unit normals containing all n_i (as column vectors), $i \in I \subset J$;

(3.7) U_I = vector in $R^{r(I)}$, where $r(I)$ is the number of constraints in I.

We assume that

(3.8) f is continuously differentiable and convex,

(3.9) X is convex and compact.

The Kuhn–Tucker conditions (Mangasarian (1969)) that $\bar{x}$ solve (3.1) are

$$\nabla f(\bar{x}) + N_J U_J = 0, \quad U_J \in R^m, \tag{3.10}$$

$$N'_J\bar{x} - b_J \leqq 0, \tag{3.11}$$

$$U_J[N'_J\bar{x} - b_J] = 0, \tag{3.12}$$

$$U_J \geqq 0. \tag{3.13}$$

However, for an inactive constraint $i, n_i\bar{x} - b_i < 0$ so (3.12) $\Leftrightarrow U_i = 0$, and thus conditions (3.10)–(3.13) can be written

$$\nabla f(\bar{x}) + N_{I(\bar{x})}U_{I(\bar{x})} = 0, \tag{3.14}$$

$$U_{I(\bar{x})} \geqq 0 \tag{3.15}$$

for a feasible point $\bar{x}$, where $U_{I(\bar{x})} \in R^{r(I(\bar{x}))}$, $r(I(\bar{x}))$ = number of constraints in $I(\bar{x})$ and $U_i = 0, i \in J - I(\bar{x})$.

We now give a result due to Rosen (1965) in a somewhat more general form which shows that an attempt to verify the Kuhn–Tucker conditions (3.14), (3.15) either leads to a feasible direction or shows that a point is optimal.

THEOREM 3 (Rosen). *Let*

$$\|v(\bar{u})\|_2^2 \equiv \min_{\substack{u_I \geqq 0 \\ u_I \in R^{r(I)}}} \|\nabla f(\bar{x}) + N_I u_I\|_2^2.$$

Then if $v(\bar{u}) \neq 0$,

$$N'_I v(\bar{u}) \geqq 0, \tag{3.16}$$

$$\nabla f(\bar{x})v(\bar{u}) = v(\bar{u})v(\bar{u}) = \|v(\bar{u})\|_2^2 > 0. \tag{3.17}$$

Note that if $I = I(\bar{x})$ and $v(\bar{u}) = 0$, then $\bar{x}$ is optimal since the Kuhn–Tucker conditions are satisfied.

Proof. Let $p(u) = \|v(u)\|_2^2$ so that $p(\bar{u}) = \|v(\bar{u})\|_2^2 = v(\bar{u})v(\bar{u})$ and set $\bar{g} = \nabla f(\bar{x})$. Then from

$$v(\bar{u}) = \bar{g} + N_I\bar{u}, \tag{3.18}$$

$$\frac{\partial p(\bar{u})}{\partial u_i} = 2v(\bar{u})n_i, \quad i \in I. \tag{3.19}$$

Thus we must have $v(\bar{u})n_i \geqq 0$ for $i \in I$, for suppose $v(\bar{u})n_j < 0$ for some j. Then we could increase u_j and decrease $p(\bar{u})$ since $\partial p/\partial u_j < 0$, which contradicts the fact that $\bar{u}$ is an optimal solution to Theorem 3. Thus (3.16) is proven.

Looking again at (3.19) we see that if $u_j > 0$ we must have $v(\bar{u})n_j = 0$, for if it were > 0 we could decrease u_j and decrease p, which would again be a contradiction. Therefore, $v(\bar{u})n_i\bar{u}_i = 0$, $i \in I$ since either $v(\bar{u})n_i = 0$ or $u_i = 0$. Thus,

$$v(\bar{u})N_I\bar{u} = 0. \tag{3.20}$$

Using (3.18) and (3.20) we have

$$0 < v(\bar{u})v(\bar{u}) = v(\bar{u})[\nabla f(\bar{x}) + N_I\bar{u}] = v(\bar{u})\nabla f(\bar{x}),$$

i.e.,

$$\nabla f(\bar{x})v(\bar{u}) > 0. \tag{3.21}$$

Thus (3.17) is proven. Q.E.D.

For convenience we now define a function $\phi : R^n \times R^1 \to R^1$ by

$$\phi(x,\varepsilon) = \phi(x, I(x,\varepsilon)) = \min_{u \geqq 0} \|\nabla f(x) + N_\varepsilon u\|_2^2, \qquad u \in R^{r(I(x,\varepsilon))}. \tag{3.22}$$

Note that by the remark following Theorem 3 if $\phi(x) = \phi(x,0) = 0$, then x is optimal.

Definition 2. A point x is *stationary* if $x \in S = \{x \in X | \phi(x) = 0\}$.

We can now state the algorithm.

The Kuhn–Tucker Algorithm. Assume $x_1 \in X$ is given and $\varepsilon > 0, p > 0$ are fixed. Set $\varepsilon_1 = \varepsilon, p_1 = p, j = 1$.

Step 1. Compute $\phi(x_j)$. If $\phi(x_j) = 0$, terminate; otherwise go to Step 2.

Step 2. Compute $\phi(x_j, \varepsilon_j) = \delta_j$. If $\delta_j \geqq p_j$, set $s_j = v_{\varepsilon_j}$, where $\phi(x_j, \varepsilon_j) = \|v_{\varepsilon_j}\|_2^2 \equiv \|\nabla f_j + N_{\varepsilon_j}u_{\varepsilon_j}\|_2^2$ and go to Step 4. If $\delta_j < p_j$, go to Step 3.

Step 3. Set $\varepsilon_j = \varepsilon_{j/2}$, $p_j = p_{j/2}$ and return to Step 2.

Step 4. Determine λ_j by $\lambda_j = \max\{\lambda | x_j - \lambda s_j \in X\}$. Set $y_j = x_j - \lambda_j s_j$. Find x_{j+1} by $f(x_{j+1}) = \min_{[x_j, y_j]} f(x)$, where $[x_j, y_j] = \{z | z = (1-\lambda)x_j + \lambda y_j, 0 \leqq \lambda \leqq 1\}$. Set $\varepsilon_{j+1} = \varepsilon$, $p_{j+1} = p$, $j + 1 \to j$ and return to Step 1.

Remark 5. The algorithm will only return to Step 2 a finite number of times since for sufficiently small ε, $\phi(x_j, \varepsilon_j) = \phi(x_j, I(x_j, \varepsilon_j)) = \phi(x_j, I(x_j, 0)) = \phi(x_j) > 0$ by Step 1. Thus once $I(x_j, \varepsilon_j) = I(x_j, 0)$, ϕ remains constant with decreasing ε and we need only halve ε and p until $\phi(x_j) \geqq p_j$.

Remark 6. Our theory also covers the case where $p = \varepsilon$, which is used by some authors (e.g., Polak (1971) for gradient projection and feasible directions).

Remark 7. It is also possible to consider alternative step sizes. For example, in Step 2 let $s_j = v_j$ as stated in the algorithm. Then let $y_j = x_j - \eta_j s_j$, where $\eta_j = \max\{1, \frac{1}{2}, \frac{1}{4}, \cdots\}$, such that $y_j \in X$. Then following Armijo (1966) we may pick $\lambda_j = \max\{1, \frac{1}{2}, \frac{1}{4}, \cdots\}$ such that $f(x_j) - f(x_j - \lambda_j y_j) \geqq -\lambda_j \nabla f(x_j) y_j$, and set $x_{j+1} = x_j - \lambda_j y_j$. A third alternative to the step size can be obtained by (Goldstein (1967)) setting $x_{j+1} = x_j - \lambda_j y_j$, where

$$\lambda_j = \begin{cases} 1 & \text{if } \gamma_j(1) \geqq \rho, \\ \hat{\lambda}_j & \text{such that } \rho \leqq \gamma_j(\hat{\lambda}_j) \leqq 1 - \rho \quad \text{if } \gamma_j(1) < \rho, \end{cases}$$

and

$$\gamma_j(\lambda) = \frac{f(x_j) - f(x_j + \lambda y_j)}{-\lambda \nabla f(x_j) y_j},$$

where $0 < \rho \leqq \frac{1}{2}$. See also Cea (1971) and Daniel (1971).

Remark 8. The difference between Rosen's (1960) gradient projection and the Kuhn–Tucker algorithm is that Rosen does not require $u_I \geqq 0$. The

Kuhn–Tucker algorithm utilizes

$$\min_{u_I \geqq 0} \|\nabla f(x) + N_I u_I\|^2$$

and Rosen uses

$$\min_{u_I} \|\nabla f(x) + N_I u_I\|_2^2.$$

The following example illustrates that each method produces a feasible but different direction.

Consider $\min \frac{1}{2}(x^2 + y^2), x, y \in R^1$ subject to

$$x - y \leq 0, \quad y \leq 2, \quad x \leq 2, \quad x \geq 0 \quad (\text{or } -x \leq 0)$$

which has the feasible region shown in Fig. 1. Consider the point $x_0 = (\frac{1}{2}, 2)$

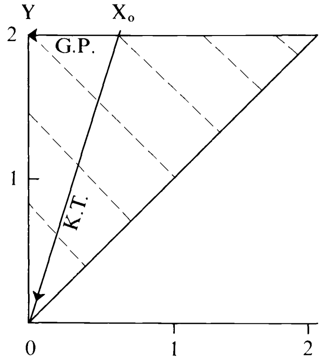


FIG. 1

with $\nabla f(x_0) = (\frac{1}{2}, 2)$. Then gradient projection requires

$$\min_{u \in R^1} \|\nabla f(x_0) + Nu\|_2^2 = \min_{u \in R^1} \left\| \begin{pmatrix} \frac{1}{2} \\ 2 \end{pmatrix} + \begin{pmatrix} 0 \\ 1 \end{pmatrix} u \right\|^2$$

$$= \min_{u \in R^1} \|(\tfrac{1}{2}, 2 + u)\|^2 \Rightarrow u = -2,$$

and we obtain the direction $(-\frac{1}{2}, 0)$ on which to minimize. For the Kuhn–Tucker algorithm we obtain $u = 0$ and the direction $(-\frac{1}{2}, -2)$ from x_0 as the minimizing direction.

Before verifying conditions (i)–(iii), we require the following theorem.

THEOREM 4. *For a fixed* $I \subset J$, $\phi(x, I) \equiv \min_{u_I \geqq 0} \|\nabla f(x) + N_I u_I\|_2^2$ *is a continuous function of* x.

A proof of this theorem is included in Schultz (1971).

We now verify that the algorithm is convergent via (i)–(iii) with $Z = f$. Condition (i) is satisfied since the algorithm only terminates if $\phi(x_j) = 0$, i.e., $x_j \in S$. Condition (ii) is satisfied by (3.17) since $-\nabla f_j v_j < 0 \Rightarrow f(x_{j+1}) < f(x_j)$. We next require the following lemma.

LEMMA. *If* $x_j \to \bar{x}, j \in K$ *and* $\phi(\bar{x}) > 0$, *then* $\varepsilon_j \not\to 0, j \in K$.

Proof. Suppose $\varepsilon_j \to 0$. Then $p_j \to 0, j \in K$, since they are halved simultaneously. This also implies $\phi(x_j, 2\varepsilon_j) \leqq 2p_j$ for sufficiently large $j, j \in K$, or halving would not have been required. Since there are only finitely many constraints, we may take a further subsequence $\{x_j\}, j \in K' \subset K$, such that $\bar{I} \equiv I(x_j, 2\varepsilon_j)$ is a fixed set $\bar{I}$ for all $j \in K'$. Now $\bar{I} \subset I(\bar{x})$ since $2\varepsilon_j \to 0$, and $\phi(x_j, 2\varepsilon_j) \to 0, j \in K'$, since $2p_j \to 0$, but this implies $\phi(\bar{x}) = 0$ since $\phi(x_j, 2\varepsilon_j) = \phi(x_j, \bar{I}) \to \phi(\bar{x}, \bar{I})$ and $\bar{I} \subset I(\bar{x})$. This contradiction proves the assertion. Q.E.D.

To verify condition (iii), suppose that $x_i \to \bar{x}, j \in K$, and $\bar{x} \notin S$, i.e., $\phi(\bar{x}) > 0$.

By Lemma 1, $\varepsilon_j \not\to 0$, $p_j \not\to 0$ so there exist $\varepsilon^* > 0$, $p^* > 0$ and a subsequence $\{x_j\}, j \in K' \subset K$, such that $\varepsilon_j \geqq \varepsilon^*$, $p_j \geqq p^*$, $j \in K'$. Thus $\phi(x_j, \varepsilon_j) \geqq p_j \geqq p^*$, and since there are only finitely many constraints, we may take a further subsequence $\{x_j\}, j \in K'' \subset K'$, such that $I(x_j, \varepsilon_j) = I$, a fixed set for all $j \in K''$. Then by the continuity of ϕ (Theorem 4), $\phi(\bar{x}, I) \geqq p^*$. Also since $\varepsilon_j \not\to 0$, $I(\bar{x}) \subset I$. Now since X is compact, y_j and λ_j lie in compact sets, so that we may take a further subsequence $\{x_j\}, j \in K''' \subset K''$, such that $x_j \to \bar{x}$, $y_j \to \bar{y}$, $\lambda_j \to \bar{\lambda}$. We claim $\bar{\lambda} > 0$, for if $\bar{\lambda} = 0$, then for each $j \in K'''$, we must have $n_i y_j - b_i = 0$ and $n_i s_j < 0$ for some $i \in J$ or λ_j could be increased. Since there are finitely many constraints, we may take a further subsequence $\{x_j\}, j \in K^{\text{iv}} \subset K'''$, such that this occurs for the same i. But then since $\lambda_j \to 0$, $y_j \to \bar{x} \Rightarrow n_i\bar{x} - b_i = 0 \Rightarrow i \in I(\bar{x}) \subset I$, which is a contradiction to Theorem 4 since $n_i s_j \geqq 0$ for all $i \in I$. Thus $\lambda_j \to \bar{\lambda} > 0$, $j \in K^{\text{iv}}$. Now $\nabla f_j(y_j - x_j) = -\lambda_j \nabla f_j s_j = -\lambda_j \phi(x_j, \varepsilon_j)$ by Theorem 4, and $-\lambda_j \phi(x_j, \varepsilon_j) \leqq -\lambda_j p^*$, $j \in K^{\text{iv}}$. Therefore since $x_j \to \bar{x}$, $y_j \to \bar{y}$ and ∇f_j is continuous, we have

$$\nabla f(\bar{x})(\bar{y} - \bar{x}) \leqq -\bar{\lambda} p^* < 0.$$

Thus there exist $1 > \hat{\lambda} > 0$ and $\bar{\varepsilon} > 0$ such that

$$f(\hat{\lambda}\bar{x} + (1 - \lambda)\bar{y}) = f(\bar{x}) - \bar{\varepsilon}$$

implies, for sufficiently large j by the continuity of f, that

$$f(x_{j+1}) \leqq f(\hat{\lambda} x_j + (1 - \hat{\lambda}) y_j) \leqq f(\bar{x}) - \bar{\varepsilon}/2 < f(\bar{x}),$$

i.e., condition (iii) is satisfied. We have proved the following theorem.

THEOREM 5. *Each limit point $\bar{x}$ of the sequence generated by the Kuhn–Tucker algorithm satisfies the Kuhn–Tucker conditions, i.e., it solves problem* (3.1).

A version of this algorithm was implemented on the UNIVAC 1108 at the University of Wisconsin by Teorey (1971). The algorithm was used to solve test problems 1 and 7 of Colville's (1968) nonlinear programming study. Problem #1 was solved in a standard time of .0057 units and problem #7 was solved in a standard time of .0202 units, whereas the fastest time for other methods was .0069 and .0234, respectively, for 17 different methods in problem 1 and 8 methods in problem 2.

REFERENCES

[1] J. ABADIE (1970), *Applications of the GRG algorithm to optimal control problems*, Integer and Nonlinear Programming, J. Abadie, ed., American Elsevier, New York.

[2] L. ARMIJO (1966), *Minimization of functions having Lipschitz continuous first partial derivatives*, Pacific J. Math., 16, pp. 1–3.

[3] J. CEA (1971), *Optimisation theorie et algorithmes*, Dunod, Paris.

[4] A. R. COLEVILLE (1968), *A comparative study of non-linear programming codes*, IBM Tech. Rep. 320–2949.

[5] J. W. DANIEL (1970), *The Approximate Minimization of Functionals*, Prentice-Hall, Englewood Cliffs, N.J.

[6] V. F. DEMJANOV (1967), *On the solution of certain minimax problems, I, II*, Kybernetica, 2 (1966), pp. 58–66; 3 (1967), pp. 62–66.

[7] R. M. ELKIN (1968), *Convergence theorems for Gauss–Seidel and other minimization algorithms*, Computer Science Tech. Rep. 68-59, University of Maryland, College Park.

[8] D. GOLDFARB (1969), *Extension of Davidon's variable metric method to maximization under linear inequality and equality constraints*, SIAM J. Appl. Math., 17, pp. 739–764.

[9] A. A. GOLDSTEIN (1967), *Constructive Real Analysis*, Harper and Row, New York.

[10] E. S. LEVITIN AND B. T. POLYAK (1966), *Constrained minimization methods*, Zh. Vychisl. Mat. i Mat. Fiz., 6, pp. 787–823 = U.S.S.R. Comput. Math. and Math. Phys., 6, pp. 1–50.

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

[12] ——— (1972), *Dual, feasible direction algorithms*, Mathematics Research Center Tech. Rep. 1173, University of Wisconsin, Madison.

[13] GARTH P. MCCORMICK (1970a), *A second order method for the linearly constrained nonlinear programming problem*, Nonlinear Programming, J. B. Rosen, O. L. Mangasarian and K. Ritter, eds., Academic Press, New York.

[14] ——— (1970b), *The variable reduction method for nonlinear programming*, Management Sci., Theory Series.

[15] J. M. ORTEGA AND W. C. RHEINBOLDT (1970), *Iterative Solution of Nonlinear Equations of Several Variables*, Academic Press, New York.

[16] E. POLAK (1971), *Computational Methods in Optimization, A Unified Approach*, Academic Press, New York.

[17] K. RITTER (1970), *A superlinearly convergent method for minimization problems with linear inequality constraints*, Mathematics Research Center Tech. Rep. 1098, University of Wisconsin, Madison.

[18] J. B. ROSEN (1960), *The gradient projection method for nonlinear programming, Part I, Linear constraints*, SIAM J. Appl. Math., 8, pp. 181–217.

[19] ——— (1965), Private communication, *Gradient projection as a least squares solution of Kuhn Tucker conditions.*

[20] H. K. SCHULTZ (1971), *General convergence conditions in non-linear programming and a Kuhn Tucker algorithm*, Computer Science Tech. Rep. 140, University of Wisconsin, Madison.

[21] T. J. TEOREY (1971), Private communication. *Nonlinear programming algorithms.*

[22] D. M. TOPKIS AND A. F. VEINOTT, JR. (1967), *On the convergence of some feasible direction algorithms for nonlinear programming*, this Journal, 5, pp. 268–279.

[23] W. I. ZANGWILL (1967), *An algorithm for the Chebyshev problem—with an application to concave programming*, Management Sci., 14, pp. 58–78.

[24] ——— (1969). *Nonlinear Programming*, Prentice-Hall, Englewood Cliffs, N.J.

[25] G. ZOUTENDIJK (1960), *Methods of Feasible Directions*, Elsevier, New York.

[26] ——— (1970), *Some algorithms based on the principle of feasible directions*, Nonlinear Programming, J. B. Rosen, O. L. Mangasarian and K. Ritter, eds., Academic Press, New York.

SIAM J. CONTROL
Vol. 11, No. 3, August 1973

STRUCTURAL INVARIANTS OF LINEAR MULTIVARIABLE SYSTEMS*

A. S. MORSE†

Abstract. This article identifies some of the structural properties of the matrix triple (C, A, B) which remain invariant under various transformation groups. The paper begins with a brief account of a recent result which states that the controllable space of (A, B) can be decomposed into a direct sum of singly-generated controllability subspaces, the dimension of each subspace being determined by one of the controllability indices of (A, B). In certain instances the component subspaces of this decomposition can be chosen so that their C-images also decompose system output space in a special way; matrix triples (C, A, B) for which such a decomposition is possible are called *prime*. If $\bar{\mathfrak{C}}$ is an appropriately defined group of system-coordinate and state-feedback transformations acting on prime triples, then the controllability indices of (A, B) determine a complete orbital invariant under $\bar{\mathfrak{C}}$.

By imbedding $\bar{\mathfrak{C}}$ in a richer transformation group $\mathfrak{C}^*$, which also includes nonphysically realizable "output-injection" transformations, it is shown that the $\mathfrak{C}^*$-orbit of any matrix triple (C, A, B) is uniquely characterized by three lists of positive integers and a list of monic polynomials called the *transmission polynomials* of (C, A, B). These lists determine a $\mathfrak{C}^*$-canonical form for (C, A, B).

Introduction. For a number of years there has been an interest in feedback invariants of linear multivariable systems and in their relation to various problems of analysis and design. In 1964 Popov [1] identified certain feedback invariants in connection with a study of stability and linear optimal control. Gilbert [2] utilized other feedback invariants in describing the structure of noninteracting control systems. More recently, Kwakernaak and Sivan [3] have used a feedback invariant, the polynomial of transmission zeros of a transfer function matrix, to describe certain asymptotic properties of the solution to the regulator problem. Of particular interest is a recent result due initially to Brunovský [4] which states that the orbit of a controllable matrix pair (A, B), under the group of state-coordinate, input-coordinate and state-feedback transformations ($\mathfrak{C}$), is uniquely characterized by a list of positive integers called the *controllability indices* of (A, B). In the present article we develop new results along these lines by studying the structural invariants of the matrix triple (C, A, B) under various transformation groups.

In § 2 we summarize Brunovský's result and also the main result of [5] which states that the controllable space of (A, B) can be decomposed into a direct sum of singly-generated controllability subspaces, the dimension of each subspace being determined by one of the controllability indices of (A, B). In certain instances, the component subspaces of this decomposition can be chosen so that their C-images also decompose system output space in a special way (§ 3); matrix triples (C, A, B) for which such a decomposition is possible are called *prime*. If $\mathfrak{C}$ is imbedded in a larger group $\bar{\mathfrak{C}}$, which also includes output-coordinate transformations, and $\bar{\mathfrak{C}}$ is regarded as a transformation group acting on prime triples (C, A, B), then the controllability indices of (A, B) determine a complete orbital invariant under $\bar{\mathfrak{C}}$.

* Received by the editors January 31, 1972.

† Department of Engineering and Applied Science, Yale University, New Haven, Connecticut 06520. This research was supported by the U.S. Air Force Office of Scientific Research under Grant 72-2211.

The investigation in § 3 is preliminary to a more general study of structural invariants of (C, A, B) (§ 4). By imbedding $\mathfrak{C}$ in a still larger group $\mathfrak{C}^*$, which also includes nonphysically realizable "output-injection" transformations, it is shown that the $\mathfrak{C}^*$-orbit of *any* matrix triple (C, A, B) is uniquely characterized by three lists of positive integers and a list of monic polynomials called the *transmission polynomials* of (C, A, B). These lists determine a $\mathfrak{C}^*$-canonical representation for (C, A, B) from which one can construct a corresponding quasi-canonical representation relative to only the smaller but more meaningful group $\mathfrak{C}$.

1. Preliminaries.

1.1. Notation. Below, script letters $\mathscr{X}, \mathscr{Y}, \cdots$ denote real vector spaces with elements $x, y, \cdots$; the vector space spanned by a single element x is written as $\mathscr{x}$; the zero space, zero vector, $\cdots$ are denoted by 0; $d(\mathscr{Y})$ is the dimension of $\mathscr{Y}$. The dual space of $\mathscr{X}$ is written as $\mathscr{X}'$; if $\mathscr{S} \subset \mathscr{X}$, then $\mathscr{S}^{\perp}$ is the annihilator of $\mathscr{S}$.

Both matrices and linear maps are defined over the real numbers and are denoted by capital italic letters $A, B, C, \cdots$; the same symbol is used for both a matrix and its map. Dual maps are written as $A', B', C', \cdots$. If $M : \mathscr{R} \to \mathscr{S}$ and $\mathscr{T} \subset \mathscr{R}$, then $M|\mathscr{T}$ denotes the restricted map $\mathscr{T} \to \mathscr{S}, t \mapsto Mt$; in case $\mathscr{S} = \mathscr{R}$ and $\mathscr{T}$ is M-invariant, the codomain of $M|\mathscr{T}$ is taken to be $\mathscr{T}$.

If k is a positive integer, $\mathbf{k} \equiv \{1, 2, \cdots, k\}$ and $\bar{\mathbf{k}} \equiv \{1, 2, \cdots, k - 1\}$; if $k_2 \geqq k_1, \mathbf{k}_2 - \mathbf{k}_1 \equiv \{k_1 + 1, k_1 + 2, \cdots, k_2\}$.

The maps $A : \mathscr{X} \to \mathscr{X}$, $B : \mathscr{U} \to \mathscr{X}$, $C : \mathscr{X} \to \mathscr{Y}$ $(d(\mathscr{X}) = n$, $d(\mathscr{U}) = m$, $d(\mathscr{Y}) = p)$ are fixed throughout and are associated with the linear system $\dot{x} = Ax + Bu$, $y = Cx$. The pair (A, B) is called *standard* if B is monic; similarly, (C, A, B) is standard if both (A, B) and (A', C') are. We write $\mathscr{B}$ = image B, $\mathscr{N}$ = kernel C and $\langle A|\mathscr{B}\rangle = \mathscr{B} + A\mathscr{B} + \cdots + A^{n-1}\mathscr{B}$ for the controllable space of (A, B).

It is assumed that the reader is familiar with the basic concepts and properties of (A, B)-invariant and controllability subspaces [6], [7]. If $\mathscr{W}$ is (A, B)-invariant, $\mathbf{F}(A, B, \mathscr{W})$ denotes the class of maps $F : \mathscr{X} \to \mathscr{U}$ satisfying $(\mathrm{A} + BF)\mathscr{W} \subset \mathscr{W}$.

1.2. System properties. Two particular subspaces, each uniquely determined by (C, A, B), will be important in the sequel. Write $\mathscr{S}$ for the largest (A, B)-invariant subspace contained in $\mathscr{N}$. It is known (cf. [6, Thm. 3.1]) that $\mathscr{S} = \mathscr{S}_n$, where

$$(1.1) \qquad \mathscr{S}_0 = \mathscr{X}, \qquad \mathscr{S}_i = \mathscr{N} \cap A^{-1}(\mathscr{B} + \mathscr{S}_{i-1}), \qquad i \in \mathbf{n},$$

and that

$$(1.2) \qquad \mathscr{S}_i \subset \mathscr{S}_{i-1}, \qquad i \in \mathbf{n}.$$

By replacing $\mathscr{N}$ with kernel B', A with A' and $\mathscr{B}$ with image C' and then dualizing (1.1) and (1.2), it is easy to verify that if

$$(1.3) \qquad \mathscr{T}_0 \equiv 0, \qquad \mathscr{T}_i \equiv A(\mathscr{N} \cap \mathscr{T}_{i-1}) + \mathscr{B}, \qquad i \in \mathbf{n},$$

then

$$(1.4) \qquad \mathscr{T}_i \supset \mathscr{T}_{i-1}, \qquad i \in \mathbf{n},$$

and $\mathscr{T} \equiv \mathscr{T}_n$ is the smallest subspace satisfying $\mathscr{B} \subset \mathscr{T}$ and $A(\mathscr{N} \cap \mathscr{T}) \subset \mathscr{T}$.

If $\mathscr{R}$ is the largest (A, B)-controllability subspace contained in $\mathscr{N}$, then it is known (cf. [6, Lemma 4.2, Thm. 4.3]) that $\mathscr{R} = \mathscr{R}_n$, where

$$\mathscr{R}_0 \equiv 0, \qquad \mathscr{R}_i \equiv (A\mathscr{R}_{i-1} + \mathscr{B}) \cap \mathscr{S}, \qquad i \in \mathbf{n}. \tag{1.5}$$

There is a simple relationship between $\mathscr{R}$, $\mathscr{S}$ and $\mathscr{T}$.

LEMMA 1.1.

$$\mathscr{R} = \mathscr{S} \cap \mathscr{T}. \tag{1.6}$$

Proof. It is enough to show that

$$\mathscr{R}_i = \mathscr{S}, \cap \mathscr{T}_i, \qquad i \in \mathbf{n}. \tag{1.7}$$

Since $\mathscr{T}_1 = \mathscr{B}$ and $\mathscr{R}_1 = \mathscr{B} \cap \mathscr{S}$, (1.7) is true for $i = 1$. Assuming it holds at i, there follows

$$\begin{aligned} \mathscr{R}_{i+1} &= (A\mathscr{R}_i + \mathscr{B}) \cap \mathscr{S} \\ &= (A(\mathscr{S} \cap \mathscr{T}_i) + \mathscr{B}) \cap \mathscr{S} \\ &= (A(\mathscr{N} \cap (A^{-1}(\mathscr{S} + \mathscr{B})) \cap \mathscr{T}_i) + \mathscr{B}) \cap \mathscr{S} \\ &= ((A(\mathscr{N} \cap \mathscr{T}_i)) \cap (\mathscr{S} + \mathscr{B}) + \mathscr{B}) \cap \mathscr{S} \\ &= (A(\mathscr{N} \cap \mathscr{T}_i) + \mathscr{B}) \cap (\mathscr{S} + \mathscr{B}) \cap \mathscr{S} \\ &= \mathscr{S} \cap \mathscr{T}_{i+1}. \end{aligned}$$

Remark 1.1. There is complete duality between the set $\{C, A, B, \mathscr{S}, \mathscr{T}\}$ and the set $\{B', A', C', \mathscr{T}^{\perp}, \mathscr{S}^{\perp}\}$. For example, by Lemma 1.1, $\mathscr{T}^{\perp} \cap \mathscr{S}^{\perp}$ is the largest (A', C')-controllability subspace contained in kernel B'.

The subspaces $\mathscr{R}$ and $\mathscr{S}$ determine a map $A_t : \mathscr{S}/\mathscr{R} \to \mathscr{S}/\mathscr{R}$ whose set of invariant factors is an important structural invariant of (C, A, B). To construct the map let $F \in \mathbf{F}(A, B, \mathscr{S})$ be arbitrary; it is known [6, Thm. 4.3] that $F \in \mathbf{F}(A, B, \mathscr{R})$. Let $\bar{A}_F$ denote the map induced in $\mathscr{X}/\mathscr{R}$ by $A + BF$ and define $A_t = \bar{A}_F|(\mathscr{S}/\mathscr{R})$. It can be shown [8, Lemma 4.1] that A_t is independent of $F \in \mathbf{F}(A, B, \mathscr{S})$. Since both $\mathscr{R}$ and $\mathscr{S}$ are uniquely determined by (C, A, B), it follows that A_t also is. The invariant factors of A_t are called the *transmission polynomials* of (C, A, B). These polynomials arise in connection with the decoupling problem [2], [6] and the regulator problem [3]. They are also related to the feedback-invariant structure of the Smith–McMillan form of the transfer matrix $C(\lambda I - A - BF)^{-1}B$; this topic will be treated in a future article.

2. Canonical structure of (A, B). Let us begin by considering the feedback invariant structure of the pair (A, B) for which a more or less complete canonical description is now known. Consider, for example, the main result of [5].

PROPOSITION 2.1. *Let (A, B) be fixed with $d(\mathscr{B}) = m$. There exist (A, B)-controllability subspaces $\mathscr{W}_i$, $i \in \mathbf{m}$, such that*

$$d(\mathscr{B} \cap \mathscr{W}_i) = 1, \qquad i \in \mathbf{m},$$

and

$$\langle A|\mathscr{B}\rangle = \mathscr{W}_1 \oplus \mathscr{W}_2 \oplus \cdots \oplus \mathscr{W}_m.$$

Of particular importance to this decomposition is the list of integers

$$\mathbf{I}(A, B) = \{d(\mathscr{W}_1), d(\mathscr{W}_2), \cdots, d(\mathscr{W}_m)\}$$

which is assumed to be ordered so that $d(\mathscr{W}_{i+1}) \geqq d(\mathscr{W}_i), i \in \overline{\mathbf{m}}$. Brunovský [4] showed that this list is uniquely determined by (A, B); in fact, he showed that $\mathbf{I}(A, B)$ can be computed directly from the list of integers $d(\mathscr{B} + A\mathscr{B} + \ + A^{i-1}\mathscr{B})$, $i \in \mathbf{n}$. The elements of $\mathbf{I}(A, B)$, called the *controllability indices* of (A, B), determine a canonical representation of (A, B) as follows.

Write $\mathfrak{C}$ for the class of all triples (T, F, G), where T and G are automorphisms of $\mathscr{X}$ and $\mathscr{U}$ respectively, and $F: \mathscr{X} \to \mathscr{U}$ is an arbitrary map. Elements of $\mathfrak{C}$ act on (A, B) according to the rule

$$(A, B) \mapsto (T(A + BF)T^{-1}, TBG). \tag{2.1}$$

It is easy to verify that $\mathfrak{C}$ admits the structure of a transformation group with composition rule defined in the obvious way [5]. The $\mathfrak{C}$-*orbit* (equivalence class) of (A, B), denoted by $\mathfrak{C}(A, B)$, is the set of all pairs $(T(A + BF)T^{-1}, TBG)$.

Proposition 2.1 states, in effect, that for any standard, controllable pair (A, B), there is in $\mathfrak{C}(A, B)$ another pair (A_c, B_c) of the form

$$A_c = \text{block diag.} [A_1, A_2, \cdots, A_m],$$

$$B_c = \text{block diag.} [B_1, B_2, \cdots, B_m],$$

where

$$A_i = \begin{bmatrix} 0 & 1 & 0 & . & . & . \\ 0 & 0 & 1 & 0 & . & . \\ . & . & 0 & 1 & & . \\ . & . & . & & & . \\ . & . & . & . & & 1 \\ . & . & . & . & . & 0 \end{bmatrix}_{n_i \times n_i}, \quad B_i = \begin{bmatrix} 0 \\ . \\ . \\ . \\ 0 \\ 1 \end{bmatrix}_{n_i \times 1},$$

and $\mathbf{I}(A, B) = \{n_1, n_2, \cdots, n_m\}$. To state that this pair is $\mathfrak{C}$-*canonical* is to state that (A_c, B_c) can be determined from any element in $\mathfrak{C}(A, B)$. For this to be possible, the function ϕ: {standard controllable pairs (A, B)} $\to$ {lists of positive integers}, $(A, B) \mapsto \mathbf{I}(A, B)$, must be an orbital invariant of $\mathfrak{C}$.[1] The following result due to Brunovský [4] states that ϕ does indeed have the required property.

PROPOSITION 2.2. *The function ϕ is a complete orbital invariant of* $\mathfrak{C}$.

We call (A_c, B_c), now identified as canonical, the *feedback canonical form* of (A, B).

Remark 2.1. Recently, Kalman [9] has made clear the connection between the controllability indices of (A, B) and the classical Kronecker invariants associated

[1] Let $\mathbf{S}$ be a set and $\hat{\mathfrak{C}}$ a group of transformations (bijections) $\mathbf{S} \to \mathbf{S}$; Let $\mathbf{I}$ be another set (e.g., lists of integers, polynomials, etc.). A function $f: \mathbf{S} \to \mathbf{I}$ is called a $\hat{\mathfrak{C}}$-*invariant* if $f = f\hat{g}$ for all $\hat{g} \in \hat{\mathfrak{C}}$; i.e., if f is "constant" on each $\hat{\mathfrak{C}}(s), s \in \mathbf{S}$. The *value* of f at s is sometimes informally referred to as a $\hat{\mathfrak{C}}$-invariant. An invariant f is *complete* if $f(s_1) = f(s)$ implies $\hat{\mathfrak{C}}(s_1) = \hat{\mathfrak{C}}(s_2)$ for all $s_1, s_2 \in \mathbf{S}$; thus $f(s)$ serves to label $\hat{\mathfrak{C}}(s)$ in the quotient set $\{\hat{\mathfrak{C}}(t): t \in \mathbf{S}\}$.

with the singular pencil $[\lambda I - A, B]$ [10]; the connection is suggested by the work of Rosenbrock [11].

Remark 2.2. If (T, F, G) is the triple which transforms (A, B) into feedback canonical form (A_c, B_c), then it is easy to see that the matrix pair $(A_c - B_cG^{-1}FT^{-1}, B_c)$ is in the quasi-canonical form discovered by Luenberger [12].

The natural next step in the study of feedback invariants is to investigate the effect of $\mathfrak{C}$ on triples (C, A, B). Elements on $\mathfrak{C}$ act on (C, A, B) in the obvious way:

$$(C, A, B) \mapsto (CT^{-1}, T(A + BF)T^{-1}, TBG).$$

Note that the function $\hat{\phi}$, defined by $(C, A, B) \mapsto \mathbf{I}(A, B)$, is an orbital invariant of $\mathfrak{C}$. However, since $\mathbf{I}(A, B)$ does not uniquely characterize (C, A, B), $\hat{\phi}$ is not complete. The (perhaps hopeless) search for a complete $\mathfrak{C}$-invariant for (C, A, B) remains a challenging and unsolved problem.

3. Prime systems. Section 2 summarizes the fact that if (A, B) is controllable, then $\mathscr{X}$ can be expressed as a direct sum of m singly-generated controllability subspaces. Below we further investigate such decompositions, but now with the additional constraint that the C-images of the component subspaces decompose $\mathscr{Y}$ in a special way. Because of the additional constraint, the proposed decomposition can only be accomplished for a restricted class of systems. We begin by describing the elementary components of the decomposition.

A singly-generated (A, B)-controllability subspace[2] $\mathscr{W}$ is called *prime* if for some $F \in \mathbf{F}(A, B, \mathscr{W})$ and $b \in \mathscr{B} \cap \mathscr{W}$,

$$C(A + BF)^{i-1}b = 0, \qquad i \in \bar{\mathbf{d}}(\mathscr{W}), \tag{3.1}$$

and

$$C(A + BF)^{d(\mathscr{W})-1}b \neq 0, \tag{3.2}$$

It is easy to show that if $\mathscr{W}$ is prime, then both (3.1) and (3.2) hold for all $F \in \mathbf{F}(A, B, \mathscr{W})$ and all nonzero $b \in \mathscr{B} \cap \mathscr{W}$. Furthermore, (3.1) and (3.2) imply that

$$C(\lambda I - A - BF)^{-1}b = \frac{1}{\alpha(\lambda)}y,$$

where $\alpha(\lambda)$ is the minimal polynomial of $(A + BF)|\mathscr{W}$ and

$$y = C(A + BF)^{d(\mathscr{W})-1}b.$$

The vector $y \in \mathscr{Y}$ is independent of $F \in \mathbf{F}(A, B, \mathscr{W})$ and is a basis for the one-dimensional subspace $C\mathscr{W}$.

A standard controllable triple (C, A, B) is called *prime* if there exist prime subspaces, $\mathscr{W}_i$ $(i \in \mathbf{m})$, satisfying

$$\mathscr{X} = \mathscr{W}_1 \oplus \mathscr{W}_2 \oplus \cdots \oplus \mathscr{W}_m \tag{3.3}$$

and

$$\mathscr{Y} = C\mathscr{W}_1 \oplus C\mathscr{W}_2 \oplus \cdots \oplus C\mathscr{W}_m. \tag{3.4}$$

[2] An (A, B) controllability subspace $\mathscr{W}$ is *singly-generated* if $d(\mathscr{B} \cap \mathscr{W}) = 1$.

If such a decomposition exists, then for any $F \in \bigcap_{i \in \mathbf{m}} \mathbf{F}(A, B, \mathscr{W}_i)$ and nonzero $b_i \in \mathscr{B} \cap \mathscr{W}_i$,

$$C(\lambda I - A - BF)^{-1} b_i = \frac{1}{\alpha_i(\lambda)} y_i, \qquad i \in \mathbf{m}. \tag{3.5}$$

Property (3.4) insures that the y_i are independent and that $\{y_1, y_2, \cdots, y_m\}$ is a basis for $\mathscr{Y}$.

Prime systems are characterized as follows.

THEOREM 3.1. *Let (C, A, B) be a standard, controllable triple. Then (C, A, B) is prime if and only if*

$$d(\mathscr{Y}) = d(\mathscr{U}), \tag{3.6}$$

$$\mathscr{S} = 0, \tag{3.7}$$

$$\mathscr{T}_i = \mathscr{B} + A\mathscr{B} + \cdots + A^{i-1}\mathscr{B}, \qquad i \in \mathbf{n}. \tag{3.8}$$

Let us write $\overline{\mathfrak{C}}$ for the group of transformations (H, T, F, G), where H is an automorphism of $\mathscr{Y}$ and T, F, G are as before. Elements of $\overline{\mathfrak{C}}$ act on (C, A, B) as follows:

$$(C, A, B) \mapsto (HCT^{-1}, T(A + BF)T^{-1}, TBG).$$

Clearly the function $\bar{\phi}$: {triples (C, A, B)} → {lists of positive integers}, $(C, A, B) \mapsto \mathbf{I}(A, B)$, is an orbital invariant of $\overline{\mathfrak{C}}$; in fact, the restriction of $\bar{\phi}$ to prime triples is even complete. Relations (3.3) and (3.4) imply that a $\overline{\mathfrak{C}}$-canonical representation for a prime triple (C, A, B) is of the form

$$A \sim \text{block diag.} [A_1, A_2, \cdots, A_m],$$
$$B \sim \text{block diag.} [B_1, B_2, \cdots, B_m],$$
$$C \sim \text{block diag.} [C_1, C_2, \cdots, C_m],$$

where

$$C_i = [1 \quad 0 \cdots]_{1 \times n_i}, \qquad A_i = \begin{bmatrix} 0 & 1 & 0 & . & . & . \\ 0 & 0 & 1 & 0 & . & . \\ . & . & 0 & 1 & & . \\ . & . & . & & & . \\ . & . & . & . & & 1 \\ . & . & . & . & . & 0 \end{bmatrix}_{n_i \times n_i}, \qquad B_i = \begin{bmatrix} 0 \\ . \\ . \\ . \\ 0 \\ 1 \end{bmatrix}_{n_i \times 1}$$

and $\mathbf{I}(A, B) = \{n_1, n_2, \cdots, n_m\}$. This structure is called the *prime canonical form* for the prime triple (C, A, B).

Remark 3.1. One possible application of the preceding concepts is as follows. Let (C, A, B) be any standard, controllable triple with $d(\mathscr{Y}) = d(\mathscr{U}) = m$. Suppose it is desired to find a decomposition

$$\mathscr{Y} = \mathscr{Y}_1 \oplus \mathscr{Y}_2 \oplus \cdots \oplus \mathscr{Y}_m \tag{3.9}$$

$(d(\mathscr{Y}_i) > 0, i \in \mathbf{m})$ with the property that, for some F and G, the triple $(C, A + BF, BG)$ is decoupled relative to (3.9) (cf. [13]); i.e., with control law $u = Fx + Gv$ applied to $\dot{x} = Ax + Bu$, the ith component of v can control output $y_i \in \mathscr{Y}_i$ without influencing $y_j \in \mathscr{Y}_j, j \neq i$. The present problem differs from previously studied decoupling problems [7], in that here the partition of output variables to be decoupled is not specified at the outset.

Write $P: \mathscr{X} \to \mathscr{X}/\mathscr{S}$ for the canonical projection, $\bar{B} = PB$ and $\bar{C}$ for the unique solution to $\bar{C}P = C$. Let $F \in \mathbf{F}(A, B, \mathscr{S})$ be fixed and write $\bar{A}$ for the map induced in $\mathscr{X}/\mathscr{S}$ by $A + BF$. Starting with the results of [6], it is not difficult to show that a decomposition of $\mathscr{Y}$ with the preceding properties exists if and only if $(\bar{C}, \bar{A}, \bar{B})$ is prime.

3.1. Decomposition procedure. We now describe the procedure for computing prime subspaces $\mathscr{W}_i, i \in \mathbf{m}$, which satisfy (3.3) and (3.4). As a first step, define the sequence of subspaces

$$(3.10) \qquad \mathscr{V}_0 = 0, \quad \mathscr{V}_1 = A^{-1}\mathscr{B}, \quad \mathscr{V}_{i+1} = \mathscr{N} \cap A^{-1}(\mathscr{B} + \mathscr{V}_i), \qquad i \in \bar{\mathbf{n}}.$$

LEMMA 3.1. *If* (3.7) *and* (3.8) *hold, then the subspaces* $\mathscr{V}_i, i \in \mathbf{n}$, *are independent.*

Now assume that (3.6)–(3.8) hold. It follows from Lemma 3.1 and (3.10) that a map $F: \mathscr{X} \to \mathscr{U}$ can be defined so that

$$(3.11) \qquad (A + BF)\mathscr{V}_i \subset \mathscr{V}_{i-1}, \qquad i \in \mathbf{n}.$$

Write $d_0 = 0$ and

$$d_i = \sum_{j \in \mathbf{i}} d(\mathscr{B} \cap \mathscr{V}_j), \qquad i \in \mathbf{n}.$$

For each $i \in \mathbf{n}$, let $\{b_j; j \in \mathbf{d}_i - \mathbf{d}_{i-1}\}$ be a basis for $\mathscr{B} \cap \mathscr{V}_i$, and define

$$(3.12) \qquad \mathscr{W}_j = \langle A + BF | \mathscr{b}_j \rangle, \qquad j \in \mathbf{d}_i - \mathbf{d}_{i-1}.$$

In this way, exactly d_n subspaces $\mathscr{W}_i$ are defined. It will be shown below that if the conditions of Theorem 3.1 are satisfied, then $d_n = m$ and the $\mathscr{W}_i$ are prime subspaces satisfying (3.3) and (3.4).

LEMMA 3.2. *If* (3.7) *and* (3.8) *hold, then*

$$(3.13) \qquad \mathscr{B} = \mathscr{B} \cap \mathscr{V}_1 \oplus \mathscr{B} \cap \mathscr{V}_2 \oplus \cdots \oplus \mathscr{B} \cap \mathscr{V}_n.$$

Proof. From (1.1), $\mathscr{V}_1 \subset \mathscr{S}_0$; if $\mathscr{V}_i \subset \mathscr{S}_{i-1}$, then

$$\mathscr{V}_{i+1} = \mathscr{N} \cap A^{-1}(\mathscr{B} + \mathscr{V}_i) \subset \mathscr{N} \cap A^{-1}(\mathscr{B} + \mathscr{S}_{i-1}) = \mathscr{S}_i.$$

Thus

$$(3.14) \qquad \mathscr{V}_i \subset \mathscr{S}_{i-1}, \qquad i \in \mathbf{n}.$$

It will now be shown that

$$(3.15) \qquad \mathscr{T}_j \cap \mathscr{S}_{i-1} = \mathscr{T}_j \cap \mathscr{S}_i + \mathscr{T}_j \cap \mathscr{V}_i, \qquad i, j \in \mathbf{n}.$$

From (1.3) and (3.8),

$$\begin{aligned} A\mathscr{T}_j &\subset A(\mathscr{B} + A\mathscr{B} + \cdots + A^{j-1}\mathscr{B}) + \mathscr{B} \\ &= \mathscr{T}_{j+1} \\ &= A(\mathscr{N} \cap \mathscr{T}_j) + \mathscr{B}, \qquad j \in \mathbf{n}. \end{aligned}$$

There follows

$$\begin{aligned} \mathscr{S}_0 \cap \mathscr{T}_j &= \mathscr{T}_j \cap (\mathscr{T}_j \cap \mathscr{N} + A^{-1}\mathscr{B}) \\ &= \mathscr{T}_j \cap \mathscr{S}_1 + \mathscr{T}_j \cap \mathscr{V}_1, \qquad j \in \mathbf{n}, \end{aligned}$$

so (3.15) is true for $i = 1$; if (3.15) holds at i, then using (1.1), (1.3), and (3.10) yields

$$\begin{aligned} A(\mathscr{T}_j \cap \mathscr{S}_i) &\subset \mathscr{T}_{j+1} \cap (\mathscr{B} + \mathscr{S}_{i-1}) \\ &= \mathscr{T}_{j+1} \cap \mathscr{S}_{i-1} + \mathscr{B} \\ &= \mathscr{T}_{j+1} \cap (\mathscr{T}_j \cap \mathscr{S}_i + \mathscr{T}_j \cap \mathscr{V}_i) + \mathscr{B} \\ &\subset \mathscr{T}_{j+1} \cap \mathscr{S}_i + \mathscr{V}_i + \mathscr{B}, \qquad j \in \mathbf{n}. \end{aligned}$$

There follows

$$\begin{aligned} \mathscr{T}_j \cap \mathscr{S}_i &= \mathscr{T}_j \cap \mathscr{S}_i \cap A^{-1}(\mathscr{T}_{j+1} \cap \mathscr{S}_i + \mathscr{V}_i + \mathscr{B}) \\ &= \mathscr{T}_j \cap \mathscr{S}_i \cap A^{-1}((A(\mathscr{N} \cap \mathscr{T}_j) + \mathscr{B}) \cap \mathscr{S}_i + \mathscr{V}_i + \mathscr{B}) \\ &= \mathscr{T}_j \cap \mathscr{S}_i \cap A^{-1}((A(\mathscr{N} \cap \mathscr{T}_j) + \mathscr{B}) \cap (\mathscr{S}_i + \mathscr{B}) + \mathscr{V}_i) \\ &= \mathscr{T}_j \cap \mathscr{S}_i \cap A^{-1}((A(\mathscr{N} \cap \mathscr{T}_j)) \cap (\mathscr{B} + \mathscr{S}_i) + \mathscr{V}_i + \mathscr{B}) \\ &= \mathscr{T}_j \cap \mathscr{S}_i \cap (\mathscr{N} \cap \mathscr{T}_j \cap A^{-1}(\mathscr{B} + \mathscr{S}_i) + A^{-1}(\mathscr{V}_i + \mathscr{B})) \\ &= \mathscr{T}_j \cap \mathscr{S}_i \cap (\mathscr{T}_j \cap \mathscr{S}_{i+1} + A^{-1}(\mathscr{V}_i + \mathscr{B})) \\ &= \mathscr{T}_j \cap \mathscr{S}_{i+1} + \mathscr{T}_j \cap \mathscr{S}_i \cap A^{-1}(\mathscr{V}_i + \mathscr{B}) \\ &= \mathscr{T}_j \cap \mathscr{S}_{i+1} + \mathscr{T}_j \cap \mathscr{S}_i \cap \mathscr{V}_{i+1} \\ &= \mathscr{T}_j \cap \mathscr{S}_{i+1} + \mathscr{T}_j \cap \mathscr{V}_{i+1}, \qquad j \in \mathbf{n}. \end{aligned}$$

Thus (3.15) is true.

Since $\mathscr{T}_1 = \mathscr{B}$, it follows from (3.15) that

$$\mathscr{B} \cap \mathscr{S}_{i-1} = \mathscr{B} \cap \mathscr{S}_i + \mathscr{B} \cap \mathscr{V}_i, \qquad i \in \mathbf{n}. \tag{3.16}$$

Thus

$$\begin{aligned} \mathscr{B} &= \mathscr{B} \cap \mathscr{S}_0 \\ &= \mathscr{B} \cap \mathscr{V}_1 + \mathscr{B} \cap \mathscr{S}_1 \\ &= \mathscr{B} \cap \mathscr{V}_1 + \mathscr{B} \cap \mathscr{V}_2 + \mathscr{B} \cap \mathscr{S}_2 \qquad (3.17) \\ &\;\;\vdots \\ &= \sum_{i \in \mathbf{n}} \mathscr{B} \cap \mathscr{V}_i + \mathscr{B} \cap \mathscr{S}_n. \end{aligned}$$

But $\mathscr{S}_n = \mathscr{S} = 0$, so

$$\mathscr{B} = \sum_{i \in \mathbf{n}} \mathscr{B} \cap \mathscr{V}_i. \tag{3.18}$$

To show that this is a direct sum, it is enough to prove Lemma 3.1.

Proof of Lemma 3.1. From (1.1), (3.10) and (3.17),

$$\begin{aligned} A\left(\mathscr{S}_i \cap \sum_{j \in \mathbf{i}} \mathscr{V}_j\right) &\subset A\mathscr{S}_i \cap A \sum_{j \in \mathbf{i}} \mathscr{V}_j \\ &\subset (\mathscr{B} + \mathscr{S}_{i-1}) \cap \left(\mathscr{B} + \sum_{j \in \mathbf{i}} \mathscr{V}_{j-1}\right) \\ &= \mathscr{B} + \mathscr{S}_{i-1} \cap \left(\mathscr{B} \cap \mathscr{S}_{i-1} + \sum_{j \in \mathbf{i}} \mathscr{V}_{j-1}\right) \\ &= \mathscr{B} + \mathscr{S}_{i-1} \cap \sum_{j \in \mathbf{i}} \mathscr{V}_{j-1}, \qquad i \in \mathbf{n}. \end{aligned}$$

Thus

$$\mathscr{S}_i \cap \sum_{j \in \mathbf{i}} \mathscr{V}_j \subset \mathscr{S}_i \cap A^{-1}\left(\mathscr{B} + \mathscr{S}_{i-1} \cap \sum_{j \in \mathbf{i}} \mathscr{V}_{j-1}\right), \qquad i \in \mathbf{n}. \tag{3.19}$$

But $\mathscr{S}_0 \cap \mathscr{V}_0 = 0$ and if $\mathscr{S}_{i-1} \cap \sum_{j \in \mathbf{i}} \mathscr{V}_{j-1} = 0$, then by (3.19),

$$\mathscr{S}_i \cap \sum_{j \in \mathbf{i}} \mathscr{V}_j \subset \mathscr{S}_i \cap A^{-1}\mathscr{B} \subset \mathscr{N} \cap A^{-1}\mathscr{B} \subset \mathscr{S} = 0.$$

There follows

$$\mathscr{V}_i \cap \sum_{j \in \mathbf{i}} \mathscr{V}_i \subset \mathscr{S}_i \cap \sum_{j \in \mathbf{i}} \mathscr{V}_j = 0, \qquad i \in \mathbf{n}, \tag{3.20}$$

so the $\mathscr{V}_i$ are independent.

Proof of Theorem 3.1. *Sufficiency.* Let F and b_j, $\mathscr{W}_j$ ($j \in \mathbf{d}_n$) be as defined in the decomposition procedure. By Lemma 3.2, $d_n = \sum_{i \in \mathbf{n}} d(\mathscr{B} \cap \mathscr{V}_i) = m$. It follows that $\{b_1, b_2, \cdots, b_m\}$ is a basis for $\mathscr{B}$. Clearly,

$$\sum_{j \in \mathbf{m}} \mathscr{W}_j = \sum_{j \in \mathbf{m}} \langle A + BF | \mathscr{b}_j \rangle = \left\langle A + BF \middle| \sum_{j \in \mathbf{m}} \mathscr{b}_j \right\rangle = \langle A + BF | \mathscr{B} \rangle = \mathscr{X}. \tag{3.21}$$

In addition, since C is epic,

$$\sum_{j \in \mathbf{m}} C\mathscr{W}_j = \mathscr{Y}. \tag{3.22}$$

Write $A_0 = A + BF$. From (3.10) and (3.11),

$$A_0^i \mathscr{V}_i \subset \mathscr{V}_0 = 0, \qquad i \in \mathbf{n}. \tag{3.23}$$

For $i > 1$, (3.10) provides

$$(3.24)\qquad \sum_{k\in\bar{\mathbf{i}}} A_0^{i-1-k}\mathscr{V}_i \subset \sum_{k\in\bar{\mathbf{i}}} \mathscr{V}_{k+1} \subset \mathscr{N}.$$

Since $b_j \in \mathscr{V}_i, j \in \mathbf{d}_i - \mathbf{d}_{i-1}$, it follows from (3.23) that

$$(3.25)\qquad A_0^i b_j = 0.$$

If $i > 1$, (3.24) implies that

$$(3.26)\qquad \sum_{k\in\bar{\mathbf{i}}} A_0^{i-1-k} b_j \subset \mathscr{N}.$$

Clearly $\mathscr{W}_j$ is spanned by the set $\{b_j, A_0 b_j, \cdots, A_0^{i-1} b_j\}$. Now if $i > 1$ and

$$(3.27)\qquad A_0^{i-1} b_j \in \sum_{k\in\bar{\mathbf{i}}} A_0^{i-1-k} b_j \in \mathscr{N},$$

then $\mathscr{W}_j \subset \mathscr{N}$; thus $\mathscr{W}_j \subset \mathscr{S} = 0$ or $b_j = 0$, which is a contradiction. Therefore (3.27) is false. It follows that $\{b_j, A_0 b_j, \cdots, A_0^{i-1} b_j\}$ is a basis for $\mathscr{W}_j$ and that $CA_0^{i-1} b_j \neq 0$. If we now write $n_j = i, j \in \mathbf{d}_i - \mathbf{d}_{i-1}, i \in \mathbf{n}$, then

$$(3.28)\qquad CA_0^{n_j-1} b_j \neq 0, \qquad j \in \mathbf{m},$$

and if $n_j > 1$,

$$(3.29)\qquad CA_0^{i-1} b_j = 0, \qquad i \in \bar{\mathbf{n}}_j, \quad j \in \mathbf{m}.$$

From (3.28) and (3.29), $d(C\mathscr{W}_j) = 1, j \in \mathbf{m}$. There follows

$$\sum_{j\in\mathbf{m}} d(C\mathscr{W}_j) = m = d(\mathscr{U}) = d(\mathscr{Y}) = d\left(\sum_{j\in\mathbf{m}} C\mathscr{W}_j\right),$$

so the $C\mathscr{W}_j$ are independent and (3.4) is true.

Since

$$C\left(\mathscr{W}_j \cap \sum_{\substack{i\in\mathbf{m}\\ i\neq j}} \mathscr{W}_i\right) \subset C\mathscr{W}_j \cap \sum_{\substack{i\in\mathbf{m}\\ i\neq j}} C\mathscr{W}_i = 0,$$

we have

$$\mathscr{W}_j \cap \sum_{\substack{i\in\mathbf{m}\\ i\neq j}} \mathscr{W}_i \subset \mathscr{N}.$$

But $\mathscr{W}_j \cap \sum_{i\in\mathbf{m}, i\neq j} \mathscr{W}_i$ is A_0-invariant and is therefore contained in $S = 0$. It follows that the $\mathscr{W}_j$ are independent and that (3.3) is true.

From the independence of the $\mathscr{W}_j$, $\mathscr{B} = \mathscr{B} \cap \mathscr{W}_1 \oplus \mathscr{B} \cap \mathscr{W}_2 \oplus \cdots \oplus \mathscr{B} \cap \mathscr{W}_{\mathbf{m}}$. Thus $m = \sum_{j\in\mathbf{m}} d(\mathscr{B} \cap \mathscr{W}_j)$; but $d(\mathscr{B} \cap \mathscr{W}_j) > 0, j \in \mathbf{m}$, so $d(\mathscr{B} \cap \mathscr{W}_j) = 1$, $j \in \mathbf{m}$. From this and (3.28), (3.29), it follows that the $\mathscr{W}_j$ are each prime.

Necessity. Let $\mathscr{W}_k, k \in \mathbf{m}$, be a set of prime subspaces satisfying (3.3) and (3.4); write $\mathscr{b}_k = \mathscr{B} \cap \mathscr{W}_k$ and $n_k = d(\mathscr{W}_k)$. Choose $F \in \bigcap_{k\in\mathbf{m}} \mathbf{F}(A, B, \mathscr{W}_k)$ and write $A_0 = A + BF$. Since $\mathscr{W}_k$ is prime, a simple computation provides

$$(3.30)\qquad \mathscr{B} + A_0 \sum_{j\in\mathbf{i}} A_0^{j-1}\mathscr{b}_k = \mathscr{B} + A_0\left(\mathscr{N} \cap \sum_{j\in\mathbf{i}} A_0^{j-1}\mathscr{b}_k\right), \qquad i \in \mathbf{n}, \quad k \in \mathbf{m}.$$

Now $\mathscr{T}_1 = \mathscr{B}$, so (3.8) holds at $i = 1$; if true at $i \geqq 1$, then with $r \equiv i + 1$

$$\begin{aligned}
\sum_{j\in\mathbf{r}} A^{j-1}\mathscr{B} &= \sum_{j\in\mathbf{r}} A_0^{j-1}\mathscr{B} \\
&= A_0 \sum_{j\in\mathbf{i}} A_0^{j-1}\mathscr{B} + \mathscr{B} \\
&= A_0 \sum_{k\in\mathbf{m}} \sum_{j\in\mathbf{i}} A_0^{j-1}\mathscr{C}_k + \mathscr{B} \\
&= A_0 \sum_{k\in\mathbf{m}} \left(\mathscr{N} \cap \sum_{j\in\mathbf{i}} A_0^{j-1}\mathscr{C}_k\right) + \mathscr{B} \\
&\subset A_0\left(\mathscr{N} \cap \left(\sum_{k\in\mathbf{m}} \sum_{j\in\mathbf{i}} A_0^{j-1}\mathscr{C}_k\right)\right) + \mathscr{B} \\
&= A(\mathscr{N} \cap \sum_{j\in\mathbf{i}} A^{j-1}\mathscr{B}) + \mathscr{B} \\
&= A\mathscr{T}_i + \mathscr{B} \\
&= \mathscr{T}_{i+1} \\
&\subset \sum_{j\in\mathbf{r}} A^{j-1}\mathscr{B}.
\end{aligned}$$

By induction, (3.8) is true.

Condition (3.6) follows immediately from (3.4). From (3.5) it is obvious that the transfer matrix $C(\lambda I - A_0)^{-1}B$ is invertible over the field of rational functions in λ. Therefore by [7, Thm. 5], $\mathscr{B} \cap \mathscr{S} = 0$. By (1.5), $\mathscr{R} = 0$. But from (3.8), $\mathscr{T} = \mathscr{T}_n = \mathscr{X}$, so by Lemma 1.1, $\mathscr{S} = 0$.

4. Canonical structure of (C, A, B). Thus far it has been shown that if $\mathfrak{C}$ is imbedded in a larger group $\bar{\mathfrak{C}}$, then it is possible to find a complete $\bar{\mathfrak{C}}$-invariant for prime systems. To identify and interpret feedback invariants for arbitrary triples (C, A, B), it proves useful to imbed $\mathfrak{C}$ in a still larger group $\mathfrak{C}^*$. The elements of $\mathfrak{C}^*$ are lists of the form (H, K, T, F, G), where $K: \mathscr{Y} \to \mathscr{X}$ is arbitrary and H, T, F, G are as before. The action of $\mathfrak{C}^*$ on (C, A, B) is defined by

$$(C, A, B) \mapsto (HCT^{-1}, T(A + BF + KC)T^{-1}, TBG).$$

Although we are not aware of any system theoretic interpretation[3] of the transformation K, all $\mathfrak{C}^*$-invariants are also $\mathfrak{C}$-invariants, so the study of the action of $\mathfrak{C}^*$ on (C, A, B) does provide additional information regarding the feedback-invariant structure of (C, A, B). It turns out that $\mathfrak{C}^*$ is sufficiently rich to enable us to identify a complete $\mathfrak{C}^*$-invariant and this means that (C, A, B) possesses a $\mathfrak{C}^*$-canonical form. If desired, this canonical form can then be used to construct a quasi-canonical representation for (C, A, B) using only transformations from $\mathfrak{C}$.

Let us right away identify some of the more obvious properties of (C, A, B) which are $\mathfrak{C}^*$-invariant. It is straightforward to show that the map A_t and the subspaces $\mathscr{S}$, $\mathscr{R}_i$, $\mathscr{T}_i$, $i \in \mathbf{n}$, defined in § 1 are all invariant relative to transformations

[3] The matrix $A + BF + KC$ does appear in a Kalman filter or observer which estimates the state of the system (C, A, B) forced by the control $u = F\tilde{x}$ ($\tilde{x}$ = estimated state). However, the connection between $\mathfrak{C}^*$-invariants and the properties of these filters has not yet been determined.

H, K, F, G. It follows from this, that the coordinate-independent properties of $A_t, \mathscr{S}, \mathscr{R}_i, \mathscr{T}_i, i \in \mathbf{n}$, are $\mathfrak{C}^*$-invariant. In other words, $d(\mathscr{S}), d(\mathscr{R}_i), d(\mathscr{T}_i), i \in \mathbf{n}$, and the invariant factors of A_t are all $\mathfrak{C}^*$-invariant.

We now proceed to determine a $\mathfrak{C}^*$-canonical representation for (C, A, B). Our approach will be to decompose $\mathscr{X}$ into a direct sum of four subspaces, each with special properties relative to (C, A, B). For this, we need some preliminary results.

Let $\bar{\bar{\mathscr{B}}}$ be any subspace such that

$$\mathscr{B} = \bar{\bar{\mathscr{B}}} \oplus \mathscr{B} \cap \mathscr{S} \tag{4.1}$$

and write $\bar{\bar{\mathscr{T}}} = \bar{\bar{\mathscr{T}}}_n$, where

$$\bar{\bar{\mathscr{T}}}_0 \equiv 0, \qquad \bar{\bar{\mathscr{T}}}_i \equiv A(\mathscr{N} \cap \bar{\bar{\mathscr{T}}}_{i-1}) + \bar{\bar{\mathscr{B}}}, \qquad i \in \mathbf{n}. \tag{4.2}$$

LEMMA 4.1.

$$\mathscr{T} = \bar{\bar{\mathscr{T}}} \oplus \mathscr{S} \cap \mathscr{T}. \tag{4.3}$$

LEMMA 4.2. *There exists a map* $K_1 : \mathscr{Y} \to \mathscr{X}$ *such that*

$$\mathscr{T}_i = \sum_{j \in \mathbf{i}} (A + K_1 C)^{j-1} \mathscr{B}, \qquad i \in \mathbf{n}, \tag{4.4}$$

$$\bar{\bar{\mathscr{T}}}_i = \sum_{j \in \mathbf{i}} (A + K_1 C)^{j-1} \bar{\mathscr{B}}, \qquad i \in \mathbf{n}. \tag{4.5}$$

LEMMA 4.3. *There exists a subspace* $\mathscr{Z}$ *and a map* $K_0 : \mathscr{Y} \to \mathscr{X}$ *such that*

$$\mathscr{X} = \mathscr{Z} \oplus (\mathscr{S} + \mathscr{T}), \tag{4.6}$$

$$\mathscr{Y} = C\mathscr{Z} \oplus C\mathscr{T}, \tag{4.7}$$

$$(A + K_0 C)\mathscr{Z} \subset \mathscr{B} + \mathscr{Z}, \tag{4.8}$$

$$(A + K_0 C)(\mathscr{S} + \mathscr{T}) \subset \mathscr{S} + \mathscr{T}. \tag{4.9}$$

The decomposition of $\mathscr{X}$ and the identification of appropriate $\mathfrak{C}^*$-invariants will be accomplished in several steps:

I. *Decomposition of* $\mathscr{S}$. Write $\mathbf{I}_1 = \{\alpha_1(\lambda), \alpha_2(\lambda), \cdots, \alpha_\mu(\lambda)\}$ for the set of transmission polynomials of (C, A, B), listed in order of increasing degree; i.e., the elements of $\mathbf{I}_1$ are the invariant factors of A_t. Thus $\alpha_\mu(\lambda)$ is the minimal polynomial of $(A + BF)|\mathscr{S} \bmod \mathscr{R}$ for all $F \in \mathbf{F}(A, B, \mathscr{S})$. Define $F_0 \in \mathbf{F}(A, B, \mathscr{S})$ so that the minimal polynomial of $(A + BF_0)|\mathscr{R}$ is coprime with $\alpha_\mu(\lambda)$. Then define

$$\mathscr{X}_1 = \mathscr{S} \cap (\ker(\alpha_\mu(A + BF_0))).$$

It follows (cf. [10, Chap. 7, Thm. 1]) that

$$\mathscr{S} = \mathscr{X}_1 \oplus \mathscr{R} \tag{4.10}$$

and that

$$(A + BF_0)\mathscr{X}_1 \subset \mathscr{X}_1. \tag{4.11}$$

In addition, the invariant factors of $(A + BF_0)|\mathscr{X}_1$ coincide with the elements of $\mathbf{I}_1$.

Noting that $\mathscr{R} = \langle A + BF_0 | \mathscr{B} \cap \mathscr{R} \rangle$, define $\mathbf{I}_2 = \mathbf{I}(A + BF_0, B_2)$, where B_2 is the insertion of $\mathscr{B} \cap \mathscr{R}$ in $\mathscr{X}$. According to Brunovský [4], $\mathbf{I}_2$ depends only on the integers $d(\mathscr{B}_2 + (A + BF_0)\mathscr{B}_2 + \cdots + (A + BF_0)^{i-1}\mathscr{B}_2)$, $i \in \mathbf{n}$. But

$$\mathscr{R}_i = \mathscr{B}_2 + (A + BF_0)\mathscr{B}_2 + \cdots + (A + BF_0)^{i-1}\mathscr{B}_2, \qquad i \in \mathbf{n},$$

and the $d(\mathscr{R}_i)$ are $\mathfrak{C}^*$-invariant. It follows that $\mathbf{I}_2$ is $\mathfrak{C}^*$-invariant as well.

II. *Decomposition of $\mathscr{Y}$.* Choose $\mathscr{Z}$ and K_0 in accordance with Lemma 4.3. Define $\mathscr{Y}_1 = C\mathscr{Z}$ and $\mathscr{Y}_2 = C\mathscr{T}$; by (4.7),

$$\mathscr{Y} = \mathscr{Y}_1 \oplus \mathscr{Y}_2. \tag{4.12}$$

Define maps $C_i : \mathscr{X} \to \mathscr{Y}_i$, $i \in \mathbf{2}$, so that

$$\begin{aligned} C_1|\mathscr{Z} &= C|\mathscr{Z}, \\ C_1|(\mathscr{S} + \mathscr{T}) &= 0, \\ C_2|\mathscr{Z} &= 0, \\ C_2|(\mathscr{S} + \mathscr{T}) &= C|(\mathscr{S} + \mathscr{T}). \end{aligned} \tag{4.13}$$

Thus

$$C = C_1 \oplus C_2 \tag{4.14}$$

and by a simple computation,

$$\ker C_1 = \mathscr{N} + \mathscr{S} + \mathscr{T}, \tag{4.15}$$

$$\ker C_2 = \mathscr{N} + \mathscr{S} + \mathscr{Z}. \tag{4.16}$$

Since C is epic, C_1 and C_2 are also.

Define $\mathbf{I}_3 = \mathbf{I}(A' + C'K_0', C_1')$. From (4.15), $\mathscr{C}_1' = \mathscr{C}' \cap \mathscr{S}^\perp \cap \mathscr{T}^\perp$, while (4.9) provides $(A' + C'K_0')(\mathscr{S}^\perp \cap \mathscr{T}^\perp) \subset \mathscr{S}^\perp \cap \mathscr{T}^\perp$. Thus by Remark 1.1, $\mathscr{S}^\perp \cap \mathscr{T}^\perp = \langle A' + C'K_0' | \mathscr{C}_1' \rangle$. Recall that $\mathbf{I}_2$ is $\mathfrak{C}^*$-invariant; by duality $\mathbf{I}_3$ is also.

By (4.8), there exists a map F_1 such that

$$(A + K_0C + BF_1)\mathscr{Z} \subset \mathscr{Z}. \tag{4.17}$$

Since $\mathscr{S}^\perp \cap \mathscr{T}^\perp \subset \ker B'$, it follows that

$$\mathscr{S}^\perp \cap \mathscr{T}^\perp = \langle A' + C'K_0' + F_1'B' | \mathscr{C}_1' \rangle \tag{4.18}$$

and that

$$\mathbf{I}(A' + C'K_0 + F_1'B', C_1') = \mathbf{I}_3. \tag{4.19}$$

Relation (4.18) is equivalent to

$$\bigcap_{i \in \mathbf{n}} \ker (C_1(A + BF_1 + K_0C)^{i-1}) = \mathscr{S} + \mathscr{T}.$$

But from (4.6), $\mathscr{Z} \cap (\mathscr{S} + \mathscr{T}) = 0$, so the pair $(C_1|\mathscr{Z}, (A + BF_1 + K_0C)|\mathscr{Z})$ is observable.

III. *Decomposition of* $\mathscr{T}$. To decompose $\mathscr{T}$, define $\bar{\mathscr{T}}$ and K_1 in accordance with (4.1)–(4.5). Since $\mathscr{R} = \mathscr{S} \cap \mathscr{T}$, (4.3) provides

$$\mathscr{T} = \bar{\mathscr{T}} \oplus \mathscr{R}. \tag{4.20}$$

Write $\bar{C} = C_2|\bar{\mathscr{T}}$, $\bar{A} = (A + K_1C)|\bar{\mathscr{T}}$ and $\bar{B}:\mathscr{B} \to \bar{\mathscr{T}}$ for the insertion of $\bar{\mathscr{B}}$ in $\bar{\mathscr{T}}$.

We claim that the standard triple $(\bar{C}, \bar{A}, \bar{B})$ is prime. To show this, write $\bar{\mathscr{N}} = \ker \bar{C}$; from (4.13), $\bar{\mathscr{N}} = \mathscr{N} \cap \bar{\mathscr{T}}$. If $\bar{\mathscr{S}}$ is the largest $(\bar{A}, \bar{B})$-invariant subspace contained in $\bar{\mathscr{N}}$, then clearly $\bar{\mathscr{S}} \subset \mathscr{S}$. But $\bar{\mathscr{S}} \subset \bar{\mathscr{T}}$ and by (4.3), $\bar{\mathscr{T}} \cap \mathscr{S} = 0$; thus

$$\bar{\mathscr{S}} = 0. \tag{4.21}$$

From (4.2) and (4.5),

$$\bar{\mathscr{T}}_0 = 0, \qquad \bar{\mathscr{T}}_i = \bar{A}(\bar{\mathscr{N}} \cap \bar{\mathscr{T}}_{i-1}) + \bar{\mathscr{B}}, \qquad i \in \mathbf{n}, \tag{4.22}$$

and

$$\bar{\mathscr{T}}_i = \sum_{j \in \mathbf{i}} \bar{A}^{j-1}\bar{\mathscr{B}}, \qquad i \in \mathbf{n}. \tag{4.23}$$

Relation (4.21) implies that $(\bar{C}, \bar{A}, \bar{B})$ is left invertible [7, Thm. 5]. Since the $\bar{\mathscr{T}}_i$ satisfy (4.22) and $\bar{\mathscr{T}} = \bar{\mathscr{T}}_n$, by duality, $(\bar{C}, \bar{A}, \bar{B})$ is right invertible; thus $(\bar{C}, \bar{A}, \bar{B})$ is invertible. But this can be possible only if

$$d(\mathscr{Y}_2) = d(\bar{\mathscr{B}}). \tag{4.24}$$

It now follows from (4.21)–(4.24) and Theorem 3.1 that $(\bar{C}, \bar{A}, \bar{B})$ is prime.

Write $\mathbf{I}_4 = \mathbf{I}(A + K_1C, \bar{B})$, where $\bar{B}$ is now the insertion of $\bar{\mathscr{B}}$ in $\mathscr{X}$. We claim that $\mathbf{I}_4$ is $\mathfrak{C}^*$-invariant. To show this, note that $\mathscr{R}$ $(\subset \mathscr{N})$ is an $(A + K_1C, B)$-controllability subspace and that

$$\mathbf{I}(A + BF_0 + K_1C, B_2) = \mathbf{I}(A + BF_0, B_2) = \mathbf{I}_2.$$

Thus from (4.20),

$$\mathbf{I}(A + K_1C, B) = \mathbf{I}_2 \cup \mathbf{I}_4 \qquad \text{(as sets)}. \tag{4.25}$$

But from (4.4), the elements of $\mathbf{I}(A + K_1C, B)$ are determined by the integers $d(\mathscr{T}_i)$, $i \in \mathbf{n}$, which are $\mathfrak{C}^*$-invariant. It follows that $\mathbf{I}(A + K_1C, B)$ is $\mathfrak{C}^*$-invariant; by (4.25), $\mathbf{I}_4$ is also.

IV. *Decomposition of* $\mathscr{X}$. From (4.3) and (4.6), $\mathscr{X} = \mathscr{L} \oplus \bar{\mathscr{T}} \oplus \mathscr{S}$; from this and (4.10) there follows

$$\mathscr{X} = \mathscr{L} \oplus \bar{\mathscr{T}} \oplus \mathscr{X}_1 \oplus \mathscr{R}.$$

Define F on $\mathscr{X}$ so that

$$F|(\mathscr{X}_1 \oplus \mathscr{R}) = F_0|(\mathscr{X}_1 \oplus \mathscr{R}),$$
$$F|\mathscr{Z} = F_1|\mathscr{Z},$$
$$F|\bar{\mathscr{T}} = 0.$$

Define K on $\mathscr{Y}$ so that

$$K|\mathscr{Y}_1 = K_0|\mathscr{Y}_1,$$
$$K|\mathscr{Y}_2 = K_1|\mathscr{Y}_2.$$

It follows that

$$(A + BF + KC)|\mathscr{X}_1 = (A + BF_0)|\mathscr{X}_1,$$
$$(A + BF + KC)|\mathscr{R} = (A + BF_0)|\mathscr{R},$$
$$(A + BF + KC)|\mathscr{Z} = (A + BF_1 + K_0C)|\mathscr{Z},$$
$$(A + BF + KC)|\bar{\mathscr{T}} = (A + K_1C)|\bar{\mathscr{T}},$$

so all properties in I–III remain unchanged if F is substituted for F_0 or F_1 and K is substituted for K_0 or K_1.

We now summarize what has been accomplished. For this, write $\mathscr{X}_2 = \mathscr{R}$, $\mathscr{X}_3 = \mathscr{Z}$, and $\mathscr{X}_4 = \bar{\mathscr{T}}$.

THEOREM 4.1. *Let (C, A, B) be a fixed, standard triple. There exist subspaces $\mathscr{X}_i (i \in \mathbf{4})$, $\mathscr{Y}_j (j \in \mathbf{2})$ and maps $F : \mathscr{X} \to \mathscr{U}$, $K : \mathscr{Y} \to \mathscr{X}$, $C_j : \mathscr{X} \to \mathscr{Y}_j (j \in \mathbf{2})$ for which the following conditions hold*:

$$\mathscr{X} = \mathscr{X}_1 \oplus \mathscr{X}_2 \oplus \mathscr{X}_3 \oplus \mathscr{X}_4,$$
$$\mathscr{Y} = \mathscr{Y}_1 \oplus \mathscr{Y}_2,$$
$$(A + BF + KC)\mathscr{X}_i \subset \mathscr{X}_i, \qquad i \in \mathbf{4},$$
$$C = C_1 \oplus C_2,$$
$$\mathscr{X}_1 \oplus \mathscr{X}_2 \oplus \mathscr{X}_3 \subset \ker C_2,$$
$$\mathscr{X}_1 \oplus \mathscr{X}_2 \oplus \mathscr{X}_4 \subset \ker C_1,$$
$$\mathscr{B} = \mathscr{B} \cap \mathscr{X}_2 \oplus \mathscr{B} \cap \mathscr{X}_4.$$

Write B_i for the insertion of $\mathscr{B} \cap \mathscr{X}_i$ in $\mathscr{X}_i (i \in \{2, 4\})$, $A_i = (A + BF + KC)|\mathscr{X}_i (i \in \mathbf{4})$, $C_i = C_{i-2}|\mathscr{X}_i (i \in \{3, 4\})$ and let $\mathbf{I}_i (i \in \mathbf{4})$ be the lists determined by I–III. *Then*

(i) *the invariant factors of A_1 coincide with the elements of $\mathbf{I}_1$,*

(ii) *(A_2, B_2) is controllable and $\mathbf{I}(A_2, B_2) = \mathbf{I}_2$,*

(iii) *(C_3, A_3) is observable and $\mathbf{I}(A_3', C_3') = \mathbf{I}_3$.*

(iv) *(C_4, A_4, B_4) is prime and $\mathbf{I}(A_4, B_4) = \mathbf{I}_4$.*

The theorem states that $\mathscr{X}$ can be decomposed into four independent subspaces, each with special properties. Relative to the triple $(C, A + BF + KC, B)$, the subspace $\mathscr{X}_1$ is both uncontrollable and unobservable while $\mathscr{X}_2$ is unobservable and controllable. The subspace $\mathscr{X}_3$ is observable but uncontrollable and $\mathscr{X}_4$ is both controllable and observable. The decomposition is reminiscent of Kalman's classical decomposition of $\mathscr{X}$ (relative to (C, A, B)) in which the underlying group consists of only coordinate transformations T.

The theorem clearly implies that any standard triple (C, A, B) is $\mathfrak{C}^*$-equivalent to a triple of the form

$$C \sim \begin{bmatrix} 0 & 0 & C_3 & 0 \\ 0 & 0 & 0 & C_4 \end{bmatrix},$$

$$A \sim \begin{bmatrix} A_1 & 0 & 0 & 0 \\ 0 & A_2 & 0 & 0 \\ 0 & 0 & A_3 & 0 \\ 0 & 0 & 0 & A_4 \end{bmatrix}, \qquad B \sim \begin{bmatrix} 0 & 0 \\ B_2 & 0 \\ 0 & 0 \\ 0 & B_4 \end{bmatrix},$$

where A_i is a $d(\mathscr{X}_i) \times d(\mathscr{X}_i)$ matrix. Property (i) shows that A_1 can be represented in the rational canonical form determined by $\mathbf{I}_1$. Property (ii) implies that (A_2, B_2) can be expressed in the feedback canonical form determined by $\mathbf{I}_2$. Similarly, property (iii) means that (A'_3, C'_3) can be written in the feedback canonical form specified by $\mathbf{I}_3$. Finally, property (iv) indicates that (C_4, A_4, B_4) can be represented in the prime canonical form determined by $\mathbf{I}_4$. In this way, a $\mathfrak{C}^*$-canonical structure for (C, A, B) results which depends only on the lists $\mathbf{I}_i\,(i \in \mathbf{4})$. We have therefore established the following result.

COROLLARY 4.1. *The function* ϕ^*: {standard triples (C, A, B)} $\rightarrow$ {ordered lists}, $(C, A, B) \mapsto (\mathbf{I}_1, \mathbf{I}_2, \mathbf{I}_3, \mathbf{I}_4)$ *is a complete orbital invariant of* $\mathfrak{C}^*$.

Remark 4.1. The preceding canonical structure is closely related to the classical canonical form for the singular pencil

$$\begin{bmatrix} \lambda I - A & B \\ C & 0 \end{bmatrix}.$$

The invariants identified above correspond to the Kronecker invariants associated with this pencil.

The $\mathfrak{C}^*$-canonical structure for (C, A, B) can be used to construct a quasi-canonical form relative to just $\mathfrak{C}$. This can be accomplished by simply "canceling out" the maps K and H which are required to transform (C, A, B) into $\mathfrak{C}^*$-canonical form. This is illustrated by the following example.

Example. Let (C, A, B) be fixed and suppose that $\mathbf{I}_1 = \{\lambda - 3, \lambda^2 - 2\lambda - 3\}$, $\mathbf{I}_2 = \{2\}, \mathbf{I}_3 = \{2\}, \mathbf{I}_4 = \{2, 3\}$. Then (C, A, B) is $\mathfrak{C}$-equivalent to the matrix triple

$$C \sim H\left[\begin{array}{c|cc|cc|cc|cc|ccc} & & & & & 0 & 1 & & & & & \\ & & & & & & & 1 & 0 & & & \\ & & & & & & & & & 1 & 0 & 0 \end{array}\right],$$

$$A \sim \left[\begin{array}{c|cc|cc|cc|cc|ccc} 3 & & & & & & x & x & & x & & \\ \hline & 0 & 1 & & & & x & x & & x & & \\ & 3 & 2 & & & & x & x & & x & & \\ \hline & & & 0 & 1 & & x & x & & x & & \\ & & & 0 & 0 & & & & & & & \\ \hline & & & & & 0 & x & x & & x & & \\ & & & & & 1 & x & x & & x & & \\ \hline & & & & & & x & x & 1 & x & & \\ & & & & & & & 0 & 0 & & & \\ \hline & & & & & & x & x & & x & 1 & 0 \\ & & & & & & x & x & & x & 0 & 1 \\ & & & & & & & & & 0 & 0 & 0 \end{array}\right], \quad B \sim \left[\begin{array}{ccc} & & \\ \hline & & \\ & & \\ \hline 0 & & \\ 1 & & \\ \hline & & \\ & & \\ \hline & 0 & \\ & 1 & \\ \hline & & 0 \\ & & 0 \\ & & 1 \end{array}\right].$$

In the above matrices, entries not specified are zero, x's denote numbers not uniquely determined by (C, A, B), and H is an invertible matrix. The $\mathfrak{C}^*$-canonical form for (C, A, B) also has the same structure, but in this case the x's are all zero and H is the identity matrix.

Proof of Lemma 4.1. It will first be shown that

$$\mathcal{S} \cap \bar{\mathcal{T}}_i = 0, \qquad i \in \mathbf{n}. \tag{4.26}$$

Since $\bar{\mathcal{T}}_i = \bar{\mathcal{B}}$, (4.26) holds at $i = 1$; suppose it is true at i. If $x \in \mathcal{S} \cap \bar{\mathcal{T}}_{i+1}$, then by (4.2), $x = At + b$ for some $t \in \mathcal{N} \cap \bar{\mathcal{T}}_i$, $b \in \bar{\mathcal{B}}$. Since $x \in \mathcal{S}$, there follows $At \in \mathcal{B} + \mathcal{S}$ or $t \in \bar{\mathcal{T}}_i \cap \mathcal{N} \cap A^{-1}(\mathcal{B} + \mathcal{S}) = \mathcal{S} \cap \bar{\mathcal{T}}_i = 0$; thus $t = 0$ and $x = b \in \bar{\mathcal{B}} \cap \mathcal{S} = 0$, so $\bar{\mathcal{T}}_{i+1} \cap \mathcal{S} = 0$. It now follows from (4.26) that

$$\mathcal{S} \cap \bar{\mathcal{T}} = 0.$$

To complete the proof, it is enough to show that

$$\mathcal{R}_i + \bar{\mathcal{T}}_i = \mathcal{T}_i, \qquad i \in \mathbf{n}. \tag{4.27}$$

For if this is so, then

$$\mathcal{T} = \bar{\mathcal{T}} \oplus \mathcal{R} = \bar{\mathcal{T}} \oplus \mathcal{T} \cap \mathcal{S}.$$

Since $\mathcal{R}_1 = \mathcal{B} \cap \mathcal{S}$, (4.27) holds at $i = 1$; if true at i, then

$$\begin{aligned} \mathcal{T}_{i+1} &= A(\mathcal{N} \cap \mathcal{T}_i) + \mathcal{B} \\ &= A(\mathcal{N} \cap (\mathcal{R}_i + \bar{\mathcal{T}}_i)) + \mathcal{B} \\ &= A(\mathcal{N} \cap \bar{\mathcal{T}}_i) + A\mathcal{R}_i + \mathcal{B} \\ &= A(\mathcal{N} \cap \bar{\mathcal{T}}_i) + (A\mathcal{R}_i + \mathcal{B}) \cap (\mathcal{B} + \mathcal{S}) \\ &= A(\mathcal{N} \cap \bar{\mathcal{T}}_i) + \bar{\mathcal{B}} + (A\mathcal{R}_i + \mathcal{B}) \cap \mathcal{S} \\ &= \bar{\mathcal{T}}_{i+1} + \mathcal{R}_{i+1}. \end{aligned}$$

Proof of Lemma 4.2. For $i \in \mathbf{n}$, define subspaces $\hat{\mathscr{T}}_i$ so that

$$\bar{\bar{\mathscr{T}}}_i = \hat{\mathscr{T}}_i \oplus \bar{\bar{\mathscr{T}}}_i \cap (\bar{\bar{\mathscr{T}}}_{i-1} + \mathscr{N}).$$

Since $\bar{\bar{\mathscr{T}}}_{i-1} \subset \bar{\bar{\mathscr{T}}}_i$, $i \in \mathbf{n}$, it follows that the subspaces $\hat{\mathscr{T}}_i$ $(i \in \mathbf{n})$ and $\mathscr{N}$ are independent. Therefore, if $\{t_i, i \in \mathbf{k}\}$ is a basis for $\sum_{i \in \mathbf{n}} \hat{\mathscr{T}}_i$, then $\{Ct_i, i \in \mathbf{k}\}$ is a basis for $C \sum_{i \in \mathbf{n}} \hat{\mathscr{T}}_i$. Define K_1 so that $K_1 C t_i = -At_i (i \in \mathbf{k})$. It follows that $(A + K_1 C) \cdot \sum_{i \in \mathbf{n}} \hat{\mathscr{T}}_i = 0$.

Observe that (4.5) holds at $i = 1$; if true for some $i \geqq 1$, then

$$\begin{aligned}
\bar{\mathscr{B}} + (A + K_1 C)\bar{\bar{\mathscr{T}}}_i &= (A + K_1 C)(\hat{\mathscr{T}}_i + \bar{\bar{\mathscr{T}}}_i \cap (\bar{\bar{\mathscr{T}}}_{i-1} + \mathscr{N})) + \bar{\mathscr{B}} \\
&= (A + K_1 C)(\hat{\mathscr{T}}_i + \bar{\bar{\mathscr{T}}}_{i-1} + \mathscr{N} \cap \bar{\bar{\mathscr{T}}}_i) + \bar{\mathscr{B}} \\
&= (A + K_1 C)\bar{\bar{\mathscr{T}}}_{i-1} + A(\mathscr{N} \cap \bar{\bar{\mathscr{T}}}_i) + \bar{\mathscr{B}} \\
&= \sum_{j \in \mathbf{i}} (A + K_1 C)^{j-1} \bar{\mathscr{B}} + \bar{\bar{\mathscr{T}}}_{i+1} \\
&= \bar{\bar{\mathscr{T}}}_i + \bar{\bar{\mathscr{T}}}_{i+1} \\
&= \bar{\bar{\mathscr{T}}}_{i+1}.
\end{aligned}$$

By induction on i, (4.5) now follows.

To show that (4.4) holds, write

$$\begin{aligned}
\mathscr{T}_{i+1} &= \bar{\bar{\mathscr{T}}}_{i+1} + \mathscr{R}_{i+1} \\
&= (A + K_1 C)\bar{\bar{\mathscr{T}}}_i + \bar{\mathscr{B}} + (A\mathscr{R}_i + \mathscr{B}) \cap \mathscr{S} \\
&= (A + K_1 C)\bar{\bar{\mathscr{T}}}_i + (A\mathscr{R}_i + \mathscr{B}) \cap (\mathscr{B} + \mathscr{S}) \\
&= (A + K_1 C)\bar{\bar{\mathscr{T}}}_i + A\mathscr{R}_i + \mathscr{B} \\
&= (A + K_1 C)\bar{\bar{\mathscr{T}}}_i + (A + K_1 C)\mathscr{R}_i + \mathscr{B} \\
&= (A + K_1 C)(\mathscr{R}_i + \bar{\bar{\mathscr{T}}}_i) + \mathscr{B} \\
&= (A + K_1 C)\mathscr{T}_i + \mathscr{B}, \qquad i \in \bar{\mathbf{n}}.
\end{aligned}$$

Relation (4.4) follows at once.

Proof of Lemma 4.3. The dual of (4.1)–(4.5) states that there is a map $F' : \mathscr{U}' \to \mathscr{X}'$ and a subspace $\bar{\mathscr{S}}^\perp \subset \mathscr{X}'$ such that

$$\mathscr{S}^\perp = \bar{\mathscr{S}}^\perp \oplus \mathscr{S}^\perp \cap \mathscr{T}^\perp, \tag{4.28}$$

$$\begin{aligned}
\mathscr{C}' &= \mathscr{C}' \cap \bar{\mathscr{S}}^\perp \oplus \mathscr{C}' \cap \mathscr{S}^\perp \cap \mathscr{T}^\perp, \\
\bar{\mathscr{S}}^\perp &= \langle A' + F'B' | \mathscr{C}' \cap \bar{\mathscr{S}}^\perp \rangle, \\
\mathscr{S}^\perp &= \langle A' + F'B' | \mathscr{C}' \rangle.
\end{aligned} \tag{4.29}$$

Recall that $\mathscr{S}^\perp \cap \mathscr{T}^\perp$ is the largest $(A' + F'B', C')$-controllability subspace in $\ker B'$. From (4.28) it is clear that $\mathbf{K}' \equiv \mathbf{K}'(A' + F'B', C', \bar{\mathscr{S}}^\perp) \cap \mathbf{K}'(A' + F'B',$

$C', \mathscr{S}^\perp \cap \mathscr{T}^\perp)$ is nonempty. Let $K'_0 \in \mathbf{K}'$ be arbitrary and write

$$\alpha(\lambda)_{K'_0} = \text{char. p.}\,(A' + F'B' + C'K'_0), \tag{4.30}$$

$$\beta(\lambda)_{K'_0} = \text{char. p.}\,(A' + F'B' + C'K'_0)|\bar{\mathscr{S}}^\perp, \tag{4.31}$$

$$\gamma(\lambda)_{K'_0} = \text{char. p.}\,(A' + F'B' + C'K'_0)|(\mathscr{S}^\perp \cap \mathscr{T}^\perp). \tag{4.32}$$

Since $\bar{\mathscr{S}}^\perp$ and $\mathscr{S}^\perp \cap \mathscr{T}^\perp$ are independent $(A' + F'B' + C'K'_0)$-invariant subspaces, it follows that $\beta_{K'_0}\gamma_{K'_0}$ divides $\alpha_{K'_0}$. Thus there exists a polynomial $\mu(\lambda)$ such that

$$\alpha_{K'_0} = \mu\beta_{K'_0}\gamma_{K'_0}. \tag{4.33}$$

Since $\mathscr{S}^\perp$ is the controllable space of $(A' + F'B', C')$, a simple computation shows that μ is independent of $K'_0 \in \mathbf{K}'$. Thus $K'_0 \in \mathbf{K}$ can be chosen so that (4.30)–(4.33) hold and, in addition, so that μ and $\beta_{K'_0}\gamma_{K'_0}$ are coprime. Having done this, define

$$\mathscr{Z}^\perp = \bar{\mathscr{S}}^\perp \oplus \ker \mu(A' + F'B' + C'K'_0).$$

It follows (cf. [10, Chap. 7, Thm. 1]) that

$$(A' + F'B' + C'K'_0)\mathscr{Z}^\perp \subset \mathscr{Z}^\perp \tag{4.34}$$

and that

$$\mathscr{Z}^\perp \oplus \mathscr{S}^\perp \cap \mathscr{T}^\perp = \mathscr{X}'. \tag{4.35}$$

In addition, from (4.29),

$$\mathscr{C}' \subset \mathscr{Z}^\perp + \mathscr{C}' \cap \mathscr{T}^\perp. \tag{4.36}$$

Since $K'_0 \in \mathbf{K}'$ and $\mathscr{S}^\perp \cap \mathscr{T}^\perp \subset \ker B'$,

$$(A' + K'_0C')(\mathscr{S}^\perp \cap \mathscr{T}^\perp) \subset \mathscr{S}^\perp \cap \mathscr{T}^\perp. \tag{4.37}$$

Note that (4.34) is equivalent to

$$(A + BF + K_0C)\mathscr{Z} \subset \mathscr{Z}$$

from which (4.8) follows. In addition, (4.35) is equivalent to (4.6) and (4.37) is equivalent to (4.9).

In general, $C\mathscr{Z} \cap C\mathscr{T} = C(\mathscr{Z} \cap (\mathscr{T} + \mathscr{N}))$; but (4.36) is equivalent to $\mathscr{Z} \cap (\mathscr{T} + \mathscr{N}) \subset \mathscr{N}$; thus $C\mathscr{Z} \cap C\mathscr{T} = 0$. Since C is epic, $\mathscr{Y} = C\mathscr{X} = C\mathscr{Z} + C(\mathscr{S} + \mathscr{T}) = C\mathscr{Z} \oplus C\mathscr{T}$.

Concluding remarks. The main result of this paper, Theorem 4.1, provides a canonical representation for (C, A, B) relative to the group $\mathfrak{C}^*$. Because of the nature of $\mathfrak{C}^*$, the result is, at present, mainly of theoretical interest. Of the $\mathfrak{C}^*$-invariants identified, it is the transmission polynomials of (C, A, B) which probably have the most significance. For example, it is plausible that the roots of these polynomials (i.e., the *transmission zeros* of (C, A, B)) coincide with the fixed eigenvalues of the inverse system for (C, A, B).

As a final point, we call attention to a recent article by Heymann [14] in which an alternative definition of a prime system can be found. Although Heymann's definition is different than the one used here, there appears to be sufficient similarity between the two concepts (and corresponding results) to justify further study.

REFERENCES

[1] V. M. Popov, *Hyperstability and optimality of automatic systems with several control functions*, Rev. Roumaine Sci. Tech. Electrotech. et Energ., 9 (1964), pp. 629–690.

[2] E. G. Gilbert, *The decoupling of multivariable systems by state feedback*, this Journal, 7 (1969), pp. 50–63.

[3] H. Kwakernaak and R. Sivan, *The maximally achievable accuracy of linear optimal regulators and linear optimal filters*, IEEE Trans. Automatic Control, AC-17 (1972), pp. 79–85.

[4] P. Brunovský, *A classification of linear controllable systems*, Kybernetika, 3 (1970), pp. 173–187.

[5] W. M. Wonham and A. S. Morse, *Feedback invariants of linear multivariable systems*, Automatica, 8 (1972), pp. 93–100.

[6] ———, *Decoupling and pole assignment in linear multivariable systems: A geometric approach*, this Journal, 8 (1970), pp. 1–18.

[7] A. S. Morse and W. M. Wonham, *Status of noninteracting control*, IEEE Trans. Automatic Control, AC-16 (1971), pp. 568–581.

[8] ———, *Decoupling and pole assignment by dynamic compensation*, this Journal, 8 (1970), pp. 317–337.

[9] R. E. Kalman, *Kronecker invariants and feedback*, Proc. Conference on Ordinary Differential Equations, Math. Research Center, Naval Research Laboratory, Washington, D.C., 1971.

[10] F. R. Gantmacher, *The Theory of Matrices, I, II*, Chelsea, New York, 1959.

[11] H. H. Rosenbrock, *State-Space and Multivariable Theory*, John Wiley, New York, 1970.

[12] D. G. Luenberger, *Canonical forms for linear multivariable systems*, IEEE Trans. Automatic Control, AC-12 (1967), pp. 290–293.

[13] A. S. Morse, *Output controllability and system synthesis*, this Journal, 9 (1971), pp. 143–148.

[14] M. Heymann, *The prime structure of linear dynamical systems*, this Journal, 10 (1972), pp. 460–469.

SIAM J. Control
Vol. 11, No. 3, August 1973

A DIFFERENT APPROACH TO THE ANALYSIS OF TRACER DATA*

A. PAPOULIS†

Abstract. The response $h(t)$ of a biological system consisting of a finite number of compartments is a sum of exponentials and can be interpreted as the Laplace transform of an impulse train $g(x)$. A method of inversion is proposed yielding $g(x)$ as a sum involving the values of $h(t)$ at a sequence of equidistant points t_i. A simple numerical technique results for determining the exponential components of $h(t)$.

1. Introduction. In tracer kinetics, it is often assumed that the biological system under consideration consists of a finite number of compartments [1]–[3]. This assumption leads to responses of the form

$$(1)\qquad h(t) = \sum_{i=1}^{N} A_i e^{-\alpha_i t}, \qquad \alpha_i \geqq 0.$$

An important problem in the study of such systems and in other applications is the determination of the constants A_i and α_i in terms of the observed response $h(t)$. This problem has been extensively treated [4]–[6]. However, most of the known methods do not take full advantage of the fact that the exponentials α_i in (1) are real. This fact leads to a representation of $h(t)$ as a unilateral Laplace transform

$$(2)\qquad h(t) = \int_0^\infty e^{-tx} g(x)\,dx$$

of a train of impulses

$$(3)\qquad g(x) = \sum_{i=1}^{N} A_i \delta(x - \alpha_i)$$

as in Fig. 1. Hence, to determine the constants A_i and α_i it suffices to find the inverse Laplace transform of $h(t)$. The generally known methods of inversion involve partial fraction expansion or integration along some contour of the complex t-plane [7]. They are, therefore, not suitable for tracer data because they require knowledge of $h(t)$ on the complex plane. What is needed here is a method of inversion expressing $g(x)$ in terms of the values of $h(t)$ on the real t-axis only. In 1955, we developed such a method [8]–[10]. The purpose of this paper is to adopt the method to the tracer data problem. The proposed computational scheme will be based on a simplification of our method developed in 1970 by the Russian scientists D. M. Lerner and G. M. Lerner [11].

We should point out that the integral representation (2) of $h(t)$ has been recognized and forms the basis of the Fourier transform method [12]–[15] proposed by Brownell and Gardner in 1960. However, their approach does not involve inversion of a Laplace transform. With the change of variables

$$t = e^z, \quad x = e^{-w},$$

* Received by the editors March 6, 1972, and in revised form September 18, 1972.

† Department of Electrical Engineering and Electrophysics, Polytechnic Institute of Brooklyn, Farmingdale, New York 11735. This work was supported by the National Science Foundation under Grant GK-31033.

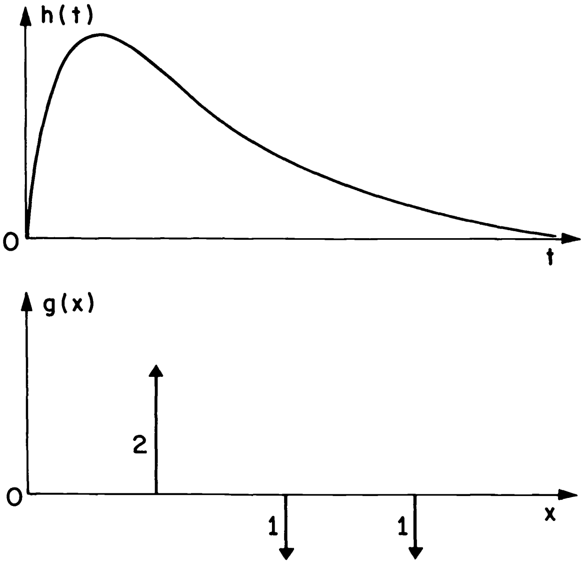


FIG. 1. *h(t): sum of exponentials; g(x) inverse Laplace transform of h(t)*

(2) is written as a convolution integral

$$e^z h(e^z) = e^{(z-e^z)} * g(e^z) \tag{4}$$

and $g(x)$ is obtained with Fourier transform techniques. As we hope to show, our approach is much simpler computationally. Furthermore, it is less sensitive to roundoff errors and errors due to noise in the data.

2. A real axis inversion of the Laplace transform. In this section, we present a method for expressing an arbitrary function $g(x)$ in terms of the values of its Laplace transform $h(t)$ at a sequence of equidistant points.

Original approach (Papoulis [8], 1955). With $\sigma > 0$ a real parameter, the transformation

$$e^{-\sigma x} = \cos\theta \tag{5}$$

maps the axis $t \geqq 0$ into the interval $0 \leqq \theta < \pi/2$ and the function $g(x)$ is transformed into

$$g\left(-\frac{1}{\sigma}\ln\cos\theta\right) = r(\theta) \tag{6}$$

as in Fig. 2.

Expanding $r(\theta)$ into an odd sine series, we obtain

$$r(\theta) = \sum_{k=0}^{\infty} C_k \sin(2k+1)\theta. \tag{7}$$

This can be done, of course, by extending the definition of $r(\theta)$ in the interval $(-\pi, \pi)$ as in Fig. 2. In the Appendix, we show that the coefficients C_k in (7) are

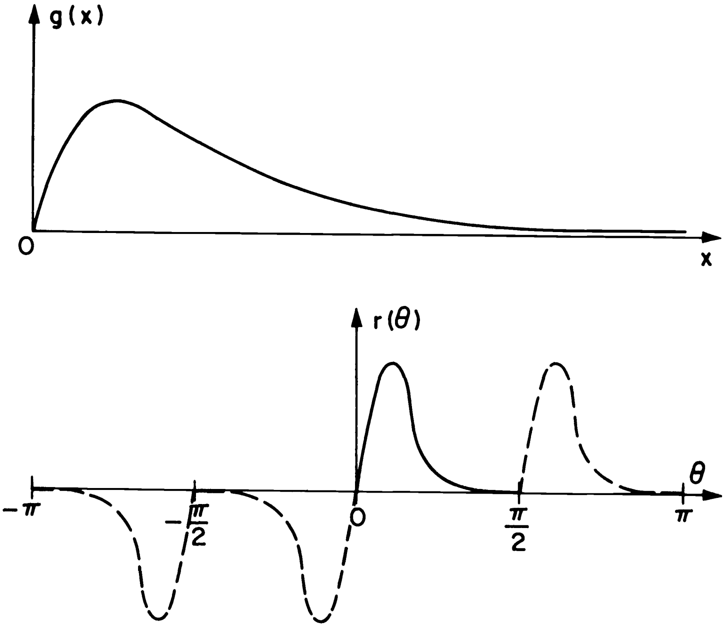


FIG. 2. $r(\theta) = g(x), e^{-\sigma x} = \cos\theta$

given by

$$C_k = \frac{4\sigma}{\pi}\sum_{m=0}^{k}(-1)^{k+m}\binom{k+m}{2m}h_m, \tag{8}$$

where

$$h_m = 4^m h(t_m), \qquad t_m = (2m+1)\sigma. \tag{9}$$

We have, thus, expressed $r(\theta)$, hence $g(x)$, in terms of the samples $h(t_m)$ of $h(t)$.

A simplified algorithm (Lerner–Lerner [11], 1970). In (7) the parameter σ is assumed constant and $r(\theta)$ is obtained by varying θ. Alternately, if θ is given a fixed value θ_0 and σ is determined as a function of x from (5), then $g(x)$ can be evaluated from $r(\theta_0)$. This approach increases the rate of convergence of (7) and it reduces the roundoff error.

Choosing for convenience $\theta_0 = \pi/4$, we obtain from (7):

$$r\left(\frac{\pi}{4}\right) = \sum_{k=0}^{\infty} C_n \sin(2k+1)\frac{\pi}{4} = \frac{1}{\sqrt{2}}\sum_{k=0}^{\infty}(-1)^{[k/2]}C_k, \tag{10}$$

where $[k/2]$ is the largest integer not exceeding $k/2$. With

$$\sigma = \ln\sqrt{2}/x \tag{11}$$

we conclude from (6) and (7) that

$$r\left(\frac{\pi}{4}\right) = g\left(-\frac{1}{\sigma}\ln\cos\frac{\pi}{4}\right) = g(x), \tag{12}$$

hence

$$g(x) = \frac{1}{\sqrt{2}} \sum_{k=0}^{\infty} (-1)^{[k/2]} C_k, \tag{13}$$

where [see (8)]

$$C_k = \frac{2 \ln 2}{\pi x} \sum_{m=0}^{k} (-1)^{k+m} \binom{k+m}{2m} h_m. \tag{14}$$

In a numerical evaluation of $g(x)$ the sum in (13) must be truncated. Retaining the first $n + 1$ terms, we thus obtain the approximation

$$g(x) \simeq g_n(x) = \frac{\sqrt{2} \ln 2}{\pi x} \sum_{m=0}^{n} (-1)^m h_m \sum_{k=m}^{n} (-1)^{[3k/2]} \binom{k+m}{2k}. \tag{15}$$

This follows by inserting (14) into (13) and changing the order of summation. We have, thus, approximated $g(x)$ by a simple sum involving the values of $h(t)$ at the points

$$t_m = \frac{(2m+1) \ln \sqrt{2}}{x}. \tag{16}$$

The preceding holds for a general $g(x)$. If $g(x)$ is an impulse train as in (3), then, as we show later, the truncated sum $g_n(x)$ exhibits principal maxima at $x = \alpha_i$ and secondary maxima (side lobes) near α_i. To distinguish between the two, it is desirable to plot not only $g_n(x)$ but also lower order approximations. For this purpose, we introduce the partial sums

$$B_k = [C_k + C_{k+1} - C_{k+2} - C_{k+3}]/\sqrt{2}. \tag{17}$$

Choosing n such that $n = 4s + 3$ where s is an integer, we obtain from (13):

$$g_n(x) = \sum_{k=0}^{s} B_k, \qquad n = 4s + 3. \tag{18}$$

From (14) and (17), we conclude (omitting details) that

$$B_k = \frac{\sqrt{2} \ln 2}{\pi x} \sum_{m=0}^{4k+3} b_{km} h_m, \tag{19}$$

where

$$\begin{gathered} b_{k0} = 0, \qquad b_{k1} = -2, \\ b_{km} = (-1)^m 2 \binom{4k+m}{2m-3} \frac{2k+1}{m-1}, \qquad m \geqq 2. \end{gathered} \tag{20}$$

The constants b_{km} are shown in Table 1.

For small x the first term

$$g_3(x) = \frac{20}{x}\left[-\frac{1}{8} h\left(\frac{3 \ln 2}{2x}\right) + h\left(\frac{5 \ln 2}{2x}\right) - h\left(\frac{7 \ln 2}{2x}\right)\right] \tag{21}$$

is often sufficient.

Truncation and roundoff errors. As is well known from the theory of Fourier series [16, p. 46], the truncated sum

$$r_n(\theta) = \sum_{k=0}^{n} C_k \sin(2k + 1)\theta^{-} \tag{22}$$

TABLE 1

b_{km}

k = m =	0	1	2	3
1	−2	−2	−2	−2
2	4	36	100	196
3	−1	−105	−825	−3185
4		112	2640	20384
5		−54	−4290	−68068
6		12	4004	136136
7		−1	−2275	−176358
8			800	155040
9			−170	−94962
10			20	40964
11			−1	−12397
12				2576
13				−350
14				28
15				−1

is the weighted average

$$r_n(\theta) = \int_0^{\pi/2} r(y)k_n(\theta - y)\,dy \tag{23}$$

of $r(\theta)$ (see (7)) with the kernel

$$k_n(\theta) = \frac{2\sin(4n + 3)\theta/2}{\pi \sin \theta/2}. \tag{24}$$

From the above, it follows that the approximation $g_n(x)$ in (15) is given by (see (6) and (11))

$$g_n(x) = r_n\left(\frac{\pi}{4}\right) = \int_0^{\pi/2} g\left(-x\frac{\ln\cos y}{\ln\sqrt{2}}\right)k_n\left(\frac{\pi}{4} - y\right)dy. \tag{25}$$

The function $k_n(\pi/4 - y)$ is shown in Fig. 3 for $n = 11$ and $n = 7$.

In a numerical evaluation of $g_n(x)$ from (15) the samples h_m must be replaced by $h_m + \varepsilon_m$, where ε_m is an error term due to the inaccuracy of measurement and to roundoff. As a result, the right side of (15) yields $g_n(x) + \varepsilon$. If we assume that the errors ε_m are uncorrelated with zero mean and variance μ, then the accumulated error ε has zero mean and variance

$$2\mu\left(\frac{\ln 2}{\pi x}\right)^2 \sum_{m=0}^{n} a_{nm}^2, \quad \text{where} \quad a_{nm} = \sum_{k=m}^{n} (-1)^{[3k/2]}\binom{k+m}{2k}. \tag{26}$$

From (25) we conclude that as n increases, the weighted average $g_n(x)$ approaches $g(x)$. However, the variance (26) of the error increases. The optimum choice of n depends on accuracy of computation and on the nature of $g(x)$.

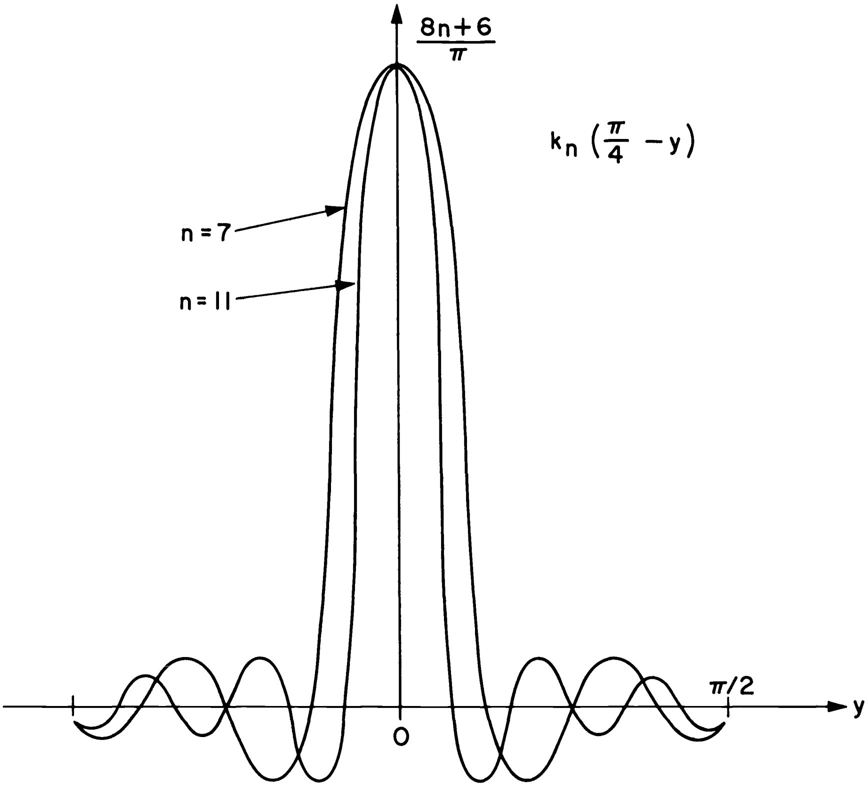


FIG. 3. *Fourier kernel $k_n(\pi/4 - y)$ for $n = 7$ and $n = 11$; linear y-scale*

3. Inversion of exponential sums. The preceding analysis permits us to recover an arbitrary signal $g(x)$ at every point x in which it is continuous. If $g(x)$ is an impulse train as in (3), then our objective is not the recovery of $g(x)$ but the determination of the constants A_i and α_i. With this in mind, we assume first that $g(x)$ consists of a single term

$$g(x) = A_i\delta(x - \alpha_i) \tag{27}$$

and we examine the corresponding $g_n(x)$ as given by (25). As it is known [17, p. 38],

$$\delta[\varphi(y)] = \sum_k \frac{\delta(y - y_k)}{|\varphi'(y_k)|}, \qquad \varphi(y_k) = 0, \tag{28}$$

where the y_k are all the real roots of the equation $\varphi(y) = 0$. Hence,

$$\begin{gathered}\delta\left(-\frac{x \ln \cos y}{\ln\sqrt{2}} - \alpha_i\right) = \frac{\ln\sqrt{2}}{x}\cot y_i\delta(y - y_i),\\ \cos y_i = \exp\left\{-\frac{\alpha_i \ln\sqrt{2}}{x}\right\}.\end{gathered} \tag{29}$$

Inserting into (25), we obtain

$$g_n(x) = \frac{A_i \ln\sqrt{2}}{x}\cot y_i k_n\left(\frac{\pi}{4} - y_i\right). \tag{30}$$

We note that if $x = \alpha_i$, then $y_i = \pi/4$, hence

$$g_n(\alpha_i) = \frac{A_i \ln\sqrt{2}}{\alpha_i}k_n(0) = \frac{A_i \ln 2}{\pi\alpha_i}(4n + 3).$$

Thus, $g_n(x)$ has a principal maximum at $x = \alpha_i$ and secondary maxima due to the side lobes of the kernel $k_n(\theta)$. As n increases, the position of the principal maximum remains fixed, whereas the secondary maxima move closer to $x = \alpha_i$. Thus, if $g(x)$ is of the form (3), then, plotting

$$g_n(x)/(4n + 3)$$

for various values of n, we can determine α_i and A_i from the position and amplitude of the stationary maxima. We note that the height of the maximum is proportional to $1/\alpha_i$.

To illustrate the method, we have carried out the computation of $g(x)$ for $h(t) = e^{-2t}$. The numerical results are plotted in Fig. 4 for $s = 0, 1, 2, 3$. Comparing the four curves, we readily recognize a single stationary maximum at $x = 2$. Sums of exponentials as in (1) lead to figures that are sums of such curves centered at α_i. If the exponents of $h(t)$ are close, then the corresponding main lobes overlap. To separate them we must increase n. An alternate approach is the following.

We first determine approximately the smallest exponent α_1. This can be done either by observing on a logarithmic scale the asymptotic slope of $h(t)$ or from a preliminary application of the inversion method. We next select a number $\alpha < \alpha_1$

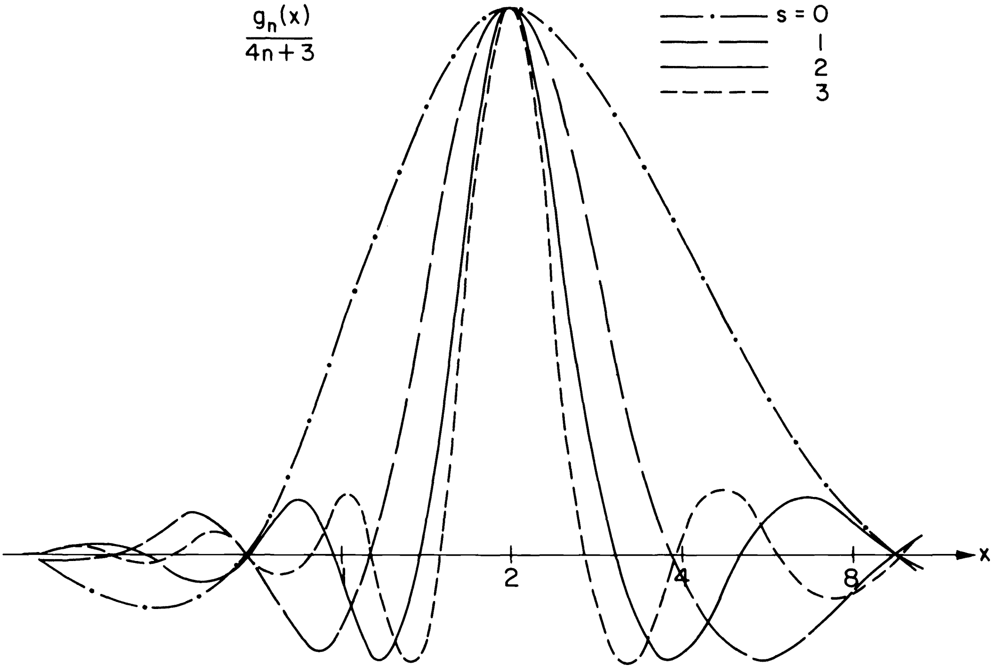


FIG. 4. *Computer output $g_n(x)/4n + 3$ of inverse Laplace transform of an exponential for $n = 3, 7, 11, 15$; logarithmic x-scale*

and form the signal

$$h_1(t) = e^{\alpha t}h(t)$$

whose exponents are $\beta_i = \alpha_i - \alpha$. If α is close to α_1, then the lowest exponent β_1 of $h_1(t)$ is small compared to the others and can be readily determined. This process can be repeated for the signal

$$h_2(t) = h_1(t) - A_1 e^{-\beta_1 t}.$$

We finally remark without elaboration that the side lobes of the kernel $k_n(\theta)$ can be reduced if the coefficients C_k in (7) are multiplied with suitable weights [18], [19].

Appendix. Introducing the transformation (5) into (1), we obtain

$$\text{(A.1)} \qquad \sigma h(t) = \int_0^{\pi/2} (\cos\theta)^{(t/\sigma - 1)} \sin\theta r(\theta)\, d\theta,$$

and with $t = (2k + 1)\sigma$, we have

$$\text{(A.2)} \qquad \sigma h[(2k + 1)\sigma] = \int_0^{\pi/2} (\cos\theta)^{2k} \sin\theta r(\theta)\, d\theta$$

for every $k \geqq 0$. But

$$(\cos\theta)^{2n} \sin\theta = \left(\frac{e^{j\theta} + e^{-j\theta}}{2}\right)^{2n} \frac{e^{j\theta} - e^{-j\theta}}{2j}.$$

Expanding the right side and collecting terms, we find that

$$\text{(A.3)}\qquad 2^{2n}(\cos\theta)^{2n}\sin\theta = \sum_{k=0}^{n}\left[\binom{2n}{k} - \binom{2n}{k-1}\right]\sin[2(n-k)+1]\theta,$$

where the bracket equals 1 for $k = 0$. We now insert the series expansion

$$r(\theta) = \sum_{k=0}^{\infty} C_k \sin(2k+1)\theta$$

into (A.2) and use (A.3). Since

$$\int_0^{\pi/2} \sin^2(2k+1)\theta\, d\theta = \frac{\pi}{4}$$

and the odd sines are orthogonal in the $(0, \pi/4)$, we conclude that

$$\text{(A.4)}\qquad \frac{4\sigma}{\pi}4^n h[(2n+1)] = \sum_{k=0}^{n}\left[\binom{2n}{k} - \binom{2n}{k-1}\right]C_{n-k}, \qquad n = 0, 1, \cdots.$$

Solving for the coefficients C_k, we obtain (8).

REFERENCES

[1] C. W. Sheppard and A. S. Householder, *Mathematical basis of the interpretation of tracer experiments in closed steady-state systems*, J. Appl. Phys., 22 (1951), pp. 510–520.

[2] M. Berman and R. Schoenfeld, *Invariants in experimental data on linear kinetics and the formulation of models*, Ibid., 27 (1956), pp. 1361–1370.

[3] J. S. Robertson, *Handbook of Physiology*, American Physical Society, Washington, D.C., 1962, pp. 617–644.

[4] M. Berman, M. F. Weiss and E. Shahn, *Formal approaches to the analysis of kinetic data in terms of linear compartmental systems*, Biophys. J., 2 (1962), pp. 289–316.

[5] J. Myhill, G. P. Wadsworth and G. L. Brownell, *Investigation of an operator method in the analysis of biological tracer data*, Ibid., 5 (1965), pp. 89–107.

[6] C. Lanczos, *Applied Analysis*, Prentice-Hall, Englewood Cliffs, N.J., 1956.

[7] D. V. Widder, *The Laplace Transform*, Princeton University Press, Princeton, N.J., 1946.

[8] A. Papoulis, *Network response in terms of behavior at imaginary frequencies*, Proc. Symposium on Modern Network Synthesis, Polytechnic Institute of Brooklyn, New York, 1955.

[9] ———, *A new method of inversion of the Laplace transform*, Quart. Appl. Math., 14 (1957), pp. 405–414.

[10] G. Doetsch, *Guide to the Applications of the Laplace Transform*, Van Nostrand, New York, 1963.

[11] D. M. Lerner and G. M. Lerner, *A simplified algorithm for the inversion of the Laplace transform*, Radiofizika, 13 (4) (1970), pp. 618–621. (In Russian.)

[12] S. M. Pizer, A. B. Ashane, A. B. Callahan and G. L. Brownell, *Concepts and Models oj Biomathematics*, F. Heinmetz, ed., Marcel Dekker, New York, 1969.

[13] D. G. Gardner, *Resolution of multi-component exponential decay curves using Fourier transforms*, Ann. N.Y. Acad. Sci., 108 (1963), pp. 195–203.

[14] G. L. Brownell and A. B. Callahan, *Transform methods for tracer data analysis*, Ibid., 108 (1963), pp. 172–194.

[15] A. B. Callahan and S. M. Pizer, *Natural Automata and Useful Simulations*, H. H. Pattee, E. A. Edelsack, L. Fein and A. B. Callahan, eds., Spartan, Washington, D.C., 1966.

[16] A. Papoulis, *The Fourier Integral and Its Applications*, McGraw-Hill, New York, 1962.

[17] ———, *Systems and Transforms with Applications in Optics*, McGraw-Hill, New York, 1968.

[18] ———, *Minimum bias windows for high resolution spectral estimates*, IEEE Trans. Information Theory, IT–19 (1973), pp. 9–12.

[19] ———, *A new class of Fourier series kernels*, IEEE Trans. Circuit Theory, CT–20 (1973).

SIAM J. CONTROL
Vol. 11, No. 3, August 1973

QUADRATIC PERFORMANCE CRITERIA IN BOUNDARY CONTROL OF LINEAR SYMMETRIC HYPERBOLIC SYSTEMS*

DAVID L. RUSSELL†

Abstract. The present article develops necessary and sufficient conditions for the solution of the following optimal control problem. We let $w(x, t), u(t)$ solve the hyperbolic mixed initial-boundary value problem

$$w(x, 0) = w_0(x),$$

$$\frac{\partial w}{\partial t} = A(x)\frac{\partial w}{\partial x} + B(x)w,$$

$$C_0 w(0, t) \equiv 0, \qquad C_1 w(1, t) \equiv Cu(t)$$

in the region $0 \leqq x \leqq 1, t \geqq 0$. For $t_1 > 0$ we seek to minimize a quadratic cost

$$J(w_0, u, t_1) = \int_0^{t_1}\left[\int_0^1 \int_0^1 w(x, t)^T W(x, \xi) w(\xi, t)\, dx\, d\xi + u(t)^T U u(t)\right] dt$$

$$+ \int_0^1 \int_0^1 w(x, t_1)^T G(x, \xi) w(\xi, t_1)\, dx\, d\xi.$$

In addition to the development of optimality conditions, the paper also presents a synthesis of the optimal solution in terms of a matrix Riccati partial differential equation. The case $t_1 = +\infty$ is also discussed.

1. Introduction. Since its introduction by Kalman and Bucy in the early 1960's [7], [8], the quadratic optimization criterion has developed to the point where it is now a standard design technique in stabilization of finite-dimensional linear systems. In the present article we shall be concerned with an extension of this theory to a certain class of infinite-dimensional systems modeled by partial differential equations.

Much work has already been done along similar lines. We cite the papers of Kim and Erzberger [9], Kim and Gajwani [10], Alvarado and Mukundan [1], Lukes and Russell [12], Datko [4], and, in particular, the book of Lions [11]. There are many other contributions—too many to list all of them. We must, therefore, be prepared to justify our addition to an already extensive literature.

In this paper we restrict our attention to systems modeled by linear symmetric hyperbolic partial differential equations in two independent variables (time, and one space dimension) with control variables appearing in the spatial boundary conditions. Such systems have not been widely treated in the above references—which are primarily concerned with parabolic equations. The form of the quadratic cost control law for hyperbolic systems has been outlined by Lions [11], but he gives no development of the theoretical foundations.

* Received by the editors February 15, 1972, and in revised form September 11, 1972.

† Departments of Mathematics and Computer Sciences, University of Wisconsin, Madison, Wisconsin 53706. This work was supported in part by the Control Theory Centre, University of Warwick, England, the Office of Naval Research under Contract NR-041-404, and the United States Army under Contract DA-31-124-ARO-D-462.

A further reason for the present paper is the fact that the very important case where the quadratic cost is defined over an infinite time interval $0 \leqq t < \infty$, which is given extensive treatment in [11] under the assumption that the system equations are parabolic, has never been given adequate mathematical treatment when those equations are hyperbolic. This case is one of the primary concerns of the present work.

2. Background. Linear symmetric hyperbolic systems in two independent variables have very special properties which have been examined in detail by a number of authors. Expository treatments are given by Courant and Hilbert [3] and Garabedian [5]. Phillips [14] and Olubummo and Phillips [13] have studied such systems in the framework of semigroups in Hilbert space. Control problems for such systems have been studied by the author in [15], [16], [17] and [18] and by Grainger [6].

We consider a system

$$\frac{\partial \tilde{w}}{\partial t} = \tilde{A}(x)\frac{\partial \tilde{w}}{\partial x} + \tilde{B}(x)\tilde{w}. \tag{2.1}$$

The $n \times n$ matrices $\tilde{A}(x)$ and $\tilde{B}(x)$ have the properties

(i) $\tilde{A}(x) \in C^1[0, 1]$;

(ii) $\tilde{B}(x) \in C[0, 1]$;

(iii) $\tilde{A}(x) = \tilde{A}^T(x)$ has eigenvalues $\lambda_1(x), \lambda_2(x), \cdots, \lambda_n(x)$ which satisfy $\lambda_1(x) \leqq \lambda_2(x) \leqq \cdots \leqq \lambda_p(x) < 0 < \lambda_{p+1}(x) \leqq \lambda_{p+2} \leqq \cdots \leqq \lambda_{p+q}(x)$, p and q being nonnegative integers with $p + q = n$, and if $\lambda_k(x) = \lambda_{k+1}(x)$ for any $x \in [0, 1]$, then $\lambda_k(x) \equiv \lambda_{k+1}(x)$, $x \in [0, 1]$.

We are concerned with n-dimensional vector solutions $\tilde{w}(x, t)$ in the region

$$R = \{(x, t)|0 \leqq x \leqq 1, 0 \leqq t < \infty\}$$

which satisfy initial conditions

$$\tilde{w}(x, 0) = \tilde{w}_0(x), \qquad x \in [0, 1], \tag{2.2}$$

and boundary conditions

$$\tilde{C}_0\tilde{w}(0, t) \equiv 0, \qquad \tilde{C}_1\tilde{w}(1, t) \equiv Cu(t), \qquad t \in [0, +\infty). \tag{2.3}$$

The meaning of equation (2.1) and the conditions (2.2), (2.3) is most easily understood if we assume $\tilde{w}_0(x)$ and $u(t)$ to be (at least piecewise) continuously differentiable, for then the solution $w(x, t)$ is easily shown [3] to have the same property. We shall see, however, that one may also discuss solutions $w(x, t)$ of (2.1), (2.2), (2.3) when $\tilde{w}_0(x)$ and $u(t)$ are only assumed square integrable.

Two examples of systems (2.1), motivated physically by the vibrating string and the vibrating beam, are cited in [17]. These are quite familiar and need not be repeated here. Another simple, and yet important, example is provided by counterflow processes in chemical engineering. Consider three pipes of sectorial cross section, as shown in Fig. 1. In pipe #1 fluid flows

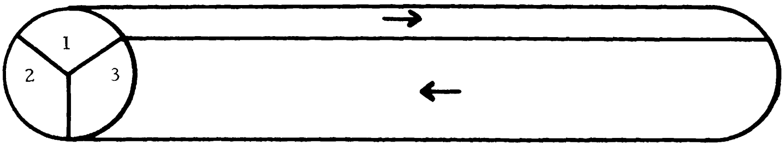


FIG. 1

from left to right with a velocity of a_1 meters per second. In pipes #2 and 3 fluid flows from right to left with velocities a_2 and a_3, respectively. The pipes are L meters long. The fluids in the pipes contain a certain dissolved substance whose concentrations are measured by functions $w_1(x, t)$, $w_2(x, t)$ and $w_3(x, t)$, $0 \leqq x \leqq L$, at a given instant t. This substance can diffuse from one pipe to another via semipermeable membranes in the walls of the pipes at rates which are proportional to the difference in concentration in the two pipes in question, with a constant depending on various physical characteristics of the fluids and membranes. If we assume that the pipes are rather long and thin and the fluid velocity is fairly high, we may neglect diffusion in the x-direction for a first approximation. Taking into account the movements of the fluids, the concentrations w_1, w_2 and w_3 obey a system of partial differential equations

$$\frac{\partial w_1}{\partial t} = -a_1\frac{\partial w_1}{\partial x} + b_{12}(w_2 - w_1) + b_{13}(w_3 - w_1),$$

$$\frac{\partial w_2}{\partial t} = a_2\frac{\partial w_2}{\partial x} + b_{21}(w_1 - w_2) + b_{23}(w_3 - w_2),$$

$$\frac{\partial w_3}{\partial t} = a_3\frac{\partial w_3}{\partial x} + b_{31}(w_1 - w_3) + b_{32}(w_2 - w_3),$$

which has the form (2.1). If we assume that the concentration w_1 can be specified at $x = 0$ while the concentrations w_2 and w_3 can be specified at $x = 1$, the boundary conditions might be given in the form

$$w_1(0, t) \equiv 0, \qquad t \in [0, \infty),$$

$$w_2(1, t) + \alpha w_1(1, t) = \beta u_1(t), \qquad t \in [0, \infty),$$

$$w_3(1, t) + \gamma w_1(1, t) = \delta u_2(t), \qquad t \in [0, \infty),$$

where u_1 and u_2 are scalar control functions. These conditions agree in form with (2.3).

The treatment of systems of the form (2.1) is facilitated by reducing the system (2.1) to a convenient normal form. Because we have insisted that the multiplicity of any eigenvalue of the symmetric matrix $\tilde{A}(x)$ should remain constant on (0, 1], we may use repeatedly a trivial modification of a result proved by Phillips [14, p. 117] to show that there is an orthogonal matrix-valued function $O(x)$ with $O \in C^1[0, 1]$ and

$$O(x)^T\tilde{A}(x)O(x) = A(x), \tag{2.4}$$

where

$$A(x) = \operatorname{diag}(\lambda_1(x), \cdots, \lambda_n(x)), \qquad x \in [0, 1]. \tag{2.5}$$

(We remark here that even if $\tilde{A}(x)$ is not symmetric, it may be possible to reduce it to the form (2.5) via a transformation $P(x)^{-1}\tilde{A}(x)P(x) = A(x)$, with P and P^{-1} in $C^1[0, 1]$. Our theory then applies with no significant change. Obviously this is possible if and only if $\tilde{A}(x)$ is symmetrizable.)

It will frequently be useful to write

$$\begin{gathered} A(x) = \begin{pmatrix} A^-(x) & 0 \\ 0 & A^+(x) \end{pmatrix}, \\ A^-(x) = \operatorname{diag}(\lambda_1(x), \cdots, \lambda_p(x)), \qquad A^+(x) = \operatorname{diag}(\lambda_{p+1}(x), \cdots, \lambda_{p+q}(x)). \end{gathered} \tag{2.6}$$

When we do so, we shall also represent vectors $w \in R^n$ in the form

$$w = \begin{pmatrix} w^- \\ w^+ \end{pmatrix}, \qquad w^- \in R^p, \qquad w^+ \in R^q. \tag{2.7}$$

We see now that if we put

$$\begin{gathered} \tilde{w}(x, t) = O(x)w(x, t), \\ B(x) = O(x)^T\tilde{B}(x) + O(x)^T\tilde{A}(x)O'(x), \end{gathered} \tag{2.8}$$

then the system (2.1) becomes

$$\frac{\partial w}{\partial t} = A(x)\frac{\partial w}{\partial x} + B(x)w. \tag{2.9}$$

The initial state is now $w_0(x) = O(x)^T\tilde{w}_0(x)$. It is clear that we still have $A \in C^1[0, 1]$, $B \in C(0, 1]$. The boundary conditions (2.3) become

$$C_0w(0, t) \equiv 0, \qquad C_1w(1, t) \equiv Cu(t) \tag{2.10}$$

with

$$C_0 = \tilde{C}_0O(0), \qquad C_1 = \tilde{C}_1O(1). \tag{2.11}$$

Using the decomposition (cf. (2.6), (2.7))

$$C_0 = (C_{0-} \quad C_{0+}), \qquad C_1 = (C_{1-} \quad C_{1+}),$$

we replace (2.10) by

$$(C_{0-} \quad C_{0+})\begin{pmatrix} w^-(0, t) \\ w^+(0, t) \end{pmatrix} \equiv 0, \tag{2.12}$$

$$(C_{1-} \quad C_{1+})\begin{pmatrix} w^-(1, t) \\ w^+(1, t) \end{pmatrix} \equiv Cu(t). \tag{2.13}$$

We may then state the following conditions.

Assumptions on the boundary conditions. The matrices C_0, C_1 have dimensions $p \times n$, $q \times n$, respectively, and the $p \times p$ matrix C_{0-} and the $q \times q$ matrix C_{1+} are both nonsingular.

Then, with

$$D_0 = -(C_{0-})^{-1}C_{0+}, \qquad D_1 = -(C_{1+})^{-1}C_{1-}, \qquad D = (C_{1+})^{-1}C,$$

(2.12) and (2.13) become

$$w^-(0,t) \equiv D_0 w^+(0,t), \tag{2.14}$$

$$w^+(1,t) \equiv D_1 w^-(1,t) + Du(t). \tag{2.15}$$

It is essential that the boundary conditions (2.10) be reducible to this form. Otherwise even uniqueness of solutions may fail. This is seen rather easily for the system

$$\frac{\partial}{\partial t}\begin{pmatrix} w^- \\ w^+ \end{pmatrix} = \begin{pmatrix} -1 & 0 \\ 0 & 1 \end{pmatrix}\frac{\partial}{\partial x}\begin{pmatrix} w^- \\ w^+ \end{pmatrix},$$

$$w^+(0,t) \equiv 0, \qquad w^-(1,t) \equiv 0.$$

For more on this, see [18].

Consider now the initial value problem

$$w(x,0) = w_0(x) \qquad (\equiv O^T(x)\tilde{w}_0(x)) \tag{2.16}$$

for the system (2.9), (2.14), (2.15). Existence and uniqueness theorems are given in [3], [5]. The proofs are based on the fact that the kth component w^k of w satisfies a first order linear ordinary differential equation along the *characteristic* curves $x = x_k(t)$ which satisfy the differential equation

$$dx_k/dt = -\lambda_k(x). \tag{2.17}$$

Indeed, we have, along such a curve,

$$\frac{d}{dt}w^k(x_k(t),t) = \sum_{j=1}^{n} b_j^k(x_k(t))w^j(x_k(t),t), \qquad k = 1,2,\cdots,n.$$

If $w_0(x)$ and $u(t)$ are continuously differentiable and satisfy the consistency conditions

$$w_0^-(0) = D_0 w_0^+(0),$$

$$w_0^+(1) = D_1 w_0^-(1) + Du(0),$$

$$A^-(0)\frac{dw_0^-}{dx}(0) + B_-^-(0)w_0^-(0) + B_+^-(1)w_0^+(0)$$

$$= D_0\left[A^+(0)\frac{dw_0^+}{dx}(0) + B_-^+(0)w_0^-(0) + B_+^+(0)w_0^+(0)\right], \tag{2.18}$$

$$A^+(1)\frac{dw_0^+}{dx}(1) + B_-^+(1)w_0^-(1) + B_+^+(1)w_0^+(1)$$

$$= D_1\left[A^-(1)\frac{dw_0^-}{dx}(1) + B_-^-(1)w_0^-(1) + B_+^-(1)w_0^+(1)\right] + D\frac{du}{dt}(0),$$

then one obtains a solution $w(x,t)$ in $C^1(R), R = \{(x,t)|0 \leqq x \leqq 1, 0 \leqq t < \infty\}$.

Less differentiability of $w_0(x)$ or $u(t)$ or failure to satisfy these consistency conditions leads to solutions with less differentiability. If we assume only that $w_0(x)$ and $u(t)$ are piecewise C^1 and do not require that any consistency conditions be satisfied, the solution $w(x, t)$ will likewise be piecewise C^1 in R. It can be shown (see [3]) that any jump discontinuity of $w^k(x, t)$ or its first order partial derivatives is propagated along a characteristic curve $x = x_k(t)$ obeying (2.17). Thus $w(x, t)$ will be of class C^1 in subregions of R bounded by such characteristic curves. A bounded subset of R will meet only finitely many such regions.

One can use the same methods to construct solutions for other types of systems. One which is of importance in this paper is obtained by replacing the boundary condition (2.15) by

$$w^+(1, t) \equiv D_1 w^-(1, t) + \int_0^1 H_1(x)w(x, t)\,dx + Du(t), \tag{2.19}$$

where H_1 is a piecewise continuous $q \times n$ matrix function on $0 \leqq x \leqq 1$. The existence, continuity and differentiability considerations are the same as before, with the consistency conditions (2.18) modified in the obvious way.

The basic estimate for solutions of (2.9), (2.14), (2.19) (includes (2.15) for $H_1(x) \equiv 0$) is provided by the following theorem.

THEOREM 2.1. *Let $w(x, t)$ be a C^1-solution of* (2.9), (2.14), (2.19) *in*

$$R = \{(x, t) | 0 \leqq x \leqq 1, 0 \leqq t < +\infty\}.$$

Then there are positive constants K_0 and K_1 such that

$$\int_0^1 \|w(x, t)\|^2\,dx \leqq K_0 e^{K_1 t}\left(\int_0^1 \|w(x, 0)\|^2\,dx + \int_0^t \|u(t)\|^2\,dt\right). \tag{2.20}$$

Proof. This requires only standard techniques after a simple preliminary transformation.

Since $-A^+(0)$ and $A^-(1)$ are both negative definite, we can find positive numbers ε_0 and ε_1 such that

$$-(A^+(0) + \varepsilon_0 D_0^T A^-(0) D_0) \leqq 0, \qquad A^-(1) + 3\varepsilon_1 D_1^T A^+(1) D_1 \leqq 0. \tag{2.21}$$

We let $e^-(x), e^+(x)$ be continuously differentiable functions with

$$\begin{aligned} \varepsilon_0 = e^-(0) \leqq e^-(x) \leqq e^-(1) = 1, &\qquad 0 \leqq x \leqq 1, \\ 1 = e^+(0) \geqq e^+(x) \geqq e^+(1) = \varepsilon_1, &\qquad 0 \leqq x \leqq 1. \end{aligned} \tag{2.22}$$

Then let

$$E(x) = \begin{pmatrix} e^-(x)I^- & 0 \\ 0 & e^+(x)I^+ \end{pmatrix}, \tag{2.23}$$

where I^- and I^+ are identity matrices of dimensions $p \times p$ and $q \times q$, respectively. Multiplying (2.8) on the left by $E(x)$ we obtain

$$E(x)\frac{\partial w}{\partial t} = E(x)A(x)\frac{\partial w}{\partial x} + E(x)B(x)w. \tag{2.24}$$

Let $R_\tau = \{(x, t)|0 \leqq x \leqq 1, 0 \leqq t \leqq \tau\}$. We compute, using the fact that w satisfies (2.24),

$$\begin{aligned}(2.25)\qquad &\int_{R_\tau}\int \left[\frac{\partial}{\partial t}(w(x,t)^T E(x)w(x,t)) - \frac{\partial}{\partial x}(w(x,t)^T E(x)A(x)w(x,t))\right] dx\,dt \\ &= \int_{R_\tau}\int w(x,t)^T(E(x)B(x) + B^T(x)E(x) - (E(x)A(x))')w(x,t)\,dx\,dt.\end{aligned}$$

Applying the divergence theorem, we have

$$\begin{aligned}&\int_0^1 w(x,\tau)^T E(x)w(x,\tau)\,dx - \int_0^1 w(x,0)^T E(x)w(x,0)\,dx \\ (2.26)\qquad &= \int_0^\tau w(1,t)^T E(1)A(1)w(1,t)\,dt - \int_0^\tau w(0,t)^T E(0)A(0)w(0,t)\,dt \\ &\quad + \int_{R_\tau}\int w(x,t)^T(E(x)B(x) + B^T(x)E(x) - (E(x)A(x))')w(x,t)\,dx\,dt.\end{aligned}$$

Using (2.14), (2.21), (2.22) and (2.23) we compute

$$\begin{aligned}-w(0,t)^T E(0)A(0)w(0,t) &= -e^-(0)w^-(0,t)^T A^-(0)w^-(0,t) \\ (2.27)\qquad &\quad -e^+(0)w^+(0,t)^T A^+(0)w^+(0,t) \\ &= -[w^+(0,t)^T(A^+(0) + \varepsilon_0 D_0^T A^-(0)D_0)w^+(0,t)] \leqq 0.\end{aligned}$$

Also, via the same reasoning but replacing (2.14) by (2.19),

$$\begin{aligned}&w(1,t)^T E(1)A(1)w(1,t) \\ &= e^-(1)w^-(1,t)^T A^-(1)w^-(1,t) \\ &\quad + e^+(1)\left[D_1 w^-(1,t) + \int_0^1 H_1(x)w(x,t)\,dx + du(t)\right]^T A^+(1) \\ &\quad \cdot\left[D_1 w^-(1,t) + \int_0^1 H_1(x)w(x,t)\,dx + Du(t)\right] \\ (2.28)\qquad &\leqq w^-(1,t)^T[A^-(1) + 3\varepsilon_1 D_1^T A^+(1)D_1]w(1,t) \\ &\quad + 3\varepsilon_1\left(\int_0^1 H_1(x)w(x,t)\,dx\right)^T A^+(1)\left(\int_0^1 H_1(x)w(x,t)\,dx\right) \\ &\quad + 3\varepsilon_1 u(t)^T D^T A^+(1)Du(t) \\ &\leqq 3\varepsilon_1\left(\lambda_n(1)\left\|\int_0^1 H_1(x)w(x,t)\,dx\right\|^2 + u(t)^T D^T A^+(1)Du(t)\right),\end{aligned}$$

the last inequality following from (2.21) and the fact that $\lambda_n(1) = \lambda_{p+q}(1)$ is the largest eigenvalue of $A^+(1)$. Noting that $H_1(x)$, $E(x)$, $B(x)$ and $(E(x)A(x))'$ are uniformly bounded on $0 \leqq x \leqq 1$, we may proceed from (2.26) and the inequalities (2.27), (2.28) to see that

$$\begin{aligned}&\int_0^1 w(x,\tau)^T E(x) w(x,\tau)\,dx - \int_0^1 w(x,0)^T E(x) w(x,0)\,dx \\ &\qquad \leqq K \int_0^\tau \int_0^1 \|w(x,t)\|^2\,dx\,dt + \hat{K}\int_0^\tau \|u(t)\|^2\,dt\end{aligned} \tag{2.29}$$

for appropriately large positive K, $\hat{K}$. From (2.22) we see that

$$w(x,t)^T E(x) w(x,t) \geqq \min(\varepsilon_0,\varepsilon_1)\|w(x,t)\|^2. \tag{2.30}$$

Combining (2.29) and (2.30) and using a familiar estimate [2, Chap. 1] we see that with $K_1 = K/\min(\varepsilon_0,\varepsilon_1)$ we have

$$\int_0^1 \|w(x,t)\|^2\,dx \leqq e^{K_1\tau}\int_0^1 \|w(x,0)\|^2\,dx + \hat{K}\int_0^\tau e^{K_1(\tau-s)}\|u(s)\|^2\,ds,$$

from which we obtain (2.20) immediately, letting $K_0 = \max(1,\hat{K})$ and then replacing τ by t. Thus Theorem 2.1 is proved.

In addition to being important in its own right, this estimate enables us to define solutions of (2.9), (2.14), (2.19) when $w_0(x)$ and $u(t)$ are assumed only square integrable. It is well known that under these circumstances one can find continuously differentiable functions $w_{0k}(x)$ and $u_k(t)$ vanishing outside of compact subintervals of $(0,1)$, $(0,\infty)$, respectively (and thus satisfying the consistency conditions (2.18)) such that

$$\lim_{k\to\infty}\|w_0 - w_{0k}\|_{L^2[0,1]} = 0,$$

$$\lim_{k\to\infty}\|u - u_k\|_{L^2[0,\infty)} = 0.$$

Now we let $w_k(x,t)$ be the continuously differentiable solution of (2.9), (2.14), (2.19), $k = 1,2,3,\cdots$. Theorem 2.1 guarantees the existence of a function $w(x,t)$ for which

$$\lim_{k\to\infty}\|w(\cdot,t) - w_k(\cdot,t)\|_{L^2[0,1]} = 0 \quad \text{for all } t \geqq 0.$$

This function $w(x,t)$ is the solution of (2.9), (2.14), (2.19) corresponding to the square integrable initial data $w_0(x)$ and the square integrable control $u(t)$. For more on such "generalized" solutions, see [3], [5] and other standard works on partial differential equations.

3. Optimality conditions. Working with our control system in the form (2.9), (2.14), (2.15), we define a quadratic cost functional

$$\begin{aligned}J(u,w_0,t_1) &= \int_0^{t_1}\left[\int_0^1\int_0^1 w(x,t)^T W(x,\xi) w(\xi,t)\,dx\,d\xi + u(t)^T U u(t)\right]dt \\ &\quad + \int_0^1\int_0^1 w(x,t_1)^T G(x,\xi) w(\xi,t_1)\,dx\,d\xi.\end{aligned} \tag{3.1}$$

Here W and G are $n \times n$ matrix functions of classes C and C^1, respectively, in the domain $0 \leqq x \leqq 1$, $0 \leqq \xi \leqq 1$, with the properties (stated for W only, but also assumed for G)

$$W(x,\xi) = W(\xi,x)^T, \tag{3.2}$$
$$\int_0^1\int_0^1 w(x)^T W(x,\xi)w(\xi)\,dx\,d\xi \geqq 0, \qquad w\in L^2[0,1].$$

Given an initial state $w_0 \in L^2[0,1]$ and a control $u \in L^2[0,t_1]$, the resulting generalized solution w of (2.9), (2.14), (2.15), (2.16) can be determined and the cost $J(u, w_0, t_1)$ computed. Our goal is to characterize the control $u_* \in L^2[0,t_1]$ which, together with the resulting response w_*, yields a minimal value for this cost. The existence of such a control u_* may be proved by methods developed in [11] and is assumed in the rest of this article.

THEOREM 3.1. *Let $u_* \in L^2[0,t_1]$ and let w_* be the resulting solution of* (2.9), (2.14), (2.15), (2.16) *in the domain $R_{t_1} = \{(x,t)|0 \leqq x \leqq 1, 0 \leqq t \leqq t_1\}$. Let v be the solution of the "adjoint" system*

$$\frac{\partial v}{\partial t} = A(x)\frac{\partial v}{\partial x} - (B^T(x) - A'(x))v - \int_0^1 W(x,\xi)w_*(\xi,t)\,d\xi \tag{3.3}$$

satisfying the terminal condition

$$v(x,t_1) = \int_0^1 G(x,\xi)w_*(\xi,t_1)\,d\xi \tag{3.4}$$

and the boundary conditions

$$v^+(0,t) = -(A^+(0))^{-1}D_0^T A^-(0)v^-(0,t), \tag{3.5}$$

$$v^-(1,t) = -(A^-(1))^{-1}D_1^T A^+(1)v^+(1,t). \tag{3.6}$$

Then u_* and w_* minimize the cost (3.1) *if and only if*

$$u_*(t) = -U^{-1}D^T A^+(1)v^+(1,t) \quad \textit{for almost all } t\in[0,t_1]. \tag{3.7}$$

Proof. Standard theorems on positive quadratic forms show that for an arbitrary control $u_1 \in L^2[0,t_1]$,

$$J(u_*, w_0, t_1) \leqq J(u_* + \varepsilon u_1, w_0, t_1)$$

for all real values of ε if and only if

$$\int_0^{t_1}\left[\int_0^1\int_0^1 w_*(x,t)^T W(x,\xi)w_1(\xi,t)\,dx\,d\xi + u_*(t)^T U u_1(t)\right]dt \tag{3.8}$$
$$+ \int_0^1\int_0^1 w_*(x,t_1)^T G(x,\xi)w_1(\xi,t_1)\,dx\,d\xi = 0,$$

where w_1 solves (2.9), (2.14), (2.15) (with $u = u_1$) and $w_1(x,0) \equiv 0$.

Consider first the case where $w(x,t)$ solves (2.9), (2.14), (2.15) with $w(x,0)$ and $u(t)$ both continuously differentiable and vanishing outside compact subintervals of (0, 1), (0, t_1], respectively, and $v(x,t)$ solves (3.3), (3.5), (3.6) with $v(x,t_1)$ continuously differentiable and vanishing outside a compact subinterval of (0, 1). Then all consistency requirements are met and both $w(x,t)$ and $v(x,t)$ are continuously differentiable in R_{t_1}. Then

$$
\begin{aligned}
\int_0^1 v(x,0)^T w(x,0)\,dx &= \int_0^1 v(x,t_1)^T w(x,t_1)\,dx \\
&\quad - \int_0^{t_1} \frac{d}{dt}\left[\int_0^1 v(x,t)^T w(x,t)\,dx\right] dt \\
&= \int_0^1 \int_0^1 v(x,t)^T w(x,t_1)\,dx \\
&\quad - \int_0^{t_1}\int_0^1 \left\{\left[A(x)\frac{\partial v(x,t)}{\partial x} - (B^T(x) - A'(x))v(x,t) \right.\right. \\
&\qquad\qquad \left. - \int_0^1 W(x,\xi)w_*(\xi,t)\,d\xi\right]^T w(x,t) \\
&\qquad\qquad \left. + v(x,t)^T\left[A(x)\frac{\partial w(x,t)}{\partial x} + B(x)w(x,t)\right]\right\} dx\,dt \qquad (3.9)\\
&= \int_0^1 v(x,t_1)^T w(x,t_1)\,dx \\
&\quad + \int_0^{t_1}\left[\int_0^1\int_0^1 w(x,t)^T W(x,\xi) w_*(\xi,t)\,dx\,d\xi\right] dt \\
&\quad + \int_0^{t_1}\int_0^1 \left\{\left[A(x)\frac{\partial v(x,t)}{\partial x} - (B^T(x) - A'(x))v(x,t)\right] w(x,t)\right. \\
&\qquad\qquad \left. + v(x,t)^T\left[A(x)\frac{\partial w(x,t)}{\partial x} + B(x)w(x,t)\right]\right\} dx\,dt.
\end{aligned}
$$

In the last integral above we employ integration by parts together with the boundary conditions (2.14), (2.15), (3.5) and (3.6) to see that it reduces to

$$
-\int_0^{t_1} v^+(1,t)^T A^+(1) Du(t)\,dt.
$$

If we take $w(x,0) = 0$ and let u converge in $L^2[0,t]$ to u_1, then $w(x,t)$ converges in $L^2[0,1]$, for each fixed value of t, to $w_1(x,t)$, as we see from Theorem 2.1. Indeed, this convergence is uniform for $t \in [0,t_1]$. We also let $v(x,t_1)$ converge in $L^2[0,1]$ to the value (3.4). From this and (3.9) we see that we can write (in some cases interchanging the roles of x and ξ)

$$
\begin{aligned}
&\int_0^{t_1}\left[\int_0^1\int_0^1 w_*(x,t)^T W(x,\xi) w_1(\xi,t)\,dx\,d\xi + u_*(t)Uu_1(t)\right] dt \\
&\quad + \int_0^1\int_0^1 w_*(x,t_1)^T G(x,\xi) w_1(\xi,t_1)\,dx\,d\xi \qquad (3.10)\\
&= \int_0^{t_1} [u_*(t)Uu_1(t) + v^+(1,t)^T A^+(1) Du_1(t)]\,dt.
\end{aligned}
$$

From (3.10) we see that (3.8) holds for all $u_1 \in L^2[0, t_1]$ if and only if (3.7) is true. The proof is complete.

We remark here that it is quite easy to obtain estimates of the type obtained in Theorem 2.1 for solutions of (3.3). We have tacitly assumed such results in the paragraph preceding (3.10), where we let $v(x, t_1)$ converge in $L^2[0, 1]$ to the value (3.4). It is not necessary to approximate $w_*(\xi, t)$ in the last term of (3.3) by a smooth function because that term is a continuous function of x and t even when $w_*(\xi, t)$ is only a generalized solution of (2.9) as explained in § 2.

Theorem 3.1 characterizes the optimal control u_* in terms of a solution w_*, v of the two-point boundary value problem consisting of the systems (2.9), (2.14), (2.15) and (3.3), (3.5), (3.6), coupled via (3.7), and the conditions (2.16) and (3.4) given at $t = 0$ and $t = t_1$, respectively. As usual in optimal control problems, such a characterization is of limited value because the solution of such a two-point boundary value problem is by no means easy. This difficulty is overcome here, as in the case of related problems for ordinary differential equations [7], [12], [4] by a decoupling process which involves the introduction of a nonlinear partial differential equation of "Riccati-type".

4. Formal derivation of the Riccati equation. The basic idea of the decoupling procedure is to represent $v(x, t)$ as a linear function of the optimal solution $w_*(x, t)$. This is plausible because the terminal state (3.4) for v and the nonhomogeneous terms in the linear equation (3.3) are both linear functions of $w_*(x, t)$. We shall seek a representation

$$v(x, t) = \int_0^1 Q(x, \xi, t) w_*(\xi, t)\, d\xi, \tag{4.1}$$

where the $n \times n$ matrix function Q has the property

$$Q(x, \xi, t) \equiv Q^T(\xi, x, t). \tag{4.2}$$

In formally deriving the partial differential equation which Q satisfies, we shall assume that Q and w_* are everywhere of class C^1. The smoothness assumption on w_* will be removed later. Under appropriate assumptions, to be discussed later, Q will be of class C^1 or at least piecewise of class C^1.

In agreement with (2.6) and (2.7) we write

$$Q = \begin{pmatrix} Q^- \\ Q^+ \end{pmatrix} = (Q_- \quad Q_+), \tag{4.3}$$

distinguishing certain rows and columns of Q, respectively.

If one replaces w in (3.9) by w_*, one finds, in much the same way as we did there, that

$$\int_0^1 v(x, 0)^T w_*(x, 0)\, dx$$

$$= J(u_*, w_0, t_1) - \int_0^{t_1} u_*(t)^T U u_*(t)\, dt$$

(*continued on page* 486)

$$\tag{4.4} -\int_0^{t_1}\int_0^1 \left\{\left[A(x)\frac{\partial v(x,t)}{\partial x} - (B^T(x) - A'(x))v(x,t)\right]^T w_*(x,t) + v(x,t)^T\left[A(x)\frac{\partial w_*(x,t)}{\partial x} + B(x)w_*(x,t)\right]\right\} dx\,dt.$$

Thus if the last two integrals in (4.4) add to zero, that formula becomes

$$\tag{4.5} J(u_*, w_0, t_1) = \int_0^1 v(x,0)^T w_*(x,0)\,dx,$$

an identity which will be very useful later. Taking the last two integrals in (4.4), adding and subtracting the term

$$\int_0^{t_1}\int_0^1\int_0^1 w_*(x,t)W(x,\xi)w_*(\xi,t)\,dx\,d\xi\,dt,$$

we require that

$$\begin{aligned} 0 = {} & \int_0^{t_1}\int_0^1 \left\{\left[A(x)\frac{\partial v(x,t)}{\partial x} - (B^T(x) - A'(x))v(x,t) - \int_0^1 W(x,\xi)w_*(\xi,t)\,d\xi\right]^T w_*(x,t) \right. \\ & \left. + v(x,t)^T\left[A(x)\frac{\partial w_*(x,t)}{\partial x} + B(x)w_*(x,t)\right]\right\} dx\,dt \\ & + \int_0^{t_1}\left\{\int_0^1\int_0^1 w_*(x,t)^T W(x,\xi)w_*(\xi,t)\,dx\,d\xi + u_*(t)^T U u_*(t)\right\} dt \\ = {} & \int_0^{t_1}\int_0^1 \left[\left(\frac{\partial v(x,t)}{\partial t}\right)^T w_*(x,t) + v(x,t)\left(A(x)\frac{\partial w_*(x,t)}{\partial x} + B(x)w_*(x,t)\right)\right] dx\,dt \\ & + \int_0^{t_1}\left\{\int_0^1\int_0^1 w_*(x,t)^T W(x,\xi)w_*(\xi,t)\,dx\,d\xi + u_*(t)^T U u_*(t)\right\} dt \\ = {} & \int_0^{t_1}\left\{\int_0^1\int_0^1 \left[w_*(x,t)^T\frac{\partial Q(x,\xi,t)}{\partial t}w_*(\xi,t) + w_*(x,t)^T Q(x,\xi,t)A(\xi)\frac{\partial w_*(\xi,t)}{\partial \xi} \right.\right. \\ & \left.\left. + w_*(x,t)^T Q(x,\xi,t)B(\xi)w_*(\xi,t) + \frac{\partial w_*(x,t)^T}{\partial x}A(x)Q(x,\xi,t)w_*(\xi,t)\right.\right. \end{aligned} \tag{4.6}$$

(*continued on page* 487)

$$+ w_*(x,t)^T B(x)^T Q(x,\xi,t) w_*(\xi,t) + w_*(x,t)^T W(x,\xi) w_*(\xi,t)\Big]\,dx\,d\xi$$

$$+ u_*(t)^T U u_*(t)\Big\}\,dt \qquad \text{(using (3.3) and then (4.1), (4.2))}.$$

We employ integration by parts to see that

$$\begin{aligned}\int_0^1 \frac{\partial w_*(x,t)^T}{\partial x} A(x) Q(x,\xi,t)\,dx &\\ = w_*(1,t)^T A(1) Q(1,\xi,t) - w_*(0,t)^T A(0) Q(0,\xi,t)&\\ - \int_0^1 w_*(x,t)^T \left(A(x)\frac{\partial Q(x,\xi,t)}{\partial x} + A'(x) Q(x,\xi,t)\right) dx,&\end{aligned}$$

and a similar result is obtained for

$$\int_0^1 Q(x,\xi,t) A(\xi) \frac{\partial w_*(\xi,t)}{\partial \xi}\,d\xi.$$

Now, using (2.14),

$$\begin{aligned} w_*(0,t)^T A(0) Q(0,\xi,t) &= w_*^-(0,t)^T A^-(0) Q^-(0,\xi,t) + w_*^+(0,t)^T A^+(0) Q^+(0,\xi,t)\\ &= w_*^+(0,t)^T (D_0^T A^-(0) Q^-(0,\xi,t) + A^+(0) Q^+(0,\xi,t)),\end{aligned}$$

and a similar calculation applies to $Q(x,0,t)A(0)w_*(0,t)$. Also, using (2.15),

$$\begin{aligned} &w_*(1,t)^T A(1) Q(1,\xi,t)\\ &= w_*^-(1,t)^T A^-(1) Q^-(1,\xi,t) + w_*^+(1,t)^T A^+(1) Q^+(1,\xi,t)\\ &= w_*^-(1,t)^T (A^-(1) Q^-(1,\xi,t) + D_1^T A^+(1) Q^+(1,\xi,t))\\ &\quad + u_*(t)^T D^T A^+(1) Q^+(1,\xi,t)\\ &= w_*^-(1,t)^T (A^-(1) Q^-(1,\xi,t) + D_1^T A^+(1) Q^+(1,\xi,t))\\ &\quad - \left(\int_0^1 w_*(x,t)^T Q_+(x,1,t)\,dx\right) A^+(1) D U^{-1} D^T A^+(1) Q^+(1,\xi,t)\end{aligned} \tag{4.7}$$

(using (3.7) and (4.1)).

A similar calculation applies to $Q(x,1,t)A(1)w_*(1,t)$.

We see then that if we impose upon Q the boundary conditions

$$Q^+(0,\xi,t) = -(A^+(0))^{-1} D_0^T A^-(0) Q^-(0,\xi,t), \tag{4.8}$$

$$Q_+(x,0,t) = -Q_-(x,0,t) A^-(0) D (A^+(0))^{-1}, \tag{4.9}$$

$$Q^-(1,\xi,t) = -(A^-(1))^{-1} D_1^T A^+(1) Q^+(1,\xi,t), \tag{4.10}$$

$$Q_-(x,1,t) = -Q_+(x,1,t) A^+(1) D_1 (A^-(1))^{-1} \tag{4.11}$$

and use the formula (derived from (3.7), (4.1))

$$u_*(t) = -U^{-1}D^TA^+(1)\int_0^1 Q^+(1,\xi,t)w_*(\xi,t)\,d\xi \tag{4.12}$$

for the control $u(t)$, then (4.6) becomes

$$\begin{aligned} 0 = \int_0^{t_1}\int_0^1\int_0^1 w_*(x,t)^T\bigg[&\frac{\partial Q(x,\xi,t)}{\partial t} - A(x)\frac{\partial Q(x,\xi,t)}{\partial x} \\ &- \frac{\partial Q(x,\xi,t)}{\partial \xi}A(\xi) + (B(x)^T - A'(x))Q(x,\xi,t) \\ &+ Q(x,\xi,t)(B(\xi) - A'(\xi)) + W(x,\xi) \\ &- Q_+(x,1,t)A^+(1)DU^{-1}D^TA^+(1)Q^+(1,\xi,t)\bigg]w_*(\xi,t)\,dx\,d\xi\,dt. \end{aligned}$$

Thus the sum of the last two integrals in (4.4) is reduced to zero if Q satisfies the boundary conditions (4.8), (4.9), (4.10), (4.11) and the partial differential equation

$$\frac{\partial Q}{\partial t} = A(x)\frac{\partial Q}{\partial x} + \frac{\partial Q}{\partial \xi}A(\xi) + P(x,\xi,Q), \tag{4.13}$$

where

$$\begin{aligned} P(x,\xi,Q(\cdot,\cdot,t)) = &-(B(x)^T - A'(x))Q(x,\xi,t) - Q(x,\xi,t)(B(\xi) - A'(\xi)) \\ &- W(x,\xi) + Q_+(x,1,t)A^+(1)DU^{-1}D^TA^+(1)Q^+(1,\xi,t). \end{aligned} \tag{4.14}$$

The terminal condition (3.4) for v, which has the form

$$v(x,t_1) = \int_0^1 G(x,\xi)w_*(\xi,t)\,d\xi,$$

will be satisfied if we require

$$Q(x,\xi,t_1) = G(x,\xi). \tag{4.15}$$

That $v(x,t)$, as given by (4.1), satisfies

$$\frac{\partial v}{\partial t} = A(x)\frac{\partial v}{\partial x} - (B^T(x) - A'(x))v - \int_0^1 W(x,\xi)w_*(\xi,t)\,d\xi$$

can be verified directly. Moreover, with our requirements on Q, (4.5) is valid and takes the form

$$J(u_*, w_*, t_1) = \int_0^1\int_0^1 w_0(x,0)^TQ(x,\xi,0)w_*(\xi,0)\,dx\,d\xi. \tag{4.16}$$

Our task, then, is to show that the system (4.13), (4.8), (4.9), (4.10), (4.11), (4.15) has, in $0 \leqq x \leqq 1, 0 \leqq \xi \leqq 1, 0 \leqq t \leqq t_1$, a solution $Q(x,\xi,t)$ sufficiently well-behaved so that the above computations are meaningful. Assuming that this can be done, the equation (4.12) will provide a "synthesis" of the optimal control $u_*(t)$. The resulting closed loop system is (2.9) with boundary conditions (2.14)

and

$$(4.17)\quad w^+(1,t) = D_1w^-(1,t) + \int_0^1 [-DU^{-1}D^TA^+(1)Q^+(1,\xi,t)]w(\xi,t)\,d\xi.$$

The boundary condition (4.17) is of the form (2.19) with $H_1(x)$ replaced by

$$(4.18)\qquad H_1(x,t) = -DU^{-1}D^TA^+(1)Q^+(1,x,t)$$

and with the term $Du(t)$ removed so the optimal closed loop system is of the type discussed in Theorem 2.1 except that H_1 depends on t. In any region where $\|H_1(x,t)\|$ can be uniformly bounded, the results of Theorem 2.1 will still hold.

5. Local existence and uniqueness of $Q(x, \xi, t)$. If we think of the matrix Q as a vector q of dimension n^2, the equation (4.13) assumes the form

$$(5.1)\qquad \frac{\partial q}{\partial t} - A_1(x)\frac{\partial q}{\partial x} - A_2(\xi)\frac{\partial q}{\partial \xi} - p(x,\xi,q) = 0,$$

where A_1 and A_2 are matrices of dimension $n^2 \times n^2$. It is easy to see that the fact that A is diagonal in (4.13) implies that both $A_1(x)$ and $A_2(\xi)$ are diagonal in (5.1). (This can be verified by observing that if Q is a matrix with only one nonzero entry then both $A(x)Q$ and $QA(\xi)$ are scalar multiples of Q if A is diagonal.) Then, obviously, A_1 and A_2 are symmetric, so (5.1) is a system of symmetric hyperbolic type.

The system is semilinear because the principal part of the equation,

$$\frac{\partial q}{\partial t} - A_1(x)\frac{\partial q}{\partial x} - A_2(\xi)\frac{\partial q}{\partial \xi},$$

is linear in q; only the function p is nonlinear in q. The nonlinearity of the equation does not in itself introduce any really novel feature. But when we examine $p(x, \xi, q) = P(x, \xi, Q)$, as given by (4.14), we see that no matter what the values of x and ξ are, $P(x, \xi, Q)$ depends in part upon *boundary* values of Q at $x = 1$ and at $\xi = 1$. This feature might be expected to cause substantial problems in proving the existence and uniqueness of solutions of (4.13).

In the present instance, the special nature of (4.13) spares us from unsurmountable difficulty. Ordinarily, hyperbolic systems with two or more space variables possess characteristic surfaces of dimension two or more. But if in an equation of the type (5.1) the matrices $A_1(x)$ and $A_2(\xi)$ can be *simultaneously diagonalized*, then one can define a system of one-dimensional characteristic curves and obtain existence and uniqueness results by integrating ordinary differential equations over these characteristic curves just as one treats existence and uniqueness problems for hyperbolic systems with a single space variable. (See [3, Chap. V] for details in this latter instance.) Existence and uniqueness proofs of this type are carried out within the framework of the C- (or L^∞-) norm. In such a setting the fact that $P(x, \xi, Q)$ depends upon values of Q at points other than (x, ξ, t) does not cause any real problems, as we shall see.

In our case $A_1(x)$ and $A_2(\xi)$ are both already in diagonal form. If we let q^i_j denote the entry in the ith row and jth column of the matrix Q and let p^i_j denote the corresponding entry of the matrix $P = P(x, \xi, Q)$, then (4.13) and (5.1) are equivalent to the following systems of n^2 equations:

$$\frac{\partial q_j^i}{\partial t} - \lambda_i(x)\frac{\partial q_j^i}{\partial x} - \lambda_j(\xi)\frac{\partial q_j^i}{\partial \xi} + p_j^i(x, \xi, Q) = 0, \tag{5.2}$$
$$i = 1, 2, \cdots, n, \quad j = 1, 2, \cdots, n.$$

We define the characteristics of the system (5.1) to be curves in the region $0 \leqq x \leqq 1, 0 \leqq \xi \leqq 1, 0 \leqq t \leqq t_1$ having a parametric representation

$$\left\{(x^i(t), \xi_j(t), t) \middle| \frac{dx^i}{dt} = -\lambda_i(x^i), \frac{d\xi_j}{dt} = -\lambda_j(\xi_j)\right\}. \tag{5.3}$$

Such a characteristic curve will be denoted by the symbol c_j^i. When it is necessary to specify a characteristic passing through a particular point (x, ξ, t), we shall write $c_j^i(x, \xi, t)$.

If we consider the restriction of q_j^i to a characteristic curve c_j^i, namely $q_j^i(x^i(t), \xi_j(t), t)$, we see that (5.2) can be expressed in the form

$$\frac{d}{dt}q_j^i(x^i(t), \xi_j(t), t) + p_j^i(x^i(t), \xi_j(t), Q) = 0, \tag{5.4}$$
$$i = 1, 2, \cdots, n, \quad j = 1, 2, \cdots, n.$$

Because we require a solution in the bounded region $0 \leqq x \leqq 1, 0 \leqq \xi \leqq 1$, $0 \leqq t \leqq t_1$, we can make no further progress without consideration of the boundary conditions (4.8)–(4.11). Attempting to take these into account leads one to the notion of a *chain* of characteristic curves.

DEFINITION 5.1. Let $0 \leqq \hat{x} \leqq 1, 0 \leqq \hat{\xi} \leqq 1, 0 \leqq \hat{t} \leqq t_1$. A *chain* to $(\hat{x}, \hat{\xi}, \hat{t})$ from the surface $t = t_0, t_1 \geqq t_0 > \hat{t}$, is a sequence

$$\pi_0, c_{j_1}^{i_1}, \pi_1, c_{j_2}^{i_2}, \pi_2, \cdots, \pi_{m-1}, c_{j_m}^{i_m}, \pi_m$$

of characteristic curves c_j^i and points $\pi \in R^3$ with the following properties:

(i) $c_{j_k}^{i_k}$ joins π_{k-1} to π_k;

(ii) $\pi_m = (\hat{x}, \hat{\xi}, \hat{t})$ and $\pi_0 = (x_0, \xi_0, t_0)$ for some x_0, ξ_0 satisfying $0 \leqq x_0 \leqq 1$, $0 \leqq \xi_0 \leqq 1$;

(iii) $\pi_k, k = 1, 2, \cdots, m - 1$, is a point on one of the surfaces $x = 0, x = 1$, $\xi = 0$ or $\xi = 1$;

(iv) if $\pi_k, k = 1, 2, \cdots, m - 1$, lies on $x = 0$ or $x = 1$, but not on $\xi = 0$ or $\xi = 1$, then $j_k = j_{k+1}$, and if π_k lies on $x = 1, 1 \leqq i_{k+1} \leqq p, p + 1 \leqq i_k \leqq n$, while if π_k lies on $x = 0, p + 1 \leqq i_k \leqq n, 1 \leqq i_k \leqq p$;

(v) if $\pi_k, k = 1, 2, \cdots, m - 1$, lies on $\xi = 0$ or $\xi = 1$ but not on $x = 0$ or $x = 1$, the requirements are as in (iv) but with the roles of x and ξ, i and j, both reversed;

(vi) if $\pi_k, k = 1, 2, \cdots, m - 1$, lies on the intersection of a surface $x = 0$ or $x = 1$ with a surface $\xi = 0$ or $\xi = 1$, then it must be possible to approximate the chain segment $\pi_{k-1}, c_{j_k}^{i_k}, \pi_k, c_{j_{k+1}}^{i_{k+1}}, \pi_{k+1}$ as closely as desired by chain segments $\pi_{k-1}, \hat{c}_{j_k}^{i_k}, \hat{\pi}_k, \bar{c}_{j_k}^{i_k}, \tilde{\pi}_k, \tilde{c}_{j_{k+1}}^{i_{k+1}}, \pi_{k+1}$, where (iv) or (v) applies at $\hat{\pi}_k$ and $\tilde{\pi}_k$.

It is clear that not all chains are created equal. A further definition is convenient.

DEFINITION 5.2. If any π_k, $k = 1, 2, \cdots, m - 1$, lies on the intersection of a surface $x = 0$ or $x = 1$ with a surface $\xi = 0$ or $\xi = 1$, then the chain in question is *degenerate.*

DEFINITION 5.3. The *network of direct dependency* of $(\hat{\pi}, \hat{\xi}, \hat{t})$, denoted NDD $(\hat{x}, \hat{\xi}, \hat{t})$, is the union of all chains from the surface $t = t_1$ to $(\hat{x}, \hat{\xi}, \hat{t})$. The set Δ of points $(\hat{x}, \hat{\xi}, \hat{t})$ for which NDD$(\hat{x}, \hat{\xi}, \hat{t})$ includes a degenerate chain will be called the *set of degenerate points.*

The term "network of direct dependency" is motivated by the observation that if we remove the term $P(x, \xi, Q)$ from our system, then (5.4) becomes

$$\frac{dq_j^i}{dt}(x^i(t), \xi_j(t), t) = 0,$$

and $Q(\hat{x}, \hat{\xi}, \hat{t})$ can then be computed directly from the values of $Q(x, \xi, 0)$ at the initial points (x, ξ, t_1) of chains in NDD$(\hat{x}, \hat{\xi}, \hat{t})$ together with the use of the boundary conditions (4.8), (4.9), (4.10), (4.11), provided $(\hat{x}, \hat{\xi}, \hat{t})$ is nondegenerate and Q is unambiguously defined at (x, ξ, t_1).

The following lemma is important in our existence and uniqueness proofs.

LEMMA 5.4. *There is a* $T_0 > 0$ *such that whenever* $\hat{t} \geqq t_0 - T_0$, *then every chain from the surface* $t = t_0$ *to a point* $(\hat{x}, \hat{\xi}, \hat{t}), 0 \leqq x \leqq 1, 0 \leqq \xi \leqq 1$, *contains at most three characteristic arcs* c_j^i, *i.e.,* $m \leqq 3$.

Proof. If there are four or more characteristic arcs in a chain, then that chain includes the sequence

$$\pi_1, c_{j_2}^{i_2}, \pi_2, c_{j_3}^{i_3}, \pi_3.$$

Assuming for the moment that the chain is nondegenerate, it is clear from (iv) and (v) of Definition 5.1 that either $i_2 = i_3$ or $j_2 = j_3$. Thus the characteristic arcs $c_{j_2}^{i_2}$ and $c_{j_3}^{i_3}$ lie on opposite sides of a plane through π_2, parallel to the t-axis with an equation $x =$ const. or $\xi =$ const., respectively. Since both π_1 and π_3 lie on boundaries $x = 0$, $x = 1$, $\xi = 0$ or $\xi = 1$ different from the one on which π_2 lies, we conclude that the combined arc length of $c_{j_2}^{i_2}$ and $c_{j_3}^{i_3}$ is at least 1. Thus the total length of the arcs in the chain is at least one. Since degenerate chains are approximable by regular chains having more arcs, we conclude that the total length of the arcs in any chain consisting of at least four characteristic curves is at least 1.

Noting the representation (5.3) of the characteristics of our system and the fact that the $\lambda_i(x)$ are uniformly bounded, we see that $|dt/d\sigma|$ (σ being arc length) is uniformly bounded away from zero along any characteristic arc. If we measure arc length starting with $\sigma = 0$ at $t = t_0$, σ increasing as t decreases, then there is a fixed $d > 0$ such that

$$dt/d\sigma \leqq -d. \tag{5.5}$$

Then, if we take $T_0 > d$, all chains from the surface $t = t_0$ to a point $(\hat{x}, \hat{\xi}, \hat{t})$ with $\hat{t} \geqq t_0 - T$ have a total arc length less than one and must, according to the above, consist of at most three characteristic arcs. This proves the lemma.

COROLLARY 5.5. *There is a nondecreasing positive integer-valued function* $M(T_0)$ *such that, for any* $T_0 > 0$, *a chain from the surface* $t = t_0$ *to a point* $(\hat{x}, \hat{\xi}, \hat{t})$

with $\hat{t} \geqq t_0 - T_0$ contains at most $M(T_0)$ characteristic arcs. Moreover,

$$M(T_0) \leqq 3[T_0/d + 1].$$

The proof is clear.

The proof of the local existence and uniqueness theorem is much the same as that of similar theorems in [3, Chap. V]. The main differences arise from the fact that we must take boundary conditions and boundary values into account more or less explicitly.

DEFINITION 5.6. The *set Σ of singular points* is the smallest subset of $0 \leqq x \leqq 1$, $0 \leqq \xi \leqq 1$, $t < t_1$ with the properties:

(i) $(\hat{x}, \hat{\xi}, \hat{t}) \in \Sigma$ if $\mathrm{NDD}(\hat{x}, \hat{\xi}, \hat{t})$, $\mathrm{NDD}(\hat{x}, 1, t)$ or $\mathrm{NDD}(1, \hat{\xi}, t)$ includes a chain from the surface $t = t_1$ for which $\pi_0 = (x, \xi, t_1)$ has the property $x = 0$, $x = 1$, $\xi = 0$ or $\xi = 1$;

(ii) $(\tilde{x}, \tilde{\xi}, \tilde{t}) \in \Sigma$ if $\mathrm{NDD}(\tilde{x}, \tilde{\xi}, \tilde{t})$, $\mathrm{NDD}(\tilde{x}, 1, t)$ or $\mathrm{NDD}(1, \tilde{\xi}, t)$ contains a chain which includes an arc lying in Σ.

A little reflection, combined with the use of (iv) and (v) in Definition 5.1, shows that the set Σ of singular points consists of a countable collection of two-dimensional C^1-surfaces S, only finitely many of which meet a region

$$\hat{R}_{T_0} = \{(x, \xi, t) | 0 \leqq x \leqq 1, 0 \leqq \xi \leqq 1, t_1 - T_0 \leqq t \leqq t_1\}$$

for $T_0 > 0$. If (x, ξ, t) lies on such a surface S, then one of the sets of vectors

$$(\lambda_i(x), 0, 1), (0, 1, 0) \quad \text{or} \quad (1, 0, 0), (0, \lambda_j(\xi), 1)$$

is a basis for the tangent space to S at (x, ξ, t), for appropriate i, j, $1 \leqq i \leqq n$, $1 \leqq j \leqq n$.

On the other hand, the set Δ of degenerate points is a countable collection of C^1-surfaces S, only finitely many of which meet $\hat{R}_{T_0}$ for fixed $T_0 > 0$, with the property that if a point $(x, \xi, t) \in S$, then, for appropriate i, j, $1 \leqq i \leqq n$, $1 \leqq j \leqq n$, the vectors $(\lambda_i(x), \lambda_j(\xi), 0)$, $(0, 0, 1)$ form a basis for the tangent space to S at (x, ξ, t).

Let $\mathcal{Q}_{T_0}$ denote the space of matrix-valued functions defined on $\hat{R}_{T_0}$ which are continuously extendable to the closure of any region of $\hat{R}_{T_0}$ whose interior does not meet $\Sigma \cup \Delta$. That is, the only possible discontinuities of a function $Q \in \mathcal{Q}_{T_0}$ are jump discontinuities across surfaces in $\Sigma \cup \Delta$. With the norm

$$\|Q\|_s = \sup_{(x,\xi,t) \in \hat{R}_T - (\Sigma \cup \Delta)} (\max_{i,j} |q^i_j(x, \xi, t)|),$$

$\mathcal{Q}_{T_0}$ becomes a Banach space. In $\mathcal{Q}_{T_0}$ we define a closed subset

$$\mathcal{N}(G, M) = \{Q \in \mathcal{Q}_{T_0} | \, \|Q\|_s \leqq M \|G\|_s\}. \tag{5.6}$$

(The meaning of $\|G\|_s$ should be clear even though G does not depend on t.)

We shall now show that the problem (4.13), (4.8), (4.9), (4.10), (4.11), (4.15) has a solution (in a sense which we shall make precise below) in $\hat{R}_{T_0}$ if $T_0 > 0$ is chosen sufficiently small. In particular, we shall assume at the start that $T_0 < d$ (cf. (5.5)).

We let $\hat{Q}(x, \xi, t)$ be an arbitrary function in the set $\mathcal{N}(G, M)$. We construct a mapping

$$F : \hat{Q} \to \tilde{Q} \tag{5.7}$$

in the following manner.

Let (x_1, ξ_1, t_1) be a point on the terminal surface $t = t_1$ with $0 \leqq x_1 \leqq 1$, $0 \leqq \xi_1 \leqq 1$. Consider a characteristic $c_j^i(x_1, \xi_1, t_1)$ through this point, i.e.,

$$c_j^i(x_1, \xi_1, t_1) = \{(x^i(t), \xi_j(t), t) | \tilde{t} \leqq t \leqq t_1\},$$

where $x^i(t)$, $\xi_j(t)$ solve

$$\frac{dx^i}{dt} = -\lambda_i(x^i), \qquad x^i(t_1) = x_1;$$

$$\frac{d\xi_j}{dt} = -\lambda_j(\xi_j), \qquad \xi_j(t_1) = \xi_1;$$

and $\tilde{t}$ is either $t_1 - T_0$ or that value of $t < t_1$ such that c_j^i meets one of the surfaces $x = 0$, $x = 1$, $\xi = 0$ or $\xi = 1$, whichever is greater. For $\tilde{t} \leqq t \leqq t_1$ we integrate (5.4) over c_j^i with Q replaced by $\hat{Q}$ in the expression $p_j^i(x, \xi, Q)$, thereby defining

$$(5.8) \qquad \tilde{q}_j^i(x^i(t), \xi_j(t), t) = g_j^i(x_1, \xi_1) + \int_t^{t_1} p_j^i(x^i(s), \xi_j(s), \hat{Q})\, ds,$$

where $G(x, \xi) = (g_j^i(x, \xi))$.

In this way we obtain all values $\tilde{q}_j^i(x, \xi, t)$, $(x, \xi, t) \in \hat{R}_{T_0}$, which can be calculated by integrating along a single characteristic arc from the set $\{(x_1, \xi_1, t_1) | 0 \leqq x_1 \leqq 1, 0 \leqq \xi_1 \leqq 1\}$. (We shall have to qualify this statement to a mild degree, but we choose not to interrupt the train of thought at this point.) If NDD(x, ξ, t) contains only chains consisting of a single characteristic arc, we assign (x, ξ, t) to the set X_1.

Next we let (x_2, ξ_2, t_2) be a point in X_1 which lies on one of the boundary surfaces $x = 0$, $x = 1$, $\xi = 0$ or $\xi = 1$. Since NDD(x_2, ξ_2, t_2) contains only chains consisting of a single characteristic arc, $\tilde{q}_j^i(x_2, \xi_2, t_2)$ is already defined for all pairs i, j such that $c_j^i(x_2, \xi_2, t_2) \cap \hat{R}_T$ lies in the region $t \geqq t_2$. Then the boundary conditions (4.8), (4.9), (4.10), (4.11) are employed to determine the values of $\hat{q}_l^k(x_2, \xi_2, t_2)$ for pairs k, l such that $c_l^k(x_2, \xi_2, t_2) \cap \hat{R}_T$ lies in the region $t \leqq t_2$. Along such a characteristic c_l^k we again define $\tilde{q}_l^k$ by integration:

$$(5.9) \quad \hat{q}_l^k(x^k(t), \xi_l(t), t) = \tilde{q}_l^k(x_2, \xi_2, t_2) + \int_t^{t_2} p_l^k(x^k(s), \xi_l(s), \hat{Q})\, ds, \qquad t_3 \leqq t \leqq t_2,$$

where t_3 is $t_1 - T_0$ or that value of t for which the characteristic c_l^k again encounters one of the surfaces $x = 0$, $x = 1$, $\xi = 0$ or $\xi = 1$, whichever is greater. (Remember that we are proceeding in the direction of *decreasing* t.)

Up to the present point, then, we have obtained all values $\tilde{q}_j^i(x, \xi, t)$, $(x, \xi, t) \in \hat{R}_{T_0}$, which can be calculated by integrating along chains consisting of at most two characteristic arcs, with the boundary conditions being used at the conjunction of the two arcs. (Again, we shall have to qualify this statement below.) If NDD(x, ξ, t) consists only of chains containing at most two characteristic arcs, we assign (x, ξ, t) to the set X_2. Clearly $X_1 \subset X_2$.

Finally we let (x_3, ξ_3, t_3) be a point in $X_2 - X_1$ which lies on one of the boundary surfaces $x = 0$, $x = 1$, $\xi = 0$ or $\xi = 1$. By virtue of our definition of X_2, $\tilde{q}_j^i(x_3, \xi_3, t_3)$ is defined if $c_j^i(x_3, \xi_3, t_3) \cap \hat{R}_{T_0}$ lies in $t \geqq t_3$. The boundary conditions are again used to define $\tilde{q}_l^k(x_3, \xi_3, t_3)$ for pairs k, l such that $c_l^k(x_3, \xi_3, t_3) \cap \hat{R}_T$

lies in $t \leqq t_3$, and we define

$$\tilde{q}_l^k(x^k(t), \xi_l(t), t) = \tilde{q}(x_3, \xi_3, t_3) + \int_t^{t_3} p_l^k(x^k(x), \xi_l(s), \hat{Q})\, ds. \tag{5.10}$$

Now we have obtained all values $\tilde{q}_j^i(x, \xi, t), (x, \xi, t) \in \hat{R}_T$, which can be calculated by integrating along chains consisting of at most three characteristic arcs, with the boundary conditions being used at the conjunction of any two arcs. Applying Lemma 5.4, we see that this set is, in fact, $\hat{R}_{T_0}$ since we have taken $T_0 < d$. Thus we have determined all $\tilde{q}_j^i(x, \xi, t), (x, \xi, t) \in \hat{R}_{T_0}$, and hence the matrix function $\tilde{Q}(x, \xi, t)$ is defined there. Thus $\tilde{Q}$ has been constructed from $\hat{Q}$ and the mapping (5.7) has been defined.

Now we state the qualification which we have been warning of all along. It is quite possible that the boundary conditions (4.8), (4.9), (4.10) and (4.11) are not satisfied by $G(x, \xi)$. If such a problem arises in connection with G, then $\tilde{Q}$ must be discontinuous along some of the intersections of the surfaces $x = 0$, $x = 1$, $\xi = 0$, $\xi = 1$ with the surface $t = t_1$. Such discontinuities propagate along surfaces in the set Σ of singular points.

At first glance it might also seem that the boundary conditions (4.8), (4.9), (4.10) and (4.11) could be inconsistent, leading to discontinuities in $\tilde{Q}$ along the intersections of the vertical planes $x = 0$, $x = 1$, $\xi = 0$ and $\xi = 1$. Such discontinuities would propagate along surfaces contained in the set Δ of degenerate points. Our construction allows for such discontinuities. We shall show in the next section, however, that the boundary conditions (4.8), (4.9), (4.10) and (4.11) are, in fact, consistent and that no such discontinuities actually arise.

Our definition of $\mathcal{N}(G, M)$ also allows $\hat{Q}$ to have discontinuities at points $(x, \xi, t) \in \hat{R}_{T_0} \cap (\Sigma \cup \Delta)$. These discontinuities in $\hat{Q}$ will produce discontinuities in $\tilde{Q}$ when a characteristic arc $(x^k(t), \xi_j(t), t)$ used in defining $\tilde{Q}$ lies in a surface along which $\hat{Q}$ is discontinuous or if $\tilde{Q}$ is discontinuous along one of the boundary arcs $(x^i(t), 1, t), (1, \xi_j(t), t)$. The sets Σ and Δ have the property that all such discontinuities induced in Q again have their loci along the surfaces of $\Sigma \cup \Delta$. We grant the skeptical reader that this assertion is not quite obvious, but a careful consideration of the way in which Σ and Δ are defined together with the form taken by the boundary conditions will, we claim, lead him to the same conclusion. A formal proof would be quite tedious and seems hardly warranted here.

The rest of the proof consists of making sure that the hypotheses of the contraction fixed-point theorem are met by the mapping F in the set $\mathcal{N}(G, M)$. We shall try to be brief. The definition (4.14) of $P(x, \xi, Q)$ shows that there are continuous positive functions $M_0 : R^1 \to R^1$, $M_1 : R^1 \to R^1$ such that

$$\|P(\cdot, \cdot, Q)\|_s \leqq M_0(\|Q\|_s)\|Q\|_s, \tag{5.11}$$

$$\|P(\cdot, \cdot, Q_1) - P(\cdot, \cdot, Q_2)\|_s \leqq M_1(\|Q_1\|_s + \|Q_2\|_s)\|Q_1 - Q_2\|_s \tag{5.12}$$

for all matrix functions Q, Q_1, Q_2 in $\mathcal{Q}_{T_0}$. Also, there is a positive number μ such that (cf. (4.8), (4.9), (4.10), (4.11))

$$\|(A^+(0))^{-1} D_0^T A^-(0)\| \leqq \mu,$$

$$\|(A^-(1))^{-1} D_1^T A^+(1)\| \leqq \mu,$$

the norm of such a matrix being defined as the maximum row sum of the absolute values of its entries. Then we let (cf. (5.7))

$$\hat{\eta}(t) = \|\hat{Q}(\cdot,\cdot,t)\|_s, \qquad \tilde{\eta}(t) = \|\tilde{Q}(\cdot,\cdot,t)\|_s \tag{5.13}$$

and let $\bar{\eta}(t, t_2, t_3)$ satisfy

$$\bar{\eta}(t_1, t_2, t_3) = \hat{\eta}(t_1) = \|G\|_s; \tag{5.14}$$

$$-\frac{d\bar{\eta}}{dt} = M_0(\hat{\eta})\hat{\eta}, \qquad t_2 < t < t_1; \tag{5.15}$$

$$\bar{\eta}(t_2-, t_2, t_3) = \mu\bar{\eta}(t_2+, t_2, t_3); \tag{5.16}$$

$$-\frac{d\bar{\eta}}{dt} = M_0(\hat{\eta})\hat{\eta}, \qquad t_3 < t < t_2; \tag{5.17}$$

$$\bar{\eta}(t_3-, t_2, t_3) = \mu\bar{\eta}(t_3+, t_2, t_3); \tag{5.18}$$

$$-\frac{d\bar{\eta}}{dt} = M_0(\hat{\eta})\hat{\eta}, \qquad t_1 - T_0 < t < t_3. \tag{5.19}$$

It is clear then that for $t_1 - T_0 \leqq t \leqq t_1$,

$$\tilde{\eta}(t) \leqq \sup_{t_1 - T_0 \leqq t_3 \leqq t_2 \leqq t_1} \{\bar{\eta}(t, t_2, t_3)\}. \tag{5.20}$$

The equations (5.15), (5.17) and (5.19) correspond to integration of (5.4) over characteristic arcs, while (5.16) and (5.18) apply at the conjunction of these arcs, which occur at one of the boundaries $x = 0$, $x = 1$, $\xi = 0$ or $\xi = 1$. It should be understood that from one to four of the last of these equations may be missing, in general. If $\hat{Q} \in \mathcal{N}(G, M)$, then

$$\hat{\eta}(t) \leqq M\|G\|_s, \qquad t_1 - T_0 \leqq t \leqq t_1,$$

and we have

$$\begin{aligned} \bar{\eta}(t_1 - T_0, t_2, t_3) \leqq {} & ((\|G\|_s + M_0(M\|G\|_s)(M\|G\|_s)(t_1 - t_2))\mu \\ & + M_0(M\|G\|_s)(M\|G\|_s)(t_2 - t_3))\mu \\ & + M_0(M\|G\|_s)(M\|G\|_s)(t_3 - (t_1 - T_0)), \end{aligned} \tag{5.21}$$

when all of the equations (5.14)–(5.19) apply. If any of the last equations are missing, as we have indicated is quite possible, the inequality (5.21) should be modified in the obvious way.

Fix any value of M satisfying $M > 1 + \mu + \mu^2$. Then by taking T_0 sufficiently small, it is clear from (5.21) that we can ensure

$$\bar{\eta}(t_1 - T_0, t_2, t_3) \leqq M\|G\|_s, \qquad t_1 - T_0 < t_3 < t_2 < t_1.$$

Then (5.20) and (5.6) show that $\tilde{Q} \in \mathcal{N}(G, M)$ and the mapping (5.7) consequently takes the closed set $\mathcal{N}(G, M)$ into itself.

The contracting property of (5.7) is established in more or less the same way. We assume $\hat{Q}_1$ and $\hat{Q}_2$ lie in $\mathcal{N}(G, M)$ and let their images under (5.7) be $\tilde{Q}_1$ and $\tilde{Q}_2$. We set

$$\hat{\zeta}(t) = \|\hat{Q}_2(\cdot,\cdot,t) - \hat{Q}_1(\cdot,\cdot,t)\|_s, \tag{5.22}$$

$$\tilde{\zeta}(t) = \|\tilde{Q}_2(\cdot,\cdot,t) - \tilde{Q}_1(\cdot,\cdot,t)\|_s, \tag{5.23}$$

and let $\bar{\zeta}(t, t_2, t_3)$ satisfy

$$\bar{\zeta}(t_1, t_2, t_3) = 0; \tag{5.24}$$

$$-\frac{d\bar{\zeta}}{dt} = M_1(2M\|G\|_s)\hat{\zeta}, \qquad t_2 < t < t_1; \tag{5.25}$$

$$\bar{\zeta}(t_2-, t_2, t_3) = \mu\bar{\zeta}(t_2+, t_2, t_3); \tag{5.26}$$

$$-\frac{d\bar{\zeta}}{dt} = M_1(2M\|G\|_s)\hat{\zeta}, \qquad t_3 < t < t_2; \tag{5.27}$$

$$\bar{\zeta}(t_2-, t_2, t_3) = \mu\bar{\zeta}(t_3+, t_2, t_3); \tag{5.28}$$

$$-\frac{d\bar{\zeta}}{dt} = M_1(2M\|G\|_s)\hat{\zeta}, \qquad t_1 - T_0 < t < t_3. \tag{5.29}$$

Again, up to four of the last of these equations may be missing, in general. Then

$$\tilde{\zeta}(t) \leqq \sup_{t_1 - T_0 \leqq t_3 \leqq t_2 \leqq t_1} \{\bar{\zeta}(t, t_2, t_3)\}, \qquad t_1 - T_0 \leqq t \leqq t_1. \tag{5.30}$$

Now it is clear that

$$\begin{aligned} \bar{\zeta}(t, t_2, t_3) \leqq [((M_1(2M\|G\|_s)(t_1 - t_2))\mu + M_1(2M\|G\|_s)(t_2 - t_3))\mu \\ + M_1(2M\|G\|_s)(t_3 - (t_1 - T_0))] \sup_{t_1 - T \leqq t \leqq t_1} \{\hat{\zeta}(t)\}. \end{aligned} \tag{5.31}$$

(Again this inequality should be modified if some of the last equations (5.25)–(5.29) are inapplicable.) Then if we choose T_0 so that

$$T_0 < [(1 + \mu + \mu^2)M_1(2M\|G\|_s)]^{-1},$$

it is clear from (5.30), (5.31), (5.22) and (5.23) that the mapping (5.7) is a contraction. Applying the contraction fixed-point theorem, we have proved the following theorem.

THEOREM 5.7. *If* $T_0 > 0$ *is chosen sufficiently small, the system* (4.13), (4.8), (4.9), (4.10), (4.11), (4.15) *has a unique solution in*

$$\hat{R}_{T_0} = \{(x, \xi, t) | 0 \leqq x \leqq 1, 0 \leqq \xi \leqq 1, t_1 - T \leqq t \leqq t_1\}$$

in the sense that the mapping (5.7) *has a unique fixed point in the set* $\mathcal{N}(G, M)$.

Remark. If M_0 and M_1 can be replaced by constants in (5.11) and (5.12), as is the case when $P(x, \xi, Q)$ is taken to be a linear function instead of nonlinear as in (4.14), then the choice of T_0 depends only upon μ, not upon $\|Q(\cdot,\cdot,t_1)\|_s$. Then the above reasoning can be used repeatedly in intervals $[t_1 - T_0, t_1]$, $[t_1 - 2T_0, t_1 - T_0]$, $[t_1 - 3T_0, t_1 - 2T_0]$, $\cdots$ to obtain a global solution in $0 \leqq x \leqq 1, 0 \leqq \xi \leqq 1, t \leqq t_1$. We shall have occasion to make use of this observation in § 7.

6. Discussion of smoothness properties of Q, w_* and v. The existence proof for $Q(x, \xi, t)$ outlined in the previous section actually constructs $Q(x, \xi, t)$ at points $(x, \xi, t) \in \hat{R}_{T_0} - (\Sigma \cup \Delta)$. It is clear upon examination that the set of such points has full measure in $\hat{R}_{T_0}$. The "initial" members of the family $\Sigma \cup \Delta$ are surfaces passing through one of the eight edges of the set $0 \leqq x \leqq 1$, $0 \leqq \xi \leqq 1$, $t \leqq t_1$. Each intersection of one of these "initial" surfaces with a two-dimensional bounding surface of this region serves as a source curve for "secondary" surfaces—and so ad infinitum. In the three-dimensional regions cut out by these surfaces, $Q(x, \xi, t)$ will be continuously differentiable, provided only that $A(x) \in C^1[0, 1]$, $B(x) \in [0, 1]$, as we have already assumed, and provided that $G(x, \xi) \in C^1$, $W(x, \xi) \in C^0$ in $0 < x < 1, 0 < \xi < 1$. The proof of this assertion follows more or less standard lines. See [3] for comparable material dealing with linear hyperbolic systems in two independent variables x and t.

Differentiability, and even continuity, across surfaces $S \in \Sigma \cup \Delta$ depends upon certain consistency conditions being satisfied at the eight edges of $0 \leqq x \leqq 1$, $0 \leqq \xi \leqq 1, t \leqq t_1$. The consistency conditions along $x = 0, t = t_1$; $x = 1, t = t_1$; $\xi = 0, t = t_1$; $\xi = 1, t = t_1$ are requirements similar in form to (2.18) on the boundary values of $G(x, \xi)$ and its first order partial derivatives. If such requirements are not met, discontinuities in Q or its partial derivatives of first order will be propagated along the surfaces in Σ.

A complete study of the consistency of the boundary conditions (4.8), (4.9), (4.10) and (4.11) along the intersections of the planes $x = 0, x = 1, \xi = 0$ and $\xi = 1$ is quite involved. The essential features of the argument, however, are quite simple.

Consider first the boundary conditions (4.8) and (4.9) applied at $x = 0$, $\xi = 0$, respectively. We need to show that these are in agreement along the line $x = 0$, $\xi = 0, t \leqq t_1$. Now the solution Q is constructed "backward" in time from $t = t_1$. Thus the "given" data at a point $(0, 0, \tau)$ is obtained from only those entries q^i_j for which the characteristic $c^i_j(0, 0, \tau)$ intersects the region $0 \leqq x \leqq 1$, $0 \leqq \xi \leqq 1$, $t \leqq t_1$ for $t > \tau$. This means $\lambda_i(0) < 0$ and $\lambda_j(0) < 0$, from which we infer that the "given" data at $(0, 0, \tau)$ is the matrix $Q^-_-(0, 0, \tau)$. Here we combine the equations (4.3) in the form

$$Q = \begin{pmatrix} Q^-_- & Q^-_+ \\ Q^+_- & Q^+_+ \end{pmatrix}.$$

By letting

$$A^+_-(0) \equiv -(A^+(0))^{-1} D^T_0 A^-(0),$$

$$A^-_+(0) = (A^+_-(0))^T,$$

the conditions (4.8) become, at $(0, 0, \tau)$,

(6.1) $$Q^+_-(0, 0, \tau) = A^+_-(0) Q^-_-(0, 0, \tau),$$

(6.2) $$Q^+_+(0, 0, \tau) = A^+_-(0) Q^-_+(0, 0, \tau),$$

while the conditions (4.9) become

(6.3) $$Q^-_+(0, 0, \tau) = Q^-_-(0, 0, \tau) A^-_+(0),$$

(6.4) $$Q^+_+(0, 0, \tau) = Q^+_-(0, 0, \tau) A^-_+(0).$$

Conditions (6.1) and (6.3) are in agreement with the fact that $Q(0,0,\tau)$ must be symmetric, for

$$\begin{aligned}(Q_-^+(0,0,\tau))^T &= (Q_-^-(0,0,\tau))^T(A_-^+(0))^T\\ &= Q_-^-(0,0,\tau)A_+^-(0) = Q_+^-(0,0,\tau).\end{aligned}$$

Substituting (6.3) into (6.2) and (6.1) into (6.4), we obtain

$$Q_+^+(0,0,\tau) = A_-^+(0)(Q_-^-(0,0,\tau)A_+^-(0))$$

and

$$Q_+^+(0,0,\tau) = (A_-^+(0)Q_-^-(0,0,\tau))A_+^-(0),$$

so that $Q_+^+(0,0,\tau)$ is unambiguously determined as a symmetric matrix. Thus conditions (4.8) and (4.9) are consistent. The consistency of (4.10) and (4.11) along $x = 1$, $\xi = 1$, $t \leqq t_1$ is established similarly.

Now let us consider (4.9) and (4.10) along the line $x = 1$, $\xi = 0$, $t \leqq t_1$. Reasoning as before, the "given" data at a point $(1,0,\tau)$ is the matrix $Q_-^+(1,0,\tau)$. Letting

$$A_+^-(1) \equiv -(A^-(1))^{-1}D_1^TA^+(1),$$

we rewrite (4.9) as

$$Q_+^-(1,0,\tau) = Q_-^-(1,0,\tau)A_+^-(0), \tag{6.5}$$

$$Q_+^+(1,0,\tau) = Q_-^+(1,0,\tau)A_+^-(0), \tag{6.6}$$

and (4.10) as

$$Q_-^-(1,0,\tau) = A_+^-(1)Q_-^+(1,0,\tau), \tag{6.7}$$

$$Q_+^-(1,0,\tau) = A_+^-(1)Q_+^+(1,0,\tau). \tag{6.8}$$

The equations (6.6) and (6.7) determine $Q_+^+(1,0,\tau)$ and $Q_-^-(1,0,\tau)$ in terms of $Q_-^+(1,0,\tau)$. Substituting (6.6) into (6.8) and (6.7) into (6.5), we obtain

$$Q_+^-(1,0,\tau) = A_+^-(1)(Q_-^+(1,0,\tau)A_+^-(0))$$

and

$$Q_+^-(1,0,\tau) = (A_+^-(1)Q_-^+(1,0,\tau))A_+^-(0),$$

and thus the conditions (4.9) and (4.10) are consistent along $x = 1$, $\xi = 0$, $t \leqq t_1$. The consistency of (4.8) and (4.11) along $x = 0$, $\xi = 1$, $t \leqq t_1$ is established similarly.

We see therefore that the boundary conditions for Q are in agreement with each other and no discontinuities in Q or its partial derivatives are induced along the intersections of the planes $x = 0$, $x = 1$, $\xi = 0$, $\xi = 1$. This means that the set Δ of degenerate points is not the carrier of any discontinuities of the solution Q. Any such discontinuities must arise from failure on the part of $G(x,\xi) = Q(x,\xi,t_1)$ to satisfy consistency conditions on the intersections of the planes $x = 0$, $x = 1$, $\xi = 0$, $\xi = 1$ with the plane $t = t_1$. In particular, if G and its partial derivative satisfy appropriate conditions, similar in basic form to (2.18) but somewhat more complicated (such conditions are certainly satisfied when $G(x,\xi) \equiv 0$, an important case), then $Q(x,\xi,t)$ will be a C^1-solution of (4.13), (4.8), (4.9), (4.10), (4.11).

Assuming $Q(x, \xi, t)$ to be of class C^1 and anticipating, for the moment, the result of the next section, that $Q(x, \xi, t)$ is defined in the whole domain $0 \leqq x \leqq 1$, $0 \leqq \xi \leqq 1$, $0 \leqq t \leqq t_1$, we can discuss the smoothness properties of the optimal solution $w_*(x, t)$ and the adjoint solution $v(x, t)$.

Let us suppose first of all that $w_0(x)$, the initial state for $w_*(x, t)$, is of class C^1. Now $w_*(x, t)$ satisfies (2.9), (2.14), (2.16) and (4.17), the latter being equivalent to (2.15) when $u = u_*$ as given by (4.12). The consistency conditions for w_* at $x = 0$ are just the first and third equations of (2.18), and the consistency conditions at $x = 1$ are obtained by substituting (4.12) (with w_* replaced by w_0) and (4.18) into the second and fourth equations of (2.18), viz:

$$w_0^+(1) = D_1 w_0^-(1) + \int_0^1 H_1(x, 0) w_0(x)\, dx,$$

$$\begin{aligned} A^+(1)\frac{dw_0^+}{dx}(1) &+ B_-^+(1)w_0^-(1) + B_+^+(1)w_0^+(1) \\ &= D_1\left[A^-(1)\frac{dw_0^-}{dx}(1) + B_-^-(1)w_0^-(1) + B_+^-(1)w_0^+(1)\right] \\ &\quad + \int_0^1 \left[\frac{\partial H_1}{\partial t}(x, 0)w_0(x) + H_1(x, 0)\left(A\frac{dw_0}{dx}(x) + Bw_0(x)\right)\right] dx. \end{aligned} \tag{6.9}$$

If these conditions are all fulfilled, then $w_*(x, t)$ will be of class C^1. If not, then $w_*(x, t)$ will be piecewise C^1. Now the first and third conditions of (2.18) will automatically be satisfied if $w_0(x)$ represents the terminal value of a solution of (2.9), (2.14), (2.15) (perhaps without any control action, i.e., $u(t) \equiv 0$) during a time interval $\tau \leqq t \leqq 0$. So these conditions might normally be expected to be satisfied. It would seem, however, that the satisfaction of the conditions (6.9) would have to be regarded as an exceptionally rare coincidence in practice. Thus piecewise C^1-solutions $w_*(x, t)$ appear to be the "nicest" solutions which one can regard as typical in the application of this theory. (Of course, if $w_0(x)$ is merely of class $L^2[0, 1]$, the same will be true of $w_*(x, t)$ for each $t \geqq 0$.)

Since $v(x, t)$ is given by (4.1) and, by (4.8) and (4.10), the columns of Q satisfy the boundary conditions (3.5), (3.6) imposed on v, and since the inhomogeneous term $-\int_0^1 W(x, \xi)w_*(\xi, t)\, d\xi$ is a continuous function of x, t even if $w_*(x, t)$ is merely of class L^2, one can conclude without difficulty that $v(x, t)$ is of class C^1 if $Q(x, \xi, t)$ is of class C^1.

Even if $G(x, \xi)$ does not satisfy the appropriate consistency conditions at $t = t_1$, the assumptions on W and G made following (3.1) guarantee that $Q(x, \xi, t)$ will be piecewise C^1 throughout the region of existence established in § 5. This is all we shall need to assume in § 7.

7. Global existence and uniqueness of $Q(x, \xi, t)$**.** We shall now show that the solution $Q(x, \xi, t)$ of (4.13), shown in § 5 to exist for $0 \leqq x \leqq 1$, $0 \leqq \xi \leqq 1$, $t_1 - T \leqq t \leqq t_1$, provided $T > 0$ is sufficiently small, actually exists for $0 \leqq t \leqq t_1$. The method used is similar to the one in [12], where ordinary differential equations of Riccati-type whose solutions are operators in Hilbert space are studied. We need first of all the following lemma, which is also useful in § 8.

LEMMA 7.1. *For an arbitrary feedback control of the form*

$$\hat{u}(t) = \int_0^1 K(x)w(x,t)\,dx, \tag{7.1}$$

with $K(x)$ *a continuous* $r \times n$ *matrix function for* $0 \leqq x \leqq 1$, *and for a cost functional* $J(w_{t_0}, u, t_0, t_1)$ *defined by*

$$\begin{aligned} J(w_{t_0}, u, t_0, t_1) = \int_{t_0}^{t_1} \left[\int_0^1 \int_0^1 w(x,t)^T W(x,\xi) w(\xi,t)\,dx\,d\xi + u(t)^T U u(t) \right] dt \\ + \int_0^1 \int_0^1 w(x,t_1)^T G(x,\xi) w(\xi,t_1)\,dx\,d\xi \end{aligned} \tag{7.2}$$

(*where it is understood now that* w *satisfies* (2.9), (2.14) *and* (2.15) *with* $w(x,t_0) \equiv w_{t_0}(x)$), *the cost is given by*

$$J(w_{t_0}, \hat{u}, t_0, t_1) = \int_0^1 \int_0^1 w_{t_0}(x)^T \hat{Q}(x,\xi,t_0) w_{t_0}(\xi)\,dx\,d\xi, \tag{7.3}$$

where $\hat{Q}(x,\xi,t)$ *satisfies* (4.15), (4.8), (4.9), (4.10), (4.11) *and* (4.13) *with* $P(x,\xi,Q(\cdot,\cdot,t))$ *replaced by*

$$\begin{aligned} \hat{P}(x,\xi,\hat{Q}(\cdot,\cdot,t)) = -W(x,\xi) - K(x)^T U K(\xi) - K(x)^T D^T A^+(1) \hat{Q}^+(1,\xi,t) \\ - \hat{Q}_+(x,1,t) A^+(1) D K(\xi). \end{aligned} \tag{7.4}$$

This lemma is proved by replacing w_* in (3.4), (4.1) and related formulas by w and replacing Q in (4.1) by $\hat{Q}$. The calculations then proceed as in (4.6), (4.7) with $\hat{u}$, as given by (7.1), replacing u_*, as given by (4.12).

The fact that the partial differential equation for $\hat{Q}$ differs from (4.13), the equation for Q, to the extent that $P(x,\xi,Q(\cdot,\cdot,t))$ is replaced by $\hat{P}(x,\xi,\hat{Q}(\cdot,\cdot,t))$ leads to considerable simplification, for now $\hat{P}(x,\xi,\hat{Q}(\cdot,\cdot,t))$ is an affine (linear plus constant) function of $\hat{Q}$ for fixed x and ξ. In this case, the results of §§ 5 and 6 apply, but the functions M_0 and M_1 of (5.11) and (5.12) can be taken to be constants. The choice of T_0 is then independent of $\|\hat{Q}(\cdot,\cdot,t_1)\|_s$. Therefore, one may repeat the arguments of the proof of Theorem 5.7 in successive intervals $[t_1 - T_0, t_1], [t_1 - 2T_0, t_1 - T_0], \cdots, [t_1 - (k+1)T_0, t_1 - kT_0], \cdots$ of the same length. Thus one covers, in a finite number of steps, the interval $[0, t_1]$, and we obtain the *global* existence and uniqueness of a solution $\hat{Q}(x,\xi,t)$. Throughout $0 \leqq x \leqq 1$, $0 \leqq \xi \leqq 1$, $0 \leqq t \leqq t_1$, $\hat{Q}(x,\xi,t)$ will have the same properties indicated for $Q(x,\xi,t)$ in § 6.

Let the function $K(x)$ be fixed, e.g., $K(x) = 0$ will suffice. Then the cost $J(w_0, u_*, t_0, t_1)$ associated with the optimal control $u_*(t)$ must be less than or equal to the cost associated with the control (7.1), by the definition of optimality. Thus we conclude from (4.16) (replacing 0 by t_0) and (7.3) that for each initial state $w_{t_0}(x) \in L^2[0,1]$ we have, provided $0 \leqq t_0 < t_1$ and $Q(x,\xi,t)$ is defined for

$0 \leqq x \leqq 1, 0 \leqq \xi \leqq 1, t_0 \leqq t \leqq t_1,$

$$\int_0^1 \int_0^1 w_{t_0}(x)^T[\hat{Q}(x, \xi, t_0) - Q(x, \xi, t_0)]w_{t_0}(\xi)\, dx\, d\xi \geqq 0. \tag{7.5}$$

Consider the diagonal entries $q_i^i, \hat{q}_i^i$ of $Q, \hat{Q}$. Any discontinuities of these functions must occur along a surface S having $(-\lambda_i(x), -\lambda_i(\xi), 1)$ as a tangent vector at all points $(x, \xi, t) \in S$. If at (x, ξ, t) the vector $(1, 1, 0)$ is also a tangent vector to S, then any linear combination $\alpha_1(1, 1, 0) + \alpha_2(-\lambda_i(x), -\lambda_i(\xi), 1)$ will likewise be a tangent vector at that point. If $x = \xi$, we can take $\alpha_1 = \lambda_i(x) = \lambda_i(\xi)$, $\alpha_2 = 1$, and we find that $(0, 0, 1)$ is a tangent vector to the surface. It follows that any surface discontinuity of q_i^i, or $\hat{q}_i^i$, which includes an arc of a line $x = \xi$, t const., must also have $(0, 0, 1)$ as a tangent vector to the surface along that arc. Then the discontinuity at such points (x, x, t) is equal to

$$\lim_{\varepsilon \to 0} (q_i^i(x - \varepsilon, x + \varepsilon, t) - q_i^i(x + \varepsilon, x - \varepsilon, t)). \tag{7.6}$$

(A similar expression is obtained for $\hat{q}_i^i$.) But since $Q(x, \xi, t) = Q^T(\xi, x, t)$, the diagonal elements q_i^i satisfy

$$q_i^i(x, \xi, t) = q_i^i(\xi, x, t), \tag{7.7}$$

and we conclude that (7.4) is zero. Thus there can be no discontinuity of q_i^i across such a surface S. All other surfaces of discontinuity of q_i^i meet a line $x = \xi$, $t =$ const. at one point at most. Therefore there are at most finitely many points along such a line where q_i^i suffers a discontinuity. The same is true of $\hat{q}_i^i$. Now we can prove the next lemma.

Lemma 7.2. *For any t in an interval $[t_0, t_1]$ such that a piecewise C^1-solution $Q(x, \xi, t)$ of (4.13), (4.8), (4.9), (4.10), (4.11), (4.15) exists in $\hat{R}_{t_0} = \{(x, \xi, t) | 0 \leqq x \leqq 1, 0 \leqq \xi \leqq 1, t_0 \leqq t \leqq t_1\}$, we have*

$$\hat{q}_i^i(x_0, x_0, t) \geqq q_i^i(x_0, x_0, t), \qquad i = 1, 2, \cdots, n, \tag{7.8}$$

for all but finitely many $x_0 \in [0, 1]$.

Proof. Let x_0, $0 \leqq x_0 \leqq 1$, be a point such that both $\hat{q}_i^i$ and q_i^i are continuous at (x_0, x_0, t). From the remarks above it is clear that this is true for all but finitely many x_0. Define an n-vector function $w_{t_0}(x)$ by

$$\begin{aligned} w_{t_0}^j(x) &\equiv 0, && 0 \leqq x \leqq 1, \quad j \neq i; \\ w_{t_0}^i(x) &\equiv 1, && x_0 - \delta \leqq x \leqq x_0 + \delta; \\ w_{t_0}^i(x) &\equiv 0, && 0 \leqq x \leqq x_0 - \delta, \quad x_0 + \delta < x \leqq 1, \end{aligned}$$

$\delta > 0$ being chosen appropriately small. Substituting this function for w_{t_0} in (7.5), and replacing t_0 by t, we have

$$\int_{x_0 - \delta}^{x_0 + \delta} \int_{x_0 - \delta}^{x_0 + \delta} (\hat{q}_i^i(x, \xi, t) - q_i^i(x, \xi, t))\, dx\, d\xi \geqq 0.$$

(Replace $x_0 - \delta$ by x_0 or $x_0 + \delta$ by x_0 if $x_0 = 0$ or $x_0 = 1$, respectively.) Since (x_0, x_0, t) is a point of continuity for both $\hat{q}_i^i$ and q_i^i, this can be true for all small δ only if (7.8) is true, which completes the proof.

LEMMA 7.3. *Let* (x_0, ξ_0, t), $0 \leqq x_0 \leqq 1$, $0 \leqq \xi_0 \leqq 1$, $t_0 \leqq t \leqq t_1$ (t_0 *as defined in Lemma* 7.2) *be a point of continuity for* q^i_j, $1 \leqq i \leqq n$, $1 \leqq j \leqq n$. *Then*

$$|q^i_j(x_0, \xi_0, t)| \leqq \sup_{\substack{0 \leqq x \leqq 1 \\ 1 \leqq l \leqq n}} \{q^l_l(x, x, t)\}. \tag{7.9}$$

Proof. The discontinuities of $q^i_j(x, \xi, t)$, for fixed t, occur along two-dimensional curves in the planar region $0 \leqq x \leqq 1, 0 \leqq \xi \leqq 1$. Thus we may approximate (x_0, ξ_0) as closely as we please by points $(\hat{x}, \hat{\xi})$ such that q^i_j, q^i_i and q^j_i are continuous at the points $(\hat{x}, \hat{\xi}, t)$, $(\hat{x}, \hat{x}, t)$ and $(\hat{\xi}, \hat{\xi}, t)$, respectively. We define

$$\begin{aligned}
&w^k_{t_0}(x) \equiv 0, && 0 \leqq x \leqq 1, && k \neq i, j;\\
&w^i_{t_0}(x) \equiv 1, && \hat{x} - \delta \leqq x \leqq \hat{x} + \delta;\\
&w^i_{t_0}(x) \equiv 0, && 0 \leqq x < \hat{x} - \delta, && \hat{x} + \delta < x \leqq 1;\\
&w^j_{t_0}(x) \equiv 1, && \hat{\xi} - \delta \leqq x \leqq \hat{\xi} + \delta;\\
&w^j_{t_0}(x) \equiv 0, && 0 \leqq x < \hat{\xi} - \delta, && \hat{\xi} + \delta < x \leqq 1.
\end{aligned}$$

From (3.1) and the positivity conditions on $W(x, \xi)$ and U, it is clear that

$$\int_0^1 \int_0^1 w_0(x)^T Q(x, \xi, t) w_0(\xi)\, dx\, d\xi \geqq 0,$$

which, in the present case, becomes

$$\begin{aligned}
&\int_{\hat{x}-\delta}^{\hat{x}+\delta} \int_{\hat{x}-\delta}^{\hat{x}+\delta} q^i_i(x, \xi, t)\, dx\, d\xi + \int_{\hat{\xi}-\delta}^{\hat{\xi}+\delta} \int_{\hat{\xi}-\delta}^{\hat{\xi}+\delta} q^j_j(x, \xi, t)\, dx\, d\xi \\
&\qquad - 2 \int_{\hat{x}-\delta}^{\hat{x}+\delta} \int_{\hat{\xi}-\delta}^{\hat{\xi}+\delta} q^i_j(x, \xi, t)\, dx\, d\xi \geqq 0.
\end{aligned} \tag{7.10}$$

Clearly (7.10) can be true for all $\delta > 0$ only if

$$2q^i_j(\hat{x}, \hat{\xi}, t) \leqq q^i_i(\hat{x}, \hat{x}, t) + q^j_j(\hat{\xi}, \hat{\xi}, t),$$

whence, using Lemma 7.2,

$$\begin{aligned}
q^i_j(\hat{x}, \hat{\xi}, t) &\leqq \sup_{\substack{0 \leqq x \leqq 1 \\ 1 \leqq l \leqq n}} \{q^l_l(x, x, t)\} \\
&\leqq \sup_{\substack{0 \leqq x \leqq 1 \\ 1 \leqq l \leqq n}} \{\hat{q}^l_l(x, x, t)\}.
\end{aligned} \tag{7.11}$$

Letting $(\hat{x}, \hat{\xi})$ approach (x_0, ξ_0) through values satisfying the above continuity requirements, and using the continuity of q^i_j at (x_0, ξ_0, t), we conclude that (7.9) holds, and the proof is complete.

We are now able to prove the following theorem.

THEOREM 7.4. *The semilinear system* (4.13), (4.8), (4.9), (4.10), (4.11), (4.15) *has a piecewise* C_1*-solution* $Q(x, \xi, t)$ *in the entire region* $0 \leqq x \leqq 1$, $0 \leqq \xi \leqq 1$, $0 \leqq t \leqq t_1$.

Proof. We already know that $\hat{Q}(x, \xi, t)$ exists throughout the indicated region and it clearly satisfies an inequality

$$\|\hat{Q}\|_s \leqq K(t_1)$$

for some constant $K(t_1)$ depending only on t_1. Then Lemma 7.3 shows that

$$\|Q\|_s \leqq K(t_1)$$

in any region $0 \leqq x \leqq 1$, $0 \leqq \xi \leqq 1$, $t_0 \leqq t \leqq t_1$, where the solution Q of (4.13) exists. This a priori estimate allows one to replace the functions M_0 and M_1 in (5.11), (5.12) by the constants $M_0(K(t_1))$ and $M_1(2K(t_1))$. One then obtains the global solution by considering successive intervals $[t_1 - (k+1)T_0, t_1 - kT_0]$ as observed in the remark at the end of § 5. This completes the proof.

With Theorem 7.4 established, we can now summarize our work so far.

THEOREM 7.5. *Consider the optimization problem of* § 3. *The unique solution* u_* *of this problem is generated by the feedback law*

$$u_*(t) = -U^{-1}D^TA^+(1)\int_0^1 Q^+(1, \xi, t)w_*(\xi, t)\,d\xi,$$

which holds almost everywhere in $0 \leqq t \leqq t_1$, $Q(x, \xi, t)$ *being the piecewise* C_1-*solution of* (4.13), (4.8), (4.9), (4.10), (4.11), (4.15). *Moreover, the minimal cost, realized with this control, is given by* (4.16).

8. Remarks concerning the optimization problem on the infinite interval. In the previous discussion we have shown that the solution $Q(x, \xi, t)$ of (4.13), (4.8), (4.9), (4.10), (4.11), (4.15) exists for $0 \leqq x \leqq 1, 0 \leqq \xi \leqq 1, 0 \leqq t \leqq t_1$. Let us identify this solution more precisely by calling it $Q_{t_1}(x, \xi, t)$. There is no restriction on t_1 other than $t_1 > 0$, so we may consider also solutions $Q_{t_2}(x, \xi, t)$ for $t_2 \neq t_1$. Since the equation, boundary conditions and terminal data are independent of t, it is clear that for $0 \leqq x \leqq 1, 0 \leqq t \leqq t_1, t_2 > t_1$,

$$Q_{t_2}(x, \xi, t + (t_2 - t_1)) \equiv Q_{t_1}(x, \xi, t). \tag{8.1}$$

It is then natural to define

$$\tilde{Q}(x, \xi, t) \equiv Q_t(x, \xi, 0). \tag{8.2}$$

We now make a further assumption, namely,

$$G(x, \xi) \equiv 0. \tag{8.3}$$

It should be noted that this guarantees that $Q(x, \xi, t)$ is of class C^1.

LEMMA 8.1. *If* (8.3) *holds, then for any vector function* $w_0 \in L^2[0, 1]$,

$$\int_0^1\int_0^1 w_0(x)^T\tilde{Q}(x, \xi, t_2)w_0(\xi)\,dx\,d\xi \geqq \int_0^1\int_0^1 w_0(x)^T\tilde{Q}(x, \xi, t_1) \cdot w_0(\xi)\,dx\,d\xi, \qquad t_2 < t_1. \tag{8.4}$$

Proof. If u_1 and u_2 are the optimal controls relative to the costs $J(w_0, u, 0, t_1)$ and $J(w_0, u, 0, t_2)$, respectively, then

$$\begin{aligned} J(w_0, u_2, 0, t_2) &\geqq J(w_0, u_2, 0, t_1) \\ &\geqq J(w_0, u_1, 0, t_1). \end{aligned} \tag{8.5}$$

(The condition (8.3) is used in establishing the first of these inequalities.) Then (8.4) follows from (8.5), (4.16) and (8.2).

COROLLARY 8.2. *If* (x_0, ξ_0, t_1) *and* (x_0, ξ_0, t_2) *are both points of continuity for* $\tilde{q}^i_j$ *(the i, jth component of* $\tilde{Q}$*), then*

$$|\tilde{q}^i_j(x_0, \xi_0, t_2) - \tilde{q}^i_j(x_0, \xi_0, t_1)| \leqq \sup_{\substack{0 \leqq x \leqq 1 \\ 1 \leqq l \leqq n}} \{\tilde{q}^l_l(x,x, t_2) - \tilde{q}^l_l(x, x, t_1)\}, \qquad t_2 > t_1.$$

The proof follows the lines of Lemma 7.3 with $Q(x, \xi, t)$ replaced by $\tilde{Q}(x, \xi, t_2) - \tilde{Q}(x, \xi, t_1)$.

THEOREM 8.3. *If* (8.3) *is assumed and if there exists a positive constant K such that*

$$\sup_{\substack{0 \leqq x \leqq 1 \\ 0 \leqq l \leqq n}} \{\tilde{q}^l_l(x, x, t)\} \leqq K, \qquad 0 \leqq x \leqq 1, \quad t > 0, \tag{8.6}$$

then there is a matrix function $Q_\infty(x, \xi)$ *with*

(i) $Q_\infty(x, \xi)^T = Q_\infty(\xi, x), 0 \leqq x \leqq 1, 0 \leqq \xi \leqq 1$;

(ii) $\int_0^1 \int_0^1 w_0(x)^T Q_\infty(x, \xi) w_0(\xi)\, dx\, d\xi \geqq 0$ *for each vector function* w_0 *in* $L^2[0, 1]$;

(iii) $Q_\infty \in C([0, 1] \otimes [0, 1])$, *and* $\lim_{t\to\infty} \|Q_\infty(\cdot, \cdot) - \tilde{Q}(\cdot, \cdot, t)\|_s = 0$.

Proof. Consider the C^1-functions $\tilde{q}^l_l(x, x, t)$, $1 \leqq l \leqq n$, $0 \leqq x \leqq 1$, $t > 0$. Lemma 7.1 implies that

$$\tilde{q}^l_l(x, x, t_2) - \tilde{q}^l_l(x, x, t_1) \geqq 0, \qquad 0 \leqq x \leqq 1, \quad 1 \leqq l \leqq n, \tag{8.7}$$

if $t_2 > t_1$. Then using (8.6), (8.7) and the monotone convergence theorem, we see that there is a nonnegative measurable function $q_{\infty,l}(x)$ $(= q^l_{\infty,l}(x, x))$ with $q_{\infty,l}(x) \leqq K, 0 \leqq x \leqq 1$, which is the pointwise limit of $\tilde{q}^l_l(x, x, t)$ as $t \to \infty$.

Fix a value of l, $1 \leqq l \leqq n$, and assume, without loss of generality, that $\lambda_l(x) > 0$, $0 \leqq x \leqq 1$. Consider a curve $\gamma^l_l(0, 0, \tau)$ passing through a point $(0, 0, \tau)$, $\tau \geqq 0$, defined by

$$\gamma^l_l(0, 0, \tau) = \{(x, x, \tau) | 0 \leqq x \leqq 1, t = t_l(x, \tau)\}$$

with

$$\frac{dt_l}{dx} = \frac{1}{\lambda_l(x)}, \qquad 0 \leqq x \leqq 1, \quad t_l(0, \tau) = \tau.$$

Using (8.2) and (5.2) we see that

$$\frac{d}{dx}\tilde{q}^l_l(x, x, t_l(x, \tau)) = \frac{1}{\lambda_l(x)} p^l_l(x, x, \tilde{Q}), \qquad 0 \leqq x \leqq 1. \tag{8.8}$$

The assumption (8.6) together with (8.2) and the first inequality of (7.11) in the proof of Lemma 7.3 provides a uniform bound for the right-hand side of (8.8). Thus the family of functions $\{q^l_l(\cdot, \cdot, t_l(\cdot, t)) | \tau \geqq 0\}$ is an equicontinuous family on [0, 1]. Combined with the boundedness already assumed, the Ascoli–Arzela theorem shows that there is a sequence $\tau_k \to \infty$ such that

$$\lim_{\tau_k \to \infty} \tilde{q}_l^l(x, x, t_l(x, \tau)) = \bar{q}_{\infty,l}(x)$$

uniformly for $0 \leqq x \leqq 1$, for some limit function $\bar{q}_{\infty,l}(x)$. But, for each τ_k,

$$\begin{aligned} \tilde{q}_l^l(x, x, t_l(x, \tau_k)) &\leqq \tilde{q}_l^l(x, x, t_l(1, \tau_k)) \\ &\leqq \tilde{q}_l^l(x, x, t_l(x, \tau_{k'})), \qquad \tau_{k'} \geqq t_l(1, \tau_k), \end{aligned}$$

and since, pointwise, $\tilde{q}_l^l(x, x, t)$ converges monotonically to $q_{\infty,l}(x)$ as $t \to \infty$, we are able to conclude that

$$\lim_{t \to \infty} \|q_{\infty,l}(\cdot) - \tilde{q}_l^l(\cdot, \cdot, t)\|_s = 0.$$

Corollary 8.2 then completes the proof of the theorem.

THEOREM 8.4. *Let the hypotheses of Theorem* 8.3 *hold. Then* $Q_\infty(x, \xi)$ *is a solution of the boundary value problem consisting of the equation*

$$A(x)\frac{\partial Q}{\partial x} + \frac{\partial Q}{\partial \xi}A(\xi) + P(x, \xi, Q) = 0 \tag{8.9}$$

and the boundary conditions (4.8), (4.9), (4.10) *and* (4.11) (*altered in the obvious way*). *The equation* (8.9) *is satisfied by* $Q_\infty(x, \xi)$ *in the sense that for any point* (x_2, ξ_2) *on a curve described by*

$$\left\{(x, \xi) | \xi = \xi_j^i(x), \frac{d\xi_j^i(x)}{dx} = \frac{\lambda_j(\xi_j^i(x))}{\lambda_i(x)}, \xi_1 = \xi_j^i(x_1)\right\}, \tag{8.10}$$

we have

$$q_{\infty,j}^i(x_2, \xi_2) = q_{\infty,j}^i(x_1, \xi_1) - \int_{x_1}^{x_2} \frac{p_j^i(x, \xi_j^i(x), Q_\infty)}{\lambda_i(x)}\,dx. \tag{8.11}$$

Proof. Let $t_1 > 0$ and let $\gamma_j^i(x_1, \xi_1, t_1)$ be a characteristic curve for $\tilde{Q}$ (cf. γ_l^l in Theorem 8.3) passing through (x_1, ξ_1, t_1). The parametric description is (for conveniently chosen t_1)

$$\begin{aligned} \gamma_j^i = \{(x^i(t), \xi_j(t), t) | t_\alpha \leqq t \leqq t_\beta, x^i(t_1) = x_1, \\ \xi_i(t_1) = \xi_1, \quad \frac{dx^i(t)}{dt} = \lambda_i(x^i(t)), \quad \frac{d\xi_i(t)}{dt} = \lambda_i(\xi_j(t))\}. \end{aligned} \tag{8.12}$$

It is clear that in (8.12), ξ_j can be written in the form $\xi_j(x^i)$ and

$$\frac{d\xi_j(x^i)}{dx^i} = \frac{\lambda_j(\xi_j(x^i))}{\lambda_i(x^i)}.$$

The curve γ_j^i will pass through (x_2, ξ_2, t_2) for some value of t_2. We may without loss of generality assume $t_2 > t_1$. (Otherwise we merely reverse the roles of the indices 1 and 2.) From (5.2) and (8.2) we have

$$\tilde{q}_j^i(x_2, \xi_2, t_2) = \tilde{q}_j^i(x_1, \xi_1, t_1) - \int_{t_1}^{t_2} p_j^i(x^i(t), \xi_j(t), \tilde{Q})\,dt.$$

Since x^i is a monotone function of t, one can change variables to obtain

$$q_j^i(x_2, \xi_2, t_2) = q_j^i(x_1, \xi_1, t_1) - \int_{x_1}^{x_2} \frac{p_j^i(x^k, \xi_j(t(x^i)), \tilde{Q})}{\lambda_i(x^i)} \, dx^i.$$

Now we let t_1 (and consequently t_2 also) tend to $+\infty$. The uniform convergence of $\tilde{Q}(x, \xi, t)$ to $Q_\infty(x, \xi)$ implied by Theorem 8.3 gives (noting that $\xi_j(t(x^i))$ can be written $\xi_j^i(x^i)$)

$$q_{\infty,j}^i(x_2, \xi_2) = q_{\infty,j}^i(x_1, \xi_1) - \int_{x_1}^{x_2} \frac{p_j^i(x^i, \xi_j^i(x^i), Q_\infty)}{\lambda_i(x^i)} \, dx^i,$$

which is equivalent to (8.11), thus completing the proof of the theorem.

In the theory of the quadratic performance criterion for ordinary differential equations [7], [12], [4], the importance of Q_∞ lies in the fact that it can be used to obtain optimal fixed-form (non-time-varying) linear feedback control policies. In the present case this is also true. One can prove the following theorem in much the same way as the comparable theorem is proved in [12].

THEOREM 8.5. *Let the hypotheses of Theorem* 8.3 *hold. Then the fixed-form linear feedback control law*

$$u_*(t) \equiv -U^{-1}D^TA^+(1)\int_0^1 Q_\infty^+(1, \xi)w_*(\xi, t)\, d\xi \tag{8.13}$$

synthesizes the optimal control u_0 for the system (2.9), (2.14), (2.15) *relative to the cost functional*

$$J(w_0, u, \infty) = \int_0^\infty \left[\int_0^1 \int_0^1 w(x, t)^T W(x, \xi) w(\xi, t)\, dx\, d\xi + u(t)^T U u(t)\right] dt.$$

As an example of a system for which Theorems 8.3–8.5 are valid and useful, we study the equation of motion for a nonhomogeneous vibrating string:

$$\rho(x)\frac{\partial^2 \tilde{w}}{\partial t^2} - \frac{\partial}{\partial x}\left(p(x)\frac{\partial \tilde{w}}{\partial x}\right) = 0, \qquad \rho \in C[0, 1], \quad p \in C^1[0, 1]. \tag{8.14}$$

We impose boundary conditions

$$\alpha_0 \frac{\partial \tilde{w}}{\partial t}(0, t) + \beta_0 \frac{\partial \tilde{w}}{\partial x}(0, t) = 0, \qquad \frac{\alpha_0}{\beta_0} \leqq 0 \quad \text{or} \quad \beta_0 = 0, \tag{8.15}$$

$$\alpha_1 \frac{\partial \tilde{w}}{\partial t}(1, t) + \beta_1 \frac{\partial \tilde{w}}{\partial x}(1, t) = u(t), \qquad \frac{\alpha_1}{\beta_1} > 0. \tag{8.16}$$

The scalar function $u(t)$ is the control variable. The condition $\alpha_0/\beta_0 \leqq 0$ or $\beta_0 = 0$ ensures that no energy enters the system at $x = 0$, while the condition $\alpha_1/\beta_1 > 0$ implies that there definitely is a loss of energy (presumably due to friction) at $x = 1$ when $u(t) \equiv 0$. If (8.14) is reduced, in the usual way (see, e.g., [18]) to a first order system

$$\frac{\partial}{\partial t}\begin{pmatrix} w^- \\ w^+ \end{pmatrix} = \begin{pmatrix} -\sqrt{p(x)/\rho(x)} & 0 \\ 0 & \sqrt{p(x)/\rho(x)} \end{pmatrix}\begin{pmatrix} w^- \\ w^+ \end{pmatrix} + \begin{pmatrix} b_-^-(x) & b_+^-(x) \\ b_-^+(x) & b_+^+(x) \end{pmatrix}\begin{pmatrix} w^- \\ w^+ \end{pmatrix} \tag{8.17}$$

for appropriate $b_-^-(x), b_+^-(x), b_-^+(x)$ and $b_+^+(x)$, then the boundary conditions (8.15), (8.16) become

$$\left(\frac{\alpha_0}{\sqrt{2\rho(0)}} - \frac{\beta_0}{\sqrt{2p(0)}}\right) w^-(0,t) + \left(\frac{\alpha_0}{\sqrt{2\rho(0)}} + \frac{\beta_0}{\sqrt{2p(0)}}\right) w^+(0,t) = 0, \tag{8.18}$$

$$\left(\frac{\alpha_0}{\sqrt{2\rho(1)}} - \frac{\beta_0}{\sqrt{2p(1)}}\right) w^-(1,t) + \left(\frac{\alpha_0}{\sqrt{2\rho(1)}} + \frac{\beta_0}{\sqrt{2p(1)}}\right) w^+(1,t) = u(t). \tag{8.19}$$

It is shown in [18] that if (8.15) and (8.16) hold there are continuous functions $k^-(x), k^+(x)$ such that if one sets

$$\hat{u}(t) = \int_0^1 [k^-(x)w^-(x,t) + k^+(x)w^+(x,t)]\,dx, \tag{8.20}$$

the resulting closed loop system has a sort of damped periodic behavior. More precisely, there is a fixed $T_1 > 0$ such that solutions of this closed loop system obey

$$\begin{pmatrix} w^-(x, t + T_1) \\ w^+(x, t + T_1) \end{pmatrix} = \gamma \begin{pmatrix} w^-(x,t) \\ w^+(x,t) \end{pmatrix}, \qquad |\gamma| < 1. \tag{8.21}$$

Rewriting (8.20) in the form

$$\hat{u}(t) = \int_0^1 K(x)w(x,t)\,dx,$$

we can compute the cost associated with the control (8.20) in the same way as we did in Lemma 7.1. For the resulting cost matrix $\hat{Q}(x,\xi,t)$ and optimal cost matrix $Q(x,\xi,t)$, we have

$$\int_0^1 \int_0^1 w_0(x)^T(\hat{Q}(x,\xi,0) - Q(x,\xi,0))w_0(\xi)\,dx\,d\xi \geqq 0.$$

As noted at the beginning of this section, this can be done for any $t_1 > 0$, so we write

$$\int_0^1 \int_0^1 w_0(x)^T(\hat{Q}_{t_1}(x,\xi,0) - Q_{t_1}(x,\xi,0))w_0(\xi)\,dx\,d\xi \geqq 0.$$

Then we have (cf. (8.2))

$$\int_0^1 \int_0^1 w_0(x)^T(\hat{Q}_{t_1}(x,\xi,0) - \tilde{Q}(x,\xi,t_1))w_0(\xi)\,dx\,d\xi \geqq 0,$$

whence

$$\hat{q}_l^l(x,x,t_1) \leqq \hat{q}_{t_1 l}^l(x,x,0), \qquad t_1 \geqq 0, \qquad l = 1, 2, \cdots, n. \tag{8.22}$$

Now the "periodicity" (8.21) implies that

$$\hat{Q}_{t_1}(x,\xi,t_1 - kT_1) = (1 + \gamma^2 + \cdots + \gamma^{2(k-1)})\hat{Q}_{t_1}(x,\xi,t_1 - T_1)$$

for all positive integers k and for any $t_1 > 0$. Since the $\hat{q}_l^l$ increase as t decreases,

$$\hat{q}_{t_1 l}^l(x,x,0) \leqq \left(\sum_{k=0}^{\infty} \gamma^{2k}\right) \hat{q}_{t_1 l}^l(x,x,t_1 - T_1) \tag{8.23}$$

for all $t_1 > 0$. Since the system (4.13) is autonomous, the right-hand side of (8.23) is independent of t_1. Setting

$$K = \frac{1}{1-\gamma^2} \sup_{\substack{0 \leqq x \leqq 1 \\ 1 \leqq l \leqq n}} \{q^l_{t_1 l}(x, x, t_1 - T_1)\},$$

we have

$$\hat{q}^l_{t_1 l}(x, x, 0) \leqq K, \qquad t_1 \geqq 0.$$

Then, via (8.22), we have (8.6), so that the principle hypothesis of Theorem 8.3 is established. Thus Theorems 8.3, 8.4 and 8.5 do indeed apply to (8.14), (8.15) and (8.16).

It seems reasonable to conjecture that results comparable to the one obtained in this example should be fairly generally obtainable whenever the uncontrolled system is exponentially damped. However, obtaining a result as specific as that of Theorems 8.3 and 8.4, where $Q_\infty(x, \xi)$ is uniformly bounded, may well be fairly special. The difficulty lies in obtaining a dominating $\hat{Q}$ which is uniformly bounded. The boundedness of $\hat{Q}$ as constructed in Lemma 8.7 has been proved making essential use of (8.21). We hope to return to this problem in a later paper.

We have reason to believe that the feedback form (8.13) for the optimal control $u_*(t)$ may be of considerable practical interest. The matrix function $-U^{-1}D^TA^+(1)Q^+_\infty(1, x)$, $0 \leqq x \leqq 1$, provides a *feedback profile*, showing what "weight" is assigned to each state component, according to its spatial location. It is, of course, impossible to measure the complete system state $w_*(x, t)$, as would be required in order to compute the integral on the right-hand side of (8.13). But the feedback profile matrix would seem to provide an indication of advantageous sensor locations. Those locations correspond to points x in whose neighborhood a component of the feedback profile matrix assumes relatively large values. In this sense we are hopeful that the present theory may indicate a useful design technique.

The matrix function $Q_\infty(x, \xi)$ can be approximated numerically if one approximates solutions $Q(x, \xi, t)$ of (4.13) by a stable numerical technique in an interval $[0, t_1]$ for t_1 sufficiently large. But this is a method which normally converges rather slowly. It would be very helpful to have available numerical procedures for attacking the partial differential equation (8.9) directly. Since the equation (8.9) together with the boundary conditions (4.8), (4.9), (4.10) and (4.11) constitute a nonlinear *hyperbolic* boundary value problem, we are entering here upon rather unexplored mathematical territory.

REFERENCES

[1] F. L. Alvarado and R. Mukundan, *An optimization problem in distributed parameter systems*, Internat. J. Control, 9 (1969), pp. 665–677.

[2] E. A. Coddington and N. Levinson, *Theory of Ordinary Differential Equations*, McGraw-Hill, New York, 1955.

[3] R. Courant and D. Hilbert, *Methods of Mathematical Physics, Part II: Partial Differential Equations*, Interscience, New York, 1962.

[4] R. Datko, *A linear control problem in an abstract Hilbert space*, J. Differential Equations, 9 (1971), pp. 346–359.

[5] P. R. GARABEDIAN, *Partial Differential Equations*, John Wiley, New York, 1964.

[6] J. J. GRAINGER, *Boundary-value control of distributed systems characterized by hyperbolic differential equations*, Doctoral thesis, Dept. of Electrical Engineering, University of Wisconsin, Madison, 1967.

[7] R. E. KALMAN, *Contributions to the theory of optimal control*, Bol. Soc. Mat. Mexicana, 5 (1960), pp. 102–119.

[8] R. E. KALMAN AND R. S. BUCY, *New results in linear prediction and filtering theory*, J. Basic Engrg. Trans. ASME Ser. D, 83 (1961), pp. 95–100.

[9] M. KIM AND H. ERZBERGER, *On the design of optimum distributed parameter systems with boundary control functions*, IEEE Trans. Automatic Control, AC-12 (1967), pp. 22–28.

[10] M. KIM AND S. H. GAJWANI, *A variational approach to optimum distributed parameter systems*, Ibid., AC-13 (1968), pp. 191–193.

[11] J. L. LIONS, *Optimal Control of Systems Governed by Partial Differential Equations*, Springer-Verlag, New York, 1971.

[12] D. L. LUKES AND D. L. RUSSELL, *The quadratic criterion for distributed systems*, this Journal, 7 (1969), pp. 101–121.

[13] A. OLUBUMMO AND R. S. PHILLIPS, *Dissipative ordinary differential operators*, J. Math. Mech., 14 (1965), pp. 929–950.

[14] R. S. PHILLIPS, *Dissipative hyperbolic systems*, Trans. Amer. Math. Soc., 86 (1957), pp. 109–173.

[15] D. L. RUSSELL, *Optimal regulation of linear symmetric hyperbolic systems with finite dimensional controls*, this Journal, 4 (1966), pp. 276–294.

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

[17] ———, *On boundary-value controllability of linear symmetric hyperbolic systems*, Mathematical Theory of Control, Academic Press, New York, 1967.

[18] ———,*Control theory of hyperbolic equations related to certain questions in harmonic analysis and spectral theory*, J. Math. Anal. Appl., 40 (1972), pp. 336–368.

SIAM J. CONTROL
Vol. 11, No. 3, August 1973

THE RITZ-GALERKIN PROCEDURE FOR PARABOLIC CONTROL PROBLEMS*

R. S. McKNIGHT† AND W. E. BOSARGE, JR.‡

Abstract. A systematic numerical approximation approach to the problem of minimizing an integral cost functional subject to a parabolic partial differential equation constraint is discussed. This problem is viewed as a standard variational minimization problem subject to nonholonomic constraints and is treated using Lagrange multipliers.

A practical computational procedure for obtaining approximate solutions to this problem is developed. Error estimates for the control, state, costate, and cost functional are established under appropriate smoothness and boundedness conditions. Also, explicit order bounds are obtained for these estimates over generalized spline interpolating spaces.

1. Introduction. The theory involved in control problems for systems governed by partial differential equations has reached a high level of sophistication. Beginning with the fundamental achievements of Butkovskii [7] and Wang [21], many investigators have explored various aspects of the subject. Advances in computational methods, however, have lagged significantly behind theoretical developments. This lag is, in part, a consequence of the computational labyrinth which surrounds numerical approximation of partial differential equations.

In the past, numerical techniques have involved various space, space/time, and time discretizations. As exemplified in the work of Sage and Chandhuri [16], [17], and Sage [18], finite difference methods usually produce simple computational algorithms. However, these schemes are unwieldy for proving convergence and require an unrealistically small increment size to insure nominal accuracy. Also, for many partial differential equations, finite difference schemes often have stringent continuity requirements and only provide modest a priori order convergence.

More recently, a few authors (see [2], [9], [11]) have studied several approximating schemes which seem to offer an attractive alternative to finite difference methods. In hopes of stimulating further interest in approximating schemes, we pursue the application of techniques in the spirit of Ritz and Galerkin methods for directly discretizing distributed control problems. These techniques, elementary in principle and application, produce solutions of high order accuracy from relatively simple computational algorithms.

A variation of the basic Ritz method is introduced in two articles by Bosarge and Johnson [3], [4]. In these articles they treat systems governed by linear ordinary differential equations subject to a quadratic cost criterion. Smith, in [19], succeeds in extending this method to systems governed by linear parabolic partial differential equations. These investigators are among the first to seriously consider Ritz's method for control problems and to demonstrate convergence using specific approximating spaces. By employing piecewise polynomial inter-

* Received by the editors February 28, 1972.

† Marathon Oil Company, Denver Research Center, Littleton, Colorado 80120.

‡ Department of Mathematical Sciences, Rice University, Houston, Texas and FTI Corporation, Houston, Texas 77027.

polating functions, they prove that the suboptimal (approximate) control converges to the optimal value with high order accuracy. The order bounds they obtain are a substantial improvement over existing finite difference schemes of equivalent computational effort.

Although the above method, known as the Ritz–Trefftz procedure, proves very suitable for problems with linear equations, it cannot be extended directly to general nonlinear systems. In a sequel to these studies, Bosarge et al. in [6] succeed in constructing a procedure genetically similar to the Ritz–Trefftz approach but applicable to nonlinear systems. That approach, known as the Ritz–Galerkin procedure, also provides a convergence result and is readily adaptable to piecewise polynomial interpolating functions.

The extension of that Ritz–Galerkin technique to systems governed by nonlinear partial differential equations is discussed in this paper. Specifically, we treat the problem of minimizing an integral cost functional subject to an equality constraint which is described by a quasi-linear parabolic differential equation under specific boundary and initial conditions.

Our main objective is to derive a priori error estimates for the difference between the optimal and approximate solutions. To accomplish this, we actually transform the problem of bounding the error expressions into a question in approximation theory. These error bounds thus provide a convenient way of proving the convergence of the Ritz–Galerkin approximations to their optimal values.

If, for example, cubic splines are used in evaluating the error bounds, we prove that the approximate control, state, and costate converge to the optimal quantities, in norm, with order $O(h^3)$. Also, the approximate cost functional converges to the optimal quantity, in absolute value, with order $O(h^6)$.

2. Problem description. Consider the second order controllable dynamical system characterized by the equation

$$\Upsilon(u, v) \equiv -v_t + \nabla \cdot A(u, v, \nabla v) + B(u, v, \nabla v) = 0, \qquad (x, t) \in \Omega \times (0, T], \tag{2.1}$$

where Ω is a closed and bounded domain in R^n. Then $v = v(x, t)$ is the *state* of the system and belongs to V_T, the *state space*; and $u = u(x, t)$ is the *control* function and belongs to the space U, of *admissible controls*.

In addition, we assume that A is uniformly elliptic in the domain $\Omega \times (0, T]$; that is, for all real vectors $p = (p_1, p_2, \cdots, p_n)$ and any $v \in V_T$ and $u \in U$ there exist positive constants π and η such that

$$\pi \sum_{i=1}^{n} p_i^2 \leqq \sum_{i,j=1}^{n} \frac{\partial A_i}{\partial v_{x_j}} p_i p_j \leqq \eta \sum_{i=1}^{n} p_i^2 . \tag{2.2}$$

We also require that the behavior of (2.1) be well-posed (in the sense of Hadamard) with respect to the initial and boundary conditions

$$\begin{aligned} v(x, 0) &= v_0(x), && x \in \Omega, \\ v(x, t) &= 0, && x \in \partial\Omega, \quad t \in (0, T]. \end{aligned} \tag{2.3}$$

We constrain V_T and U to be real Hilbert spaces over $\Omega \times (0, T]$ such that

$$V_T \equiv \{W_2^{r,s}(\Omega \times (0, T]), R\},^1$$
$$U \equiv \{W_2^{r,s}(\Omega \times (0, T]), R\},^1$$

for integers $r, s \geqq 1$.

We restrict the operator $\Upsilon(u, v)$, for practical reasons, to map $U \times V_T$ into V_T. Along with this governing equation, a cost functional is prescribed by

$$J(u) = \int_0^T \int_\Omega G(u, v)\, dx\, dt, \tag{2.4}$$

which determines the "cost" of moving the system under the control action u. Here $G(u, v)$ is a real continuous function on $\Omega \times (0, T]$.

We now formally state the optimal control problem.

Problem 1. Determine the control $u^* \in U$ such that if v^* is the solution of (2.1) with $u = u^*$, then the cost functional $J(u^*)$ is minimum for all admissible $u \in U$.

Remarks. It is tacitly assumed that: (a) A, B, and G are continuously (Fréchet) differentiable with respect to their arguments, inducing sufficient continuity on a solution; (b) the control, u, is unconstrained in the space U; and (c) the terminal time, T, is fixed.

3. Lagrangian formulation. Problem 1 represents abstractly the "minimization of a functional over a Hilbert space subject to a nonholonomic equality constraint". The classical approach to solving problems of this form is with the use of Lagrange multipliers. For this purpose consider the function $\lambda(x, t)$ and define the Lagrangian, L, as

$$L(u, v; \lambda) = J(u) + \int_0^T \langle \lambda, \Upsilon(u, v)\rangle\, dt, \tag{3.1}$$

where $\lambda \in V_T^*$, the dual space of V_T. (The inner product notation $\langle\, ,\rangle$ is used extensively throughout this paper without specific reference to the underlying Hilbert space. In most cases it refers to $L_2(\Omega)$.)

In view of this definition, the theory of Lagrange multipliers provides the following alternate definition of Problem 1.

Problem 1′. Given the nonlinear equation (2.1) and cost functional (2.4), determine the functions u^*, v^*, and λ^* such that the Lagrangian is extremized; that is,

$$L(u^*, v^*; \lambda^*) = \max_{\lambda \in V_T} \min_{\substack{u \in U \\ v \in V_T}} L(u, v; \lambda). \tag{3.2}$$

Remarks. The function λ actually belongs to V_T^* but, since V_T is a Hilbert space, V_T^* is identical to V_T. At the optimum values, we have $L(u^*, v^*; \lambda^*) = J(u^*)$, which we henceforth abbreviate by J^*.

[1] $W_q^{m,n}(\Omega \times (0, T])$, for m and n integers, is the Banach space consisting of the elements of $L_{q,q}(\Omega \times (0, T])$ having generalized derivatives of the form $D_x^\alpha D_t^\beta$. The norm is given by

$$\|w\|_{W_q^{m,n}} \equiv \left(\sum_{\substack{|\alpha| \leq m \\ |\beta| \leq n}} \int_0^T \int_\Omega \|D_x^\alpha D_t^\beta w(x, t)\|_{L_{q,q}}^q\, dx\, dt \right)^{1/q}.$$

Since the functions u, v, and λ are unconstrained, an equivalent formulation of Problem 1′ is given by the following.

Problem 1″. Under the same assumptions as Problem 1′, determine the functions u^*, v^*, and λ^* such that

$$\delta_u L(u^*, v^*; \lambda^*)(\,\cdot\,) = 0, \tag{3.3}$$

$$\delta_v L(u^*, v^*; \lambda^*)(\,\cdot\,) = 0, \tag{3.4}$$

$$\delta_\lambda L(u^*, v^*; \lambda^*)(\,\cdot\,) = 0, \tag{3.5}$$

where the above symbols denote the partial Fréchet derivatives of L by u, v, and λ, respectively, and $(\,\cdot\,)$ refers to admissible variations in the spaces U, V_T, and V_T, respectively (denoted by Δ_u and Δ_v).

Assuming the existence and uniqueness of a solution to Problem 1, the equivalence of these three problems is easily deduced from standard arguments of Lagrange multiplier theory and the calculus of variations.

We now present the appropriate Lagrange multiplier theorem which provides a necessary condition for an extremum of L.

THEOREM 3.1. *Let u^* and v^* be optimal for Problem* 1. *Then there exists a function $\lambda \in V_T$ which satisfies the following relations*:

$$\Lambda(u, v; \lambda) \equiv G_u - A_u^T \nabla \lambda + B_u \lambda = 0, \tag{3.6}$$

$$\Gamma(u, v; \lambda) \equiv \lambda_t + G_v - A_v^T \nabla \lambda + B_v \lambda + \nabla \cdot (A_{v_x}^T \nabla \lambda - B_{v_x}^T \lambda) = 0, \tag{3.7}$$

subject to the terminal and boundary conditions

$$\begin{aligned} &\lambda(x, T) = 0, \qquad x \in \Omega, \\ &\lambda(x, t) = 0, \qquad x \in \partial\Omega, \qquad t \in (0, T], \end{aligned} \tag{3.8}$$

where A, B, and G, and their gradients are evaluated at (u^*, v^*).

Proof. The proof follows classical arguments in the calculus of variations (Tzafestas [20] presented a thorough explanation of the necessary conditions which is applicable to nonlinear equations).

We now turn our attention towards specifying a sufficient condition for a local extremum. To find such a condition, we must first introduce a canonical expression for the second derivative of the Lagrangian.

Define the Hamiltonian function, H, by

$$H(u, v; \lambda) = \int_\Omega G\,dx + \langle \lambda, \nabla \cdot A + B \rangle. \tag{3.9}$$

As a result of the problem's assumptions, H is continuous in u, v, λ, v_x, and λ_x and has continuous partial derivatives with respect to each of these arguments. Substituting the above definition into (3.1), we obtain

$$L(u, v; \lambda) = \int_0^T [H(u, v; \lambda) - \langle \lambda, v_t \rangle]\,dt. \tag{3.10}$$

The second derivative is formally given by

$$\delta^2 L(u, v; \lambda) \cdot [(\alpha_1, \beta_1), (\alpha_2, \beta_2)] = \int_0^T \delta^2 H(u, v; \lambda) \cdot [(\alpha_1, \beta_1), (\alpha_2, \beta_2)]\,dt,$$

where $\alpha_1, \alpha_2 \in \Delta_u$ and $\beta_1, \beta_2 \in \Delta_v$. The operator $\delta^2 H$ is a symmetric bilinear form which is evaluated at the triple (u, v, λ). Since we have imbedded our problem in a Hilbert space, the operator $\delta^2 H$ may be represented by a matrix form using inner product notation. Set

$$H''(u, v; \lambda) = \begin{bmatrix} H_{uu}(u, v; \lambda) & H_{uv}(u, v; \lambda) \\ H_{vu}(u, v; \lambda) & H_{vv}(u, v; \lambda) \end{bmatrix},$$

where, for example, H_{vu} denotes the second Fréchet derivative of H with respect to v and u (this matrix is commonly called the Hessian). Introducing this form, we obtain

$$\delta^2 L(u, v; \lambda) \cdot [(\alpha_1, \beta_1), (\alpha_2, \beta_2)] = \int_0^T \left\langle H''(u, v; \lambda)\begin{pmatrix}\alpha_1 \\ \beta_1\end{pmatrix}, \begin{pmatrix}\alpha_2 \\ \beta_2\end{pmatrix}\right\rangle dt. \tag{3.11}$$

From the classical theory of the calculus of variations, we know that the "strong positivity" of the second variation is a sufficient condition for a local minimum (e.g., see [12]). Following the example set for systems governed by ordinary differential equations (see Lee and Markus [12]), we enforce an appropriate positivity condition given by

$$\int_0^T \left\langle H''(u, v; \lambda^*)\begin{pmatrix}\alpha \\ \beta\end{pmatrix}, \begin{pmatrix}\alpha \\ \beta\end{pmatrix}\right\rangle dt \geqq \sigma\|\alpha\|^2_{L_{2,2}}, \tag{3.12}$$

where $\alpha \in \Delta_u$, $\beta \in \Delta_v$, and $u \in N(u^*)$, $v \in N(v^*)$ (neighborhoods of the optimum) and σ is a positive constant.

It is now a simple matter to demonstrate the saddle point behavior of the Lagrangian.

THEOREM 3.2. *Suppose condition* (3.12) *holds at a local extremum. Then the Lagrangian possesses a degenerate saddle point at* (u^*, v^*, λ^*); *that is,*

$$L(u^*, v^*; \lambda) = L(u^*, v^*; \lambda^*) \leqq L(u, v; \lambda^*), \tag{3.13}$$

where $u \in N(u^*)$, $v \in N(v^*)$, *and* $\lambda \in N(\lambda^*)$.

Proof. The left equality follows directly from noting that the pair (u^*, v^*) must satisfy (2.1) exactly. The right-hand inequality is a result of the strong positivity condition. Expressing L in a Taylor expansion (see Dieudonné [10]) about the optimum, we have

$$\begin{aligned} L(u, v; \lambda^*) = {} & L(u^*, v^*; \lambda^*) + \delta L(u^*, v^*; \lambda^*) \cdot [(u - u^*), (v - v^*)] \\ & + \int_0^1 (1 - \theta) \int_0^T \left\langle H''(\underset{\sim}{u}, \underset{\sim}{v}; \lambda^*)\begin{pmatrix}u - u^* \\ v - v^*\end{pmatrix}, \begin{pmatrix}u - u^* \\ v - v^*\end{pmatrix}\right\rangle dt\, d\theta, \end{aligned}$$

where $\underset{\sim}{u} = \theta u + (1 - \theta)u^*$, $\underset{\sim}{v} = \theta v + (1 - \theta)v^*$, and $0 < \theta < 1$. The second term on the right-hand side is zero by virtue of the necessary conditions. Applying condition (3.12), we have

$$L(u, v; \lambda^*) \geqq L(u^*, v^*; \lambda^*) + \sigma\|u - u^*\|^2_{L_{2,2}}.$$

Equation (3.13) is now apparent since the norm is positive.

This local positivity condition along with the regularity of the point (u^*, v^*) ensures that the saddle-point condition is also sufficient for optimality (see Luenberger [14]).

COROLLARY 3.3. *A sufficient condition for the Lagrangian, $L(u, v; \lambda^*)$, to have a local minimum at (u^*, v^*), given that the first variation vanishes there, is that the second variation satisfy a strong positivity condition in some neighborhood of (u^*, v^*).*

Proof. The proof is a direct consequence of the theorem.

We conclude this section by summarizing the key assumptions.

(A1) Equation (2.1) is well-posed with respect to the state space, V_T, and the set of admissible controls U.

(A2) The Lagrangian is extremized at the triple (u^*, v^*, λ^*) and satisfies the standard necessary conditions

$$\Lambda(u^*, v^*; \lambda^*) = 0, \tag{3.14}$$

$$\Gamma(u^*, v^*; \lambda^*) = 0, \tag{3.15}$$

$$\lambda^*(x, T) = 0, \quad x \in \Omega, \qquad \lambda^*(x, t) = 0, \quad x \in \partial\Omega, \quad t \in (0, T],$$

$$\Upsilon(u^*, v^*) = 0, \tag{3.16}$$

$$v^*(x, 0) = v_0(x), \quad x \in \Omega, \qquad v^*(x, t) = 0, \quad x \in \partial\Omega, \quad t \in (0, T].$$

(A3) The second variation of the Lagrangian is strongly positive in some bounded convex neighborhood of (u^*, v^*, λ^*); that is, there exist neighborhoods $N(u^*) \subset U$, $N(v^*) \subset V_T$ and $N(\lambda^*) \subset V_T$ such that the operator H'' satisfies

$$\int_0^T \left\langle H''(\underset{\sim}{u}, \underset{\sim}{v}; \lambda)\begin{pmatrix} u - \underset{\sim}{u} \\ v - \underset{\sim}{v} \end{pmatrix}, \begin{pmatrix} u - \underset{\sim}{u} \\ v - \underset{\sim}{v} \end{pmatrix} \right\rangle dt \geqq \sigma \| u - \underset{\sim}{u} \|^2_{L_{2,2}}, \tag{3.17}$$

where $\underset{\sim}{u} \in N(u^*)$, $\underset{\sim}{v} \in N(v^*)$, $\lambda \in N(\lambda^*)$, and $\sigma > 0$.

4. Finite-dimensional subspaces. Suppose that the optimal quantities u^*, v^*, and λ^* can be approximated by functions which represent admissible comparison functions. These comparison functions are such that they are expressible as finite linear combinations of a set of basis functions. Let the sets D_u, D_v, and D_λ be admissible solutions for the functions u, v, and λ, respectively. Then, define the sets C, S, and M as

$$C = \{u | u \in U \cap N(u^*) \cap D_u\},$$

$$S = \{v | v \in V_T \cap N(v^*) \cap D_v\},$$

$$M = \{\lambda | \lambda \in V_T \cap N(\lambda^*) \cap D_\lambda\},$$

where $N(u^*)$, $N(v^*)$, and $N(\lambda^*)$ are the neighborhoods prescribed in Theorem 3.2. Next, let $C_N(w)$, $S_N(w)$ and $M_N(w)$ be finite-dimensional subspaces of C, S, and M, respectively, which are spanned by a set of basis functions, $\{w_i(x, t)\}$, $i = 1, \cdots, l$, where $l = \prod_{k=1}^{n+1} n_k$, $N = \{n_k | k = 1, \cdots, n + 1\}$, and the n_k's are the number of basis functions associated with the kth coordinate. (This definition avoids multiple indexing and tensor products—both of which would obscure the conceptual simplicity of this approach.) Symbolically this means

$$C_N(w) = \left\{ u_N | u_N(x,t) = \sum_{i=1}^{l} a_i w_i(x,t), a_i \in R \right\},$$

$$S_N(w) = \left\{ v_N | v_N(x,t) = \sum_{i=1}^{l} b_i w_i(x,t), b_i \in R \right\},$$

$$M_N(w) = \left\{ \lambda_N | \lambda_N(x,t) = \sum_{i=1}^{l} c_i w_i(x,t), c_i \in R \right\}.$$

Finally, we choose the basis functions so that the sets

$$C_N = \{u_N | u_N \in C_N(w) \cap \bar{C}\},$$
$$S_N = \{v_N | v_N \in S_N(w) \cap \bar{S}\},$$
$$M_N = \{\lambda_N | \lambda_N \in M_N(w) \cap \bar{M}\}$$

are nonvoid (the bar notation above C, S, and M indicates the closures of these sets). The finite-dimensional spaces C_N, S_N, and M_N are thus closed and bounded sets which contain the optimum. Therefore, unique best approximations to the optimum quantities u^*, v^*, and λ^* exist in these spaces (see Cheney [8]). Define

$$\|\hat{u} - u^*\| \equiv \inf_{u_N \in C_N} \|u_N - u^*\|,$$
$$\|\hat{v} - v^*\| \equiv \inf_{v_N \in S_N} \|v_N - v^*\|,$$
$$\|\hat{\lambda} - \lambda^*\| \equiv \inf_{\lambda_N \in M_N} \|\lambda_N - \lambda^*\|,$$

where $\hat{u}$, $\hat{v}$, $\hat{\lambda}$ are the "best norm-approximations" to u^*, v^*, and λ^*, respectively. (Specific norms will be designated at the appropriate times in our development.)

5. The Ritz–Galerkin approximation problem. Upon selection of the basis functions, the inner products and time integrals in the approximation problems may be carried out. What remains is an algebraic system which is characterized by the residual (expansion) coefficients (that is, a_i, b_i, c_i).

In this section we abstractly define the residuals for the full approximation procedure. First, the residual for Problem 1 is given by the following definition.

Problem 2 (Ritz residual problem). Under assumptions (A1)–(A3), determine the control $\bar{u} \in C_N$ such that

$$J(\bar{u}) = \min_{u \in C_N} J(u) \tag{5.1}$$

subject to the constraints,

$$\int_0^T \langle w_i(\cdot, t), \Upsilon(u, v)\rangle \, dt = 0, \tag{5.2}$$

$$\langle w_i(\cdot, 0), v(\cdot, 0) - v_0\rangle = 0 \tag{5.3}$$

for $i = 1, \cdots, l$, and $v \in S_N$.

Directly applying the Ritz method to Problem 1′, we obtain the following result.

Problem 2′. Given the nonlinear system (2.1) and cost functional (2.4), determine the functions $\bar{u}$, $\bar{v}$, and $\bar{\lambda}$ such that the Lagrangian is extremized; that is,

$$L(\bar{u}, \bar{v}; \bar{\lambda}) = \max_{\lambda \in M_N} \min_{\substack{u \in C_N \\ v \in S_N}} L(u, v; \lambda). \tag{5.4}$$

Remarks. At this suboptimum point, $(\bar{u}, \bar{v}, \bar{\lambda})$, we note that $L(\bar{u}, \bar{v}; \bar{\lambda}) = J(\bar{u})$. For simplicity, we designate $J_N = J(\bar{u})$.

Since Problem 1″ is essentially in functional equation form, we apply Galerkin's method to obtain a companion problem to the one above.

Problem 2″ (Galerkin residual problem). Subject to assumptions (A1)–(A3), determine the functions $\bar{u}$, $\bar{v}$, and $\bar{\lambda}$ which satisfy equations (5.2) and (5.3), and

$$\int_0^T \langle w_i(\cdot, t), \Lambda(u, v; \lambda)\rangle \, dt = 0, \tag{5.5}$$

$$\int_0^T \langle w_i(\cdot, t), \Gamma(u, v; \lambda)\rangle \, dt = 0, \tag{5.6}$$

$$\langle w_i(\cdot, T), \lambda(\cdot, T)\rangle = 0, \tag{5.7}$$

for $i = 1, \cdots, l$, and $(u, v, \lambda) \in P_N (= C_N \times S_N \times M_N)$.

Since the above problems yield identical algebraic equations to solve, we corporately refer to them as the *Ritz–Galerkin residual problem.* This problem definition leads us to the finite-dimensional analogue of Theorem 3.2.

THEOREM 5.1. *Let* $(\bar{u}, \bar{v}, \bar{\lambda})$ *be the solution of the Ritz–Galerkin residual problem for an arbitrary* (*but fixed*) *subspace* P_N. *Then the Lagrangian has a degenerate saddle-point at* $(\bar{u}, \bar{v}, \bar{\lambda})$; *that is,*

$$L(\bar{u}, \bar{v}; \lambda) = L(\bar{u}, \bar{v}; \bar{\lambda}) \leqq L(u, v; \bar{\lambda}) \tag{5.8}$$

for any $u \in C_N$, $v \in S_N$, *and* $\lambda \in M_N$.

Proof. The proof is almost identical to that of Theorem 3.2.

6. Energy and variational inequalities. In preparation for the main results, we first present several lemmas which are essential to the development. The first two are consequences of energy inequalities and the other two yield bounds for the first variations of the Lagrangian. The letters C and K are used as generic constants and are not necessarily the same upon each occurrence.

LEMMA 6.1. *Let* $(\bar{u}, \bar{v})$ *be the solution of the Ritz–Galerkin residual problem. Then there exists a positive constant* C *such that*

$$\|\bar{v} - v^*\|_{W_2^{1,0}} \leqq C[\|\bar{u} - u^*\|_{L_{2,2}} + \|v - v^*\|_{W_2^{1,1}}], \tag{6.1}$$

where $v \in S_N$.

Proof. The proof uses arguments similar to those in [11] and is therefore omitted.

LEMMA 6.2. *Let $(\bar{u}, \bar{v}, \bar{\lambda})$ be the solution of the Ritz–Galerkin residual problem. Then there exists a positive constant C such that*

$$\|\bar{\lambda} - \lambda^*\|_{W_2^{1,0}} \leqq C[\|\bar{u} - u^*\|_{L_{2,2}} + \|\bar{v} - v^*\|_{W_2^{1,0}} + \|\lambda - \lambda^*\|_{W_2^{1,1}}], \tag{6.2}$$

where $\lambda \in M_N$.

Proof. The proof of (6.2) parallels that given in the previous lemma.

LEMMA 6.3. *Let (u, v, λ) be an arbitrary triple in P_N, and (u^*, v^*, λ^*) be the solution of Problem 1′. Then there exists a positive constant K such that*

$$\delta_u L(u^*, v^*; \lambda) \cdot (\alpha) \leqq K\|\lambda - \lambda^*\|_{W_2^{1,0}}\|\alpha\|_{L_{2,2}}, \tag{6.3}$$

where $\alpha \in \Delta_u$.

Proof. The Fréchet differential of L with respect to u with increment α is given by

$$\delta_u L(u, v; \lambda) \cdot (\alpha) = \int_0^T \langle G_u - A_u^T \nabla\lambda + B_u\lambda, \alpha\rangle\, dt.$$

From this equation we have,

$$\begin{aligned}\delta_u L(u^*, v^*; \lambda) \cdot (\alpha) &= \int_0^T \langle -A_u^{*T}\nabla(\lambda - \lambda^*) + B_u^*(\lambda - \lambda^*), \alpha\rangle\, dt \\ &\leqq k_1\|\nabla(\lambda - \lambda^*)\|_{L_{2,2}}\|\alpha\|_{L_{2,2}} + k_2\|\lambda - \lambda^*\|_{L_{2,2}}\|\alpha\|_{L_{2,2}} \\ &\leqq K(\|\nabla(\lambda - \lambda^*)\|_{L_{2,2}} + \|\lambda - \lambda^*\|_{L_{2,2}})\|\alpha\|_{L_{2,2}}.\end{aligned}$$

The term in parentheses is the norm in $W_2^{1,0}$. This completes the proof.

LEMMA 6.4. *Let (u, v, λ) be an arbitrary triple in P_N, and (u^*, v^*, λ^*) the solution of Problem 1′. Then there exist positive constants K and K' such that*

$$\delta_v L(u^*, v^*; \lambda) \cdot (\beta) \leqq K\|\lambda - \lambda^*\|_{W_2^{1,0}}\|\beta\|_{W_2^{1,1}} \tag{6.4}$$

and

$$\delta_v L(u^*, v^*; \lambda) \cdot (\beta) \leqq K'\|\lambda - \lambda^*\|_{W_2^{1,1}}\|\beta\|_{W_2^{1,0}}, \tag{6.5}$$

where $\beta \in \Delta_v$.

Proof. The Fréchet differential of L with respect to v with increment β is

$$\delta_v L(u, v; \lambda) \cdot (\beta) = \int_0^T \langle \lambda_t + G_v - A_v^T\nabla\lambda + B_v\lambda, \beta\rangle\, dt,$$

$$\begin{aligned}\delta_v L(u^*, v^*; \lambda) \cdot (\beta) = \int_0^T &[\langle(\lambda - \lambda^*)_t - A_v^{*T}\nabla(\lambda - \lambda^*) + B_v^*(\lambda - \lambda^*), \beta\rangle \\ &- \langle A_{v_x}^{*T}\nabla(\lambda - \lambda^*) - B_{v_x}^{*T}(\lambda - \lambda^*), \nabla\beta\rangle]\, dt.\end{aligned}$$

After applying the Cauchy–Schwarz inequality, we see that

$$\begin{aligned}\delta_v L(u^*, v^*; \lambda) \cdot (\beta) \leqq\ & k_1(\|(\lambda - \lambda^*)_t\|_{L_{2,2}} + \|\nabla(\lambda - \lambda^*)\|_{L_{2,2}} + \|\lambda - \lambda^*\|_{L_{2,2}})\|\beta\|_{L_{2,2}} \\ &+ k_2(\|\lambda - \lambda^*\|_{L_{2,2}} + \|\nabla(\lambda - \lambda^*)\|_{L_{2,2}})\|\nabla\beta\|_{L_{2,2}},\end{aligned}$$

which produces equation (6.5) upon selection of the dominant constant. Equation (6.4) is obtained by first integrating the term with $(\lambda - \lambda^*)_t$ by parts and then simplifying.

7. A priori error estimates. We are now ready to present the bounds for the errors committed in approximating the control.

THEOREM 7.1. *Let $(\bar{u}, \bar{v}, \bar{\lambda})$ be the solution of the Ritz–Galerkin residual problem in the space P_N. Then there exists a constant C such that*

$$\|\bar{u} - u^*\|_{L_{2,2}} \leqq C[\|u - u^*\|_{L_{2,2}} + \|v - v^*\|_{W_2^{1,1}} + \|\lambda - \lambda^*\|_{W_2^{1,1}}], \tag{7.1}$$

where $(u, v, \lambda) \in P_N$ is an arbitrary (but fixed) element.

Proof. By recalling the result of Theorem 5.1, we have $J_N \leqq L(u, v; \bar{\lambda})$. Expanding by Taylor's formula, we obtain

$$\begin{aligned} J_N \leqq{} & L(u^*, v^*; \bar{\lambda}) + \delta_u L(u^*, v^*; \bar{\lambda}) \cdot (u - u^*) + \delta_v L(u^*, v^*; \bar{\lambda}) \cdot (v - v^*) \\ & + \int_0^1 (1 - \theta) \int_0^T \left\langle H''(\underset{\sim}{u}, \underset{\sim}{v}, \bar{\lambda}) \begin{pmatrix} u - u^* \\ v - v^* \end{pmatrix}, \begin{pmatrix} u - u^* \\ v - v^* \end{pmatrix} \right\rangle dt\, d\theta, \end{aligned} \tag{7.2}$$

where $\underset{\sim}{u} = \theta u + (1 - \theta)u^*$ and $\underset{\sim}{v} = \theta v + (1 - \theta)v^*$. Since $\bar{\lambda} \in M_N$, we can apply Lemmas 6.3 and 6.4 to deduce

$$\begin{aligned} J_N \leqq{} & J^* + K_1\|\bar{\lambda} - \lambda^*\|_{W_2^{1,0}}\|u - u^*\|_{L_{2,2}} + K_2\|\bar{\lambda} - \lambda^*\|_{W_2^{1,0}}\|v - v^*\|_{W_2^{1,1}} \\ & + K_3(\|u - u^*\|^2_{L_{2,2}} + \|v - v^*\|^2_{W_2^{1,0}}). \end{aligned}$$

Applying Lemma 6.2 to the terms in $\bar{\lambda} - \lambda^*$, and Lemma 6.1 to the subsequent terms in $\bar{v} - v^*$, we conclude that,

$$\begin{aligned} J_N \leqq{} & J^* + K'\|\bar{u} - u^*\|_{L_{2,2}}[\|u - u^*\|_{L_{2,2}} + \|v - v^*\|_{W_2^{1,1}}] \\ & + K''[(\|u - u^*\|_{L_{2,2}} + \|v - v^*\|_{W_2^{1,1}})(\|v - v^*\|_{W_2^{1,1}} + \|\lambda - \lambda^*\|_{W_2^{1,1}}) \\ & + \|u - u^*\|^2_{L_{2,2}} + \|v - v^*\|^2_{W_2^{1,0}}]. \end{aligned} \tag{7.3}$$

A corresponding lower bound for J_N can be obtained by using the equality part of (5.8). By applying Taylor's formula to $J_N = L(\bar{u}, \bar{v}; \bar{\lambda})$ about (u^*, v^*, λ), we have

$$\begin{aligned} J_N ={} & L(u^*, v^*; \lambda) + \delta_u L(u^*, v^*; \lambda) \cdot (\bar{u} - u^*) + \delta_v L(u^*, v^*; \lambda) \cdot (\bar{v} - v^*) \\ & + \int_0^1 (1 - \theta) \int_0^T \left\langle H''(\underset{\sim}{u}, \underset{\sim}{v}; \lambda) \begin{pmatrix} \bar{u} - u^* \\ \bar{v} - v^* \end{pmatrix}, \begin{pmatrix} \bar{u} - u^* \\ \bar{v} - v^* \end{pmatrix} \right\rangle dt\, d\theta, \end{aligned}$$

where $\underset{\sim}{u} = \theta\bar{u} + (1 - \theta)u^*$, $\underset{\sim}{v} = \theta\bar{v} + (1 - \theta)v^*$, and $0 < \theta < 1$. Employing Lemmas 6.3 and 6.4 and, this time, the strong positivity condition, we obtain

$$\begin{aligned} J_N \geqq{} & J^* - K_1\|\lambda - \lambda^*\|_{W_2^{1,0}}\|\bar{u} - u^*\|_{L_{2,2}} - K_2'\|\lambda - \lambda^*\|_{W_2^{1,1}}\|\bar{v} - v^*\|_{W_2^{1,0}} \\ & + \sigma\|\bar{u} - u^*\|^2_{L_{2,2}}. \end{aligned} \tag{7.4}$$

Using Lemma 6.1 on the term with K_2', we see that

$$\begin{aligned} J_N \geqq{} & J^* - K_1'\|\bar{u} - u^*\|_{L_{2,2}}(\|\lambda - \lambda^*\|_{W_2^{1,0}} + \|\lambda - \lambda^*\|_{W_2^{1,1}}) \\ & - K_2'\|\lambda - \lambda^*\|_{W_2^{1,1}}\|v - v^*\|_{W_2^{1,1}} + \sigma\|\bar{u} - u^*\|^2_{L_{2,2}}. \end{aligned}$$

Combining (7.3) and (7.4), we arrive at (after some inequality manipulation)

$$\begin{aligned}\bar{K}[\|u - u^*\|^2_{L_{2,2}} + \|\lambda - \lambda^*\|^2_{W_2^{1,1}} + \|v - v^*\|^2_{W_2^{1,1}}]\\ \geqq -K_4\|\bar{u} - u^*\|_{L_{2,2}}[\|u - u^*\|_{L_{2,2}} + \|v - v^*\|_{W_2^{1,1}} + \|\lambda - \lambda^*\|_{W_2^{1,1}}]\\ + \sigma\|\bar{u} - u^*\|^2_{L_{2,2}}.\end{aligned}$$

Completing the indicated square, we arrive at

$$\begin{aligned}&K'[\|u - u^*\|_{L_{2,2}} + \|v - v^*\|_{W_2^{1,1}} + \|\lambda - \lambda^*\|_{W_2^{1,1}}]^2\\ &\quad\geqq \sigma\{\|\bar{u} - u^*\|_{L_{2,2}} - (K_4/2\sigma)[\|u - u^*\|_{L_{2,2}}\| + \|v - v^*\|_{W_2^{1,1}} + \|\lambda - \lambda^*\|_{W_2^{1,1}}]\}^2.\end{aligned}$$

Taking the square root of the above inequality, we obtain the desired result.

From this fundamental error bound, it is a simple matter to show that similar bounds hold for the state and costate. These results are summarized in the following corollary.

COROLLARY 7.2. *Let the conditions of Theorem* 7.1 *hold. Then there exist positive constants C' and C'' such that*

$$\|\bar{v} - v^*\|_{W_2^{1,0}} \leqq C'\varepsilon_N, \tag{7.5}$$

$$\|\bar{\lambda} - \lambda^*\|_{W_2^{1,0}} \leqq C''\varepsilon_N, \tag{7.6}$$

where

$$\varepsilon_N = \|u - u^*\|_{L_{2,2}} + \|v - v^*\|_{W_2^{1,1}} + \|\lambda - \lambda^*\|_{W_2^{1,1}} \tag{7.7}$$

and $(u, v, \lambda) \in P_N$.

Proof. The proof of inequality (7.5) is obtained by direct substitution of (7.1) into (6.1). Likewise, inequality (7.6) is obtained by substituting (7.5) and (7.1) into (6.2).

THEOREM 7.3. *Let $(\bar{u}, \bar{v}, \bar{\lambda})$ be the solution of the Ritz–Galerkin residual problem in the space P_N. Then there exist positive constants K and K' such that*

$$J^* - K\eta_N^2 \leqq J_N \leqq J^* + K'\varepsilon_N^2, \tag{7.8}$$

where

$$\eta_N = \|\lambda - \lambda^*\|_{W_2^{1,1}} + \|v - v^*\|_{W_2^{1,1}}, \tag{7.9}$$

ε_N *is given by* (7.7), *and* $(u, v, \lambda) \in P_N$.

Proof. The right-hand inequality is obtained by direct substitution of (7.1) into (7.3) and by applying several norm inequalities. The left-hand inequality is deduced from (7.4). If we complete the square, we have

$$\begin{aligned}J_N \geqq J^* - \underline{K}[\|\lambda - \lambda^*\|_{W_2^{1,0}} + \|\lambda - \lambda^*\|^2_{W_2^{1,1}}] - \frac{K_2'}{2}\|v - v^*\|^2_{W_2^{1,1}}\\ + \sigma[\|\bar{u} - u^*\|_{L_{2,2}} - \underline{k}(\|\lambda - \lambda^*\|_{W_2^{1,0}} + \|\lambda - \lambda^*\|_{W_2^{1,1}})]^2.\end{aligned}$$

We arrive at the desired result by dropping the squared term since it is always positive.

8. Interpretation of the estimates. For a specific application one usually knows something about the structure of the solution and is able to make use of this knowledge to obtain error estimates. However, if only the solution's smoothness is assumed, we can approximate no more accurately than with piecewise polynomial functions.

In addition, several practical reasons exist for selecting these functions as a basis set. Computational experience has shown that not only do they converge rapidly to the exact solution on a coarse mesh, but also, are relatively insensitive to roundoff errors. Also, for piecewise polynomial functions defined using a patch basis, the algebraic system's sparseness properties are important.

Of particular significance are the spaces of spline and Hermite interpolates which exhibit attractive approximating properties and yield high-order error estimates for the solution to our problems.

We first present the results for single-dimensioned space and then extend them to multidimensional spaces.

DEFINITION 8.1 Let F_m^s ($s, m \geqq 1$, both integers) be an m-dimensional space of piecewise polynomials of fixed order $s - 1$ in a single variable. The F_m^s is characterized by the properties:

(P1) There exists a linear operator L_m such that

$$L_m : C^s(I) \to F_m^s.$$

(P2) For all $f \in C^s(I)$,

$$\|L_m f - f\|_{W_2^p} = O(m^{-s+p}), \qquad 0 \leqq p \leqq s.$$

As noted earlier, two spaces which exhibit this property are the spline and Hermite interpolating spaces.

Next, consider the closed rectangle on ψ in R^n defined by,

$$\psi = \{(I_1, I_2, \cdots, I_n) | a_i \leqq I_i \leqq b_i, i = 1, \cdots, n\},$$

where each $a_i, b_i \in R$. We then introduce a mesh on ψ as follows:

$$\pi_i : a_i = x_0^i < x_1^i < \cdots < x_{r_i}^i = b_i, \qquad i = 1, \cdots, n.$$

Then the composite mesh is given by,

$$\bar{\pi} = \bigtimes_{i=1}^{n} \pi_i .$$

Denote the largest interval in each mesh π_i by $h_i = \|\pi_i\|$; i.e.,

$$h_i = \max_{j=1,\cdots,r_i} |x_{j+1}^i - x_j^i|, \qquad i = 1, \cdots, n.$$

DEFINITION 8.2. Let F_M^S be the space of multivariate piecewise polynomials of dimension m over the partition $\bar{\pi}$ denoted by the tensor product

$$F_M^S = \bigtimes_{i=1}^{n} F_{m_i}^{s_i},$$

where $M = \{m_i | i = 1, \cdots, n\}$, $S = \{s_i | i = 1, \cdots, n\}$.

Reinterpretation of the two salient properties gives:
(P1′) There exists an M-linear operator L_M such that

$$L_M : C^S(\psi) \to F^S_M.$$

(P2′) For all $f \in C^S(\psi)$,

$$\|L_M f - f\|_{W_2^r} = \sum_{i=1}^{n} O(h_i^{s_i - r}).$$

To simplify notation, we define the maximum order symbology.

DEFINITION 8.3. Let h and S be sets such that $h = \{h_1, h_2, \cdots, h_n\}$ and $S = \{s_1, s_2, \cdots, s_m\}$. Then define $O(h^S)$ by

$$O(h^S) \equiv \sum_{j=1}^{n} O(h_j^{s_j}),$$

which as a consequence yields

$$O(h^{S-r}) = \max_{i=1,\cdots,n} \{O(h_i^{s_i - r})\},$$

where $r < \min_i \{s_i\}$.

For the case of splines and Hermites, there is a simple relationship between the parameters h_i and the dimensional parameters m_i. In spline spaces of cubic order, $S^4_{m_i}$, we have $m_i = r_i + 3$, where r_i is equal to the measure of the interval divided by h_i. For cubic Hermites $H^4_{m_i}$, we have $m_i = 2r_i + 3$, where r_i is the same as before. For this basic relationship between h and m we have the following interpolation result.

LEMMA 8.4. *Let $\hat{u}$, $\hat{v}$, and $\hat{\lambda}$ be the spline (Hermite) interpolates of u^*, v^* and λ^* respectively. Then the following order bounds are valid for $S = \{s_1, s_2, \cdots, s_n\}$, $s_j \geqq 1$, $r \leqq s_j$:*

$$\|\hat{u} - u^*\|_{W_2^r} \leqq O(h^{S-r}), \tag{8.1}$$

$$\|\hat{v} - v^*\|_{W_2^r} \leqq O(h^{S-r}), \tag{8.2}$$

$$\|\hat{\lambda} - \lambda^*\|_{W_2^r} \leqq O(h^{S-r}), \tag{8.3}$$

where $\hat{u} \in C^S_M$, $\hat{v} \in S^S_m$, and $\hat{\lambda} \in M^C_M$.

We now interpret each of the error estimates derived in the previous section with respect to a general spline interpolating space.

THEOREM 8.5. *Let $(\bar{u}, \bar{v}, \bar{\lambda})$ be the solution of the Ritz–Galerkin residual problem over the spline space P^S_N. Then we can conclude that*

$$\|\bar{u} - u^*\|_{L_{2,2}} \leqq O(h^{S-1}), \tag{8.4}$$

$$\|\bar{v} - v^*\|_{W_2^{1,0}} \leqq O(h^{S-1}), \tag{8.5}$$

$$\|\bar{\lambda} - \lambda^*\|_{W_2^{1,0}} \leqq O(h^{S-1}), \tag{8.6}$$

where $S = \{s_i | i = 1, \cdots, n + 1\}$ is the set of degrees of the approximating polynomials and h is the mesh norm.

Proof. These results are obtained by substituting the appropriate order bounds from (8.1)–(8.3) into inequalities (7.1), (7.5), and (7.6). That is, we choose the triple $(u, v, \lambda) \in P_N^S$ to be the spline interpolates of (u^*, v^*, λ^*).

THEOREM 8.6. *Let $(\bar{u}, \bar{v}, \bar{\lambda})$ be the solution of the Ritz–Galerkin residual problem over the spline space P_N^S. Then*

$$J^* - O(h^{2(S-1)}) \leqq J_N \leqq J^* + O(h^{2(S-1)}). \tag{8.7}$$

Proof. As above, the result is established by direct substitution of the order bounds (8.1)–(8.3) into inequality (7.8).

Remark. Observe that in this procedure the mesh norm h is determined from a total discretization of both the state space, Ω, and a finite time interval, $(0, T]$. Consequently, the set S possesses the degrees of the approximating polynomials in both x and t. Thus, for this full procedure, there is little distinction between the spatial and time variables.

9. Conclusions. In this paper we presented a practical and theoretically well-founded computational procedure for determining suboptimal controls for distributed systems. A priori error estimates were derived for approximate solutions. These error estimates were obtained in terms of fixed elements in some arbitrary finite-dimensional approximating space. Complete generality of these bounds was thus obtained so that a variety of approximating functions could be used in the implementation of the procedure.

By interpreting these estimates in spaces of piecewise polynomials, we established a measure of their rate of convergence. Although, theoretically, any order of accuracy can be attained, practical limitations do exist. For piecewise polynomial basis functions of higher degree than quintics, even on a coarse mesh, the number of residual parameters is prohibitively large. Therefore, to be competitive with finite difference schemes, piecewise cubic functions must suffice.

The essential concept in this procedure was the reduction of the original spatially distributed dynamical system to a more tractable "lower-dimensional" system. For many problems, approximation methods exhibit several advantages over finite difference schemes. First, unlike discrete solutions which must appear in tabular form, approximation schemes deliver a continuous solution. Second, a careful selection of the approximating functions can induce useful numerical structure into the system, e.g., symmetry and sparseness of matrices. Finally, for certain piecewise polynomial approximating functions a much coarser mesh spacing is required than for discrete methods to obtain comparable accuracy.

REFERENCES

[1] F. ALVARADO AND R. MUKUNDAN, *An optimization problem in distributed parameter systems*, Internat. J. Control, 9 (1969), no. 6, pp. 665–677.

[2] G. BIRKHOFF, M. H. SCHULTZ AND R. S. VARGA, *Piecewise Hermite interpolation in one and two variables with applications to partial differential equations*, Numer. Math., 11 (1968), pp. 232–256.

[3] W. E. BOSARGE, JR. AND O. G. JOHNSON, *Direct method approximation to the state regulator problem using a Ritz–Trefftz suboptimal control*, IEEE Trans. Automatic Control, AC-15 (1970), pp. 627–631.

[4] ———, *Error bounds of high order accuracy for the state regulator problem via piecewise polynomial approximations*, this Journal, 9 (1971), pp. 15–28.

[5] W. E. Bosarge, Jr., O. G. Johnson and C. L. Smith, *A direct method approximation to the linear parabolic regulator problem over multivariate spline bases*, Houston Scientific Center Rep., IBM Pub. No. 320.2401, Houston, Tex., 1970.

[6] W. E. Bosarge, Jr., O. G. Johnson, R. S. McKnight and W. P. Timlake, *The Ritz–Galerkin procedure for nonlinear control problems*, SIAM J. Numer. Anal., 10 (1973), pp. 94–111.

[7] A. G. Butkovskii, *Theory of Optimal Control of Distributed Parameter Systems*, Moscow, 1966 (in Russian); English transl., American Elsevier, New York, 1969.

[8] E. W. Cheney, *Introduction to Approximation Theory*, McGraw-Hill, New York, 1966.

[9] P. G. Ciralet, M. H. Schultz and R. J. Varga, *Numerical methods of high-order accuracy for non-linear boundary value problems, I. One dimensional problem*, Numer. Math., 9 (1967), pp. 394–430.

[10] J. Dieudonné, *Foundations of Modern Analysis*, Academic Press, New York, 1960.

[11] J. Douglas and T. Dupont, *Galerkin methods for parabolic systems*, SIAM J. Numer. Anal., 7 (1970), pp. 575–626.

[12] E. B. Lee and L. Markus, *Foundations of Optimal Control Theory*, John Wiley, New York, 1967.

[13] J. L. Lions, *Optimal Control of Systems Governed by Partial Differential Equations*, Springer-Verlag, New York, 1971.

[14] D. G. Luenberger, *Optimization by Vector Space Methods*, John Wiley, New York, 1969.

[15] S. S. Prabhu and I. McCausland, *Time optimal control of linear diffusion processes using Galerkin method*, Proc. IEE, 117 (1970), pp. 1398–1404.

[16] A. P. Sage and S. P. Chaudhuri, *Discretization schemes and the optimal control of distributed parameter system*, Proc. of Asilomar Conference on System and Circuits, 1967.

[17] ———, *Gradient and quasi-linearization computational techniques for distributed parameter systems*, Internat. J. Control, 6 (1907), no. 1, pp. 81–98.

[18] A. P. Sage, *Optimum System Control*, Prentice-Hall, New York, 1968.

[19] C. L. Smith, *A direct method approximation to the linear parabolic regulator control problem using multivariate splines*, Doctoral thesis, Rice University, Houston, Tex., 1971.

[20] S. G. Tzafestas, *Optimal distributed parameter control using classical variational theory*, Internat. J. Control, 12 (1970), no. 4, pp. 593–608.

[21] P. K. C. Wang, *Control of distributed parameter systems*, Advances in Control Systems, vol. I, C. T. Leondes, ed., Academic Press, New York, 1964.

SIAM J. Control
Vol. 11, No. 3, August 1973

LOWER SEMICONTINUITY OF MULTIVALUED LINEARIZATION MAPPINGS*

STEPHEN M. ROBINSON† and ROBERT R. MEYER‡

Abstract. Many results in mathematical programming require lower semicontinuity of the multivalued function obtained from a constraint set by replacing the functions defining the set by their linearizations about a point. In this paper we give a simple sufficient condition, involving the gradients of the active linearized constraints, for this property to hold. We further show that this is the weakest possible condition which uses only first order information at the point in question.

1. Introduction. Let $g(x)$ and $h(x)$ be differentiable functions from an open set $H \subset \mathbb{R}^n$ into $\mathbb{R}^m$ and $\mathbb{R}^q$ respectively. The constraint sets of many mathematical programming problems can be written in the form,

$$\{x|g(x) \leqq 0, h(x) = 0\}, \tag{1}$$

for suitable g and h. If we define a multivalued mapping T from H into $\mathbb{R}^n$ by

$$Tx := \{y|Lg(x, y) \leqq 0, Lh(x, y) = 0\},$$

where

$$Lg(x, y) := g(x) + g'(x)(y - x)$$

and Lh is similarly defined, then the set Tx (which could be empty) is often referred to as the linearization of the constraint set (1) about x. Such linearized constraint sets play an important role in a number of algorithms for solving nonlinear optimization problems, and the continuity properties of the mapping T are usually central to the analysis of these algorithms [3], [4], [6], [8]. If g and h are continuously differentiable, it is easy to show that the restriction of T to closed subsets of H is a closed mapping (i.e., its graph is closed; see [1, p. 111]). A property of more interest, but less easy to prove, is lower semicontinuity of T: we say T is *lower semicontinuous* at $\bar{x} \in H$ if for each open set $Q \subset \mathbb{R}^n$ such that $T\bar{x} \cap Q \neq \varnothing$, there is an open neighborhood $U(\bar{x})$ in H such that for each $x \in U$ we have $Tx \cap Q \neq \varnothing$. The lower semicontinuity of T is of fundamental importance in the analysis of algorithms which employ the linearized set Tx as an approximation to (1) [3], [4]. It is therefore of interest to have a simple and easily-applied criterion for determining whether T is lower semicontinuous at a particular point $\bar{x}$. In this paper we give such a criterion, which generalizes earlier results of Meyer [3], [4]; further, we show that this criterion is in fact the best possible if only first order information at $\bar{x}$ (i.e., only values of $g(\bar{x})$, $h(\bar{x})$, $g'(\bar{x})$ and $h'(\bar{x})$) is used.

2. Sufficient conditions for lower semicontinuity. For a fixed $\bar{x}$ and a fixed $\bar{y} \in T\bar{x}$, we let $A \subset \{1, \cdots, m\}$ be defined by $A := \{i|Lg_i(\bar{x}, \bar{y}) = 0\}$, and $I := \{1, \cdots, m\}\backslash A$. We denote by $g_A(x)$ the vector formed from those functions

* Received by the editors April 25, 1972, and in revised form October 5, 1972.

† Mathematics Research Center, University of Wisconsin, Madison, Wisconsin 53706. The work of this author was sponsored by the United States Army under Contract DA-31-124-ARO-D-462.

‡ Shell Development Company, Houston, Texas 77001. Now at Computer Sciences Department and Mathematics Research Center, University of Wisconsin, Madison, Wisconsin 53706.

$g_i(x)$ for which $i \in A$, and define $g_I(x)$ in a similar way; note that this partition of g depends upon $\bar{y}$ as well as $\bar{x}$. The *linearized constraints active at* $\bar{y}$ are then defined to be $Lg_A(\bar{x}, \cdot)$ and $Lh(\bar{x}, \cdot)$. Finally, for given positive integers p and r we say that a pair of matrices $B(p \times n)$ and $D(r \times n)$ satisfies the *LI* (linear independence) *condition* if

$$u^T B + v^T D = 0 \quad \text{and} \quad u \geqq 0$$

implies

$$u = 0 \quad \text{and} \quad v = 0.$$

It is shown in [7, Thm. 1.1.9] that this condition is equivalent to the solvability of the system

$$Bx \leqq b,$$
$$Dx = d$$

for every right-hand side $(b, d) \in \mathbb{R}^{p+r}$.

The following sufficient condition for lower semicontinuity of T at a point $\bar{x}$ is a generalization of the positive linear independence condition of [3] (for inequalities alone) and of the condition given in [4].

THEOREM 1. *Let g and h be continuously differentiable functions from an open set $H \subset \mathbb{R}^n$ into $\mathbb{R}^m$ and $\mathbb{R}^q$ respectively, and let T be defined as previously. Then the mapping T is lower semicontinuous at $\bar{x} \in H$ if one of the following holds*:

(a) *No linearized constraint is active at any point in $T\bar{x}$.*

(b) *There is a point $\bar{y} \in T\bar{x}$ at which some linearized constraint is active and such that the pair $[g'_A(\bar{x}), h'(\bar{x})]$ satisfies the LI condition.*

Proof. Part (a) is intended to cover certain trivial cases, and we will consider it first. If $T\bar{x} = \emptyset$ then T is trivially lower semicontinuous at $\bar{x}$. If $T\bar{x} \neq \emptyset$ but no linearized constraint is ever active, then for any $\bar{y} \in T\bar{x}$ we have $Lg(\bar{x}, \bar{y}) < 0$ (there can be no equality constraints), and since Lg is continuous in x, for x in H and close enough to $\bar{x}$ we have $Lg(x, \bar{y}) < 0$. Thus $\bar{y} \in Tx$, so T is lower semicontinuous at $\bar{x}$. (This case arises only when $g(\bar{x}) < 0$, $g'(\bar{x}) = 0$, and $T\bar{x} = \mathbb{R}^n$.)

For part (b), assume that at least one linearized constraint is active at some point $\bar{y}$ of $T\bar{x}$, and that for that $\bar{y}$ the pair $[g'_A(\bar{x}), h'(\bar{x})]$ satisfies the LI condition. We shall construct a point $\hat{y} \in T\bar{x}$ at which only the linearized equality constraints (if there are any) are active. Since the LI condition implies that the gradients to these constraints are linearly independent, it will follow that the gradients to all the linearized constraints active at $\hat{y}$ are linearly independent; by [4, Thm. 1.1] the mapping T must then be lower semicontinuous at $\bar{x}$.

If no linearized inequality constraints are active at $\bar{y}$, let $\hat{y} := \bar{y}$; otherwise consider the system

$$g'_A(\bar{x})z \leqq -e,$$
$$h'(\bar{x})z = 0,$$

where e is a vector with 1 in every component. This system is solvable because of the LI condition. For sufficiently small positive λ, the point $\hat{y} := \bar{y} + \lambda z$ satisfies $Lg_A(\bar{x}, \hat{y}) < 0$ (by the condition on z), $Lg_I(\bar{x}, \hat{y}) < 0$ (because $Lg_I(\bar{x}, \bar{y}) < 0$) and $Lh(\bar{x}, \hat{y}) = 0$, so that $\hat{y}$ is the required point. This completes the proof.

3. Nonexistence of a better first order criterion. In this section we prove that the condition given in Theorem 1 is the best possible if only first order information at $\bar{x}$ is considered. Before doing so, we prove a lemma which simplifies considerably the problem of deciding whether or not the LI condition holds in Tx.

LEMMA 1. *Let $g(x)$, $h(x)$, $\bar{x}$, and T be as in Theorem* 1. *Then exactly one of the following alternatives holds*:

(a) *The LI condition is satisfied everywhere in $T\bar{x}$.*

(b) *The condition is satisfied nowhere in $T\bar{x}$.*

We say that (a) *holds vacuously if no linearized constraint is active anywhere in $T\bar{x}$ (in particular, if $T\bar{x} = \varnothing$).*

Proof. Obviously (a) and (b) cannot both be satisfied. Suppose (a) does not hold; that is, $T\bar{x} \neq \varnothing$ and there is some $\bar{y} \in T\bar{x}$ such that the pair $[g'_A(\bar{x}), h'(\bar{x})]$ does not satisfy the LI condition. We shall prove that (b) is true. Let $\hat{y}$ be an arbitrary point in $T\bar{x}$. Since the LI condition does not hold at $\bar{y}$, there must exist $u \geqq 0$ and v such that

$$u^T g'_A(\bar{x}) + v^T h'(\bar{x}) = 0,$$

with u and v not both zero. In fact, u must be semipositive, since otherwise the rows of $h'(\bar{x})$ would be linearly dependent and (b) would then hold. Let $C \subset A$ be the set of indices corresponding to positive elements of u; let $u_C > 0$ and $g'_C(\bar{x})$ be formed from u and $g'_A(\bar{x})$ in the obvious way. We have

$$Lg_C(\bar{x}, \hat{y}) \leqq 0, \qquad Lh(\bar{x}, \hat{y}) = 0$$

and

$$Lg_C(\bar{x}, \bar{y}) = 0, \qquad Lh(\bar{x}, \bar{y}) = 0.$$

Subtracting, we find that

$$g'_C(\bar{x})(\hat{y} - \bar{y}) \leqq 0$$

and

$$h'(\bar{x})(\hat{y} - \bar{y}) = 0.$$

But,

$$u_C^T g'_C(\bar{x})(\hat{y} - \bar{y}) + v^T h'(\bar{x})(\hat{y} - \bar{y}) = 0,$$

and since $u_C > 0$ it follows that $g'_C(\bar{x})(\hat{y} - \bar{y}) = 0$. Hence, $Lg_C(\bar{x}, \hat{y}) = Lg_C(\bar{x}, \bar{y}) + g'_C(\bar{x})(\hat{y} - \bar{y}) = 0$, so each linearized inequality constraint whose index is in C is active at $\hat{y}$. The existence of u_C and v then implies that the LI condition is not satisfied at $\hat{y}$. Since $\hat{y}$ was arbitrary, alternative (b) must be true. This completes the proof.

THEOREM 2. *Let $g(x)$, $h(x)$, $\bar{x}$, and T be as in Theorem* 1. *Suppose that at some point $\bar{y} \in T\bar{x}$ the pair $[g'_A(\bar{x}), h'(\bar{x})]$ does not satisfy the LI condition. Then there are quadratic functions $\tilde{g}(x)$ and $\tilde{h}(x)$, such that*

$$\tilde{g}(\bar{x}) = g(\bar{x}), \qquad \tilde{g}'(\bar{x}) = g'(\bar{x}),$$

and

$$\tilde{h}(\bar{x}) = h(\bar{x}), \qquad \tilde{h}'(\bar{x}) = h'(\bar{x}),$$

but the multivalued function defined by

$$\tilde{T}x := \{y | L\tilde{g}(x, y) \leqq 0, L\tilde{h}(x, y) = 0\}$$

is not lower semicontinuous at $\bar{x}$.

Proof. Let $\bar{y}$ be as in the hypothesis of the theorem. Let b and d be vectors, with $d \in \mathbb{R}^q$ and b having the same number of components as g_A (note that A might be empty; in that case b is not used). We do not specify at this point how b and d are to be chosen, since different choices are required in various cases. We define

$$\tilde{g}_A(x) := g_A(\bar{x}) + g'_A(\bar{x})(x - \bar{x}) - \|x - \bar{x}\|^2 b,$$

$$\tilde{g}_I(x) : = g_I(\bar{x}) + g'_I(\bar{x})(x - \bar{x}),$$

and

$$\tilde{h}(x) := h(\bar{x}) + h'(\bar{x})(x - \bar{x}) - \|x - \bar{x}\|^2 d,$$

where $\|\cdot\|$ is the Euclidean norm. Let $\tilde{T}$ be defined as in the statement of the theorem; then if x is an arbitrary point and $y \in \tilde{T}x$, we must have, in particular,

$$L\tilde{g}_A(x, y) \leqq 0, \qquad L\tilde{h}(x, y) = 0.$$

When written out, these become

$$g_A(\bar{x}) + g'_A(\bar{x})(x - \bar{x}) - \|x - \bar{x}\|^2 b + [g'_A(\bar{x}) - 2b(x - \bar{x})^T](y - x) \leqq 0$$

and

$$h(\bar{x}) + h'(\bar{x})(x - \bar{x}) - \|x - \bar{x}\|^2 d + [h'(\bar{x}) - 2d(x - \bar{x})^T](y - x) = 0,$$

or

$$g'_A(\bar{x})(y - \bar{y}) \leqq b[2(x - \bar{x})^T(y - \bar{x}) - \|x - \bar{x}\|^2]$$

and

$$h'(\bar{x})(y - \bar{y}) = d[2(x - \bar{x})^T(y - \bar{x}) - \|x - \bar{x}\|^2], \tag{2}$$

where we have used the fact that $Lg_A(\bar{x}, \bar{y})$ and $Lh(\bar{x}, \bar{y})$ vanish. We note for future reference that

$$2(x - \bar{x})^T(y - \bar{x}) - \|x - \bar{x}\|^2 = \|y - \bar{x}\|^2 - \|y - x\|^2.$$

In order to prove the result, we have to consider three cases, which are enumerated below. The most general case is treated first.

Case 1. $T\bar{x} \neq \{\bar{x}\}$. By Lemma 1, the LI condition does not hold anywhere in $T\bar{x}$; therefore since $T\bar{x} \neq \{\bar{x}\}$ we can find a point $\bar{y} \neq \bar{x}$ at which the pair $[g'_A(\bar{x}), h'(\bar{x})]$ does not satisfy the LI condition. Choose b and d to be vectors such that the system

$$g'_A(\bar{x})z \leqq b,$$

$$h'(\bar{x})z = d$$

has no solution. Then we have from (2) that for $x \in \mathbb{R}^n$ and $y \in \tilde{T}x$,

$$g'_A(\bar{x})(y - \bar{y}) \leqq b[2(x - \bar{x})^T(y - \bar{x}) - \|x - \bar{x}\|^2]$$

and

$$(3) \qquad h'(\bar{x})(y - \bar{y}) = d[2(x - \bar{x})^T(y - \bar{x}) - \|x - \bar{x}\|^2].$$

Let x be restricted to be of the form $x = \bar{x} + \lambda(\bar{y} - \bar{x})$, for $\lambda > 0$, and choose $\varepsilon = \|\bar{y} - \bar{x}\|/2 > 0$. Then for $\|y - \bar{y}\| < \varepsilon$,

$$\begin{aligned} 2(y - \bar{x})^T(x - \bar{x}) - \|x - \bar{x}\|^2 &= 2\lambda(y - \bar{y})^T(\bar{y} - \bar{x}) + (2\lambda - \lambda^2)\|\bar{y} - \bar{x}\|^2 \\ &> -2\lambda(\|\bar{y} - \bar{x}\|/2)\|\bar{y} - \bar{x}\| + (2\lambda - \lambda^2)\|\bar{y} - \bar{x}\|^2 \\ &= \lambda(1 - \lambda)\|\bar{y} - \bar{x}\|^2, \end{aligned}$$

and the latter quantity is positive for $0 < \lambda < 1$, so that we may divide both sides of (3) by $2(x - \bar{x})^T(y - \bar{x}) - \|x - \bar{x}\|^2$. The resulting contradiction shows that there are x's arbitrarily close to $\bar{x}$ for which an open ε-ball about $\bar{y}$ is excluded from $\tilde{T}x$, so that $\tilde{T}$ is not lower semicontinuous at $\bar{x}$.

Case 2. $T\bar{x} = \{\bar{x}\}$ and at least one linearized inequality constraint is active at $\bar{x}$. Let

$$(4) \qquad \alpha^* := \min\{\alpha | g'_A(\bar{x})y \leqq \alpha e, h'(\bar{x})y = 0, \|y\| = 1\}.$$

If the minimum is not positive, then for sufficiently small positive λ we have $\bar{x} + \lambda\bar{y} \in T\bar{x}$, where $\bar{y}$ is any feasible point of (4) with $\alpha \leqq 0$, and this contradicts the assumption that $T\bar{x} = \{\bar{x}\}$. Hence, there is a positive α^* such that for any nonzero vector $v \in \mathbb{R}^n$ such that $h'(\bar{x})v = 0$, we have for some i depending on v, $g'_i(\bar{x})v \geqq \alpha^*\|v\|$. Now let x be such that $0 < \|x - \bar{x}\| < \frac{1}{2}\alpha^*$; choose $b = e$ and $d = 0$. If $y \in \tilde{T}x$, then from (2), using $\bar{y} = \bar{x}$, we have

$$\begin{aligned} &g'_A(\bar{x})(y - \bar{x}) \leqq e[2(x - \bar{x})^T(y - \bar{x}) - \|x - \bar{x}\|^2], \\ &h'(\bar{x})(y - \bar{x}) = 0. \end{aligned}$$

Since $x \neq \bar{x}$ the above inequality shows that $y - \bar{x} \neq 0$; hence for some $i \in A$ we have,

$$\begin{aligned} \alpha^*\|y - \bar{x}\| \leqq g'_i(\bar{x})(y - \bar{x}) &\leqq 2(x - \bar{x})^T(y - \bar{x}) - \|x - \bar{x}\|^2 \\ &\leqq 2\|x - \bar{x}\|\,\|y - \bar{x}\| \\ &< \alpha^*\|y - \bar{x}\|, \end{aligned}$$

a contradiction. Hence, for $0 < \|x - \bar{x}\| < \frac{1}{2}\alpha^*$ we have $\tilde{T}x = \varnothing$, and so $\tilde{T}$ is not lower semicontinuous at $\bar{x}$.

Note that in the treatment of this case there was no need to invoke (explicitly) the fact that the pair $[g'_A(\bar{x}), h'(\bar{x})]$ did not satisfy the LI condition.

Case 3. $T\bar{x} = \{\bar{x}\}$ and no linearized inequalities are active at $\bar{x}$.

Choose d to be a vector such that

$$h'(\bar{x})z = d$$

has no solution; such a vector exists since the rows of $h'(\bar{x})$ are linearly dependent.

Let $x \in \mathbb{R}^n$ be arbitrary, and let y be any point in $\tilde{T}x$ (if such a point exists). As before, we have $L\tilde{h}(x, y) = 0$, or,

$$h'(\bar{x})(y - \bar{x}) = (\|y - \bar{x}\|^2 - \|y - x\|^2)d. \tag{5}$$

If the right-hand side of (5) is not zero, we can divide by the quantity in parentheses to obtain a contradiction; therefore, we have

$$h'(\bar{x})(y - \bar{x}) = 0.$$

However, since $T\bar{x} = \{\bar{x}\}$ the columns of $h'(\bar{x})$ must be linearly independent, and so we have $y = \bar{x}$. Finally, substituting $y = \bar{x}$ in (5) we obtain, as before, $x = \bar{x}$. Thus $\tilde{T}x = \emptyset$ unless $x = \bar{x}$, so $\tilde{T}$ is not lower semicontinuous at $\bar{x}$. This completes the proof of Theorem 2.

4. Effect of nonlinear constraints. With certain modifications, Theorem 1 can be made to apply to the more general case in which nonlinear convex constraints are included in the linearization mapping. Specifically, let C be a convex set in $\mathbb{R}^n$, and denote by ri C and aff C, respectively, the relative interior and affine hull of C [5]. Let D be an $r \times n$ matrix of rank r and d a point in $\mathbb{R}^r$ such that

$$\operatorname{aff} C = \{y | Dy = d\}.$$

Instead of T, we shall consider the mapping $Sx := Tx \cap C$.

THEOREM 3. *Let $g(x)$, $h(x)$ and H be as in Theorem 1. Then the mapping S is lower semicontinuous at $\bar{x} \in H$ if either of the following holds*:

(a) *No linearized constraint is active at any point in $S\bar{x}$.*

(b) *There is a point $\bar{y} \in T\bar{x} \cap$ ri C such that at $\bar{y}$ we have*

$$u^T g'_A(\bar{x}) + v^T h'(\bar{x}) + w^T D = 0 \quad \textit{and} \quad u \geqq 0$$
$$\textit{implies} \quad u = 0, \quad v = 0, \quad w = 0.$$

Proof. Part (a) is trivial. For part (b), we first remark that if C is a singleton then aff $C = C$ and the matrix D is $n \times n$. In this case, the hypothesis implies that both $g'_A(\bar{x})$ and $h'(\bar{x})$ are vacuous, so for all points x sufficiently close to $\bar{x}$ we have $\bar{y} \in Tx$ and therefore $\bar{y} = Sx$ (with equality since $C = \{\bar{y}\}$), so S is lower semicontinuous at $\bar{x}$. Therefore, assume that C contains at least one line segment, so that the dimension of aff C is not less than one. Let Q be any open set with $Q \cap S\bar{x} \neq \emptyset$, and let $y \in Q \cap S\bar{x} = Q \cap C \cap Tx$. It is a basic property of convex sets [5, Thm. 6.1] that for any $\lambda \in (0, 1]$ the point $y + \lambda(\bar{y} - y)$ lies in ri C. Choose $\tilde{\lambda} > 0$ so small that $\tilde{y} := y + \tilde{\lambda}(\bar{y} - y) \in Q$. Let $V(\tilde{y})$ be an open neighborhood of $\tilde{y}$ such that $V(\tilde{y}) \subset Q$ and $V(\tilde{y}) \cap \operatorname{aff} C \subset$ ri C. By Theorem 1, the multivalued mapping,

$$Rx := Tx \cap \operatorname{aff} C = \{y | Lg(x, y) \leqq 0, Lh(x, y) = 0, Dy = d\},$$

is lower semicontinuous at $\bar{x}$. Therefore, since $V(\tilde{y}) \cap R\bar{x} \neq \emptyset$, there is an open neighborhood $U(\bar{x})$ such that for any $x \in U$ we have $Rx \cap V(\tilde{y}) \neq \emptyset$. Let

$\hat{y} \in Rx \cap V(\bar{y})$; then

$$\begin{aligned}\hat{y} \in Tx \cap [\operatorname{aff} C \cap V(\bar{y})] &\subset Tx \cap \operatorname{ri} C \cap V(\bar{y})\\ &\subset Tx \cap C \cap Q\\ &= Sx \cap Q,\end{aligned}$$

and so S is lower semicontinuous at $\bar{x}$. This completes the proof.

COROLLARY 3.1. *Let $g(x)$, $h(x)$, $\bar{x}$, H, C, and S be as in Theorem 3. Suppose C can be written in the form*

$$C = \{y \mid f(y) \leqq 0, Dy = d\},$$

where f is a differentiable function on $\mathbb{R}^n$. If for some $\bar{y} \in S\bar{x}$ we have

$$u^T g'_A(\bar{x}) + v^T h'(\bar{x}) + w^T D + r^T f'_B(\bar{y}) = 0 \quad \textit{and} \quad u \geqq 0, \quad r \geqq 0$$
$$\textit{implies} \quad u = 0, \quad v = 0, \quad w = 0, \quad r = 0,$$

where $f'_B(\bar{y})$ is the matrix whose rows are the gradients of those components of f which are zero at $\bar{y}$, then S is lower semicontinuous at $\bar{x}$.

Proof. The hypothesis guarantees that there is some $z \in \mathbb{R}^n$ such that

$$\begin{aligned} g'_A(\bar{x})z &< 0,\\ h'(\bar{x})z &= 0,\\ f'_B(\bar{y})z &< 0,\\ Dz &= 0.\end{aligned}$$

For sufficiently small positive λ, the point $\hat{y} := \bar{y} + \lambda z$ satisfies $Lg(\bar{x}, \hat{y}) < 0$, $Lh(\bar{x}, \hat{y}) = 0$, $f(\hat{y}) < 0$, and $D\hat{y} = d$. Hence, there is some open neighborhood $V(\hat{y})$ such that

$$V(\hat{y}) \cap \operatorname{aff} C \subset V(\hat{y}) \cap \{y \mid Dy = d\} \subset C.$$

Therefore $\hat{y} \in \operatorname{ri} C$, and since also $\hat{y} \in T\bar{x}$ we can apply Theorem 3 to conclude that S is lower semicontinuous at $\bar{x}$, which proves the corollary.

The following analogue of Lemma 1 holds in the present case.

THEOREM 4. *Let $g(x)$, $h(x)$, $\bar{x}$, and H be as in Theorem 1, and let C, $f(x)$ and S be as in Corollary 3.1. Suppose also that $f(x)$ is a pseudoconvex function (i.e., each component of $f(x)$ is pseudoconvex). Then exactly one of the following alternatives is true:*

(a) *The extended LI condition,*

$$u^T g'_A(\bar{x}) + v^T h'(\bar{x}) + w^T D + r^T f'_B(\bar{y}) = 0 \quad \textit{and} \quad u \geqq 0, \quad r \geqq 0$$
$$\textit{implies} \quad u = 0, \quad v = 0, \quad w = 0, \quad r = 0,$$

holds at every $\bar{y} \in S\bar{x}$ (where, as before, we say the condition holds vacuously if no constraint is binding anywhere in $S\bar{x}$).

(b) *The extended LI condition holds nowhere in $S\bar{x}$.*

Proof. As in Lemma 1, (a) and (b) cannot both hold. Suppose (a) does not. Proceeding as in the proof of Lemma 1, we find that for some $\bar{y} \in S\bar{x}$ there are subsets $E \subset A$, $F \subset B$, of which at most one can be empty, such that

$$u_E^T g_E'(\bar{x}) + v^T h'(\bar{x}) + w^T D + r_F^T f_F'(\bar{y}) = 0, \tag{6}$$

with $u_E > 0$, $r_F > 0$, v and w unrestricted. Suppose the extended LI condition holds at a point $\tilde{y} \in S\bar{x}$. Reasoning as in Corollary 3.1, we can construct a point $\hat{y} \in S\bar{x}$ such that $Lg(\bar{x}, \hat{y}) < 0$ and $f(\hat{y}) < 0$. Since f is pseudoconvex, $f_F(\hat{y}) < 0 = f_F(\bar{y})$ implies

$$f_F'(\bar{y})(\hat{y} - \bar{y}) < 0. \tag{7}$$

Also, by an argument similar to that used in Lemma 1, we have

$$\begin{aligned} g_E'(\bar{x})(\hat{y} - \bar{y}) &< 0, \\ h'(\bar{x})(\hat{y} - \bar{y}) &= 0, \\ D(\hat{y} - \bar{y}) &= 0. \end{aligned} \tag{8}$$

Combining (6), (7) and (8) yields a contradiction, so the point $\tilde{y}$ cannot exist. Thus the extended LI condition holds nowhere in $S\bar{x}$, which proves the theorem.

We conclude with an extension of Theorem 1 to the case in which no convexity assumptions are made about those functions which are not linearized. In the absence of such assumptions, the feasible region may contain points at which the LI condition is satisfied as well as points at which it fails. However, if we do assume that the LI condition holds at all points in the feasible region, then the following theorem states that the mapping is lower semicontinuous. This result is included more for the sake of theoretical completeness than for any practical computational value.

THEOREM 5. *Let $g(x), h(x), \bar{x}$, and H be as in Theorem 1. Let $f(x)$ and $t(x)$ be differentiable functions from $\mathbb{R}^n$ into $\mathbb{R}^k$ and $\mathbb{R}^l$ respectively. Suppose in addition that all first partial derivatives of t are continuous on $\mathbb{R}^n$. Define a multivalued mapping by*

$$Px := \{y | Lg(x, y) \leqq 0, Lh(x, y) = 0, f(y) \leqq 0, t(y) = 0\}.$$

Suppose that at each point $\bar{y} \in P\bar{x}$ we have

$$\begin{gathered} u^T g_A'(\bar{x}) + v^T h'(\bar{x}) + w^T f_B'(\bar{y}) + r^T t'(\bar{y}) = 0 \\ \text{and} \\ u \geqq 0, \quad w \geqq 0 \\ \text{implies} \quad u = 0, \quad v = 0, \quad w = 0, \quad r = 0, \end{gathered} \tag{9}$$

where A and B are the index sets corresponding to the components active at $\bar{y}$ of $Lg(\bar{x}, \cdot)$ and $f(\cdot)$ respectively. Then P is lower semicontinuous at $\bar{x}$.

As before, the condition (9) is understood to hold vacuously at $\bar{y}$ if no constraints are active there.

Proof. Let Q be an open set with $P\bar{x} \cap Q \neq \emptyset$, and choose any $\bar{y} \in P\bar{x} \cap Q$. By assumption, (9) holds at $\bar{y}$. If at least one of the index sets A and B is nonempty, then we can use [2, Lemma 11.2.1] in a sufficiently small open neighborhood of $\bar{y}$ to find a point $\hat{y} \in P\bar{x} \cap Q$ with $Lg(\bar{x}, \hat{y}) < 0$ and $f(\hat{y}) < 0$. On the other hand, if

A and B are both empty we can take $\hat{y} = \bar{y}$. In either case, there exists an open neighborhood $V(\hat{y}) \subset Q$ in which the inequality constraints are inactive. Since (9) implies that the gradients to all of the equality constraints at $\hat{y}$ are linearly independent, we can apply the implicit function theorem as in [4, Thm. 1.1] to find an open neighborhood $W(\bar{x})$ such that for any $x \in W$ there is some $y \in V$ such that $Lg(x, y) < 0$, $Lh(x, y) = 0$, $f(y) < 0$, and $t(y) = 0$. Hence, $y \in Px \cap Q$, so P is lower semicontinuous at $\bar{x}$. This completes the proof.

REFERENCES

[1] C. Berge, *Topological Spaces*, Macmillan, New York, 1963.

[2] O. L. Mangasarian, *Nonlinear Programming*, McGraw-Hill, New York, 1969.

[3] R. R. Meyer, *The solution of non-convex optimization problems by iterative convex programming*, Doctoral thesis, The University of Wisconsin, Madison, 1968.

[4] ———, *The validity of a family of optimization methods*, this Journal, 8 (1970), pp. 41–54.

[5] R. T. Rockafellar, *Convex Analysis*, Princeton University Press, Princeton, 1970.

[6] J. B. Rosen, *Iterative solution of nonlinear optimal control problems*, this Journal, 4 (1966), pp. 223–244.

[7] J. Stoer and C. Witzgall, *Convexity and Optimization in Finite Dimensions. I*, Springer-Verlag, New York, 1970.

[8] W. I. Zangwill, *Nonlinear Programming: A Unified Approach*, Prentice-Hall, Englewood Cliffs, N.J., 1969.

SIAM J. CONTROL
Vol. 11, No. 3, August 1973

A DUAL METHOD FOR OPTIMAL CONTROL PROBLEMS WITH INITIAL AND FINAL BOUNDARY CONSTRAINTS*

O. PIRONNEAU AND E. POLAK†

Abstract. This paper presents two new algorithms belonging to the family of dual methods of centers. The first can be used for solving fixed time optimal control problems with inequality constraints on the initial and terminal states. The second one can be used for solving fixed time optimal control problems with inequality constraints on the initial and terminal states and with affine instantaneous inequality constraints on the control. Convergence is established for both algorithms. Qualitative reasoning indicates that the rate of convergence is linear.

1. Introduction. The construction of optimal control algorithms is often hampered by two difficulties. The first is due to the fact that the cost function usually has a gradient only in L_∞, while the convergence of the algorithm must be studied in L_2, since control sequences constructed by an optimization algorithm are not likely to converge in L_∞. The second difficulty stems from the fact that "primal-type" subproblems, such as those resulting from a direct application of methods of centers or feasible directions, cannot be solved directly and usually require some sort of "dualization." Both of these sources of difficulty are taken into account in the dual method presented in this paper.

The algorithm in this paper may be classified as a dual method of centers. It has the very nice feature that it is implementable essentially without recourse to heuristics, since both its direction finding and step size finding procedures are finite, in the sense that they require only a finite number of function evaluations per iteration. Since a closely related algorithm presented in [6] converges linearly on finite-dimensional problems, it is reasonably certain that the algorithm presented in this paper also converges linearly on problems in $\mathbb{R}^n$. However, since certain sets, used in the proofs in [6], loose their compactness in general Banach spaces, the proof of rate of convergence given in [6] cannot be extended to general Banach spaces. In spite of this, there are heuristic reasons which lead us to believe that the algorithms presented in this paper do converge linearly at least on a class of optimal control problems with linear dynamics and convex costs.

Basically, the algorithm in this paper is designed for problems with inequality end constraints and without constraints on the control. However, as we shall show, simple constraints on the controls can be handled by means of a minor modification of the algorithm.

The only other algorithms which solve the problems for which our algorithm was designed are based on the use of penalty functions. Since penalty function methods suffer considerably from "ridge paralysis" when the penalty becomes high, we expect our method to be superior in moderate to high precision situations.

* Received by the editors May 25, 1971, and in revised form May 20, 1972.

† Department of Electrical Engineering and Computer Sciences and the Electronics Laboratory, University of California, Berkeley, California 94720. This research was supported in part by the Joint Services of Electronics Program under Grant AFOSR-68-1488 and in part by the National Aeronautics and Space Administration under Grant NGL-05-003-016.

2. Optimality and convergence. Because of the peculiar nature of optimal control problems, which necessitates the simultaneous use of both the L_2- and the L_∞-norms on a space of regulated functions [3], we need the following abstract structure and accompanying theorems.

Let V be a linear space, let $\|\cdot\|_1$ be a norm on V and let $\langle\cdot,\cdot\rangle_2$ be a scalar product on V, such that $\mathscr{B}_1 = \{V, \|\cdot\|_1\}$ is a Banach space and $\mathscr{B}_2 = \{V, \langle\cdot,\cdot\rangle_2\}$ is a subspace of a Hilbert space. Let $\|\cdot\|_2$ be the norm induced by $\langle\cdot,\cdot\rangle_2$ on $\mathscr{B}_2$ (i.e., $\|z\|_2 = \langle z, z\rangle_2^{1/2}$).

Assumption 2.1. There exists a $C > 0$ such that $\|z\|_2 \leqq C\|z\|_1$ for all $z \in V$.

Now consider the problem

$$\min\{f^0(z) | f^j(z) \leqq 0, j = 1, 2, \cdots, m\}, \tag{2.2}$$

where $f^j: V \to \mathbb{R}^1$ for $j = 0, 1, 2, \cdots, m$.

Assumptions 2.3.

(i) The functions $f^j(\cdot), j = 0, 1, 2, \cdots, m$, are Fréchet differentiable on $\mathscr{B}_1$, with the Fréchet derivative at $\bar{z}$ being denoted by $f^j_z(\bar{z})(\cdot), j = 0, 1, 2, \cdots, m$.

(ii) The restrictions to $\{z \in \mathscr{B}_2 | \|z\|_1 < M\}$ of the functions $f^j(\cdot)$, $j = 0, 1, 2, \cdots, m$, are continuous for any $M \in (0, \infty)$ (i.e., they are continuous in $\|\cdot\|_2$ on $\{z \in V | \|z\|_1 < M\}$).

(iii) There exist functions $\nabla f^j: V \to V, j = 0, 1, 2, \cdots, m$, with the following properties: (a) the ∇f^j are continuous on $\mathscr{B}_1$, (b) the ∇f^j have continuous restrictions on $\{z \in \mathscr{B}_2 | \|z\|_1 < M\}$ for any $M \in (0, \infty)$, (c) the ∇f^j satisfy

$$f^j_z(\bar{z})(h) = \langle \nabla f^j(\bar{z}), h\rangle_2, \qquad j = 0, 1, 2, \cdots, m, \tag{2.4}$$

for any $\bar{z}$, h in $\mathscr{B}_1$.

THEOREM 2.5. *Suppose that $\hat{z} \in \mathscr{B}_1$ is a solution of* (2.1) (*i.e.*, $f^j(\hat{z}) \leqq 0$ *for* $j = 1, 2, \cdots, m$, *and* $f^0(\hat{z}) = \min\{f^0(z) | f^j(z) \leqq 0, j = 1, 2, \cdots, m\}$). *Then there exist multipliers* $\mu^0(\hat{z}) \geqq 0, \mu^1(\hat{z}) \geqq 0, \cdots, \mu^m(\hat{z}) \geqq 0$ *such that*

$$\sum_{j=0}^{m} \mu^j(\hat{z}) \nabla f^j(\hat{z}) = 0, \tag{2.6}$$

$$\mu^j(\hat{z}) f^j(\hat{z}) = 0 \quad \textit{for } j = 1, 2, \cdots, m, \tag{2.7}$$

and

$$\sum_{j=0}^{m} \mu^j(\hat{z}) = 1. \tag{2.8}$$

Theorem 2.5 is a straightforward generalization of the well-known F. John condition of optimality [4]. It can be proved in essentially the same manner as the F. John condition (see the proof of Theorem 3.5.11 in [2]).

DEFINITION 2.9. Let the set of *feasible* points $\Omega \subset \mathscr{B}_1$ be defined by

$$\Omega = \{z \in \mathscr{B}_1 | f^j(z) \leqq 0, j = 1, 2, \cdots, m\}, \tag{2.10}$$

and let the set of *desirable* points $\Delta \subset \Omega$ be the set of points $\hat{z} \in \Omega$ for which there exist multipliers $\mu^j(\hat{z}), j = 0, 1, \cdots, m$, which satisfy (2.6)–(2.8).

Thus, Δ is the set of feasible points which satisfy the optimality condition (2.5). Since in general it is not possible to identify points in Ω which are optimal for (2.2), the best we can hope to achieve is to compute a desirable point.

The algorithm which we shall describe in the next section uses a map A: $\Omega \to 2^{\Omega}$ and is of the following form.

ALGORITHM MODEL 2.11.

Step 0. Compute a $z_0 \in \Omega$, and set $i = 0$.

Step 1. Compute a $y \in A(z_i)$.

Step 2. If $f^0(y) < f^0(z_i)$, set $z_{i+1} = y$, set $i = i + 1$, and go to Step 1; else, set $\hat{z} = z_i$, and stop.

The convergence properties of our algorithm are summarized by the following result.

THEOREM 2.12. *Suppose that Assumption* 2.3(ii) *is satisfied, that for every* $M > 0$, $\Omega_M \triangleq \{z \in \Omega | \|z\|_1 < M\}$, *and that for every* $z \in \Omega$, $z \notin \Delta$, *there exist an* $\varepsilon(z) > 0$ *and a* $\delta(z) < 0$ *such that for every* $M > \|z\|_1$,

$$f^0(z'') - f^0(z') \leqq \delta(z) \tag{2.13}$$

for all $z' \in \{z' \in \Omega_M | \|z' - z\|_2 \leqq \varepsilon(z)\}$, *for all* $z'' \in A(z')$.

Suppose that $\{z_i\}$ *is a sequence generated by Algorithm model* 2.11. *If* $\{z_i\}$ *is finite, then its last element* $\hat{z}$ *is in* Δ. *If* $K \subset \{0, 1, 2, \cdots\}$ *is an infinite subset and* $z^* \in \Omega$ *is such that either* (i) $\lim_{i \in K} \|z_i - z^*\|_1 = 0$ *or* (ii) $\lim_{i \in K} \|z_i - z^*\|_2 = 0$ *and* $\|z_i\|_1 < M$ *for some* $M > 0$ *and all* $i \in K$, *then* $z^* \in \Delta$.

We omit a proof of this theorem since it follows directly from Theorem 1.3.10 in [7] and the Assumption 2.1.

With the preliminaries out of the way, we can now get down to the task of establishing a specific algorithm for finding points in the set Δ.

3. A dual method of centers. For the algorithm below to make sense, we need the following additional hypothesis, as is usual in conjunction with methods of centers and methods of feasible directions (see §§ 4.2 and 4.3 in [7]).

Assumption 3.1. The set $\tilde{\Omega} = \{z \in \mathscr{B}_1 | f^j(z) < 0, j = 1, 2, \cdots, m\}$ is not empty.[1]

ALGORITHM 3.2. ($\beta \in (0, 1)$ is a step size parameter.)

Step 0. Compute a $z_0 \in \Omega$, and set $i = 0$.

Step 1. Compute $\mu(z_i) = (\mu^0(z_i), \mu^1(z_i), \cdots, \mu^m(z_i))^T \in \mathbb{R}^{m+1}$ to be a solution of the quadratic programming problem

$$\phi(z_i) \triangleq \max\left\{ \sum_{j=1}^{m} \mu^j f^j(z_i) - \frac{1}{2}\left\| \sum_{j=0}^{m} \mu^j \nabla f^j(z_i) \right\|_2^2 \,\middle|\, \cdot \sum_{j=0}^{m} \mu^j = 1, \mu \geqq 0 \right\}. \tag{3.3}$$

Step 2. If $\phi(z_i) = 0$, set $\hat{z} = z_i$, and stop; else, set

$$h(z_i) = -\sum_{j=0}^{m} \mu^j(z_i) \nabla f^j(z_i) \tag{3.4}$$

and go to Step 3.

Step 3. Compute the smallest nonnegative integer $k(z_i)$ such that

$$\theta(\beta^{k(z_i)}, h(z_i), z_i) - \tfrac{1}{2}\beta^{k(z_i)}\phi(z_i) \leqq 0, \tag{3.5}$$

[1] When $\tilde{\Omega}$ is empty, the algorithm below stops at z_0 and hence is useless.

where $\theta:\mathbb{R}^1 \times \mathscr{B}_1 \times \mathscr{B}_1 \to \mathbb{R}^1$ is defined by

$$\theta(\lambda, h, z) = \max\{f^0(z + \lambda h) - f^0(z); f^j(z + \lambda h), j = 1, 2, \cdots, m\}. \tag{3.6}$$

Step 4. Set $z_{i+1} = z_i + \beta^{k(z_i)}h(z_i)$, set $i = i + 1$, and go to Step 1.

The following result is obvious.

PROPOSITION 3.7. *Let $\phi:\mathscr{B}_1 \to \mathbb{R}^1$ be defined as in* (3.3) *and let $z \in \Omega$ be arbitrary. Then* $\phi(z) \leqq 0$, and $\phi(z) = 0$ *if and only if* $z \in \Delta$.

LEMMA 3.8. *Suppose that $z_i \in \Omega$ is such that $\phi(z_i) \neq 0$, and let $h(z_i)$ be defined as in* (3.4). *Then*

$$\begin{aligned}\max\{\langle\nabla f^0(z_i), h(z_i)\rangle_2; f^j(z_i) + \langle\nabla f^j(z_i), h(z_i)\rangle_2,\quad j = 1, 2, \cdots, m\}\\ \leqq \phi(z_i) - (1/2)\|h(z_i)\|_2^2 < 0.\end{aligned} \tag{3.9}$$

This lemma follows directly from the fact that the dual of (3.3) is

$$\begin{aligned}\phi(z_i) = \min\{(1/2)\|h\|^2 + \max\{\langle\nabla f^0(z_i), h\rangle; f^j(z_i) + \langle\nabla f^j(z_i), h\rangle,\\ j = 1, 2, \cdots, m\}\}.\end{aligned} \tag{3.10}$$

COROLLARY 3.11. *Suppose that $z_i \in \Omega$ is such that $\phi(z_i) < 0$. Then there exists an integer $k(z_i) \geqq 0$ such that* (3.5) *holds.*

Proof. This corollary follows directly from the definition of a Fréchet differential, (2.4) and the fact that by (3.9), $\langle\nabla f^0(z_i), h(z_i)\rangle_2 \leqq \phi(z_i)$, and $\langle\nabla f^j(z_i), h(z_i)\rangle \leqq \phi(z_i)$ for all $j \in \{1, 2, \cdots, m\}$ such that $f^j(z_i) = 0$.

THEOREM 3.12. *Let $\{z_i\}$ be a sequence generated by Algorithm* 3.2 *in the process of searching the set Ω for a point in Δ* (see Definition 2.9), *and suppose that Assumptions* 2.1, 2.3 *and* 3.1 *are satisfied. Then, either $\{z_i\}$ is finite and its last point $\hat{z} \in \Delta$, or $\{z_i\}$ is infinite, in which case any $z^* \in \Omega$ satisfying either $\lim_{i\in K}\|z_i - z^*\|_1 = 0$ or $\lim_{i\in K}\|z_i - z^*\|_2 = 0$, where K is an infinite subset of $\{0, 1, 2, \cdots\}$, also satisfies $z^* \in \Delta$, provided there exists an $M \in (0, \infty)$ such that $\|z_i\|_1 \leqq M$ for all $i \in K$.*

Proof. Algorithm 3.2 is of the form of Algorithm 2.11, with $A(\cdot)$ defined as follows. Let $S:\Omega \to 2^V$ be defined by

$$S(z) = \left\{ -\sum_{j=0}^{m} \mu^j \nabla f^j(z) \middle| \mu \geqq 0,\ \sum_{j=0}^{m} \mu^j = 1;\right.$$
$$\left. \sum_{j=1}^{m} \mu^j f^j(z) - \frac{1}{2}\left\|\sum_{j=0}^{m} \mu^j \nabla f^j(z)\right\|_2^2 = \phi(z)\right\}. \tag{3.13}$$

Then $A:\Omega \to 2^\Omega$ is given by

$$A(z) = \{z' = z + \beta^{k(z,h)}h | h \in S(z)\}, \tag{3.14}$$

where $k(z, h)$ is the smallest nonnegative integer which satisfies (3.5) for $z_i = z$, $h(z_i) = h$ and $k(z_i) = k(z, h)$. (Since by (3.6) and (3.5), $f^j(z') \leqq (1/2)\beta^{k(z,h)}\phi(z) \leqq 0$ for $j = 1, 2, \cdots, m$, it is clear that $A(\cdot)$ maps Ω into 2^Ω.) To complete our proof, we only need to show that (2.13) is satisfied by the maps $f^0(\cdot)$ and $A(\cdot)$, as defined in (3.14). Since this is straightforward, we omit it.

4. An application to an optimal control problem. We shall now show that the algorithm, presented in the preceding section, can be used to solve the following problem:

$$\min\left\{\int_{t_0}^{t_f} h^0(x(t),u(t),t)\,dt \,\middle|\, \frac{d}{dt}x(t) = h(x(t),u(t),t),\right.$$

$$(4.1)\qquad t\in[t_0,t_f];\; g_0(x(t_0)) \leqq 0,\quad g_f(x(t_f)) \leqq 0;$$

$$\left. u\in L_\infty^s[t_0,t_f]\right\},$$

where $h^0:\mathbb{R}^n \times \mathbb{R}^s \times [t_0,t_f] \to \mathbb{R}^1$, $h:\mathbb{R}^n \times \mathbb{R}^s \times [t_0,t_f] \to \mathbb{R}^n$, $g_0:\mathbb{R}^n \to \mathbb{R}^{m_1}$, g_f: $\mathbb{R}^n \to \mathbb{R}^{m_2}$, and $L_\infty^s[t_0,t_f]$ is the space of equivalence classes of essentially bounded integrable functions from $[t_0,t_f]$ into $\mathbb{R}^s$.

We must begin by transcribing problem (4.1) into the form of problem (2.2). Therefore, let $V = \{(\xi,u)|\xi\in\mathbb{R}^n,\ u\in L_\infty^s[t_0,t_f]\}$, let the norm $\|\cdot\|_1: V\to\mathbb{R}^1$ be defined by

$$(4.2)\qquad \|(\xi,u)\|_1^2 = |\xi|^2 + \operatorname*{ess\,sup}_{t\in[t_0,t_f]} |u(t)|^2,$$

where $|\cdot|$ denotes the Euclidean norm, and finally, let the scalar product $\langle\cdot,\cdot\rangle_2$ on V be defined by

$$(4.3)\qquad \langle(\xi,u),(\xi',u')\rangle_2 = \langle\xi,\xi'\rangle + \int_{t_0}^{t_f}\langle u(t),u'(t)\rangle\,dt,$$

where $\langle\cdot,\cdot\rangle$ denotes the Euclidean scalar product. Then we see that the space $\mathscr{B}_1 = \{V,\|\cdot\|_1\}$ is a Banach space and the space $\mathscr{B}_2 = \{V,\langle\cdot,\cdot\rangle_2\}$ is a subspace of a Hilbert space. Furthermore, setting $\|\cdot\|_2 = \sqrt{\langle\cdot,\cdot\rangle_2}$, it is not difficult to show that there exists a $C\in(0,\infty)$ such that $\|\cdot\|_2 \leqq C\|\cdot\|_1$. Next, let $x(t,\xi,u)$, $t\in[t_0,t_f]$, denote the solution of the differential equation

$$(4.4)\qquad \frac{d}{dt}x = h(x,u,t),\quad x(t_0) = \xi,\quad t\in[t_0,t_f],$$

corresponding to a given $(\xi,u)\in V$. Then we define the functions $f^0: V\to\mathbb{R}^1$, $f_1: V\to\mathbb{R}^{m_1}$ and $f_2:\mathbb{R}^{m_2}$ as follows:

$$(4.5)\qquad f^0(\xi,u) = \int_{t_0}^{t_f} h^0(x(t,\xi,u),u(t),t)\,dt,$$

$$(4.6)\qquad f_1(\xi,u) = g_0(\xi),$$

$$(4.7)\qquad f_2(\xi,u) = g_f(x(t_f,\xi,u)).$$

With the above definitions, problem (4.1) can be written as follows, setting $z = (\xi,u)$:

$$(4.8)\qquad \min\{f^0(z)|f_1(z)\leqq 0, f_2(z)\leqq 0\},$$

i.e., it can be written in the form (2.2).

Assumptions 4.9.

(i) For every $(\xi, u) \in V$, the solution $x(\cdot, \xi, u)$ of (4.4) exists and is unique.

(ii) The functions h^0 and h are continuously differentiable in x and in u, and h^0, h, $\partial h^0/\partial x$, $\partial h^0/\partial u$, $\partial h/\partial x$, $\partial h/\partial u$ are piecewise continuous in t.

(iii) The functions g_0 and g_1 are continuously differentiable.

(iv) The set $\{z = (\xi, u) \in V \mid f_1(z) < 0, f_2(z) < 0\}$ is not empty.

LEMMA 4.10. *Suppose that Assumptions* 4.9(i)–4.9(iii) *are satisfied. Then the functions f^0, f_1 and f_2, defined in* (4.5)–(4.7), *are Fréchet differentiable on $\mathscr{B}_1$, with their differentials f^0_z, f_{1z}, f_{2z}, defined as follows*:

$$f^0_z(z')(h) = \langle \nabla f^0(z'), h\rangle_2, \tag{4.11}$$

$$f^i_{kz}(z')(h) = \langle \nabla f^i_k(z'), h\rangle_2, \qquad i = 1, 2, \cdots, m_k, \quad k = 1, 2, \tag{4.12}$$

where, for $z' = (\xi', u')$,

$$\begin{aligned}\nabla f^0(z') = (-q_0(t_0, \xi', u'), &-\frac{\partial h}{\partial u}(x(\cdot, \xi', u'), u'(\cdot), \cdot)^T \\ &\cdot q_0(\cdot, \xi', u') + \frac{\partial h^0}{\partial u}(x(\cdot, \xi', u'), u'(\cdot), \cdot)^T),\end{aligned} \tag{4.13}$$

$$\nabla f^i_1(z') = \left(\frac{\partial g^i_0}{\partial x}(\xi')^T, 0\right), \qquad i = 1, 2, \cdots, m_1, \tag{4.14}$$

and

$$\begin{aligned}\nabla f^i_2(z') = (-q_i(t_0, \xi', u'), -\frac{\partial h}{\partial u}(x(\cdot, \xi', u'), u'(\cdot), \cdot)^T q_i(\cdot, \xi', u')),& \\ i = 1, 2, \cdots, m_2,&\end{aligned} \tag{4.15}$$

with $q_i(t, \xi', u')$, $i = 0, 1, 2, \cdots, m_2$ defined (*with $\delta_{i0} = 0$ if $i \neq 0$*) *by*

$$\begin{aligned}\frac{d}{dt} q_i(t, \xi', u') = &-\left[\frac{\partial h}{\partial x}(x(t, \xi', u'), u'(t), t)\right]^T q_i(t, \xi', u') \\ &+ \delta_{i0}\left[\frac{\partial h^0}{\partial x}(x(t, \xi', u'), u'(t), t)\right]^T, \\ t \in [t_0, t_f], &\quad q_0(t_f, \xi', u') = 0, \\ q_i(t_f, \xi', u') = &-\left[\frac{\partial g^i_f}{\partial x}(x(t_f, \xi', u'))\right]^T.\end{aligned} \tag{4.16}$$

COROLLARY 4.17. *For any $M > 0$, the functions $f^0, f_1, f_2, \nabla f^i_1$, $i = 1, 2, \cdots, m_1$, and ∇f^i_2, $i = 1, 2, \cdots, m_2$, have continuous restrictions on $\{z \in \mathscr{B}_2 \wedge \|z\|_1 < M\}$.*

This lemma and the corollary follow directly from Theorem 10.7.1 in [3] and from Theorem A1 in [5]. We therefore omit their proofs.

Thus we see that the functions f^0, f_1 and f_2 satisfy the Assumptions 2.3. We now show that the set Δ defined in Definition 2.9, with $f^j = f^j_1$, for $j = 1, 2, \cdots, m_1$, and $f^{j+m_1} = f_2^j$, for $j = 1, 2, \cdots, m_2$, is the set of initial states and controls for which the Pontryagin-maximum-principle-in-differential-form[2] is satisfied.

PROPOSITION 4.18. *Let Ω and Δ be defined as in Definition 2.9, with $f^j = f_1^j$, for $j = 1, 2, \cdots, m_1$, and $f_1^{j+m_1} = f_2^j$, for $j = 1, 2, \cdots, m_2$, and let $m = m_1 + m_2$. If $(\hat{\xi}, \hat{u}) \in \Delta$, then there exists a multiplier function $\hat{\lambda}$: $[t_0, t_f] \to \mathbb{R}^n$, and a scalar $\hat{\lambda}^0 \leqq 0$ such that*

$$\begin{aligned}\frac{d}{dt}\hat{\lambda}(t) = & -\frac{\partial h}{dx}(x(t, \hat{\xi}, \hat{u}), \hat{u}(t), t)^T \hat{\lambda}(t) \\ & + \hat{\lambda}^0 \frac{\partial h^0}{\partial x}(x(t, \hat{\xi}, \hat{u}), \hat{u}(t), t)^T, \qquad t \in [t_0, t_f],\end{aligned} \tag{4.19}$$

$$\hat{\lambda}(t_0) = \frac{\partial g_0(\hat{\xi})^T}{\partial x} \hat{v}_0, \tag{4.20}$$

$$\hat{\lambda}(t_f) = \frac{\partial g_f(x(t_f, \hat{\xi}, \hat{u}))^T}{\partial x} \hat{v}_f, \tag{4.21}$$

where $\hat{v}_0 \geqq 0$, $\hat{v}_f \geqq 0$ are such that $(\hat{v}_0, \hat{v}_f) \neq 0$,

$$\langle \hat{v}_0, g_0(\hat{\xi}) \rangle = \langle \hat{v}_f, g_f(x(t_f, \hat{\xi}, \hat{u})) \rangle = 0, \tag{4.22}$$

and

$$\frac{\partial}{\partial u}[\hat{\lambda}^0 h^0(x(t, \hat{\xi}, \hat{u}), \hat{u}(t), t) + \langle \hat{\lambda}(t), h(\hat{x}(t, \hat{\xi}, \hat{u}), \hat{u}(t), t) \rangle] \equiv 0. \tag{4.23}$$

(*Since $(\hat{\xi}, \hat{u}) \in \Delta \subset \Omega$, we must have* $g(\hat{\xi}) \leqq 0$ and $g_f(x(t_f, \hat{\xi}, \hat{u})) \leqq 0$ *by definition.*)

For the problem (4.1), Algorithm 3.2 assumes the following specific form.

ALGORITHM 4.24. (Solves (4.1); $\beta \in (0, 1)$ is a step size parameter.)

Step 0. Compute $(\xi_0, u_0) \in \mathscr{B}_1$ such that $g_0(\xi_0) \leqq 0$ and $g_f(x(t_f, \xi_0, u_0)) \leqq 0$, and set $i = 0$.

Comment. The Algorithm 4.24 can be used to compute such an $(\xi_0, u_0(\cdot))$ by solving the problem

$$\left\{ \min \int_{t_0}^{t_f} x^0(t)\, dt \,\middle|\, \begin{aligned} & \frac{d}{dt}\underline{x} = \underline{h}(x, u, t); \\ & g_0^j(x(t_0)) - x^0(t_0) \leqq 0, j = 1, 2, \cdots, m_1; \\ & g_f^j(x(t_f)) - x^0(t_f) \leqq 0, j = 1, 2, \cdots, m_2 \end{aligned} \right\}, \tag{4.25}$$

where $\underline{x} = (x^0, x)$ and $\underline{h} = (0, h)$, and for which an initial point $(\underline{\tilde{\xi}}_0, \tilde{u}_0(\cdot))$, $\underline{\tilde{\xi}}_0 = (\tilde{x}_0^0, \tilde{\xi}_0)$, can be chosen as follows: let $\tilde{\xi}_0, \tilde{u}_0(\cdot)$ be arbitrary, and let

$$\tilde{x}_0^0 = \max \{g_0^j(\tilde{\xi}_0), j = 1, 2, \cdots, m_1; g_f^j(x(t_f, \tilde{\xi}_0, u)), j = 1, 2, \cdots, m_2\}.$$

Since the optimal value of (4.25) is strictly negative, a $(\xi_0, u_0(\cdot))$ for Step 0 above can be computed by means of a finite number of iterations.

[2] In the maximum principle in differential form, the condition of maximum on the Hamiltonian is replaced by the condition $(\partial H(\hat{x}, \hat{u}, \hat{\psi}, t)/\partial u)\delta u \leqq 0$ for all admissible δu and for almost all $t \in [t_0, t_f]$.

Step 1. For $z_i = (\xi_i, u_i)$, compute $\nabla f^0(z_i)$, $\nabla f_1^j(z_i)$, $j = 1, 2, \cdots, m_1$, $\nabla f_2^j(z_i)$, $j = 1, 2, \cdots, m_2$, according to (4.13)–(4.16).

Step 2. Compute $\mu^0(z_i)$, $\mu_1^j(z_i)$, $j = 1, 2, \cdots, m_1$, $\mu_2^j(z_i)$, $j = 1, 2, \cdots, m_2$, as a solution of

$$\phi(z_i) = \max\left\{ \sum_{j=1}^{m_1} \mu_1^j g_0^j(\xi_i) + \sum_{j=1}^{m_2} \mu_2^j g_f^j(x(t_f, \xi_i, u_i)) - \frac{1}{2}\left\| \mu^0 \nabla f^0(z_i) + \sum_{j=1}^{m_1} \mu_1^j \nabla f_1^j(z_i) + \sum_{j=1}^{m_2} \mu_2^j \nabla f_2^j(z_i) \right\|_2^2 \;\middle|\; \mu^0 + \sum_{j=1}^{m_1} \mu_1^j + \sum_{j=1}^{m_2} \mu_2^j = 1, \mu^0 \geqq 0, \mu_1^j \geqq 0, j = 1, 2, \cdots, m_1, \mu_2^j \geqq 0, j = 1, 2, \cdots, m_2 \right\}, \tag{4.26}$$

where $\|z\|_2^2 = \langle z, z\rangle_2$ is defined as in (4.3).

Step 3. If $\phi(z_i) = 0$, set $\hat{\xi} = \xi_i$, $\hat{u}(\cdot) = u_i(\cdot)$ and stop; else, go to Step 4 (see (4.13)–(4.16)).

Step 4. Set

$$\omega_i = \mu^0(z_i) q_0(t_0, \xi_i, u_i) - \sum_{j=1}^{m_1} \mu_1^j(z_i) \frac{\partial g_0^j(\xi_i)^T}{\partial x} + \sum_{l=1}^{m_2} \mu_2^l(z_i) q_l(t_0, \xi_i, u_i), \tag{4.27}$$

$$v_i(\cdot) = \mu^0(z_i)\left[\frac{\partial h}{\partial u}(x(\cdot, \xi_i, u_i), u_i(\cdot), \cdot)^T \cdot q_0(\cdot, \xi_i, u_i) - \frac{\partial h^0}{\partial u}(x(\cdot, \xi_i, u_i), u(\cdot), \cdot)^T\right] + \sum_{j=1}^{m_2} \mu_2^j(z_i) \frac{\partial h}{\partial u}(x(\cdot, \xi_i, u_i), u_i(\cdot), \cdot)^T q_j(\cdot, \xi_i, u_i), \tag{4.28}$$

and go to Step 5.

Step 5. Compute the smallest integer k, such that

$$\max\left\{ \int_{t_0}^{t_f} [h^0(x(t, \xi_i + \beta^k \omega_i, u_i + \beta^k v_i), u_i + \beta^k v_i, t) - h^0(x(t, \xi_i, u_i), u_i, t)]\, dt, g_0^j(\xi_i + \beta^k \omega_i), j = 1, 2, \cdots, m_1; g_f^j(x(t_f, \xi_i + \beta^k \omega_i, u_i + \beta^k v_i)), j = 1, 2, \cdots, m_2 \right\} - (\beta^k/2)\phi(z_i) \leqq 0. \tag{4.29}$$

Step 6. Set $\xi_{i+1} = \xi_i + \beta^k \omega_i$, set $u_{i+1}(\cdot) = u_i(\cdot) + \beta^k v_i(\cdot)$, set $i = i + 1$, and go to Step 1.

The following result is obvious.

PROPOSITION 4.30. *Theorem* 3.12 *holds for Algorithm* 4.24, *with the set* Δ *defined as the set of feasible points* $(\hat{\xi}, \hat{u}) \in \mathscr{B}_1$ *satisfying the Pontryagin-maximum-principle-in-differential-form for problem* (4.1).

5. An extension to problems with instantaneous constraints on the control. We shall now show that Valentine-type transformations [9] can be used to adapt Proposition 4.30 for the solution of the following optimal control problem:

$$\text{(5.1)}\quad \min\left\{\int_{t_0}^{t_f} h^0(x(t),u(t),t)\,dt \,\middle|\, \frac{d}{dt}x(t) = h(x(t),u(t),t),\right.$$
$$t\in[t_0,t_f]\,;\, g_0(x(t_0)) \leqq 0,\, g_f(x(t_f)) \leqq 0\,;\, u\in L_\infty^s[t_0,t_f]\,;$$
$$\left. b^k \leqq \langle a_k, u(t)\rangle \leqq c^k,\, k = 1,2,\cdots,r, \text{ for all } t\in[t_0,t_f]\right\},$$

where h^0, h, g_0, g_f are as in (4.1); $a_k \in \mathbb{R}^s$ for $k = 1,2,\cdots,r$; $c_k \in \mathbb{R}^1$ for $k = 1,2,\cdots,r$, $b_k \in \mathbb{R}^1$ for $k = 1,2,\cdots,r'$, $r' \leqq r$, and $b_k = -\infty$ for $k = r'+1,\cdots,r$.

Assumptions 5.2.

(i) We shall assume that (4.9) is satisfied.

(ii) The vectors a_k, $k = 1,2,\cdots,r$, are linearly independent.

(iii) There exists a control $\bar{u} \in L_\infty^s[t_0,t_f]$ and an initial state $\xi \in \mathbb{R}^n$ such that $g_0(\xi) \leqq 0$, $g_f(x(t_f,\xi,u)) \leqq 0$, and $b^k < \langle a_k, \bar{u}(t)\rangle < c^k$ for $k = 1,2,\cdots,r$ and all $t\in[t_0,t_f]$.

To apply the Valentine trick, we must use certain substitutions for the inequalities on the control. Thus, consider the constraints

$$\text{(5.3a)}\qquad b^k \leqq \langle a_k, u(t)\rangle \leqq c^k, \quad k = 1,2,\cdots,r', \quad t\in[t_0,t_f],$$

$$\text{(5.3b)}\qquad \langle a_k, u(t)\rangle \leqq c^k, \quad k = r'+1,\cdots,r, \quad t\in[t_0,t_f].$$

Suppose that $u \in L_\infty^s[t_0,t_f]$ satisfies (5.2) and (5.3). Then we can associate with this u functions $v^k:[t_0,t_f] \to \mathbb{R}^1$, $k = 1,2,\cdots,r'$, and $w^k:[t_0,t_f] \to \mathbb{R}^1$, $k = 1,2,\cdots,r-r'$, such that

$$\text{(5.4)}\qquad \cos v^k(t) = \frac{2}{c^k - b^k}\langle a_k, u(t)\rangle - \frac{c^k + b^k}{c^k - b^k}, \quad k = 1,2,\cdots,r', \quad t\in[t_0,t_f],$$

$$\text{(5.5)}\qquad (w^k(t))^2 = c^{k+r'} - \langle a_{k+r'}, u(t)\rangle, \quad k = 1,2,\cdots,r-r', \quad t\in[t_0,t_f].$$

We shall now use these functions to construct a problem which is equivalent to (5.1). Let A^T be the $s \times r$ matrix whose columns are $(2/(c^k - b^k))a_k$, $k = 1,2,\cdots,r'$, and $-a_k$, $k = r'+1, r'+2,\cdots,r$. Then its transpose, A, is an $r \times s$ matrix which can be partitioned as follows: $A = [A' \mid A'']$ (rearranging the components of $u(\cdot)$, if necessary), where A'' is an $r \times r$ nonsingular matrix. We partition u similarly, i.e., we set $u = (u', u'')$, with $u'' \in \mathbb{R}^r$ and $u' \in \mathbb{R}^{s-r}$. Then, if u, v^k, w^k satisfy (5.4) and (5.5), we obtain

$$\text{(5.6)}\qquad \begin{aligned} u''(t) = A''^{-1}\Bigg[&(\cos v^1(t),\cdots,\cos v^{r'}(t),\omega^1(t)^2,\cdots,\omega^{r-r'}(t)^2)^T \\ &+ \left(\frac{c^1+b^1}{c^1-b^1},\cdots,\frac{c^{r'}+b^{r'}}{c^{r'}-b^{r'}}, -c^{r'u},\cdots,-c^r\right)^T\Bigg] \\ &- A''^{-1}A'u'(t) \triangleq u''(u'(t),v(t),w(t)). \end{aligned}$$

Let $\bar{u} = (u', v, w)$, let

$$\bar{h}^0(x, \bar{u}, t) = h^0(x, (u', u''(u', v, w)), t),$$

$$\bar{h}(x, \bar{u}, t) = h(x, (u', u''(u', v, w)), t)$$

and consider the problem

$$\text{(5.7)}\qquad \min\left\{ \int_{t_0}^{t_f} \bar{h}^0(x, \bar{u}, t)\, dt \middle| \dot{x} = \bar{h}(x, \bar{u}, t), t \in [t_0, t_f]; \right.$$
$$\left. g_0(x(t_0)) \leqq 0; g_f(x(t_f)) \leqq 0; \bar{u} \in L_\infty^s[t_0, t_f] \right\},$$

where all the quantities are as in (5.1) and (5.6). It is trivial to show that if $(\hat{\xi}, \hat{u}', \hat{v}, \hat{w})$ is any optimal solution of (5.7), $(\hat{\xi}, (\hat{u}', \cos \hat{v}, \hat{w}^2))$ is also optimal for problem (5.1). However, not all the points $(\tilde{\xi}, \tilde{u}', \tilde{v}, \tilde{w})$ which satisfy the Pontryagin principle for problem (5.7) result in a pair $(\tilde{\xi}, \tilde{u}) \triangleq (\tilde{\xi}, (\tilde{u}' \cos \tilde{v}, \tilde{w}^2))$ which satisfy the maximum principle for problem (5.1). Hence, although Algorithm 4.24 is directly applicable to problem (5.7), it is desirable to modify it so as to prevent convergence to points which do not satisfy the optimality conditions for problem (5.1). This can be done by modifying the step length rule in Step 5 of Algorithm 4.24. To explain how this is done, without loss of generality, we assume that there is only one constraint of the form (5.3a) and only one constraint of the form (5.3b), i.e., we assume that (5.3a) and (5.3b) have the following specific form:

$$\text{(5.8)}\qquad -1 \leqq u^{s-1}(t) \leqq 1, \qquad u^s(t) \geqq 0, \quad t \in [t_0, t_f],$$

and that the remaining components of $u(t)$ are unconstrained. For this specific case, given z_i, Algorithm 4.24 computes the following feasible direction $(\omega_i, v_i(\cdot))$:

$$\text{(5.9)}\qquad \omega_i = \mu^0(z_i)p_i(t_0) - \sum_{k=1}^{m_1} \mu_1^k(z_i)\frac{\partial g_0}{\partial x}(\xi_i)^T + \sum_{l=1}^{m_2} \mu_2^l(z_i)q_{l,i}(t_0),$$

$$\text{(5.10)}\qquad v_i^T = (v_i'^T, v_i^{s-1}, v_i^s)$$

with

$$v_i'(\cdot) = \mu^0(z_i)\left[\frac{\partial h}{\partial u'}(x_i(\cdot), u_i(\cdot), \cdot)^T p_i(\cdot) - \frac{\partial h^0}{\partial u'}(x_i(\cdot), u_i(\cdot), \cdot)^T\right]$$
$$+ \sum_{l=1}^{m_2} \mu_2^l(z_i)\frac{\partial h}{\partial u'}(x_i(\cdot), u_i(\cdot), \cdot)^T q_{l,i}(\cdot),$$

$$v_i^{s-1}(\cdot) = -\sin v_i(\cdot)\left\{\mu^0(z_i)\left[\frac{\partial h}{\partial u^{s-1}}(x_i(\cdot), u_i(\cdot), \cdot)^T q_0(\cdot)\right.\right.$$
$$\left.\left. - \frac{\partial h^0}{\partial u^{s-1}}(x_i(\cdot), u_i(\cdot), \cdot)^T\right] + \sum_{l=1}^{m_2} \mu_2^l(z_i)\frac{\partial h}{\partial u^{s-1}}(x_i(\cdot), u_i(\cdot), \cdot)^T q_{l,i}(\cdot)\right\},$$

$$v_i^s(\cdot) = 2w_i(\cdot)\left\{\mu^0(z_i)\left[\frac{\partial h}{\partial u^s}(x_i(\cdot), u_i(\cdot), \cdot)^T p_i(\cdot)\right.\right.$$
$$\left.\left. - \frac{\partial h^0}{\partial h^s}(x_i(\cdot), u_i(\cdot), \cdot)^T\right] + \sum_{l=1}^{m_2} \mu_2^l(z_i)\frac{\partial h}{\partial u^s}(x_i(\cdot), u_i(\cdot), \cdot)^T q_{l,i}(\cdot)\right\},$$

where the $q_{l,i}(\cdot)$, $l = 0, 1, 2, \cdots, m_2$, are computed by solving (4.16) for $\xi' = \xi_i$, $x' = u_i$ and h, h^0 as in (5.1).[3] The new step length rule can be stated as a substitute Step 5, which, for the problem in hand, becomes:

Step 5′. Compute the smallest integer k_i satisfying

$$\begin{aligned}\max\Bigg\{\int_{t_0}^{t_f} &[h^0(x(t, \xi_i + \beta^{k_i}\omega_i, (u_i' + \beta^{k_i}v_i', \cos(v_i + \beta^{k_i}v_i^{s-1}), (w_i + \beta^{k_i}v_i^s)^2)),\\ &(u_i'(t) + \beta^{k_i}v_i'(t), \cos(v_i(t) + \beta^{k_i}v_i^{s-1}(t)), (w_i(t) + \beta^{k_i}v_i^s(t))^2, t)\\ &- h^0(x_i(t), u_i(t), t)]\,dt, g_0(\xi_i + \beta^{k_i}\omega_i), g_f(x(t_f, \xi_i + \beta^{k_i}\omega_i,\\ &(u_i' + \beta^{k_i}v_i', \cos(v_i + \beta^{k}v_i^{s-1}), (w_i + \beta^{k_i}v_i^s)^2)))\Bigg\} - (\beta^{k_i}/2)\phi(z_i) \leqq 0,\end{aligned} \tag{5.11}$$

$$\beta^{k_i} \leqq \left(2 \times \operatorname*{ess\,sup}_{t\in[0,1]}\left\{\max\left\{\left|\frac{v_i^{s-1}(t)}{\sin v_i(t)}\right|, -\frac{v_i^s(t)}{w_i(t)}\right\}\right\}\right)^{-1}. \tag{5.12}$$

LEMMA 5.13. *The conclusions of Theorem* 3.12 *remain true for the modified Algorithm* 4.24 *with respect to problem* (5.7).

Proof. The modified algorithm differs from Algorithm 4.24 only in the additional bound (5.12) on β^k. Now, since this bound, in turn, has a denominator which can be bounded from above by a continuous function of (ξ_i, u_i', v_i, w_i), i.e., since

$$\begin{aligned}\max&\left\{\left|\frac{v_i^{s-1}(t)}{\sin v_i(t)}\right|, -\frac{v_i^s(t)}{w_i(t)}\right\}\\ &\leqq \sum_{k=s-1}^{s}\left\{\left|q_{0,i}(t)^T\frac{\partial h}{\partial u^k}(x_i(t), u_i(t), t) - \frac{\partial h^0}{\partial u^k}(x_i(t), u_i(t), t)\right|\right.\\ &\qquad\left. + \sum_{l=1}^{m_2}\left|q_{l,i}(t)^T\frac{\partial h}{\partial u^k}(x_i(t), u_i(t), t)\right|\right\},\end{aligned} \tag{5.14}$$

it follows from arguments essentially duplicating the proof of Theorem 3.12 that the conditions of Theorem 2.12 are satisfied.

LEMMA 5.15. *Let* $\{(\xi_i, u_i', v_i, w_i)\}_{i=0}^{\infty}$ *be a sequence generated by Algorithm* 4.24, *modified to use Step* 5′ *instead of Step* 5 *in the process of solving problem* (5.7). *Suppose K is an infinite subset of the positive integers such that* $\lim_{i\in K}\|(\xi_i, u_i', v_i, w_i) - (\hat{\xi}, \hat{u}', \hat{v}, \hat{w})\|_2 = 0$ *and* $\sup_{i\in K}\|(\xi_i, u_i', v_i, w_i)\|_1 < \infty$. *Furthermore, let* K' *be an infinite subset of* K, *such that* $\lim_{i\in K'}\{(\mu^0(z_i), \mu_1(z_i), \mu_2(z_i)\} = \{\hat{u}^0, \hat{u}_1, \hat{u}_2\}$, *where* $z_i = (\xi_i, (u_i', v_i, w_i))$ *and* $\mu^0(z_i)$, $\mu_1(z)$ *and* $\mu_2(z_i)$ *are defined as in Step* 2 *of Algorithm* 4.24 (*for problem* (5.7)). *Then,*

$$\begin{aligned}&\left\{\frac{\partial h^0}{\partial u^{s-1}}(\hat{x}(t), \hat{u}(t), t) - \frac{\partial h}{\partial u^{s-1}}(\hat{x}(t), \hat{u}(t), t)^T\right.\\ &\qquad\left.\cdot(\hat{\mu}^0\hat{q}(t) + \sum_{l=1}^{m_2}\hat{\mu}_2^l\hat{q}_l(t))\right\}(-1)^k \leqq 0,\end{aligned} \tag{5.16a}$$

[3] The form of (5.10) is due to the fact that we have used the chain rule to calculate the needed partial derivatives.

for almost all $t \in \{t | \hat{v}(t) = k\pi\}$, $k = 0, 1$, *and*

$$\text{(5.16b)} \qquad \frac{\partial h^0}{\partial u^s}(\hat{x}(t), \hat{u}(t), t) - \frac{\partial h}{\partial u^s}(\hat{x}(t), \hat{u}(t), t)^T \left(\hat{\mu}_0 \hat{q}_0(t) + \sum_{l=1}^{m_2} \hat{\mu}_2^l \hat{q}_l(t) \right) \geqq 0$$

for almost all $t \in \{t | \hat{w}(t) = 0\}$, *where* $\hat{u} = (\hat{u}', \cos \hat{v}, \hat{\omega}^2)$, $\hat{x}(t) \equiv x(t, \hat{\xi}, \hat{u})$, *and* $\hat{q}_l$, $l = 0, 1, \cdots, m_2$, *are defined by* (4.16) *for* $u'(t) \equiv \hat{u}(t), x(t, \xi', u') \equiv \hat{x}(t), \xi' = \hat{\xi}$.

Proof. Let $H: \mathbb{R}^n \times \mathbb{R}^s \times \mathbb{R}^n \times \mathbb{R}^1 \to \mathbb{R}^1$ be defined by

$$\text{(5.17)} \qquad H(x, u, \psi, t) = -h^0(x, u, t) + \langle \psi, h(x, u, t) \rangle .$$

Then, from (5.10) and the instructions in Step 6 of Algorithm 4.24, we find that for $t \in [t_0, t_f]$, and with

$$\text{(5.18)} \qquad \psi_i(t) = \mu^0(z_i) q_{0,i}(t) + \sum_{l=1}^{m_2} \mu_2^l(z_i) q_{l,i}(t), \qquad t \in [t_0, t_f],$$

we must have

$$\text{(5.19)} \qquad v_{i+1}(t) = v_i(t) - \beta^{k_i} \sin v_i(t) \frac{\partial H}{\partial u^{s-1}}(x_i(t), u_i(t), \psi_i(t), t),$$

$$\text{(5.20)} \qquad w_{i+1}(t) = w_i(t) + 2\beta^{k_i} w_i(t) \frac{\partial H}{\partial u^s}(x_i(t), u_i(t), \psi_i(t), t).$$

Now, the purpose of the bound (5.12) on β^{k_i} was to ensure that for almost all $t \in [t_0, t_f]$,

$$\text{(5.21)} \qquad 0 < \tfrac{1}{2} v_i(t) \leqq v_{i+1}(t) \leqq \tfrac{1}{2}(\pi + v_i(t)) < \pi,$$

$$\text{(5.22)} \qquad \tfrac{1}{2} w_i(t) \leqq w_{i+1}(t).$$

Since we must have $\lim_{i \in K'} v_i(t) = \hat{v}(t)$ for almost all $t \in [t_0, t_f]$, suppose that $t' \in [t_0, t_f]$ is such that $\lim_{i \in K'} v_i(t') = \hat{v}(t') = k\pi$, with $k = 0$ or 1. It now follows from (5.18), since $\beta^{k_i} < 1$, and since $(\partial H / \partial u^{s-1})(x_i(t'), u_i(t'), \psi_i(t'), t')$ is bounded for $i \in K'$, that we must also have

$$\text{(5.23)} \qquad \lim_{i \in K'} v_{i+1}(t') = k\pi,$$

for almost all t' such that $\lim_{i \in K'} v_i(t') = k\pi$.[4]

By construction, $v_0(t) \in (0, \pi)$ for all $t \in [t_0, t_f]$, and hence, $\sin v_0(t') > 0$. Let K'' be an infinite subsequence of the positive integers defined by $K'' = K' \cup \{i | (i-1) \in K'\}$. It now follows from (5.21) and (5.23) that $\{(-1)^{k+1}(k\pi - v_i(t))\}_{i \in K''}$ is a strictly positive sequence which converges to zero, and hence there must exist an infinite subsequence $K''' \subset K'$ such that

$$\text{(5.24)} \qquad (-1)^{k+1}(k\pi - v_{i+1}(t')) < (-1)^{k+1}(k\pi - v_i(t')) \quad \text{for all } i \in K''' .$$

[4] Note that $\{t | \lim_{i \in K} v_{i+1}(t) \neq \lim_{i \in K} v_i(t)\}$ is a null set.

Combining (5.24) with (5.18), and recalling that $\sin v_i(t') > 0$ for all i, we conclude that

$$(5.25)\qquad (-1)^k \frac{\partial H}{\partial u^{s-1}}(x_i(t'), u_i(t'), \psi_i(t'), t) \geqq 0 \quad \text{for all } i \in K'''.$$

It now follows from the continuity of $\partial H/\partial u^{s-1}$ that

$$(5.26)\qquad (-1)^k \frac{\partial H}{\partial u^{s-1}}(\hat{x}(t'), \hat{u}(t'), \hat{\psi}(t'), t') \geqq 0,$$

where $\hat{\psi}(t') = \hat{\mu}^0 \hat{p}(t') + \sum_{l=1}^{m_2} \hat{\mu}_2^l \hat{q}_l(t')$. This establishes (5.16a); (5.16b) can be established in a similar way.

The following result is now obvious.

COROLLARY 5.27. *Let $\{(\xi_i, u_i)\}$ be a sequence generated by the modified Algorithm* 4.24 *in solving problem* (5.1) *by means of* (5.7). *Then, either $\{(\xi_i, u_i)\}$ is finite and its last element satisfies the Pontryagin-maximum-principle-in-differential-form, or $\{(\xi_i, u_i)\}$ is infinite and every pair of points $(\hat{\xi}, \hat{u})$ which satisfies for some $K \subset \{0, 1, 2, \cdots\}$ either* (i) $\lim_{i \in K} \|(\xi_i, u_i) - (\hat{\xi}, \hat{u})\|_1 = 0$, *or* (ii) $\lim_{i \in K} \|(\xi_i, u_i) - (\hat{\xi}, \hat{u})\|_2 = 0$ *and* $\sup_{i \in K} \|(\xi_i, u_i)\|_1 < \infty$, *also satisfies the Pontryagin-maximum-principle-in-differential-form.*

6. Experimental results. To illustrate the behavior of our algorithm, we have applied it to two problems. The first was,

$$(6.1)\qquad \min \frac{1}{2}\int_0^2 \left(\left\| x(t) - \left[\frac{t}{2}(x_f - x_0) + x_0 \right] \right\|^2 + \|u\|^2 \right) dt,$$

where

$$x_0 = (10, 10, 10, 10)^T, \quad x_f = (2, 2, 2, 2)^T,$$

with $x(t) \in \mathbb{R}^4$, $u(t) \in \mathbb{R}^2$ and

$$(6.2)\qquad \begin{pmatrix} \dot{x}^1 \\ \dot{x}^2 \\ \dot{x}^3 \\ \dot{x}^4 \end{pmatrix} = \begin{pmatrix} -1 & 0 & 0 & 0 \\ 0 & -2 & 0 & 0 \\ 0 & 0 & -3 & 0 \\ 0 & 0 & 0 & -4 \end{pmatrix} \begin{pmatrix} x^1 \\ x^2 \\ x^3 \\ x^4 \end{pmatrix} + \begin{pmatrix} 5 & 0 \\ 0 & 5 \\ 10 & 0 \\ 0 & 10 \end{pmatrix} \begin{pmatrix} u^1 \\ u^2 \end{pmatrix}$$

and with the constraints

$$(6.3)\qquad \|x(0) - x_0\|^2 \leqq 1, \quad \|x(2) - x_f\|^2 \leqq 1.$$

Then, we added the constraint

$$(6.4)\qquad |u^1(t)| \leqq 1, \quad |u^2(t)| \leqq 1, \qquad t \in [0, 2],$$

to obtain the second problem.

The program we used was designed for general situations and did not exploit the structure of (6.1) and (6.2). The integration was performed using the Euler–Cauchy method, with an initial step size of 1/32. The algorithm was programmed to increase the integration precision on demand, according to the scheme outlined in [8]. This adaptive integration scheme was used so as to reduce computer

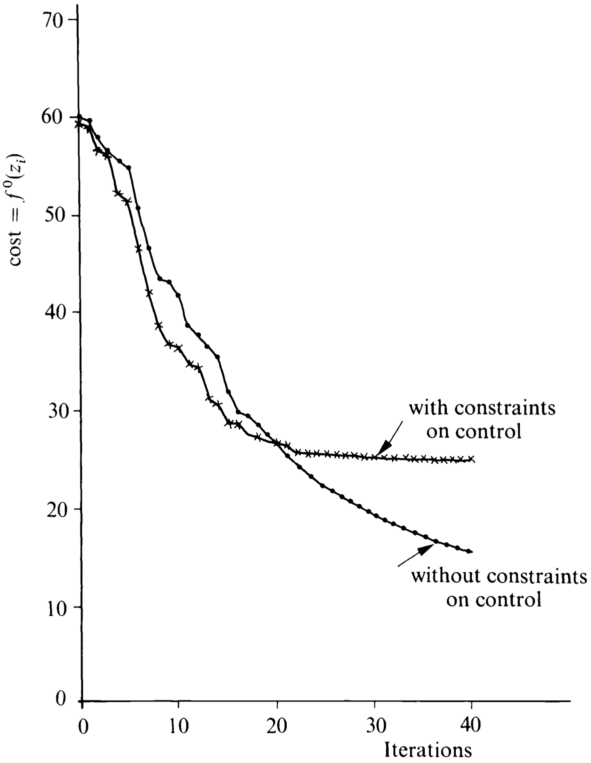


FIG. 1. *Solution of the problems* (6.1)–(6.3) *and* (6.1)–(6.4)

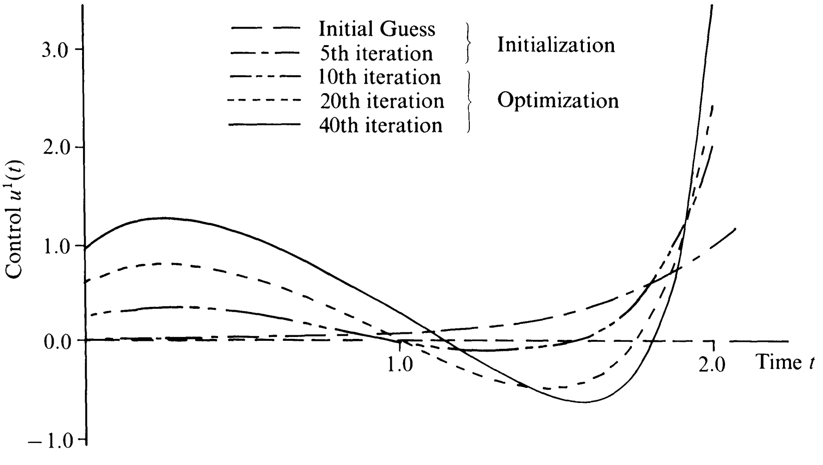


FIG. 2. *Solution of the problem* (6.1)–(6.3). *The 5th iterate of initialization cycle is feasible and is used as first iterate of minimization cycle.*

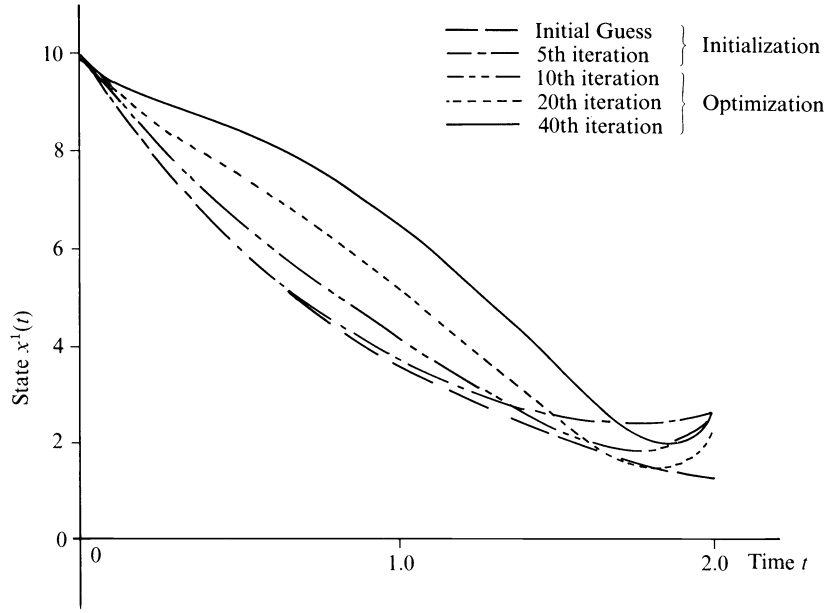


FIG. 3. *Solution of the problem* (6.1)–(6.3)

time. The computations were carried out on a CDC 6400 computer and were stopped after 40 iterations. (The first problem required about 20% less time than the second.) In each case, the program required fewer than 5 iterations to compute an initial feasible solution $(\xi_0, u(\cdot))$ and $(\xi_0, \bar{u}(\cdot))$, respectively. The results of the computation are shown graphically in the accompanying figures.

7. Conclusion. In the form stated, Algorithm 4.24 is readily implementable either by using fixed precision integration or adaptive integration schemes such as those described in [8]. Our experience has been that the adaptive precision schemes are invariably considerably faster than fixed precision schemes. Although we have not had the opportunity to confirm the following conjecture by experiment, we feel reasonably sure that our algorithm will outperform its rivals, the various penalty methods, at least in the case when the constraints are such as to cause "ridge paralysis" in the penalty methods.

In modifying Algorithm 4.24 in § 5, a substitution formula (Valentine's trick) was used to extend Algorithm 4.24 to problems with affine instantaneous inequality constraints and, in addition, a perturbation method was added to ensure that convergence was possible only to points satisfying both a first and a second order optimality condition for the derived problem (5.7). This eliminated points which satisfy a first order condition for (5.7) but not for (5.1). Industrial experience with the substitution formula, used in conjunction with simpler algorithms, indicates that it performs quite well; our own experimental results confirm this view.

Acknowledgment. We wish to thank Dr. R. Klessig for his helpful comments, Mr. H. Mukai for preparing the computer code and performing a number of experiments for us, and Dr. H. J. Kelly for sharing with us his experience in the use of the Valentine transformation.

REFERENCES

[1] C. Berge, *Topologival Spaces*, Macmillan, New York, 1963.

[2] M. Canon, C. Cullum and E. Polak, *Theory of Optimal Control and Mathematical Programming*, McGraw-Hill, New York, 1970.

[3] I. Dieudonné, *Foundation of Modern Analysis*, Academic Press, New York, 1969.

[4] F. John, *Extremum problems with inequalities as side conditions*, Studies and Essays, Courant Anniversary Volume, K. O. Friedrichs, O. W. Neugebauer and J. J. Stokers, eds., Interscience, New York, 1948, pp. 187–204.

[5] R. Klessig and E. Polak, *An adaptive algorithm for unconstrained optimization with an application to optimal control*, this Journal, 11 (1973), pp. 80–94.

[6] O. Pironneau and E. Polak, *On the rate of convergence of some methods of centers*, Memo. ERL-M296, University of California, Berkeley, 1971.

[7] E. Polak, *Computational Methods in Optimization*, Academic Press, New York, 1971.

[8] ———, *On a class of numerical methods with an adaptive integration subprocedure for optimal control problems*, Proc. 4th IFIP Colloquium on Optimization, Santa Monica, Calif., 1971.

[9] F. A. Valentine, *The problem of Lagrange with differential inequalities as added side condition*, Contributions to the Calculus of Variables 1933–1937, Department of Mathematics, University of Chicago, University of Chicago Press, Chicago.

SIAM J. CONTROL
Vol. 11, No. 4, November 1973

ON STATE REALIZATION AND CAUSALITY DECOMPOSITION FOR NONLINEAR SYSTEMS*

ROMANO M. DESANTIS†

Abstract. While most of the research on the subject has either concentrated on the formulation and analysis of a state realization, or has dealt exclusively with causality decomposition, in this paper attention is focused on the interconnections between these two types of studies. More specifically, the following basic questions are considered: Does the knowledge of the causality structure of a system supply any information about the system state structure? Can the state structure tell something about the causality structure? The proposed answer can in part be summarized as follows. If a system is represented by the sum of strongly causal, strongly anticausal, memoryless, and strongly crosscausal parts, then strongly anticausal and memoryless components have nothing to do with a state realization of the system. On the other hand, any state realization of the system defines uniquely its strongly causal and strongly crosscausal parts. These and other relevant results are stated in quite a general context. They are applicable, for example, to systems which can be time-invariant or time-variant, linear or nonlinear, discrete-time or continuous-time or hybrid, finite state or infinite state.

1. Introduction. A common feature of most of the developments concerned with the notion of state and those associated with the notion of causality is that they are essentially based on the time related properties of systems. This is natural if one considers that, for instance, the state of a system at a given time gives the description of how past and present configuration of the input to the system is going to affect the future configuration of the output. On the other hand, causality related concepts such as memorylessness or crosscausality can be viewed as basic tools to characterize the various modes by which past, present and future configuration of the input to the system affect the present configuration of the output.

In spite of this conceptual link between state and causality developments, most of the studies on the subject have been carried out independently. They have either concentrated on the formulation and analysis of state concepts (as, for example, [2], [9], [15], [19]) or they have dealt exclusively with causality concepts (as, for example, [3], [10], [11], [13]). As a consequence, little has been done to clarify relevant interconnections such as those, for instance, that one intuitively expects to exist between the state related structure of a system, and its causality related structure. In particular, for example, the following natural question arises: "Does the knowledge of the causality structure of a system supply any information about the system state structure?" Conversely one also wonders whether the state structure might tell something about the causality structure.

The present study is addressed to questions of the above type, and it can be viewed as an extension of some preliminary work recently done by Saeks [12]. In particular, while this latter reference is confined to linear systems in Hilbert spaces, here nonlinear systems in nonlinear spaces are considered. This explains in part some of the unconventional features of the development, namely, the adoption of a

* Received by the editors May 11, 1972, and in revised form November 11, 1972.

† Division d'Automatique, Dèpartement de Gènie Electrique, Ecole Polytechnique, Montrèal, Quebec, Canada. This research was supported in part by the United States Air Force Office of Scientific Research, under Grant 72-2328, and by the Canadian Research Council, under Grant A-8244.

group-valued function space setting and a somewhat unusual formulation of the concept of state.

The development is organized as follows. The objective of the first two sections is to make the paper self-contained. In particular, § 2 describes the mathematical framework in which the study is embedded, and § 3 reports the most relevant causality definitions and results which will be used. The basic notion of state and associated concepts are introduced and discussed in §§ 4 and 5. Though these sections can be considered as a part of the preliminary setting for the intended study, in view of the originality of the formulation and the generality of the results, they should be of interest in themselves. The results in § 6 provide the main contribution of the development. These results give mathematical rigor to the heuristic idea that whenever a system has something special with respect to its state structure, then it has also something special with respect to its causality structure, and conversely. The paper is finally completed by § 7 which offers a critical overview of what has been presented.

2. Mathematical preliminaries.[1] Given a group G and a linearly ordered set v, the space $S[G, v]$ is defined as follows:

$$S[G, v] = \{x | x \text{ is a mapping } v \to G\}.$$

In words: if $x \in S[G, v]$, then x associates to every $t \in v$ a well-defined element $x(t) \in G$. The space $S[G, v]$ comes equipped with a family of truncation operators P^t, P_t. These operators act on $S[G, v]$ and are defined as follows: If x, y, w belong to $S[G, v]$ and $y = P^t x$, $w = P_t x$, then

$$y(s) = \begin{cases} x(s) & \text{for } s \leqq t, \\ \varnothing & \text{otherwise}, \end{cases}$$

$$w(s) = \begin{cases} x(s) & \text{for } t \leqq s, \\ \varnothing & \text{otherwise}, \end{cases}$$

where the symbol $\varnothing$ indicates the null element of G. The space $S[G, v]$ together with the family of truncation operators will be called a *group resolution space* (in short GRS).

The family of operators $dP(t): S[G, v] \to S[G, v]$ is defined as follows: if x and y belong to $S[G, v]$ and $y = dP(t)x$, then $y(s) = x(s)$ if $s = t$ and $y(s) = \varnothing$ otherwise. Clearly $dP(t) = P^t P_t = P_t P^t$. For later use the notations $\bar{P}^t$ and $\bar{P}_t$ will indicate the following operators:

$$\bar{P}^t = P^t - dP(t), \qquad \bar{P}_t = P_t - dP(t).$$

Occasionally $dP(t)x$, $\bar{P}^t x$ and $\bar{P}_t x$ are referred to respectively as *present*, *past* and *future* of x at *time* t.

A GRS is given a group structure in the following way. If $x, y, z \in S[G, v]$, then z is the sum of x and y, $z = x + y$, when $z(s) = x(s) + y(s)$ for all $s \in v$. Similarly z is the *difference* between x and y, $z = y - x$, if $y(s) = z(s) + x(s)$. By virtue of these

[1] Most of the material appearing in this section is taken directly from [3] and is reported here to render the paper self-contained.

definitions, every $x \in S[G, v]$ can be represented as follows: $x = \sum_{t \in v} dP(t)x$. This sum of infinite elements is well-defined since $dP(t)x(s) \neq \varnothing$ only when $s = t$, and if the convention is adopted that in G, the sum of infinite times, the element $\varnothing$ is still $\varnothing$. With this understanding the following additional representations will also be used:

$$P^t x = \sum_{s \leqq t} dP(s)x, \qquad P_t x = \sum_{s \geqq t} dP(s)x.$$

A *system* is given by an operator T mapping $S[G, v]$ into $S[G, v]$. If $y, x \in S[G, v]$ and $x = Ty$, then y and x are respectively called the *input* and the *output* of the system. Consider now systems (operators) T, T', T''. The system T is the *sum* of T' and T'', $T = T' + T''$ if for every $x \in S[G, v]$ the following relation holds: $Tx = T'x + T''x$. T is the *difference* between T' and T'', $T = T' - T''$, if $Tx = T'x - T''x$. T is the *composition* of T' and T'', $T = T'T''$, if for every $x \in S[G, v]$ the following relation holds: $Tx = T'(T''x)$. T is *unbiased* if $T[\varnothing] = \varnothing$, where $\varnothing$ indicates the element x in $S[G, v]$ such that $x(s) = \varnothing$ (the null element in G) for all $s \in v$. In the sequel only unbiased operators will be considered. T is *weakly additive* if it has the following property:

$$Tx = TP^t x + T\bar{P}_t x = T\bar{P}^t x + TP_t x$$

for all $x \in S[G, v]$ and every $t \in v$.

3. Basic causality concepts. As in the previous section, the notation T is used to indicate an (unbiased) system and x and y indicate two elements in $S[G, v]$. The basic causality concepts to be considered in the sequel are as follows.

DEFINITION 3.1. T is *causal* (anticausal) if

$$P^t Tx = P^t Ty \quad \text{when } P^t x = P^t y \qquad (P_t Tx = P_t Ty \quad \text{when } P_t x = P_t y).$$

DEFINITION 3.2. T is *memoryless* if it is simultaneously causal and anticausal.

DEFINITION 3.3. T is *strongly causal* or *strongly anticausal* if T is respectively causal or anticausal and

$$T\,dP(t)x = \varnothing.$$

DEFINITION 3.4. T is *strictly causal* (strictly anticausal) if

$$P^t Tx = P^t Ty \quad \text{when } P^t x = P^t y \qquad (P_t Tx = P_t Ty \quad \text{when } P_t x = P_t y).$$

DEFINITION 3.5. T is *crosscausal* if

$$P_t Tx = \varnothing \quad \text{when } P_t x = \varnothing, \quad P^t Tx = \varnothing \quad \text{when } P^t x = \varnothing, \quad \text{and} \quad T\,dP(t)x = \varnothing.$$

DEFINITION 3.6. T is *strongly crosscausal* if

$$P_t Tx = \varnothing \quad \text{when } \bar{P}_t x = \varnothing, \quad \text{and} \quad P^t Tx = \varnothing \quad \text{when } \bar{P}^t x = \varnothing.$$

For a relevant discussion about the properties of the above concepts the interested reader is referred to [3] or [4]. Here, for brevity, we limit ourselves to reporting the following useful results.

LEMMA 3.1 [3]. *The following statements are equivalent*:

(a) *T is causal.*

(b) *$P^t T = P^t TP^t$ for all $t \in v$.*

(c) *T has the following representation*: $T = \sum_{s \in v} dP(s)TP^s$.

LEMMA 3.2 [3]. *Every system can be uniquely decomposed into the sum of strongly causal, strongly anticausal, strongly crosscausal and memoryless parts.*

LEMMA 3.3 [4]. *Every system can be uniquely decomposed into the sum of strictly causal, strictly anticausal, crosscausal, and memoryless parts.*

In analogy with the procedure adopted in [3], in the sequel we will use the following alphabetic code:

A = anticausal,	$\bar{C}$ = strictly causal,
C = causal,	X = crosscausal,
M = memoryless,	$\underline{X}$ = strongly crosscausal,
$\bar{A}$ = strictly anticausal,	$\underline{A}$ = strongly anticausal,
	$\underline{C}$ = strongly causal.

4. The concept of state. In view of the rich and authoritative literature on the subject, the task of formalizing a concept of state may appear, at first, as a simple routine exercise. This task appears no longer simple, however, if one considers the constraint of satisfying the following two partially competing requirements: the concept has to be of sufficient generality to encompass the variety of systems of interest, and, at the same time, it has to be maneuverable enough so as to lead the investigation to meaningful results. A state formulation of the type proposed by Mesarovich [9], or Zadeh [15], for example, would, no doubt, be quite adequate in regard to generality. Such a procedure, however, appears to lack the maneuverability requirement.

A formulation of state via a state realization format, as recently proposed by Saeks [12] (see also [16]), is another potential alternative. Translated in the framework of GRS, this formulation becomes the following: a state realization of a system T on $S[G, v]$ is a triple (ψ, S, ζ), where S is a set (state set) and ψ, ζ are families of mappings, indexed by $t \in v$, such that

$$\psi(t): S[G, v] \to S, \quad \psi(t) = \psi(t)\bar{P}^t; \qquad \zeta(t): S \to S[G, v], \quad \zeta(t) = \bar{P}_t\zeta(t)$$

and

$$\zeta(t)\psi(t) = \bar{P}_t T P^t \quad \text{for every } t \in v.$$

This formulation, though satisfactory in a linear systems context, is not, however, general enough as to accommodate the nonlinear framework under consideration.

A shortcoming of the above definition resides in the easily verifiable fact that, in the case of nonlinear systems, it does not satisfy the "consistency conditions" which are usually associated to a meaningful concept of state (see, for instance, Arbib [1], Mesarovich [9] or Zadeh [15]). As a consequence, it can happen that the knowledge of the state of a system T at a certain time, t, may not be sufficient to give a full description of the behavior of T after t. This shortcoming, however, can be eliminated by introducing the following generalization. In stating this generalization, S_1, S_2 are sets and $S = S_1 \times S_2$ plays the role of the state set. Similarly $\psi_1, \psi_2, \zeta_1, \zeta_2$ are parametrized families of mappings, while $\psi = (\psi_1, \psi_2)$ and $\zeta = (\zeta_1, \zeta_2)$ play the role of input-to-state and state-to-output mappings. The proposed format definition is then the following.

DEFINITION 4.1. A *state realization* of a system T is a triple (ψ, S, ζ) such that:

(i) $S = S_1 \times S_2$ is the *state set*;

(ii) $\psi = (\psi_1, \psi_2)$ is such that the domain of $\psi_i(t)$ is $P^tS[G, v]$, and the range of $\psi_i(t)$ is contained in S_i, $i = 1, 2, t \in v$;

(iii) $\zeta = (\zeta_1, \zeta_2)$ is such that the domain of $\zeta_i(t)$ is S_i, the range of $\zeta_1(t)$ is contained in $\bar{P}_tS[G, v]$, and the range of $\zeta_2(t)$ is contained in the space of operators on $\bar{P}_tS[G, v]$;

(iv) for every $x, y \in S[G, v]$ the following relations are satisfied:

$$\zeta_1(t)\psi_1(t)P^tx = \bar{P}_tTP^tx,$$

$$\zeta_2(t)(\psi_2(t)[P^tx])\bar{P}_ty = \delta T_{P^tx}[\bar{P}_ty],$$

where the symbol $\delta T_x[y]$ stands for $T[x + y] - Tx - Ty$.

Note that, according to the above definition, a state realization of a system can be viewed as given by two components. The first component, (ψ_1, S_1, ζ_1), reflects the influence of the past of the input over the future of the output with the constraint that the future of the input is null. The second component, (ψ_2, S_2, ζ_2), indicates how the past of the input influences the effect of the future of the input over the future of the output. Thus, if at a given $t_0 \in v$ the system T has the initial state (s_1, s_2), then the behavior of T after t_0 can be described as follows:

If $y, x \in \bar{P}_{t_0}S[G, v]$ and $x = \bar{P}_{t_0}Ty$, then

$$x = \zeta_1(t_0)s_1 + \zeta_2(t_0)(s_2)y + \bar{P}_{t_0}Ty.$$

This formulation is clearly a natural extension of that adopted by Saeks, and the main difference between the two consists of the appearance of (ψ_2, S_2, ζ_2), the second component of the state. This inclusion is essential in that it allows our definition to satisfy the abovementioned consistency conditions. On the other hand, as it will become clear from Proposition 6.1, the concept of state proposed in [12] satisfies these conditions only in the particular case in which the system is "separable" (in particular: linear). This explains why Saeks did not need a format of the present generality.

By the substitution of the term "state" with "costate" and by a subsequent substitution of symbols of the type indicated in the forthcoming principle of causal duality, Definition 4.1 provides a concept which is dual to that of a state realization. As this concept will also be useful, the following additional definition is introduced.

DEFINITION 4.2. A *costate realization* of a system T is a triple (β, Ω, γ) such that:

(i) $\Omega = \Omega_1 \times \Omega_2$ is the *costate* set;

(ii) $\beta = (\beta_1, \beta_2)$ is such that the domain of $\beta_i(t)$ is $P_tS[G, v]$, and the range of $\beta_i(t)$ is contained in Ω_i, $i = 1, 2, t \in v$;

(iii) $\gamma = (\gamma_1, \gamma_2)$ is such that the domain of $\gamma_i(t)$ is Ω_i, the range of $\gamma_1(t)$ is contained in $\bar{P}^tS[G, v]$, and the range of $\gamma_2(t)$ is contained in the space of operators on $\bar{P}^tS[G, v]$;

(iv) for every $x, y \in S[G, v]$ the following relations are satisfied:

$$\beta_1(t)\gamma_1(t)P_tx = \bar{P}^tTP_tx,$$

$$\beta_2(t)(\gamma_2(t)[P_tx])\bar{P}^ty = \delta T_{P_tx}[\bar{P}^ty].$$

For later use, the section is concluded with the following statements.

PROPOSITION 4.1. *Let (ψ', S', ζ') and (ψ'', S'', ζ'') be two state realizations for T' and T'' respectively. Then an admissible state realization for $T = T' + T''$ is given by the triple (ψ, S, ζ), where $\psi_i(t) = (\psi_i'(t), \psi_i''(t))$, $S_i = S_i' \times S_i''$ and $\zeta_i(t)s_i = \zeta_i'(t)s_i' + \zeta_i''(t)s_i''$, where $(s_i', s_i'') \in S_i' \times S_i''$ for $i = 1, 2$, and all $t \in v$.*

PROPOSITION 4.2 (Principle of causal duality). *Let a statement or equality be phrased using relations involving the causality alphabet $\mathscr{A} = \{A, C, M, X, \bar{A}, \bar{C}, \underline{A}, \underline{C}, \underline{X}\}$, the families of projection operators $P^t, P_t, \bar{P}^t, \bar{P}_t$, the families of mappings ψ, β, ζ, γ, and the sets S and Ω. Then this statement or equality remains valid if the following interchange in symbols occurs:*

$$\begin{array}{llll}
P^t \to P_t, & \bar{P}^t \to \bar{P}_t, & P_t \to P^t, & \bar{P}_t \to \bar{P}^t, \\
C \to A, & \underline{C} \to \underline{A}, & A \to C, & \underline{A} \to \underline{C}, \\
\bar{C} \to \bar{A}, & \bar{A} \to \bar{C}, & \underline{X} \to \underline{X}, & X \to X, \\
\psi \to \beta, & \zeta \to \gamma, & \beta \to \psi, & \gamma \to \zeta, \\
S \to \Omega, & \Omega \to S. & &
\end{array}$$

5. State related concepts and properties. The structure of systems with respect to the notion of state can be diverse. One way to describe the variety of situations which can occur is to start by introducing the concepts of controllability, observability, minimality, equivalence and separability. For details on the genesis of these concepts the interested reader is referred to [1], [5], [7].

DEFINITION 5.1. The state realization (ψ, S, ζ) is *completely controllable* with respect to the first (second) component if $\psi_1(t)$ $(\psi_2(t))$ is onto.

DEFINITION 5.2. The state realization (ψ, S, ζ) is *completely observable* with respect to the first (second) component if $\zeta_1(t)$ $(\zeta_2(t))$ is one-to-one.

DEFINITION 5.3. The state realization is *minimal* with respect to the first (second) component if it is completely controllable and completely observable with respect to the first (second) component.

DEFINITION 5.4. A state realization is *strictly minimal* (strictly completely controllable, strictly completely observable) if it is minimal (completely controllable, completely observable) with respect to both first and second components.

DEFINITION 5.5. Two state realizations (ψ, S, ζ) and (ψ', S', ζ') are *strictly equivalent* if there exists a unique family of invertible mappings K indexed by $t \in v$ such that the following conditions are satisfied:

$$\psi(t) = K(t)\psi'(t) \quad \text{and} \quad \zeta(t) = \zeta'(t)K^{-1}(t).$$

DEFINITION 5.6. A system T with state realization (ψ, S, ζ) is *separable* if for all $t \in v$ and $x \in S[G, v]$ one has

$$\bar{P}_t T x = \zeta_1(t)s_1 + \bar{P}_t T \bar{P}_t x, \quad \text{where } s_1 = \psi_1(t)x.$$

In the sequel it will be said that a state realization (ψ, S, ζ) has its first component given by the triple $(\varnothing, \varnothing, \varnothing)$, if S_1 has only one element, indicated by $\varnothing$, and $\zeta_1(t)$ maps this element into the zero element of $\bar{P}_t S[G, v]$. Similarly, it will be said that a state realization (ψ, S, ζ) has its second component given by the triple

$(\varnothing, \varnothing, \varnothing)$ if S_2 has only one element (again indicated by $\varnothing$) and $\zeta_2(t)$ maps this element into the null operator on $\bar{P}_tS[G, \nu]$. Finally, a state realization is given by the triple $(\varnothing, \varnothing, \varnothing)$ if both its first and second components are given by the triple $(\varnothing, \varnothing, \varnothing)$.

PROPOSITION 5.1. *A system is separable if and only if it admits a state representation whose second component is given by the triple* $(\varnothing, \varnothing, \varnothing)$.

Note that our formulation of state and state related concepts is more general than that usually encountered in the technical literature. To illustrate some of its powerful features we introduce the following relevant results. To begin with, Proposition 5.2 gives a connection between the mathematical concept of minimal state realization and the intuitive notion of a state set which contains only "essential" elements.

PROPOSITION 5.2. *Suppose that the completely observable* (ψ, S, ζ) *and the strictly minimal* $(\underline{\psi}, S, \underline{\zeta})$ *are admissible realizations of a given system. Then there exists a unique family of mappings* K *taking* S *into* $\underline{S}$ *and such that*

$$\underline{\psi}(t) = K(t)\psi(t) \quad \textit{and} \quad \zeta(t) = \underline{\zeta}(t)K(t).$$

The next proposition shows that all strictly minimal state realizations of a system are strictly equivalent. This is a generalization of a similar well-known result in the context of linear dynamical systems [7], [15]. A result of this type was also given by Saeks for linear systems in Hilbert space [12].

PROPOSITION 5.3. *All strictly minimal state realizations are strictly equivalent.*

An additional property of a strictly minimal state realization is that it allows one to define the notion of a "transition operator." The proof is illustrative of the techniques used to obtain all these types of results and is therefore given in detail.

THEOREM 5.1. *Suppose that* (ψ, S, ζ) *is a strictly minimal state realization, and for* $q \leqq t$, $t, q \in \nu$, *assume that no input to the system occurs between* q *and* t. *Then there exists a well-defined mapping* $\Phi(t, q)$ *such that*

$$v(t) = \Phi(t, q)v(q),$$

where v *is the state of the system.*

Proof. To every element $t \in \nu$ and each element $v \in S$, associate $U(t)$, a subset of $P^tS[G, \nu]$, such that $U(t)$ is maximal with respect to the following property: if $u \in U(t)$, then $v = \psi(t)u$. This defines the family of mappings $M(t): S \to \mathbb{P}\{S[G, \nu)\}$, where $\mathbb{P}\{S[G, \nu]\}$ indicates the power set of $S[G, \nu]$. Conversely, to every subset $U(t) \subseteq P^tS[G, \nu]$ with the property that there exists a $v \in S$ such that

$$v = \psi(t)u, \quad \text{for any } u \in U(t),$$

associate the element v. This defines a family of mappings

$$[M(t)]^{-1} : \mathbb{P}\{S[G, \nu]\} \to S.$$

Let $t, q \in \nu$, $t \geqq q$, and suppose that $\underline{u}$ is an input to T such that $[P^t - P^q]\underline{u} = \varnothing$. Consider the states $v(t)$ and $v(q)$ given by the following equations:

$$v(t) = \psi(t)P^t\underline{u}, \qquad v(q) = \psi(q)P^q\underline{u}. \tag{5.1}$$

In what follows it is shown that

$$v(t) = [M(t)]^{-1}M(q)v(q), \tag{5.2}$$

where $\Phi(t, q) = [M(t)]^{-1}M(q)$ is the mapping whose existence was to be proved.

The proof of equation (5.2) is equivalent to the proof of the following equation:

$$M(q)v(q) \subseteq M(t)v(t).$$

Suppose that $u \in U(q) = M(q)v(q)$. From the definition of state realization and equation (5.1) we have the following:

$$\zeta(q)v(q) = \begin{bmatrix} \zeta_1(q)v_1(q) \\ \zeta_2(q)v_2(q) \end{bmatrix} = \begin{bmatrix} \bar{P}_q T P^q \underline{u} \\ \bar{P}_q \delta T_{P^q \underline{u}}[\ ] \end{bmatrix} = \begin{bmatrix} \bar{P}_q T P^q u \\ \bar{P}_q \delta T_{P^q u}[\ ] \end{bmatrix}$$

for all $u \in U(q)$. It follows that

$$\zeta(t)v(t) = \begin{bmatrix} \zeta_1(t)v_1(t) \\ \zeta_2(t)v_2(t) \end{bmatrix} = \begin{bmatrix} \bar{P}_t T P^t \underline{u} \\ \bar{P}_t \delta T_{P^t \underline{u}}[\ ] \end{bmatrix}$$

$$= \begin{bmatrix} \bar{P}_t T P^q \underline{u} \\ \bar{P}_t \delta T_{P^q \underline{u}}[\ ] \end{bmatrix} = \begin{bmatrix} \bar{P}_t \bar{P}_q T P^q u \\ \bar{P}_t \bar{P}_q \delta T_{P^q u}[\ ] \end{bmatrix} = \zeta(t)\psi(t)P^q u \quad \text{for all } u \in U(q).$$

From the above equation and the fact that $\zeta(t)$ is one-to-one we have

$$v(t) = \psi(t)P^q u \quad \text{for all } u \in U(q)$$

and therefore $U(q) \subseteq U(t)$. The proof is thus complete.

By identifying the operator $\Phi(t, q)$ as the *transition operator* of the state decomposition, Theorem 5.1 can be viewed as a far reaching generalization of a similar well-known result for linear dynamical systems [15]. This proposition is also an extension of a perhaps less familiar result about the properties of a state decomposition for a linear and bounded system in a Hilbert resolution space [12]. The extent of these connections is further illustrated by the following corollary.

COROLLARY 5.1. *Let $\Phi(t, q)$ be the transition operator of a strictly minimal state realization (ψ, S, ζ). Then*

$$\Phi(t, t) = I,$$
$$\Phi(t, r) = \Phi(t, q)\Phi(q, r), \qquad t \geqq q \geqq r.$$

6. Connections between causality and state concepts. We begin with a connection among the concepts of separability, weak additivity and causality. This connection creates a link among various apparently independent developments such as those, for instance, in [1], [3] and [6].

THEOREM 6.1. *If a system is weakly additive, then it is separable. Conversely, if a system is causal and separable, then it is also weakly additive.*

Proof. Suppose that T is weakly additive. By Proposition 5.1, to prove that T is separable it is sufficient to show that it has a state decomposition whose second component is given by the triple $(\varnothing, \varnothing, \varnothing)$. To this purpose, choose a triple (ψ, S, ζ), where $\psi = (\psi, \varnothing)$, $S = (S_1, \varnothing)$, $\zeta = (\zeta_1, \varnothing)$, and (ψ_1, S_1, ζ_1) is any admissible first component for a state decomposition of T. To prove that (ψ, S, ζ) is admissible it is sufficient to show that $(\psi_2, S_2, \zeta_2) = (\varnothing, \varnothing, \varnothing)$ satisfies the requirements of Definition 3.1. That this is indeed the case follows from the relation

$$\bar{P}_t \delta T_{P^t u}[\bar{P}_t x] = \bar{P}_t T[\bar{P}_t x + P^t u] - \bar{P}_t T \bar{P}_t x - \bar{P}_t T P^t u = \varnothing$$

which is valid for all $\bar{P}_t x \in \bar{P}_t S[G, v]$ and $P^t u \in P^t S[G, v]$.

Suppose now that T is causal and separable. If T were not weakly additive, then there would exist a $t \in v$ and $x \in S[G, v]$ such that

$$Tx \neq TP^t x + T\bar{P}_t x.$$

As T is causal this inequality can hold if and only if

$$\bar{P}_t Tx \neq \bar{P}_t TP^t x + \bar{P}_t T\bar{P}_t x.$$

This latter inequality cannot hold, however, because, from the separability of T, there exists a state representation (ψ, S, ζ) such that

$$\bar{P}_t Tx = \zeta_1(t)s_1 + \bar{P}_t T\bar{P}_t x,$$

where $s_1 = \psi_1(t)P^t x$ and $\zeta_1(t)\psi_1(t)P^t x = \bar{P}_t TP^t x$. One can then conclude that T is weakly additive.

The next theorem formalizes the intuitive idea that if a system is anticausal, then its state space should be vacuous. This is a natural extension of a similar result in [12].

THEOREM 6.2. *If a system is anticausal, then it admits a state realization given by the triple* $(\varnothing, \varnothing, \varnothing)$. *Conversely, if the triple* $(\varnothing, \varnothing, \varnothing)$ *is an admissible state realization, then the system must be anticausal.*

Proof. Suppose that T is anticausal. Then for every element $u \in S[G, v]$ and $t \in v$ we have $\bar{P}_t Tu = \bar{P}_t T\bar{P}_t u$ and therefore $\bar{P}_t TP^t u = \varnothing$. Similarly, for any $x \in \bar{P}_t S[G, v]$ and $t \in v$ we have

$$\bar{P}_t \delta T_{P^t u}[x] = \bar{P}_t T[P^t u + \bar{P}_t x] - \bar{P}_t TP^t u - \bar{P}_t T\bar{P}_t x = \varnothing$$

and, from here, $\bar{P}_t \delta T_P t_u = \varnothing$. Choose now the triple (ψ, S, ζ) such that $S = (\varnothing, \varnothing)$, $\psi(t) = (\varnothing, \varnothing)$ and $\zeta(t) = (\varnothing, \varnothing)$. In view of the above equations it is trivial to verify that this triple satisfies the requirements of Definition 4.1 and therefore it is an admissible state realization for T. Conversely, suppose that the triple $(\varnothing, \varnothing, \varnothing)$ is an admissible state realization for T. Then from Definition 4.1, we have

$$\bar{P}_t TP^t = \varnothing \quad \text{and} \quad \bar{P}_t T[P^t u + \bar{P}_t x] - \bar{P}_t TP^t u - \bar{P}_t T\bar{P}_t x = \varnothing.$$

From this it follows that $\bar{P}_t T = \bar{P}_t T\bar{P}_t$ and applying the principle of causal duality to Lemma 3.1 we can conlude that T is anticausal.

COROLLARY 6.1. *A necessary and sufficient condition for a system to be memoryless is that the triple* $(\varnothing, \varnothing, \varnothing)$ *be simultaneously an admissible state and costate realization.*

COROLLARY 6.2. *If a system is anticausal, then it admits a strictly minimal state realization.*

When a system is crosscausal or when it is given by the sum of crosscausal and anticausal parts, then Theorem 6.2 has the following counterpart. The proof utilizes a line of reasoning already seen, and is omitted for brevity.

THEOREM 6.3 [4]. *A system is given by the sum of a crosscausal plus an anticausal part if and only if it admits a state realization with the first component given by the triple* $(\varnothing, \varnothing, \varnothing)$.

COROLLARY 6.3. *A crosscausal system admits a state realization which is minimal with respect to its first component.*

Using Proposition 4.1, the foregoing results can be applied in more general situations. This is illustrated by the following.

THEOREM 6.4. *Every system has a state realization with the property that the first component depends only on the strictly causal part.*

Proof. Applying Lemma 3.3, T can be represented as follows:

$$T = T_{\bar{A}} + T_{\bar{C}} + T_M + T_X.$$

Suppose that $(\psi_\alpha, S_\alpha, \zeta_\alpha)$, $\alpha \in \{\bar{A}, \bar{C}, M, X\}$, are state realizations for $T_{\bar{A}}, T_{\bar{C}}, T_X$ and T_M respectively. By Theorems 6.1 and 6.2, $(\psi_\alpha, S_\alpha, \zeta_\alpha)$, $\alpha \in \{\bar{A}, M, X\}$, can be chosen in such a way that the first component is given by the triple $(\varnothing, \varnothing, \varnothing)$. A state representation (ψ, S, ζ) for T can now be obtained by the procedure indicated in Proposition 4.1. The first component of this state realization depends clearly only on $T_{\bar{C}}$.

THEOREM 6.5. *If the strongly causal part of a system is weakly additive, then there exists a state realization such that the first component depends only on the strictly causal part and the second component depends only on the strongly crosscausal part.*

Proof. Consider again the representation $T = T_{\bar{A}} + T_{\bar{C}} + T_M + T_X$. By the application of Theorems 6.2 and 6.3 it is possible to find state realizations for $T_{\bar{A}}$, T_M and $T_{\bar{C}}$ with the property that the second component is given by the triple $(\varnothing, \varnothing, \varnothing)$. The rest of the proof can be obtained through an argument identical to that used in the proof of Theorem 6.4.

THEOREM 6.6. *A necessary and sufficient condition for a system to admit a state decomposition with a first component given by the triple* $(\varnothing, \varnothing, \varnothing)$ *is that the strictly causal part is null.*

To help motivate the next result observe that a system can clearly have many state realizations, and, conversely a state realization can correspond simultaneously to many systems. The natural question arises then about what must be in common to two systems in order that they can both have the same state realization.

THEOREM 6.7. *A state realization of a system defines uniquely its strongly causal and strongly crosscausal parts.*

Proof. We have to show that if two systems T and T' admit an identical state realization (ψ, S, ζ), then $T_{\underline{C}} = T'_{\underline{C}}$ and $T_{\underline{X}} = T'_{\underline{X}}$, where $T_{\underline{C}}, T_{\underline{X}}$, and $T'_{\underline{C}}, T'_{\underline{X}}$ are strongly causal and strongly crosscausal parts respectively of T and T'. By the application of Proposition 4.1, a state realization of $T - T'$ is given by $(\psi, \tilde{S}, \tilde{\zeta})$, where using the notation of Proposition 4.1:

$$\tilde{\psi}_i(t) = (\psi_i(t), \psi_i(t)), \qquad \tilde{S}_i \subseteq S_i \times S_i;$$

and $\tilde{\zeta}(t)$ is defined as follows: if $\tilde{v}_i = (v_i, v_i) \in S_i$, then $\tilde{\zeta}(t)\tilde{v}_i = \zeta_i v_i - \zeta_i v_i$. Clearly $(\tilde{\psi}, \tilde{S}, \tilde{\zeta})$ has the property that $\tilde{\zeta}_i(t)\tilde{\psi}_i(t) = \varnothing$, for $i = 1, 2$. From Definition 4.1 it follows that $\bar{P}_t(T - T')P^t = \varnothing$ and $\bar{P}_t\delta T_{P^t x}[\ ] = \varnothing$ for every $x \in S[G, v]$ and $t \in v$. This implies that the triple $(\varnothing, \varnothing, \varnothing)$ is an admissible state decomposition for $T - T'$. From Theorem 6.2 it follows that $T - T'$ must be anticausal. This means that $T_{\underline{C}} + T_{\underline{X}} - T'_{\underline{C}} - T'_{\underline{X}} = \varnothing$. From Lemma 3.2 it then follows that $T_{\underline{C}} = T'_{\underline{C}}$ and $T_{\underline{X}} = T'_{\underline{X}}$.

Interesting aspects of the above result are illustrated by the following corollaries.

COROLLARY. *A state and a costate realization of a system define uniquely its strongly causal, strongly anticausal and strongly crosscausal parts.*

COROLLARY. *If a system is given by the sum of strongly causal and strongly crosscausal parts, then it is completely defined by any one of its state realizations.*

COROLLARY. *Suppose that (ψ, S, ζ) and (ψ', S', ζ') are state realizations of T and T'. If (ψ, S, ζ) and (ψ', S', ζ') are strictly equivalent, then T and T' have identical strongly causal and strongly crosscausal parts.*

Note that the above statements provide results which in the context of linear dynamical systems are well known. In particular, for instance, from the last corollary one obtains that two strictly equivalent dynamical systems have the same weighting pattern [7].

Summary. The paper considers systems defined in a group-valued function space, and gives various interconnections between causality related structure on the one hand and state related structure on the other. To do this the investigation utilizes a causality decomposition setting of the kind proposed in [3], and a modified state realization format of the type adopted in [12]. The significance of our state realization format is illustrated by the generalization of a number of results previously confined to the case of linear systems. In this regard, for example, Theorem 5.1 shows that the concept of a transition operator, usually associated to linear dynamical systems, is meaningful in a much wider, nonlinear, nondynamical, and noncausal context.

The main results of the paper are contained in § 6. More specifically, Theorem 6.1 establishes a connection between the concept of separability, and the concepts of causality and weak additivity. Theorems 6.2–6.7 are all addressed to clarify the connection between state realization and causality decomposition. In particular, Theorem 6.2 states that a system is anticausal if and only if it admits a trivial state realization. Similar results for the case of systems with a more involved causality structure are given by Theorems 6.3 and 6.6. For example, Theorem 6.6 states that a system has a null strictly causal part if and only if it admits a state realization with a trivial first component.

Theorems 6.4, 6.5 and 6.7 provide further clarifications about what causality related information might be obtained from a state realization. In particular, using Theorem 6.7 it follows that from the knowledge of a state realization of the system one can reconstruct its strongly causal, strictly causal and strongly crosscausal parts, Moreover, if the causal part of the system is weakly additive, then, by Theorem 6.5, the strongly causal and the strongly crosscausal parts of the system are uniquely defined respectively by the first and the second component of the state realization.

REFERENCES

[1] M. ARBIB, *Automata theory and control theory—A rapprochement*, Automatica, 3 (1966), pp. 161–189.

[2] A. V. BALAKRISHNAN, *State space theory of linear time-varying systems*, System Theory, L. A. Zadeh and E. Polak, eds., McGraw-Hill, New York, 1969, pp. 95–126.

[3] R. M. DESANTIS and W. A. PORTER, *On time-related properties of nonlinear systems*, SIAM J. Appl. Math., 24 (1973), pp. 188–206.

[4] R. M. DESANTIS, *Causality structure of engineering systems*, Ph.D. thesis, The University of Michigan, Ann Arbor, 1971.

[5] G. EVANGELISTI, *Controllability and Observability*, C.I.M.E. Edizioni Cremonese, Roma, 1969.

[6] A. GERSHO, *Nonlinear systems with a restricted additivity property*, IEEE Trans. Circuit Theory, CT-16 (1969), pp. 150–154.

[7] R. E. KALMAN, *Mathematical description of linear dynamic systems*, this Journal, 1 (1963), pp. 159–192.

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

[9] M. MESAROVICH, *Foundations for a general systems theory*, Views on General Systems Theory, M. Mesarovich, ed., John Wiley, New York, 1964, pp. 1–24.

[10] W. A. PORTER AND C. L. ZAHM, *Basic concepts in system theory*, Tech. Rep. SEL44, The University of Michigan, Ann Arbor, 1969.

[11] R. SAEKS, *Causality in Hilbert space*, SIAM Rev., 12 (1970), pp. 357–383.

[12] ———, *State in Hilbert space*, Ibid., 15 (1973). Also Tech. Memo. EE-6912-2, University of Notre Dame, Notre Dame, Ind., 1969.

[13] T. G. WINDEKNECHT, *Mathematical systems theory: Causality*, Math. Systems Theory J., 2 (1947), pp. 279–288.

[14] A. W. WYMORE, *A Mathematical Theory of Systems Engineering: The Elements*, John Wiley, New York, 1967.

[15] L. A. ZADEH AND C. A. DESOER, *Linear Systems Theory*, McGraw-Hill, New York, 1963.

[16] H. P. ZEIGER, *Cascade decomposition of automata using covers*, The Algebraic Theory of Machine Languages and Semigroups, M. Arbib, ed., Academic Press, New York, 1968.

SIAM J. CONTROL
Vol. 11, No. 4, November 1973

MULTIPERSON CONTROLLED DIFFUSIONS*

STANLEY R. PLISKA†

Abstract. A multiperson controlled diffusion process is formulated as both a zero sum, two-person game and a nonzero sum, N-person game. Necessary and sufficient conditions are provided for a control to be a solution of these games. The value of a game, if it exists, is shown to be the solution of a second order, nonlinear differential equation. Existence theorems are included. Discounted cost, undiscounted cost, and average cost criteria are considered.

Introduction. This paper generalizes the concept of a controlled one-dimensional diffusion process by allowing the process to be controlled by N persons. If the process is controlled by two persons with opposite objectives, then the problem of optimally controlling this process may be viewed as a zero sum, two-person game. On the other hand, if the process is controlled by $N \geqq 2$ persons with possibly different objectives, then the problem of optimally controlling this process may be viewed as a nonzero sum, N-person game.

The results in this paper are intimately connected with those for single-person controlled diffusions (see Pliska [8] and Mandl [7]). In addition, minimax problems in the theory of diffusions have been treated by Girsanov [6]. The multiperson controlled diffusion process is formulated in the following section, the zero sum, two-person game problem is discussed in the succeeding three sections, and the nonzero sum, N-person game problem is treated in the final three sections. Both discounted and undiscounted costs are considered for both game problems, and existence theorems are provided. A major result of this paper is that the value of a zero sum, two-person game is the unique solution of a differential equation.

1. The multiperson controlled diffusion process. The multiperson controlled diffusion process is formulated as in Pliska [8], only taking into account the multiple number of controllers. Consider a diffusion process with state space S, a compact interval $[r_0, r_1]$ of the real line E, which is controlled by N persons (integer $N \geqq 2$). For each $i = 1, 2, \cdots, N$, some positive integer n_i, and some compact set $K_i \subset E^{n_i}$, the ith person's control is a vector-valued function on S with range K_i. Let A_s^i be a point-to-set map from S into K_i such that A_s^i is piecewise continuous in s in the Hausdorff metric and for each $s \in S$ the set A_s^i is a nonempty compact subset of K_i. Each time the process is observed in state s, the ith person chooses an action a_i from the set A_s^i. The set M_i of admissible controls for the ith person consists of all piecewise continuous functions $a_i(s)$ on S with range in K_i such that the action $a_i(s) \in A_s^i$ for each $s \in S$.

Let $M = M_1 \times M_2 \times \cdots \times M_N$, $K = K_1 \times \cdots \times K_N$, $a(s) = (a_1(s), \cdots, a_N(s))$ and $A_s = (A_s^1, \cdots, A_s^N)$, so that M is the set of admissible controls, a function $a(s)$ is an admissible control if and only if $a(s) \in M$, and $a(\cdot) \in M$ implies $a(s) \in A_s$ for each $s \in S$. Throughout this paper it should be clear from the context whether the

* Received by the editors July 6, 1972.

† Department of Industrial Engineering and Management Sciences, Northwestern University, Evanston, Illinois 60201. This research was supported by the National Science Foundation under Grant GK-188339 and the Office of Naval Research under Contract N0014-67-A-0112-0050.

letter a denotes an admissible control $a = a(\cdot) \in M$ or an admissible action $a \in A_s$ for some $s \in S$. The map A_s is characterized in Pliska [8]. We assume $M \neq \varnothing$ hereafter without further mention.

The definition of a multiperson controlled diffusion process is a slight generalization of Mandl's [7, p. 157] controlled diffusion process. Let $d(s, a)$ be a continuous, positive real-valued function on $S \times K$. Then for $a(\cdot) \in M$ the piecewise continuous function $d(s, a(s))$ is the diffusion coefficient of the process. Similarly, let $b(s, a)$ be a continuous real-valued function on $S \times K$ so that $b(s, a(s))$ is the drift coefficient of the diffusion process.

Following Mandl, with a given control $a(\cdot) \in M$, the multiperson controlled diffusion process is completely specified by the generalized classical differential operator

$$D \equiv d(s, a(s))\frac{d^2}{ds^2} + b(s, a(s))\frac{d}{ds}$$

together with Feller's [4], [5] boundary condition

$$\kappa_j v(r_j) + \theta_j\left(v(r_j) - \int_S v(s)\, d\mu_j(s)\right) - (-1)^j \pi_j v'(r_j)$$
$$+ \sigma_j(Dv)(r_j) = 0, \qquad j = 0, 1,$$

where $v(s)$ is some function whose second derivative is piecewise continuous on S. At each boundary r_0, r_1 the four nonnegative parameters $\kappa_j, \sigma_j, \pi_j$ and θ_j, at least one of which must be positive, correspond respectively to the phenomena of absorption, adhesion, reflection and instantaneous return. Corresponding to θ_j is the probability distribution function $\mu_j(s)$, where $\int_{(r_0, r_1)} d\mu_j(s) = 1$. This boundary condition is interpreted more fully in Pliska [8].

The multiperson controlled diffusion process generates costs according to its sample path and control (Mandl [7, p. 148]). With the zero sum, two-person game problem, exactly one stream of costs is generated, as is the case with the single-person controlled diffusion (Pliska [8]). But with the nonzero sum, N-person game problem, exactly N streams of costs will be generated, with one stream corresponding to each controller. The costs of a multiperson controlled diffusion will be formulated below for the N-person problem, but it should be borne in mind that the formulation for the zero sum problem is exactly the same, except that the subscript i relating the cost streams with the controllers will be dropped.

Each cost stream is comprised of the same three types of costs that were specified in Mandl [7] and Pliska [8]. The continuous movement cost for the ith person is defined by the bounded, continuous real-valued function $c_i(s, a)$ on $S \times K_1 \times \cdots \times K_N$; let $c(s, a)$ denote the N-component vector of these functions. The cost for the ith person due to instantaneous returns from the boundary r_j is expressed by the real-valued function $v_{ji}(s)$ on S, which is integrable with respect to $\mu_j(s)$; let $v_j(s)$ denote the vector of these functions. Finally, the cost for the ith person due to the termination (absorption) of the process at boundary r_j is λ_{ji}, and λ_j denotes the vector of these costs.

If $C_i(t)$ is the total of the ith person's costs generated by the process up through time t, and $C(t) = (C_1(t), \cdots, C_N(t))$ is the vector of these costs, then the N-component vector

$$v(s) = E_s \int_0^\infty e^{-\lambda t}\, dC(t)$$

denotes the conditional expectation of the total discounted costs of the process given an initial state s, a control $a(\cdot) \in M$, and a discount factor $e^{-\lambda}$, $\lambda > 0$. From Mandl [7, p. 149] we have the following result.

THEOREM 1. *The vector of expected discounted costs corresponding to $a(\cdot) \in M$ is the unique function $v(s)$ on S such that $v'(s)$ is continuous,*

$$d(s, a(s))v''(s) + b(s, a(s))v'(s) - \lambda v(s) + c(s, a(s)) = 0 \tag{1}$$

holds for every $s \in (r_0, r_1)$ which is a continuity point of $a(s)$, and

$$\begin{aligned}(\theta_j + \kappa_j)v(r_j) - \theta_j \int_S (v(s) + v_j(s))\, d\mu_j(s) - (-1)^j \pi_j v'(r_j)\\ + \sigma_j(\lambda v(r_j) - c(r_j, a(r_j))) - \kappa_j \lambda_j = 0, \qquad j = 0, 1.\end{aligned} \tag{2}$$

If the process is nonconservative and neither boundary is purely adhesive, that is, if

$$\kappa_0 + \kappa_1 > 0, \quad \kappa_j + \pi_j + \theta_j > 0, \qquad j = 0, 1,$$

then by Mandl [7, p. 152] the vector $v(s) = E_s C(\infty)$ of the expected total undiscounted costs is finite and is the unique solution of (1) and (2) for $\lambda = 0$. If the process is conservative ($\kappa_0 + \kappa_1 = 0$), then the total undiscounted costs may be infinite. The vector $\theta = (\theta_1, \cdots, \theta_N)$ in the following theorem, which is an immediate consequence of Mandl [7, pp. 152–157, 168], can be interpreted as the vector of mean costs per unit time.

THEOREM 2. *Let $\kappa_0 = \kappa_1 = 0$ and assume at least one boundary is not purely adhesive, that is, $\pi_0 + \theta_0 + \pi_1 + \theta_1 > 0$. If $v(s, \lambda)$ is the vector of expected discounted costs corresponding to $\lambda > 0$ and some $a(\cdot) \in M$, then*

$$\lim_{\lambda \downarrow 0} \lambda v(s, \lambda) = \theta \quad \text{and} \quad \lim_{\lambda \downarrow 0} \frac{d}{ds} v(s, \lambda) = w(s),$$

where θ is some vector independent of the state s, and $w(s)$ is some absolutely continuous vector-valued function on S. Moreover,

$$P(\lim_{t \to \infty} t^{-1} C(t) = \theta) = 1,$$

and (θ, w) is the unique pair satisfying

$$d(s, a(s))w'(s) + b(s, a(s))w(s) - \theta + c(s, a(s)) = 0 \tag{3}$$

for every $s \in (r_0, r_1)$ which is a continuity point of $a(s)$, and

$$\begin{aligned}\theta_j \int_S \left\{ \int_{r_j}^s w(y)\, dy + v_j(s) \right\} d\mu_j(s) + (-1)^j \pi_j w(r_j)\\ + \sigma_j(c(r_j, a(r_j)) - \theta) = 0, \qquad j = 0, 1.\end{aligned} \tag{4}$$

2. The zero sum, two-person game problem. In this and the following two sections we consider diffusion processes which are controlled by two persons ($N = 2$), but which generate single streams of costs. The persons have opposite objectives; the first wants to minimize the costs while the other wants to maximize the costs. Note that this zero sum game can be regarded as a special case of the nonzero sum game problem for $N = 2$ by letting the second player's costs equal the negative of the first player's.

We shall consider a single stream of costs and therefore omit the subscript on all cost symbols. For any particular problem, player 1, who operates the first control component, endeavors to choose a control $a_1(\cdot) \in M_1$ so as to minimize the expected costs generated by the process. Player 2, who operates the second control component, endeavors to choose a control $a_2(\cdot) \in M_2$ so as to maximize the expected costs of the process. By a solution to this game is meant some admissible control which is a saddle point of the expected cost function. Thus, if player 1 unilaterally deviates from this optimal control, then the expected costs cannot be decreased but they may increase. Similarly, player 2 can unilaterally only decrease the expected costs.

The following two sections provide results respectively for the discounted cost case and the undiscounted cost case. The method for solving a problem is basically the same in each case. A differential equation is solved and the solution is used to determine the saddle point of a function with respect to all admissible controls. If this saddle point exists, then it is used to obtain an optimal control, that is, a solution to the zero sum, two-person game.

3. The zero sum problem with discounted costs. Let $v(s, a_1, a_2) = v(s)$ denote the expected discounted cost of a process corresponding to the admissible control $a = (a_1, a_2) \in M$. Then $v(s, a_1, a_2)$ will be the unique solution of (1) and (2). The control $\tilde{a} \in M$ is said to be optimal if for all $a_1 \in M_1$, all $a_2 \in M_2$, and all $s \in S$ we have

$$v(s, \tilde{a}_1, a_2) \leqq v(s, \tilde{a}_1, \tilde{a}_2) \leqq v(s, a_1, \tilde{a}_2)$$

in which case $v(s, \tilde{a}_1, \tilde{a}_2)$ is said to be the value of the game. We shall later prove that the value of a game, if it exists, is provided by the following.

THEOREM 3. *There exists a unique solution $v(s)$ to*

$$\begin{aligned} v''(s) + \min_{a_1 \in A_s^1} \max_{a_2 \in A_s^2} \{d(s, a_1, a_2)^{-1}[b(s, a_1, a_2)v'(s) \\ - \lambda v(s) + c(s, a_1, a_2)]\} = 0 \end{aligned} \tag{5}$$

satisfying

$$\begin{aligned} (\theta_j + \kappa_j)v(r_j) - \theta_j \int_S (v(s) + \nu_j(s))\, d\mu_j(s) - (-1)^j \pi_j v'(r_j) \\ + \sigma_j(\lambda v(r_j) - \gamma_j) - \kappa_j \lambda_j = 0, \qquad j = 0, 1, \end{aligned} \tag{6}$$

where

$$\gamma_j = \min_{a_1 \in A_{r_j}^1} \max_{a_2 \in A_{r_j}^2} c(r_j, a_1, a_2), \qquad j = 0, 1.$$

Before proving this theorem, some notation will be introduced and a number of preliminary lemmas will be proved.

Define:

$$\alpha(s, a_1, a_2) = d(s, a_1, a_2)^{-1},$$

$$\beta(s, a_1, a_2) = b(s, a_1, a_2)d(s, a_1, a_2)^{-1},$$

$$\gamma(s, a_1, a_2) = c(s, a_1, a_2)d(s, a_1, a_2)^{-1},$$

$$g_1(s, v_1, v_2) = v_2,$$

$$g_2(s, v_1, v_2) = -\min_{a_1 \in A_s^1} \max_{a_2 \in A_s^2} \{\beta(s, a_1, a_2)v_2 - \lambda\alpha(s, a_1, a_2)v_1 + \gamma(s, a_1, a_2)\},$$

$$g(s, v_1, v_2) = \begin{pmatrix} g_1(s, v_1, v_2) \\ g_2(s, v_1, v_2) \end{pmatrix}.$$

We have the following result from Berge [1, pp. 115–116].

LEMMA 4. *Let X and Y be compact topological spaces. If f is a lower (upper) semicontinuous numerical function on $X \times Y$ and Γ is a lower (upper) semicontinuous mapping of X into Y such that for each x, $\Gamma x \neq \phi$, then the numerical function h defined by*

$$h(x) = \sup \{f(x, y) | y \in \Gamma x\}$$

is lower (upper) semicontinuous on X.

LEMMA 5. *If A_s^1 and A_s^2 are continuous at $\bar{s}$, and $(\bar{v}_1, \bar{v}_2)$ are arbitrary, then $g(s, v_1, v_2)$ is continuous in (s, v_1, v_2) at $(\bar{s}, \bar{v}_1, \bar{v}_2)$.*

Proof. For $i = 1, 2$ let $C_i \subset E$ denote a compact set containing an open neighborhood of $\bar{v}_i$. With the notation of Lemma 4, identify X with $S \times K_1 \times C_1 \times C_2$, Y with K_2, f with $\beta(s, a_1, a_2)v_2 - \lambda\alpha(s, a_1, a_2)v_1 + \gamma(s, a_1, a_2)$, and Γ with A_s^2. Conclude by Lemma 4 that the numerical function

$$\max_{a_2 \in A_s^2} \{\beta(s, a_1, a_2)v_2 - \lambda\alpha(s, a_1, a_2)v_1 + \gamma(s, a_1, a_2)\}$$

is continuous at $(\bar{s}, \bar{v}_1, \bar{v}_2)$ for all $a_1 \in K_1$. Repeating this reasoning in a similar manner, conclude that $g_2(s, v_1, v_2)$, and hence, trivially, $g(s, v_1, v_2)$, are continuous in (s, v_1, v_2) at $(\bar{s}, \bar{v}_1, \bar{v}_2)$.

In the following lemma, we use the norm

$$\|g(s, v_1, v_2)\| = \max \{\sup_{s \in S} |g_1(s, v_1, v_2)|, \sup_{s \in S} |g_2(s, v_1, v_2)|\}.$$

LEMMA 6. *The function $g(s, v_1, v_2)$ is Lipschitzian with respect to (v_1, v_2), that is, for some positive constant L not depending on s, v_1, or v_2,*

$$\|g(s, v_1, v_2) - g(s, \bar{v}_1, \bar{v}_2)\| \leqq L\|(v_1, v_2) - (\bar{v}_1, \bar{v}_2)\|$$

for every $s \in S$ and every pair (v_1, v_2), $(\bar{v}_1, \bar{v}_2)$.

Proof. Let s, (v_1, v_2), and $(\bar{v}_1, \bar{v}_2)$ be arbitrary, and without loss of generality assume $g_2(s, \bar{v}_1, \bar{v}_2) \leqq g_2(s, v_1, v_2)$. Suppose $\bar{a}_1 \in A_s^1$ is such that

$$-g_2(s, v_1, v_2) = \max_{a_2 \in A_s^2} \{\beta(s, \bar{a}_1, a_2)v_2 - \lambda\alpha(s, \bar{a}_1, a_2)v_1 + \gamma(s, \bar{a}_1, a_2)\},$$

and suppose $\bar{a}_2 \in A_s^2$ is such that

$$\max_{a_2 \in A_s^2} \{\beta(s, \bar{a}_1, a_2)\bar{v}_2 - \lambda\alpha(s, \bar{a}_1, a_2)v_1 + \gamma(s, \bar{a}_1, a_2)\}$$
$$= \beta(s, \bar{a}_1, \bar{a}_2)\bar{v}_2 - \lambda\alpha(s, \bar{a}_1, \bar{a}_2)\bar{v}_1 + \gamma(s, \bar{a}_1, \bar{a}_2).$$

Then

$$\begin{aligned} -g_2(s, \bar{v}_1, \bar{v}_2) &= \min_{a_1 \in A_s^1} \max_{a_2 \in A_s^2} \{\beta(s, a_1, a_2)\bar{v}_2 - \lambda\alpha(s, a_1, a_2)\bar{v}_1 + \gamma(s, a_1, a_2)\} \\ &\leqq \max_{a_2 \in A_s^2} \{\beta(s, \bar{a}_1, a_2)\bar{v}_2 - \lambda\alpha(s, \bar{a}_1, a_2)\bar{v}_1 + \gamma(s, \bar{a}_1, a_2)\} \end{aligned}$$

so that

$$\begin{aligned} &g_2(s, v_1, v_2) - g_2(s, \bar{v}_1, \bar{v}_2) \\ &\leqq \max_{a_2 \in A_s^2} \{\beta(s, \bar{a}_1, a_2)\bar{v}_2 - \lambda\alpha(s, \bar{a}_1, a_2)\bar{v}_1 + \gamma(s, \bar{a}_1, a_2)\} \\ &\qquad - \max_{a_2 \in A_s^2} \{\beta(s, \bar{a}_1, a_2)v_2 - \lambda\alpha(s, \bar{a}_1, a_2)v_1 + \gamma(s, \bar{a}_1, a_2)\} \\ &\leqq \beta(s, \bar{a}_1, \bar{a}_2)\bar{v}_2 - \lambda\alpha(s, \bar{a}_1, \bar{a}_2)\bar{v}_1 + \gamma(s, \bar{a}_1, \bar{a}_2) \\ &\qquad - \{\beta(s, \bar{a}_1, \bar{a}_2)v_2 - \lambda\alpha(s, \bar{a}_1, \bar{a}_2)v_1 + \gamma(s, \bar{a}_1, \bar{a}_2)\} \\ &\leqq C_1|\bar{v}_2 - v_2| + \lambda C_2|\bar{v}_1 - v_1|, \end{aligned}$$

where

$$C_1 = \max_{\substack{s \in S \\ a_1 \in K_1 \\ a_2 \in K_2}} |\beta(s, a_1, a_2)| \quad \text{and} \quad C_2 = \max_{\substack{s \in S \\ a_1 \in K_1 \\ a_2 \in K_2}} |\alpha(s, a_1, a_2)|.$$

Thus the desired result follows with $L = \max\{1, C_1 + \lambda C_2\}$.

In subsequent lemmas we use $v(s, u_1, u_2)$ to denote the solution of (5) on S satisfying $v(r_0, u_1, u_2) = u_1$ and $v'(r_0, u_1, u_2) = u_2$ where, of course, $v'(s, u_1, u_2) = (\partial/\partial s)v(s, u_1, u_2)$. This is not to be confused with the notation at the beginning of this section. It should be clear from the context whether the second and third arguments of $v(s, u_1, u_2)$ are initial conditions or admissible controls.

LEMMA 7. *For $u_1, u_2 \in (-\infty, \infty)$, equation* (5) *has a unique solution $v(s, u_1, u_2)$ on S satisfying $v(r_0, u_1, u_2) = u_1$ and $v'(r_0, u_1, u_2) = u_2$.*

Proof. It suffices to show that the equation

$$\frac{d}{ds}\begin{pmatrix} v_1 \\ v_2 \end{pmatrix} = g(s, v_1, v_2)$$

has a unique solution on S satisfying $v_1(r_0) = u_1$ and $v_2(r_0) = u_2$ because then $v(s, u_1, u_2) = v_1(s)$. By differential equation theory and the piecewise continuity of A_s, Lemmas 5, 6 and a result in Edwards [3, pp. 153–155] imply the result.

LEMMA 8. *The functions $v(s, u_1, u_2)$ and $v'(s, u_1, u_2)$ are continuous, strictly increasing functions of u_2 with limits $\pm\infty$ as $u_2 \to \pm\infty$ for each fixed $s \in (r_0, r_1]$ and each fixed $u_1 \in (-\infty, \infty)$.*

Proof. We first show that the function $v'(s, u_1, \cdot)$ is strictly increasing. Suppose not, so that for some $u_1 \in (-\infty, \infty)$, $s_0 \in (r_0, r_1]$, and pair $u_2 < \bar{u}_2$, say, we have $v'(s_0, u_1, u_2) = v'(s_0, u_1, \bar{u}_2)$ and $v'(s, u_1, u_2) < v'(s, u_1, \bar{u}_2)$ for all $s \in [r_0, s_0)$. It follows that $v''(s, u_1, u_2) > v''(s, u_1, \bar{u}_2)$ for some $s < s_0$ in every neighborhood of s_0 and $v(s_0, u_1, u_2) < v(s_0, u_1, \bar{u}_2)$. But since $\alpha(s, a_1, a_2) > 0$ and by continuity we have $v''(s, u_1, u_2) \leqq v''(s, u_1, \bar{u}_2)$ for all $s < s_0$ in some neighborhood of s_0, a contradiction. Thus $u_2 < \bar{u}_2$ must imply $v'(s, u_1, u_2) < v'(s, u_1, \bar{u}_2)$, in which case $v(s, u_1, u_2) < v(s, u_1, \bar{u}_2)$, for each $s \in (r_0, r_1]$. The continuity of $v(s, u_1, u_2)$ and $v'(s, u_1, u_2)$ with respect to u_2 follows by standard differential equation theory.

To show the limiting behavior of $v(s, u_1, u_2)$ and its derivative, it suffices to consider $u_2 \to \infty$ and $u_1 < 0$; the situation with $u_1 \geqq 0$ reduces to this case and the proof with $u_2 \to -\infty$ is similar. For arbitrary $u_1 < 0$, $s_0 \in (r_0, r_1]$, and $L \in (0, \infty)$, it suffices to show $v'(s_0, u_1, u_2) \geqq L$ for some $u_2 \in (-\infty, \infty)$. To do this, we consider the differential equation

$$y''(s) = -C_\beta y'(s) + \lambda C_2 y(s) - C_\gamma, \tag{7}$$

where

$$C_\beta = \max_{\substack{s \in S \\ a_1 \in K_1 \\ a_2 \in K_2}} |\beta(s, a_1, a_2)|, \qquad C_\gamma > \max_{\substack{s \in S \\ a_1 \in K_1 \\ a_2 \in K_2}} |\gamma(s, a_1, a_2)|,$$

and $C_2 > 0$. Now if $y(s)$ is a solution of (7) with $y(r_0) = u_1 \in (-\infty, \infty)$ and $(dy/ds)(r_0) = u_2 \in (-\infty, \infty)$, then it is easy to verify that $y'(s) \to \infty$ as $u_2 \to \infty$ for each $s \in S$. In particular, with

$$C_2 = \underline{C} = \min_{\substack{s \in S \\ a_1 \in K_1 \\ a_2 \in K_2}} \alpha(s, a_1, a_2),$$

there exists some constant $L_1 \geqq L$ such that if $p \in [r_0, (r_0 + s_0)/2]$, then the solution to (7) satisfying $y(p) = 0$ and $y'(p) \geqq L_1$ will be such that $y'(s) \geqq L$ for all $s \in [p, s_0]$. Also, with

$$C_2 = \bar{C} = \max_{\substack{s \in S \\ a_1 \in K_1 \\ a_2 \in K_2}} \alpha(s, a_1, a_2),$$

there exists some constant u_2 such that the solution to (7) satisfying $y(r_0) = u_1$ and $y'(r_0) = u_2$ will be such that $y(\bar{s}) = 0$ for some $\bar{s} \in [r_0, (r_0 + s_0)/2]$ and $y'(s) \geqq L_1$ for all $s \in [r_0, \bar{s}]$; let $\bar{y}(s)$ denote this solution on $[r_0, \bar{s}]$.

We now claim that $v'(s_0, u_1, u_2) \geqq L$ because $v(s, u_1, u_2)$ is bounded below by appropriate solutions of (7). For example, suppose $\tilde{s} \in [r_0, \bar{s}]$ is such that $0 \geqq v(\tilde{s}, u_1, u_2) \geqq \bar{y}(\tilde{s})$ and $v'(\tilde{s}, u_1, u_2) = y'(\tilde{s})$. Then for some $(a_1, a_2) \in A_{\tilde{s}}$,

$$\begin{aligned} v''(\tilde{s}, u_1, u_2) - \bar{y}''(\tilde{s}) = {} & [C_\beta - \beta(s, a_1, a_2)]\bar{y}'(\tilde{s}) + \lambda[\alpha(s, a_1, a_2) - \bar{C}]v(\tilde{s}, u_1, u_2) \\ & + \lambda\bar{C}[v(\tilde{s}, u_1, u_2) - \bar{y}(\tilde{s})] + [C_\gamma - \gamma(\tilde{s}, a_1, a_2)]. \end{aligned}$$

Note the last term on the right-hand side is positive and the others are nonnegative so $v''(\tilde{s}, u_1, u_2) > \bar{y}''(\tilde{s})$. By continuity, $v''(s, u_1, u_2) > \bar{y}''(\tilde{s})$ for all s in some neighborhood of $\tilde{s}$. In particular, if we let $\tilde{s} = r_0$, then it becomes apparent that $v'(s, u_1, u_2) = \bar{y}'(s)$ is impossible with $v(s, u_1, u_2) \leqq 0$ for $s \in (r_0, \bar{s}]$. Thus $v'(s, u_1, u_2) > \bar{y}'(s)$ for each such s and there exists some $p \in (r_0, \bar{s}]$ such that $v(p, u_1, u_2) = 0$ and $v'(p, u_1, u_2) \geqq L_1$.

Now let $\underline{y}(s)$ be the solution to (7) with $\underline{y}(p) = 0$, $\underline{y}'(p) = v'(p, u_1, u_2)$, and $C_2 = \underline{C}$, and note that $\underline{y}'(s) \geqq L$ for all $s \in [p, s_0]$. By comparing $v(s, u_1, u_2)$ with $\underline{y}(s)$ as we did with $\bar{y}(s)$, we conclude the desired result.

LEMMA 9. *For fixed u_1 the function*

$$v(r_1, u_1, u_2) - \int_S v(s, u_1, u_2)\, d\mu_1(s)$$

is continuous and strictly increasing in u_2 and diverges to $\pm\infty$ as $u_2 \to \pm\infty$.

Proof. Let $u_2 < u_2'$ and $\bar{s} \in [r_0, r_1]$ be arbitrary. By Lemma 8, $v'(s, u_1, u_2)$ is increasing in u_2, so $v(r_1, u_1, u_2') - v(\bar{s}, u_1, u_2') > v(r_1, u_1, u_2) - v(\bar{s}, u_1, u_2)$. The function in this lemma is just a convex combination of the right-hand side of this inequality, so by this inequality the function in Lemma 9 is increasing in u_2.

Since $v'(s, u_1, u_2) \to \pm\infty$ as $u_2 \to \pm\infty$, we have for any $\bar{s} \in (r_0, r_1)$ that $v(r_1, u_1, u_2) - v(\bar{s}, u_1, u_2) \to \pm\infty$ as $u_2 \to \pm\infty$. Thus, by the convex combination argument, the function in this lemma has the same limits. Continuity follows from the continuity of $v(s, u_1, u_2)$.

LEMMA 10. *Let $\beta(s)$, $\alpha(s)$ and $\gamma(s)$ be measurable real-valued functions on S with $|\beta(s)| \leqq C_\beta < \infty$, $\alpha(s) \geqq C_2 > 0$, and $\gamma(s) \geqq 0$, and suppose for some $\bar{s} \in [r_0, r_1)$, $u_1 > 0$, and $u_2 \geqq 0$, the function $v(s)$ is a solution to*

$$v''(s) = \beta(s)v'(s) + \alpha(s)v(s) + \gamma(s)$$

satisfying $v(\bar{s}) = u_1$ and $v'(\bar{s}) = u_2$. Then for all $s \in (\bar{s}, r]$ we have $v(s) > 0$ and $v'(s) > 0$.

Proof. Suppose there is a smallest $s_0 \geqq \bar{s}$ such that $v'(s_0) = 0$. Then $v(s_0) > 0$, and by continuity, $v''(s) > 0$ in some neighborhood of s_0. Hence $v'(s) < 0$ for all large enough $s < s_0$, contradicting $u_2 \geqq 0$ and the definition of s_0.

Proof of Theorem 3. Denote

$$N_j = \theta_j \int_S v_j(s)\, d\mu_j(s) + \sigma_j \gamma_j + \kappa_j \lambda_j, \qquad j = 0, 1.$$

By Lemma 7 it suffices to show that $v(s, u_1, u_2)$, the unique solution of (5) with $v(r_0, u_1, u_2) = u_1$ and $v'(r_0, u_1, u_2) = u_2$, satisfies

$$\begin{aligned} &(\lambda\sigma_j + \theta_j + \kappa_j)v(r_j, u_1, u_2) - \theta_j \int_S v(s, u_1, u_2)\, d\mu_j(s) \\ &\qquad - (-1)^j \pi_j v'(r_j, u_1, u_2) = N_j, \qquad j = 0, 1, \end{aligned} \tag{8}$$

for unique values of u_1 and u_2. There are two cases.

Case 1. $\theta_0 = \pi_0 = 0$. By (8), $u_1 = N_0/(\lambda\sigma_0 + \kappa_0)$. By Lemmas 8 and 9, the left-hand side of (8) for $j = 1$ increases continuously and strictly from $-\infty$ to ∞ as u_2 increases from $-\infty$ to ∞. Hence (8) for $j = 1$ is satisfied by a unique value of u_2.

Case 2. $\theta_0 + \pi_0 > 0$. For $P \in (-\infty, \infty)$ denote

$$f_P(s, v_1, v_2) = -\min_{a_1 \in A_s^1} \max_{a_2 \in A_s^2} \{\beta(s, a_1, a_2)v_2 - \lambda\alpha(s, a_1, a_2)v_1 + P\gamma(s, a_1, a_2)\}$$

and let $v_P(s, u_1, u_2)$ denote the unique solution to

$$v_P''(s, u_1, u_2) = f_P(s, v_P(s, u_1, u_2), v_P'(s, u_1, u_2)) \tag{9}$$

and

$$v_P(r_0, u_1, u_2) = u_1, \qquad v_P'(r_0, u_1, u_2) = u_2. \tag{10}$$

Then by differential equation theory, $v_P(s, u_1, u_2)$ and $v_P'(s, u_1, u_2)$ are continuous in (P, s, u_1, u_2). We seek to show that $v_1(s, u_1, u_2)$ satisfies boundary condition (8) for a unique choice of the pair u_1, u_2.

We can rewrite (8) for $j = 0$ and general P as

$$\theta_0 \int_0 v_P(s, u_1, u_2)\, d\mu_0(s) + \pi_0 u_2 = \phi(u_1), \tag{11}$$

where

$$\phi(u_1) = (\lambda\sigma_0 + \theta_0 + \kappa_0)u_1 - N_0. \tag{12}$$

We first show that for each $u_1 \in (-\infty, \infty)$ and $P \in (-\infty, \infty)$ there exists a unique $u_2 \equiv u_2(P, u_1)$ satisfying (11). But this follows from Lemma 8, because then the left-hand side of (11) is continuous and strictly increasing in u_2 with limits $\pm\infty$ as $u_2 \to \pm\infty$. Note that, since both sides of (11) are continuous in (P, u_1, u_2), the function $u_2(P, u_1)$ is continuous in (P, u_1).

It remains to show that $v(s, u_1) \equiv v_1(s, u_1, u_2(1, u_1))$ satisfies (8) with $j = 1$ for a unique value of u_1, that is, there is a unique u_1 for which

$$(\lambda\sigma_1 + \theta_1 + \kappa_1)v(r_1, u_1) - \theta_1 \int_S v(s, u_1)\, d\mu_1(s) + \pi_1 v'(r_1, u_1) = N_1. \tag{13}$$

We first show that some $u_1 \in (-\infty, \infty)$ satisfies (13). Since the left-hand side of (13) is continuous in u_1, it suffices to show that the left-hand side of (13) diverges to $\pm\infty$ as $u_1 \to \pm\infty$. We discuss only the case $u_1 \to +\infty$ since the other is similar.

We show this result by considering the limit of $Pv(s, P^{-1})$ as $P \downarrow 0$. To this end, for $P > 0$ denote $u(P) = Pu_2(1, P^{-1})$ and $\psi(P) = P\phi(P^{-1})$. Now $Pv_1(s, P^{-1}, u_2) = v_P(s, 1, Pu_2)$, so $Pv(s, P^{-1}) = v_P(s, 1, u(P))$. In view of this and (11), $u_2 = u(P)$ is the unique number satisfying

$$\theta_0 \int_S v_P(s, 1, u_2)\, d\mu_0(s) + \pi_0 u_2 = \psi(P). \tag{14}$$

Since $\psi(P)$ has a limit as $P \downarrow 0$, which we denote by $\psi(0)$, equation (14) has a unique solution $\bar{u}_2$ for $P = 0$. Since $v_P(s, 1, u_2)$ and $\psi(P)$ are continuous in (P, s, u_2), it follows that $u(P)$ is continuous in P and has the limit $\bar{u}_2$ as $P \downarrow 0$; we denote $u(0) = \bar{u}_2$. In summary, $Pv(s, P^{-1}) \to v_0(s, 1, u(0))$ as $P \downarrow 0$.

We are now in a position to show that the left-hand side of (13) diverges to $\pm\infty$ as $u_1 \to \pm\infty$. If $(dv_0/ds)(r_0, 1, u(0)) \geqq 0$, then Lemma 10 applies and $v'(s, 1, u(0)) > 0$ for each $s \in (r_0, r_1]$. On the other hand, if $v_0'(r_0, 1, u(0)) < 0$, then by (14),

$$\theta_0 \int_S v_0(s, 1, u(0))\, d\mu_0(s) = \lambda\sigma_0 + \kappa_0 + \theta_0 - \pi_0 v_0'(r_0, 1, u(0)) \geqq \theta_0.$$

Now $\theta_0 > 0$, for if not, then $\pi_0 > 0$ and by (14), $v_0'(r_0, 1, u(0)) = (\lambda\sigma_0 + \kappa_0)/\pi_0 > 0$, a contradiction. Thus for at least one $s_1 \in (r_0, r_1)$ where $d\mu_0(s_1) > 0$ we must have

$v_0(s_1, 1, u(0)) \geqq 1 = v_0(r_0, 1, u(0))$. It follows for some $s_0 \in [r_0, s_1]$ that $v_0(s_0, 1, u(0)) > 0$ and $v_0'(s_0, 1, u(0)) \geqq 0$. Applying Lemma 10, we conclude for all $s \in (s_0, r_1]$ that $v_0(s, 1, u(0)) > 0$ and $v_0'(s, 1, u(0)) > 0$.

Let Y denote the left-hand side of (13) with $v_0(s, 1, u(0))$ substituted for $v(s, u_1)$. We have by the preceding arguments that $v_0(r_1, 1, u(0)) > 0$ and $v_0'(r_1, 1, u(0)) > 0$. Moreover, for any $s \in [r_0, r_1)$ we have $v_0(r_1, 1, u(0)) > v_0(s, 1, u(0))$, so if $\theta_1 > 0$, then

$$\theta_1\left[v_0(r_1, 1, u(0)) - \int_S v_0(s, 1, u(0))\, d\mu_1(s)\right] > 0,$$

in which case $Y > 0$. Letting $u_1 \to \infty$ in the left-hand side of (13), we have

$$\begin{aligned}\lim_{u_1 \to \infty}&\left\{(\lambda\sigma_1 + \theta_1 + \kappa_1)v(r_1, u_1) - \theta_1\int_S v(s, u_1)\, d\mu_1(s) + \pi_1 v'(r_1, u_1)\right\}\\ &= \lim_{P \downarrow 0}\frac{1}{P}\Big\{(\lambda\sigma_1 + \theta_1 + \kappa_1)v_p(s, 1, u(P)) - \theta_1\int_S v_p(s, 1, u(P))\, d\mu_1(s)\\ &\qquad\qquad + \pi_1 v_p'(r_1, 1, u(P))\Big\}\\ &= \lim_{P \downarrow 0}\frac{Y}{P} = +\infty.\end{aligned}$$

Thus, by the remarks following equation (13) there exists some $u_1 \in (-\infty, \infty)$ which satisfies (13); it remains to show that this u_1 is unique.

Suppose there exist two numbers $C_0 < C_1$ and corresponding solutions $v_0(s)$ and $v_1(s)$ of (5), (6) such that $v_0(r_0) = C_0$ and $v_1(r_0) = C_1$. Let the Borel measurable function $a_1(s)$ from S into K_1 be such that $a_1(s) \in A_s^1$ and

$$\begin{aligned}\min_{a_1 \in A_s^1}\max_{a_2 \in A_s^2}&\{\beta(s, a_1, a_2)v_0'(s) - \lambda\alpha(s, a_1, a_2)v_0(s) + \gamma(s, a_1, a_2)\}\\ &= \max_{a_2 \in A_s^2}\{\beta(s, a_1(s), a_2)v_0'(s) - \lambda\alpha(s, a_1(s), a_2)v_0(s) + \gamma(s, a_1(s), a_2)\}\end{aligned}$$

for each $s \in S$. Let the Borel measurable function $a_2(s)$ from S into K_2 be such that $a_2(s) \in A_s^2$ and

$$\begin{aligned}\max_{a_2 \in A_s^2}&\{\beta(s, a_1(s), a_2)v_1'(s) - \lambda\alpha(s, a_1(s), a_2)v_1(s) + \gamma(s, a_1(s), a_2)\}\\ &= \beta(s, a_1(s), a_2(s))v_1'(s) - \lambda\alpha(s, a_1(s), a_2(s))v_1(s) + \gamma(s, a_1(s), a_2(s))\end{aligned}$$

for each $s \in S$. Then

$$\begin{aligned}0 &= v_1''(s) + \min_{a_1 \in A_s^1}\max_{a_2 \in A_s^2}\{\beta(s, a_1, a_2)v_1'(s) - \lambda\alpha(s, a_1, a_2)v_1(s) + \gamma(s, a_1, a_2)\}\\ &\leqq v_1''(s) + \max_{a_2 \in A_s^2}\{\beta(s, a_1(s), a_2)v_1'(s) - \lambda\alpha(s, a_1(s), a_2)v_1(s) + \gamma(s, a_1(s), a_2)\}\\ &= v_1''(s) + \beta(s, a_1(s), a_2(s))v_1'(s) - \lambda\alpha(s, a_1(s), a_2(s))v_1(s) + \gamma(s, a_1(s), a_2(s))\end{aligned}$$

and similarly,

$$v_0''(s) + \beta(s, a_1(s), a_2(s))v_0'(s) - \lambda\alpha(s, a_1(s), a_2(s))v_0(s) + \gamma(s, a_1(s), a_2(s)) \leqq 0.$$

Defining the Borel measurable function

$$\psi(s) = \frac{d^2}{ds^2}(v_1 - v_0)(s) + \beta(s, a_1(s), a_2(s))\frac{d}{ds}(v_1 - v_0)(s)$$
$$- \lambda\alpha(s, a_1(s), a_2(s))(v_1 - v_0)(s),$$

we see that $\psi(s) \geqq 0$. Letting $v(s) = v_1(s) - v_0(s)$, we see that $v(s)$ is a solution to

$$v''(s) = -\beta(s, a_1(s), a_2(s))v'(s) + \lambda\alpha(s, a_1(s), a_2(s))v(s) + \psi(s)$$

satisfying (by subtracting boundary conditions (6))

$$(15)\quad (\lambda\sigma_j + \theta_j + \kappa_j)v(r_j) - \theta_j\int_S v(s)\,d\mu_j(s) - (-1)^j\pi_j v'(r_j) = 0, \qquad j = 0, 1,$$

as well as $v(r_0) = C_1 - C_0$.

It remains to show that $v(s)$ cannot simultaneously satisfy all three of these boundary conditions. Assume that $v(r_0) = C_1 - C_0$ and (15) holds for $j = 0$. Let $s_0 = \min\{s \in S | v(s) = C_1 - C_0 \text{ and } v'(s) \geqq 0\}$. Note that s_0 exists because if $v'(r_0) \geqq 0$, then $s_0 = r_0$, whereas if $v'(r_0) < 0$, then by (15),

$$\theta_0\int_S v(s)\,d\mu_0(s) = (\lambda\sigma_0 + \theta_0 + \kappa_0)v(r_0) - \pi_0 v'(r_0) \geqq \theta_0 v(r_0).$$

Thus $\theta_0 > 0$ and for some $s_1 \in (r_0, r_1)$ with $d\mu_0(s_1) > 0$ we have $v(s_1) \geqq v(r_0)$ in which case $s_0 \in (r_0, s_1]$. By Lemma 10 we have $v(s) > 0$ and $v'(s) > 0$ for all $s \in (s_0, r_1]$. In particular, the left-hand side of (15) is positive for $j = 1$, and Theorem 3 is proved.

The following theorem provides a necessary and sufficient (saddle point) condition for an admissible control to be a solution to the zero sum, two-person game. We now revert to the original notation, where $v(s, a_1, a_2)$ denotes the expected discounted cost of a process corresponding to the control $a = (a_1, a_2) \in M$.

THEOREM 11. *Let $\hat{v}(s)$ be the unique solution of* (5), (6). *A control $\hat{a} = (\hat{a}_1, \hat{a}_2) \in M$ is optimal if and only if*

$$(16)\quad \begin{aligned} &\max_{a_2 \in A_s^2}\{d(s, \hat{a}_1(s), a_2)^{-1}[b(s, \hat{a}_1(s), a_2)\hat{v}'(s) - \lambda\hat{v}(s) + c(s, \hat{a}_1(s), a_2)]\} \\ &= d(s, \hat{a}_1(s), \hat{a}_2(s))^{-1}[b(s, \hat{a}_1(s), \hat{a}_2(s))\hat{v}'(s) - \lambda\hat{v}(s) + c(s, \hat{a}_1(s), \hat{a}_2(s))] \\ &= \min_{a_1 \in A_s^1}\{d(s, a_1, \hat{a}_2(s))^{-1}[b(s, a_1, \hat{a}_2(s))\hat{v}'(s) - \lambda\hat{v}(s) + c(s, a_1, \hat{a}_2(s))]\} \end{aligned}$$

for every $s \in S$ which is a continuity point of $\hat{a}$ and, for $j = 0$ and $j = 1$, $\sigma_j > 0$ implies that

$$(17)\quad \begin{aligned} \max_{a_2 \in A_{r_j}^2} c(r_j, \hat{a}_1(r_j), a_2) &= c(r_j, \hat{a}_1(r_j), \hat{a}_2(r_j)) \\ &= \min_{a_1 \in A_{r_j}^1} c(r_j, a_1, \hat{a}_2(r_j)). \end{aligned}$$

Moreover, if $\hat{a}$ is optimal, then $\hat{v}(s) = v(s, \hat{a}_1, \hat{a}_2)$ is the value of the game.

Proof. Suppose (16) and (17) are true. By the theory of saddle points we have

$$\min_{a_1\in A^1_s}\max_{a_2\in A^2_s}\{d(s,a_1,a_2)^{-1}[b(s,a_1,a_2)\hat{v}'(s)-\lambda\hat{v}(s)+c(s,a_1,a_2)]\}$$
$$=d(s,\hat{a}_1(s),\hat{a}_2(s))^{-1}[b(s,\hat{a}_1(s),\hat{a}_2(s))\hat{v}'(s)-\lambda\hat{v}(s)+c(s,\hat{a}_1(s),\hat{a}_2(s))]$$

and, if $\sigma_j>0$,

$$\min_{a_1\in A^1_{r_j}}\max_{a_2\in A^2_{r_j}} c(r_j,a_1,a_2)=c(r_j,\hat{a}_1(r_j),\hat{a}_2(r_j)).$$

Substituting these in (5) and (6), we see that $\hat{v}(s)$ is also the unique solution of (1) and (2), that is, $\hat{v}(s)=v(s,\hat{a}_1,\hat{a}_2)$.

With $\hat{a}_2(s)$ fixed, in a similar manner we see that $\hat{v}(s)$ is the unique solution in Pliska [8, Theorem 1], that is, $\hat{v}(s)$ is the minimal expected discounted cost for an ordinary optimal control problem involving the control $a_1(s)$. In view of (16) and (17), we have by Pliska [8, Theorem 1] that $v(s,\hat{a}_1,\hat{a}_2)\leqq v(s,a_1,\hat{a}_2)$ for each $a_1\in M_1$ and each $s\in S$. Similarly, $v(s,\hat{a}_1,a_2)\leqq v(s,\hat{a}_1,\hat{a}_2)$ for all $a_2\in M_2$. Hence $\hat{a}$ is optimal and $\hat{v}(s)$ is the value of the game.

Conversely, suppose $\hat{a}$ is an optimal control. First we shall show that

$$d(s,\hat{a}_1(s),\hat{a}_2(s))^{-1}[b(s,\hat{a}_1(s),\hat{a}_2(s))v'(s,\hat{a}_1,\hat{a}_2)-\lambda v(s,\hat{a}_1,\hat{a}_2)$$
$$+c(s,\hat{a}_1(s),\hat{a}_2(s))]$$
$$=\max_{a_2\in A^2_s}\{d(s,\hat{a}_1(s),a_2)^{-1}[b(s,\hat{a}_1(s),a_2)v'(s,\hat{a}_1,\hat{a}_2)-\lambda v(s,\hat{a}_1,\hat{a}_2)$$
$$+c(s,\hat{a}_1(s),a_2)]\}$$
$$\geqq\min_{a_1\in A^1_s}\max_{a_2\in A^2_s}\{d(s,a_1,a_2)^{-1}[b(s,a_1,a_2)v'(s,\hat{a}_1,\hat{a}_2)-\lambda v(s,\hat{a}_1,\hat{a}_2)$$
$$+c(s,a_1,a_2)]\}$$
$$\geqq\max_{a_2\in A^2_s}\min_{a_1\in A^1_s}\{d(s,a_1,a_2)^{-1}[b(s,a_1,a_2)v'(s,\hat{a}_1,\hat{a}_2)-\lambda v(s,\hat{a}_1,\hat{a}_2)$$
$$+c(s,a_1,a_2)]\}$$
$$\geqq\min_{a_1\in A^1_s}\{d(s,a_1,\hat{a}_2(s))^{-1}[b(s,a_1,\hat{a}_2(s))v'(s,\hat{a}_1,\hat{a}_2)-\lambda v(s,\hat{a}_1,\hat{a}_2)$$
$$+c(s,a_1,\hat{a}_2(s))]\}$$
$$=d(s,\hat{a}_1(s),\hat{a}_2(s))^{-1}[b(s,\hat{a}_1(s),\hat{a}_2(s))v'(s,\hat{a}_1,\hat{a}_2)-\lambda v(s,\hat{a}_1,\hat{a}_2)$$
$$+c(s,\hat{a}_1(s),\hat{a}_2(s))].$$

Now $v(s,\hat{a}_1,\hat{a}_2)=\inf_{a_1\in M_1}v(s,a_1,\hat{a}_2)$, so by Pliska [8, Theorem 1], the last equality is true; similarly, the first one is true. The inequalities are true by saddle point theory, so all are equalities. Similarly, if $\sigma_j>0$, then

$$c(r_j,\hat{a}_1(r_j),\hat{a}_2(r_j))=\max_{a_2\in A^2_{r_j}}c(r_j,\hat{a}_1(r_j),a_2)$$
$$=\min_{a_1\in A^1_{r_j}}\max_{a_2\in A^2_{r_j}}c(r_j,a_1,a_2)=\min_{a_1\in A^1_{r_j}}c(r_j,a_1,\hat{a}_2(r_j)).$$

Substituting these equalities into (1) and (2), we see that $v(s,\hat{a}_1,\hat{a}_2)$ is the unique solution of (5) and (6), that is, $v(s,\hat{a}_1,\hat{a}_2)=\hat{v}(s)$. Substituting $\hat{v}(s)$ for $v(s,\hat{a}_1,\hat{a}_2)$ in the above equalities yields (16), and Theorem 11 is proved.

A diffusion process two-person, zero sum game problem can be solved in principle as follows. First, obtain the solution $v(s)$ to (5) and (6). Second, consider the map Γ from S into $K_1 \times K_2$ such that $(a_1, a_2) \in \Gamma(s)$ if and only if $a_1 \in A_s^1$, $a_2 \in A_s^2$, and (a_1, a_2) is a saddle point of $d(s, a_1, a_2)^{-1}[b(s, a_1, a_2)v'(s) - \lambda v(s) + c(s, a_1, a_2)]$. If $\sigma_j > 0$, $j = 0, 1$, then redefine $\Gamma(r_j)$ so that $(a_1, a_2) \in \Gamma(r_j)$ if and only if $a_1 \in A_{r_j}^1$, $a_2 \in A_{r_j}^2$, and (a_1, a_2) is a saddle point of $c(r_j, a_1, a_2)$. Note that $\Gamma(s) = \varnothing$ is possible for some $s \in S$, in which case the game is without solution. On the other hand, if $\Gamma(s) \neq \varnothing$ for each $s \in S$, then $v(s)$ is the value of the game, even if it cannot be attained. Finally, endeavor to choose a piecewise continuous function $a(\cdot)$ such that $a(s) \in \Gamma(s)$ for each $s \in S$.

The following result is a sufficient condition for (16) and (17) to be satisfied by $v(\cdot)$ and some Borel measurable control $a(\cdot)$, that is, for the map mentioned above $\Gamma(s) \neq \varnothing$ for each $s \in S$. The real-valued function $h(\cdot)$ on the compact, convex set $C \subset E^n$ is said to be quasi-convex if $\{z \in C | h(z) \leqq \alpha\}$ is convex for each $\alpha \in E$. This function is quasi-concave if $-h(\cdot)$ is quasi-convex. Corollary 13 is an immediate consequence of Theorem 12 which, in turn, follows easily from a minimax theorem by Sion [10].

THEOREM 12. *Let $v(s)$ be the unique solution of* (5), (6) *and suppose A_s^i is convex for each $s \in S$, $i = 1, 2$. Then there exists some Borel measurable control $a(s) = (a_1(s), a_2(s))$, with $a_1(s) \in A_s^1$ and $a_2(s) \in A_s^2$ for each $s \in S$, which satisfies* (16) *and* (17) *provided that*

$$d(s, a_1, a_2)^{-1}[b(s, a_1, a_2)v'(s) - \lambda v(s) + c(s, a_1, a_2)]$$

and $\sigma_j c(r_j, a_1, a_2)$, $j = 0, 1$, are quasi-convex in $a_1 \in A_s^1$ for each $a_2 \in A_s^2$ and $s \in S$ and are quasi-concave in $a_2 \in A_s^2$ for each $a_1 \in A_s^1$ and $s \in S$.

COROLLARY 13. *Let $v(s)$ be the unique solution of* (5), (6) *and suppose A_s^1 is convex for each $s \in S$, $i = 1, 2$. Suppose $d(s, a_1, a_2)$ is constant with respect to (a_1, a_2), $b(s, a_1, a_2)$ is affine with respect to (a_1, a_2), $c(s, a_1, a_2)$ is convex in a_1 for each $a_2 \in A_s^2$, and $c(s, a_1, a_2)$ is concave in a_2 for each $a_1 \in A_s^1$, all for each $s \in S$. Then* (16) *and* (17) *are satisfied by some Borel measurable control $(a_1(s), a_2(s))$ with $a_1(s) \in A_s^1$ and $a_2(s) \in A_s^2$.*

Example.

$$\begin{aligned} &0 < r_0 < r_1, && d(s, a_1, a_2) = A > 0, \\ &A_s^1 = \{a_1 \in E | |a_1| \leqq z_1 s\}, \quad z_1 > 0, && b(s, a_1, a_2) = a_1 a_2 / s, \\ &A_s^2 = \{a_2 \in E | |a_2| \leqq z_2 s\}, \quad z_2 > 0, && c(s, a_1, a_2) = C. \end{aligned}$$

r_0 boundary condition: $v'(r_0) = 0$ (reflection);

r_1 boundary condition: $v(r_1) = \lambda_1$ (absorption with cost λ_1).

For any value of s, $v(s)$, or $v'(s)$ we have

$$\begin{aligned} &\min_{a_1 \in A_s^1} \max_{a_2 \in A_s^2} A^{-1}\left[\frac{a_1 a_2}{s} v'(s) - \lambda v(s) + C\right] \\ &= \max_{a_2 \in A_s^2} \min_{a_1 \in A_s^1} A^{-1}\left[\frac{a_1 a_2}{s} v'(s) - \lambda v(s) + C\right] \\ &= A^{-1}[C - \lambda v(s)], \end{aligned}$$

so $a(s) = (0, 0)$ is the optimal control for all $s \in S$. Therefore, the value of the game is given by (1), (2) to be $v(s) = C_1 e^{tx} + C_2 e^{-tx} + C/\lambda$, where $t = \sqrt{\lambda/A}$, $C_1 = e^{tr_1}(\lambda_1 - C/\lambda)/(e^{2tr_1} + e^{2tr_0})$, and $C_2 = e^{tr_1}(\lambda_1 - C/\lambda)/(e^{2t(r_1 - r_0)} + 1)$.

4. The zero sum problem with undiscounted costs. The zero sum, two-person diffusion process game problem with undiscounted costs will be one of two types, depending on whether the boundary conditions are conservative or nonconservative. The results in this section parallel those of § 3, and, consequently, they will be brief. The conservative case will be treated in the second half of this section. For the purposes of this section, the boundary conditions are said to be nonconservative if at least one boundary is absorbing and neither boundary is purely adhesive, that is,

$$\kappa_0 + \kappa_1 > 0, \quad \kappa_j + \pi_j + \theta_j > 0, \qquad j = 0, 1.$$

Let $v(s, a_1, a_2) = v(s)$ denote the expected undiscounted cost of a nonconservative process corresponding to the admissible control $a = (a_1, a_2) \in M$. Then $v(s, a_1, a_2)$ will be the unique solution of (1), (2) with $\lambda = 0$. The control $a \in M$ is said to be optimal if for all $a_1 \in M_1$, all $a_2 \in M_2$, and all $s \in S$ we have

$$v(s, \hat{a}_1, a_2) \leqq v(s, \hat{a}_1, \hat{a}_2) \leqq v(s, a_1, \hat{a}_2),$$

in which case $v(s, \hat{a}_1, \hat{a}_2)$ is said to be the value of the game. It will subsequently be proved that the value of a game, if it exists, is provided by the following result whose proof is a generalization of one by Mandl [7, pp. 158–167].

THEOREM 14. *With nonconservative boundary conditions, there exists a unique solution $v(s)$ to*

$$(18) \quad v''(s) + \min_{a_1 \in A_s^1} \max_{a_2 \in A_s^2} \{d(s, a_1, a_2)^{-1}[b(s, a_1, a_2)v'(s) + c(s, a_1, a_2)]\} = 0$$

satisfying

$$(19) \quad \begin{aligned} &(\theta_j + \kappa_j)v(r_j) - \theta_j \int_S (v(s) + v_j(s))\, d\mu_j(s) \\ &\quad - (-1)^j \pi_j v'(r_j) - \sigma_j \gamma_j - \kappa_j \lambda_j = 0, \qquad j = 0, 1, \end{aligned}$$

where

$$\gamma_j = \min_{a_1 \in A_{r_j}^1} \max_{a_2 \in A_{r_j}^2} c(r_j, a_1, a_2), \qquad j = 0, 1.$$

Proof. Lemma 7 does not depend on $\lambda > 0$, so for every $u_1, u_2 \in (-\infty, \infty)$, equation (18) has a unique solution $v(s)$ satisfying $v(r_0) = u_1$ and $v'(r_0) = u_2$. For fixed u_1 and u_2 denote $w(s, u_2) = v'(s)$ and note that $w(s, u_2)$ is independent of u_1 since it is the solution of a first order differential equation under the initial condition $w(r_0, u_2) = u_2$. Writing for $j = 0, 1$:

$$N_j = \theta_j \int_S v_j(s)\, d\mu_j(s) + \sigma_j \gamma_j + \kappa_j \lambda_j,$$

we have that (19) is equivalent to

$$\kappa_0 u_1 - \theta_0 \int_S \int_{r_0}^{s} w(t, u_2)\,dt\,d\mu_0(s) - \pi_0 u_2 = N_0,$$

$$\kappa_1 u_1 + (\theta_1 + \kappa_1) \int_{r_0}^{r_1} w(t, u_2)\,dt - \theta_1 \int_S \int_{r_0}^{s} w(t, u_2)\,dt\,d\mu_1(s) \tag{20}$$
$$+ \pi_1 w(r_1, u_2) = N_1.$$

Eliminating u_1 from (20), we obtain the equation for u_2:

$$\kappa_0(\theta_1 + \kappa_1) \int_{r_0}^{r_1} w(t, u_2)\,dt$$
$$+ \kappa_1 \theta_0 \int_S \int_{r_0}^{s} w(t, u_2)\,dt\,d\mu_0(s) + \kappa_1 \pi_0 u_2 + \kappa_0 \pi_1 w(r_1, u_2) \tag{21}$$
$$- \kappa_0 \theta_1 \int_S \int_{r_0}^{s} w(t, u_2)\,dt\,d\mu_1(s) = \kappa_0 N_1 - \kappa_1 N_0.$$

It remains to show that (21) is solved by a unique value of u_2, since then u_1 can be obtained from (20). By Lemma 8, $w(s, u_2)$ is continuous and strictly increasing in u_2 and $w(s, u_2) \to \pm\infty$ as $u_2 \to \pm\infty$, in which case the left-hand side of (21) has these same properties (see Mandl [7, p. 163]). Hence (21) has a unique solution and Theorem 14 is proved.

THEOREM 15. *With undiscounted costs and nonconservative boundary conditions, let $v(s)$ be the unique solution of* (18), (19). *A control $a = (a_1, a_2) \in M$ is optimal if and only if*

$$\max_{a_2 \in A_S^2} \{d(s, a_1(s), a_2)^{-1}[b(s, a_1(s), a_2)v'(s) + c(s, a_1(s), a_2)]\}$$
$$= d(s, a_1(s), a_2(s))^{-1}[b(s, a_1(s), a_2(s))v'(s) + c(s, a_1(s), a_2(s))] \tag{22}$$
$$= \min_{a_1 \in A_S^1} \{d(s, a_1, a_2(s))^{-1}[b(s, a_1, a_2(s))v'(s) + c(s, a_1, a_2(s))]\}$$

for each $s \in S$ which is a continuity point of $a(s)$, and, for $j = 0$ and $j = 1$, $\sigma_j > 0$ implies

$$\max_{a_2 \in A_{r_j}^2} c(r_j, a_1(r_j), a_2) = c(r_j, a_1(r_j), a_2(r_j)) \tag{23}$$
$$= \min_{a_1 \in A_{r_j}^1} c(r_j, a_1, a_2(r_j)).$$

Moreover, if $a(s)$ is optimal, then $v(s) = v(s, a_1, a_2)$ is the value of the game

The proof is essentially identical to that for Theorem 11, so it will be omitted. A diffusion process zero sum, two-person game problem in the undiscounted cost, nonconservative process case can be solved, in principle, in the same manner as with the discounted cost case. Moreover, there exist sufficient conditions analogous to those of Theorem 12 and Corollary 13 for equations (22) and (23) to be satisfied by some Borel measurable control $a(s)$.

Example.

$$\begin{aligned} S &= [0,1], & d(s,a_1,a_2) &= A > 0,\\ A_s^1 &= [-Y,Y] \subset E, \quad Y>0, & b(s,a_1,a_2) &= a_1a_2,\\ A_s^2 &= [-Z,Z] \subset E, \quad Z>0, & c(s,a_1,a_2) &= C. \end{aligned}$$

r_0 boundary condition: $v(r_0) = \lambda_0$ (absorption with cost λ_0);
r_1 boundary condition: $v'(r_1) = 0$ (reflection).

For any value of s or $v'(s)$ we have

$$\min_{a_1\in A_s^1}\max_{a_2\in A_s^2}\{A^{-1}[a_1a_2v'(s) + C]\} = \max_{a_2\in A_s^2}\min_{a_1\in A_s^1}\{A^{-1}[a_1a_2v'(s) + C]\} = A^{-1}C,$$

so $a(s) = (0,0)$ is the optimal control for all $s\in S$. Therefore, the value of the game is given by (1) and (2) to be $v(s) = -\frac{1}{2}(C/A)s^2 + (C/A)s + \lambda_0$.

We now discuss the other type of undiscounted cost problem, the conservative case. For the purposes of this section, the boundary conditions are said to be conservative if neither boundary is absorbing and at least one boundary is not purely adhesive, that is,

$$\kappa_0 = \kappa_1 = 0, \qquad \pi_0 + \theta_0 + \pi_1 + \theta_1 > 0.$$

Let $\Theta(a_1,a_2) = \Theta$ denote the mean cost per unit time of such a process corresponding to the admissible control $a = (a_1,a_2)\in M$. Then $\Theta(a_1,a_2)$ is the unique number to which there exists a solution to (3) and (4). The control $\hat{a}\in M$ is said to be optimal if for all $a_1\in M_1$ and all $a_2\in M_2$ we have

$$\Theta(\hat{a}_1,a_2) \leqq \Theta(\hat{a}_1,\hat{a}_2) \leqq \Theta(a_1,\hat{a}_2),$$

in which case $\Theta(\hat{a}_1,\hat{a}_2)$ is said to be the value of the game. The following result characterizes the value of a game.

THEOREM 16. *With conservative boundary conditions, there exists a unique number Θ such that the equation*

$$(24)\quad w'(s) + \min_{a_1\in A_s^1}\max_{a_2\in A_s^2}\{d(s,a_1,a_2)^{-1}[b(s,a_1,a_2)w(s) - \Theta + c(s,a_1,a_2)]\} = 0$$

has a solution $w(s)$ satisfying

$$(25)\quad \theta_j\left[\int_S\left[\int_{r_j}^s w(y)\,dy + v_j(s)\right]d\mu_j(s) + (-1)^j\pi_jw(r_j) + \sigma_j(\gamma_j - \Theta) = 0, \qquad j = 0,1,$$

where

$$\gamma_j = \min_{a_1\in A_{r_j}^1}\max_{a_2\in A_{r_j}^2} c(r_j,a_1,a_2).$$

Proof. This proof is rather similar to that for Theorem 3, so it will only be sketched. By Lemma 7, for every $u_2,\Theta\in(-\infty,\infty)$ there exists a unique solution $w(s,u_2,\Theta)$ to (24) satisfying $w(r_0,u_2,\Theta) = u_2$. By Lemma 8, $w(s,u_2,\Theta)$ is continuous and strictly increasing in u_2 and $w(s,u_2,\Theta)\to\pm\infty$ as $u_2\to\pm\infty$. It follows that the left-hand side of (25) with $w(s,u_2,\Theta)$ substituted for $w(s)$ is continuous and strictly increasing (decreasing) in u_2 and diverges to $\pm\infty$ $(\mp\infty)$ as $u_2\to\pm\infty$ for $j = 0$ $(j = 1)$. Thus, if boundary r_j is purely adhesive, then $\Theta = \gamma_j$ and u_2 can be determined uniquely from the other boundary condition.

On the other hand, if neither boundary is purely adhesive, then to every Θ there exists a unique number $u_2 = u_2(\Theta)$ such that $w(s, \Theta) \equiv w(s, u_2(\Theta), \Theta)$ satisfies (25) for $j = 0$. It remains to show that $w(s, \Theta)$ satisfies (25) for $j = 1$ with a unique value of Θ. Consider $\Theta^{-1}w(s, \Theta)$ for $\Theta > 0$. It can be shown as with Theorem 3 that $\Theta^{-1}w(s, \Theta) \to \bar{w}(s)$ as $\Theta \to \infty$ for all $s \in S$, where $\bar{w}(s)$ is the solution to

$$\bar{w}'(s) = -\min_{a_1 \in A_s^1} \max_{a_2 \in A_s^2} \{d(s, a_1, a_2)^{-1}[b(s, a_1, a_2)\bar{w}(s) - 1]\}$$

satisfying

$$\theta_0 \int_S \int_{r_0}^{s} \bar{w}(y)\, dy\, d\mu_0(s) + \pi_0 \bar{w}(r_0) - \sigma_0 = 0.$$

After showing that

$$\theta_1 \int_S \int_{r_0}^{s} \bar{w}(y)\, dy\, d\mu_1(s) - \pi_1 \bar{w}(r_1) - \sigma_1 < 0,$$

we conclude that the left-hand side of (25) for $j = 1$ with $w(s, \Theta)$ substituted for $w(s)$ diverges to $\mp\infty$ as $\Theta \to \pm\infty$. By continuity, $w(s, \Theta)$ satisfies (25) for $j = 1$ with some value Θ. This solution Θ is unique; otherwise a contradiction can be derived as was done with Theorem 3 to show the unicity of $v(r_0)$.

THEOREM 17. *With undiscounted costs and conservative boundary conditions, let Θ be the unique number such that* (24) *has a solution $w(s)$ satisfying* (25). *A control $a = (a_1, a_2) \in M$ is optimal if and only if*

$$\begin{aligned} &\max_{a_2 \in A_s^2} \{d(s, a_1(s), a_2)^{-1}[b(s, a_1(s), a_2)w(s) - \Theta + c(s, a_1(s), a_2)]\} \\ (26) \qquad &= d(s, a_1(s), a_2(s))^{-1}[b(s, a_1(s), a_2(s))w(s) - \Theta + c(s, a_1(s), a_2(s))] \\ &= \min_{a_1 \in A_s^1} \{d(s, a_1, a_2(s))^{-1}[b(s, a_1, a_2(s))w(s) - \Theta + c(s, a_1, a_2(s))]\} \end{aligned}$$

for every $s \in S$ which is a continuity point of $a(s)$, and, for $j = 0$ and $j = 1$, $\sigma_j > 0$ implies

$$\begin{aligned} &\max_{a_2 \in A_{r_j}^2} c(r_j, a_1(r_j), a_2) = c(r_j, a_1(r_j), a_2(r_j)) \\ (27) \qquad & \\ &\qquad = \min_{a_1 \in A_{r_j}^1} c(r_j, a_1, a_2(r_j)). \end{aligned}$$

Moreover, if $a(s)$ is optimal, then $\Theta = \Theta(a_1, a_2)$ is the value of the game.

This proof is essentially identical to that for Theorem 11, so it will be omitted. A diffusion process zero sum, two-person game problem in the undiscounted cost, conservative process case can, in principle, be solved in the same manner as with the discounted cost case. Moreover, there exist sufficient conditions analogous to those of Theorem 12 and Corollary 13 for equations (26) and (27) to be satisfied by some Borel measurable control $a(s)$.

Example.

$$\begin{aligned} S &= [0, 1], & d(s, a_1, a_2) &= A > 0, \\ A_s^1 &= [-Y, Y] \subset E, \quad Y > 0, & b(s, a_1, a_2) &= a_1 a_2, \\ A_s^2 &= [-Z, Z] \subset E, \quad Z > 0, & c(s, a_1, a_2) &= Cs. \end{aligned}$$

Suppose both boundary conditions are pure reflection. For any values of s, $w(s)$, and Θ we have

$$\min_{a_1 \in A_s^1} \max_{a_2 \in A_s^2} \{A^{-1}[a_1 a_2 w(s) - \Theta + Cs]\}$$
$$= \max_{a_2 \in A_s^2} \min_{a_1 \in A_s^1} \{A^{-1}[a_1 a_2 w(s) - \Theta + Cs]\} = A^{-1}[Cs - \Theta],$$

so $a(s) = (0, 0)$ is the optimal control for all $s \in S$. Therefore, the value of this game is given by (3), (4) to be $\Theta = C/2$, with $w(s) = (\Theta/A)s - \frac{1}{2}(C/A)s^2$.

5. The nonzero sum. N-person game problem. The remainder of this paper describes a class of controlled diffusion processes whose control problems can be viewed as nonzero sum, N-person games. We consider the multiperson controlled diffusion process of § 1; these processes are controlled by N persons and generate N streams of costs ($N \geqq 2$). Controller $i\,(i = 1, \cdots, N)$, who operates the ith control, endeavors to choose a control $a_i \in M_i$ so as to minimize the costs of the ith cost stream generated by the process. A game situation exists by virtue of the fact that the cost to the ith person is influenced by the actions of the other players.

The optimality criterion used for these processes is that of a Nash equilibrium point. If an admissible expected cost is defined to be the expected cost corresponding to some admissible control, then the solution to this game will be some admissible control whose corresponding expected cost is a Nash equilibrium point with respect to all admissible costs. This if player i unilaterally deviates from his component of this optimal control, then his expected costs will either be unchanged or increased. The adoption of the Nash equilibrium point optimality criterion is made in recognition of the fact that a variety of meritorious optimality criteria exist for nonzero sum, N-person game problems. In particular, a "prisoner's dilemma" situation might exist where the players would gain by deviating from the Nash equilibrium point solution in a cooperative manner.

The following two sections provide results respectively for the discounted cost case and the undiscounted cost case. The main result of each section is a necessary and sufficient condition for a control to be optimal. In addition, a method based upon the theory of differential games is provided for solving a diffusion process nonzero sum, N-person game problem. This method is substantially the same as a method used for solving an ordinary diffusion process optimal control problem. The optimizing solution of an equation is substituted into a differential equation whose solution, in turn, is used to obtain the optimal control.

To minimize ambiguity, the following terminology is used. The ith control $a_i \in M_i$ is operated by the ith player and is generally a vector-valued function on S. The *control* $a = (a_1, \cdots, a_N)$ is the vector consisting of the N players' controls.

6. The nonzero sum problem with discounted costs. Let $v(s, a) = v(s, a_1, \cdots, a_N)$ denote the expected discounted cost of a process corresponding to the admissible control $a \in M$, and let $v_i(s, a)$ be its ith component, $i = 1, \cdots, N$. Then $v(s, a)$ will be the unique solution of (1), (2). The control $\hat{a} \in M$ is said to be optimal, that is, a solution of the game if it is a Nash equilibrium point of the expected discounted cost functions; that is,

$$v_i(s, \hat{a}) \leqq v_i(s, \hat{a}_1, \cdots, \hat{a}_{i-1}, a_i, \hat{a}_{i+1}, \cdots, \hat{a}_N)$$

for all $s \in S$, all $a_i \in M_i$, and each $i = 1, \cdots, N$. In this case $v(s, \hat{a})$ is said to be a value of the game.

To simplify our notation, define

$$\operatorname*{val}_{z \in Z} g(z) \equiv \{g(z) | z \text{ is a Nash equilibrium point of } g \text{ on } Z\},$$

where $Z_i \subset E$ for $i = 1, \cdots, N, Z \equiv Z_1 \times \cdots \times Z_N$, and the function $g : Z \to E^N$. The main result of this section is the following.

THEOREM 18. *A control $\hat{a} \in M$ is optimal if and only if for each $s \in S$ which is a continuity point of $\hat{a}(s)$,*

$$\begin{aligned} &d(s, \hat{a}(s))^{-1}[b(s, \hat{a}(s))v'(s) - \lambda v(s) + c(s, \hat{a}(s))] \\ &\quad \in \operatorname*{val}_{a \in A_s} \{d(s, a)^{-1}[b(s, a)v'(s) - \lambda v(s) + c(s, a)]\}, \end{aligned} \tag{28}$$

where $v(s) \equiv v(s, \hat{a})$, and

$$\sigma_j(c(r_j, a(r_j)) - \gamma_j) = 0, \qquad j = 0, 1, \tag{29}$$

where

$$\gamma_j \in \operatorname*{val}_{a \in A_{r_j}} c(r_j, a), \qquad j = 0, 1.$$

Proof. Let $\hat{a}$ be optimal. For arbitrary i let $\hat{a}_1, \cdots, \hat{a}_{i-1}, \hat{a}_{i+1}, \cdots, \hat{a}_N$ be fixed so that

$$v_i(s, \hat{a}) = \inf_{a_i \in M_i} v_i(s, \hat{a}_1, \cdots, \hat{a}_{i-1}, a_i, \hat{a}_{i+1}, \cdots, \hat{a}_N)$$

is the minimal expected discounted cost of an optimal control problem and $\hat{a}_i$ is one of its optimal controls. By Pliska [8, Theorem 1] we have

$$\begin{aligned} &d(s, \hat{a}(s))^{-1}[b(s, \hat{a}(s))v_i'(s) - \lambda v_i(s) + c_i(s, \hat{a}(s))] \\ &\quad = \min_{a_i \in A^i_s} \{d(s, \hat{a}_1(s), \cdots, a_i, \cdots, \hat{a}_N(s))^{-1}[b(s, \hat{a}_1(s), \cdots, a_i, \cdots, \hat{a}_N(s))v_i'(s) \\ &\qquad - \lambda v_i(s) + c_i(s, \hat{a}_1(s), \cdots, a_i, \cdots, \hat{a}_N(s))]\} \end{aligned} \tag{30}$$

for each $s \in S$ which is a continuity point of $\hat{a}_i(s)$ and

$$\sigma_j(c_i(r_j, \hat{a}(r_j)) - \gamma_{ij}) = 0, \qquad j = 0, 1, \tag{31}$$

where

$$\gamma_{ij} = \min_{a_i \in A^i_{r_j}} c_i(r_j, \hat{a}_i(r_j), \cdots, a_i, \cdots, \hat{a}_N(r_j)), \qquad j = 0, 1.$$

Since i is arbitrary, (28) and (29) must be true.

Conversely, suppose (28) and (29) hold and let i be arbitrary. Now (30) and (31) must hold, so by Theorem 1 of Pliska [8] we see that $v_i(s, \hat{a})$ is the minimal expected discounted cost of the ordinary optimal control problem: minimize $v_i(s, \hat{a}_1, \cdots, a_i, \cdots, \hat{a}_N)$ subject to $a_i \in M_i$. Since i is arbitrary, $\hat{a}$ defines a Nash equilibrium point for this game.

Theorem 18 is substantially different from Theorem 11 for the zero sum, two-person game situation in one respect. In each case the necessary and sufficient condition is a function of the solution to a differential equation. With Theorem 18, this solution is explicitly a function of some control $a \in M$, whereas in the case of

Theorem 11 the corresponding differential equation solution is explicitly independent of any control $a \in M$. Thus, given a control $a \in M$, one can determine $v(s, a)$ with Theorem 1 and then ascertain whether $a(s)$ is optimal with Theorem 18. Conversely, an optimal control $a(s)$ will satisfy (28) and (29). However, Theorem 18 does not provide an explicit procedure for solving the diffusion process nonzero sum, N-person game problem.

The following computational procedure is based upon a method devised by Starr and Ho [11] as well as Case [2] for solving nonzero sum differential games. Let $t = (t_1, \cdots, t_N)$ and $u = (u_1, \cdots, u_N)$ and define $g(s, t, u, a): S \times E^{2N} \times K \to E^N$ by

$$g(s, t, u, a) \equiv d(s, a)^{-1}[b(s, a)u - \lambda t + c(s, a)].$$

Now consider the point-to-set map $\Gamma: S \times E^{2N} \to K$ defined by

$$\Gamma(s, t, u) = \{a \in A_s | g(s, t, u, a) \in \operatorname*{val}_{a \in A_s} g(s, t, u, a)\}.$$

If $\sigma_j > 0$ for $j = 0$ or $j = 1$, then redefine $\Gamma(r_j, t, u)$ so that

$$\Gamma(r_j, t, u) = \{a \in A_{r_j} | c(r_j, t, u, a) \in \operatorname*{val}_{a \in A_{r_j}} c(r_j, t, u, a)\}.$$

If $\Gamma(s, t, u) \neq \varnothing$ for each (s, t, u), then choose a function $a(s, t, u)$ with $a(s, t, u) \in \Gamma(s, t, u)$ for each (s, t, u), substitute $a(s, v(s), v'(s))$ for $a(s)$ in (1) and (2), and solve for $v(s)$. If $v(s)$ exists, then it is a value of the game. If $a(s) = a(s, v(s), v'(s))$ is piecewise continuous, then it is an optimal control and $v(s) = v(s, a)$.

Note that this procedure may break down in three different ways: $\Gamma(s, t, u)$ may not exist, $v(s)$ may not exist, and $a(s, v(s), v'(s))$ may not be piecewise continuous. The reason why $v(s)$ may fail to exist, although $a(s, t, u)$ does, is that the Euclidean norm of $g(s, t, u, a(s, t, u))$ may fail to be continuous on $S \times E^{2N}$. Since most differential equation theory existence theorems specify some form of continuity requirement, counterexamples can be easily constructed. The following proposition serves to characterize Γ.

PROPOSITION 19. *If A_s is continuous on S, then Γ is a closed map on $S \times E^{2N}$.*

Proof. For each $i = 1, \cdots, N$, define the point-to-set map $\Gamma_i: S \times E^{2N} \times K \to K_i$ by

$$\Gamma_i(s, t, u, a) = \operatorname*{arg\,min}_{a_i \in A_s^i} g_i(s, t, u, a).$$

The continuity of $g_i(s, t, u, a)$ implies that Γ_i is a closed map, so its graph $D_i = \{(s, t, u, a) | a_i \in \Gamma_i(s, t, u, a)\}$ is a closed set. The map Γ is thus closed because its graph $D_1 \cap \cdots \cap D_N$ is closed.

PROPOSITION 20. *Suppose each component i of $g(s, t, u, a)$ is quasi-convex in $a_i \in K_i$ for each $s \in S$, each $a_j \in K_j$ $(j = 1, \cdots, i - 1, i + 1, \cdots, N)$, and each $t, u \in E^N$, and assume A_s^i is convex for each $s \in S$ and $i = 1, \cdots, N$. Then $\Gamma(s, t, u) \neq \varnothing$ for each $(s, t, u) \in S \times E^{2N}$.*

The proof of Proposition 20 is omitted because it follows easily from Rosen [9] and Sion [10]. If Proposition 20 holds and the Nash equilibrium point is unique for each (s, t, u), then the closed map $\Gamma(s, t, u)$ is simply a piecewise continuous function. This observation leads to the following existence theorem. Following

Rosen [9], the function $g:E^N \to E^N$ is said to be diagonally strictly convex for $a \in K$ if for each $a^0, a^1 \in K$ we have

$$(a^1 - a^0)^T f(a^0) + (a^0 - a^1)^T f(a^1) < 0,$$

where $f(a) = (\partial g_1/\partial a_1, \cdots, \partial g_N/\partial a_N)^T$. A sufficient condition that $g(a)$ be diagonally strictly convex is that the symmetric matrix $[F(a) + F^T(a)]$ be positive definite for $a \in K$, where $F(a)$ is the Jacobian with respect to a of $f(a)$ (Rosen [9]).

THEOREM 21. *Assume A_s is continuous on S with A_s^i convex for each $s \in S$ and $i = 1, \cdots, N$. Also assume $d(s, a)$ is constant in a, $b(s, a)$ is affine in a, and $c(s, a)$ is diagonally strictly convex in a, all for each $s \in S$. Then a solution exists for this diffusion process game.*

Proof. The function $g(s, t, u, a)$ is strictly diagonally convex, so by Rosen [9] there exists a unique Nash equilibrium point for each (s, t, u), that is, by the above remarks, $\Gamma(s, t, u)$ is a continuous function on $S \times E^{2N}$. By differential equation theory and the arguments of § 3, there exists a solution $v(s)$ to (1), (2) with $\Gamma(s, v(s), v'(s))$ substituted for $a(s)$. Hence a solution of the game is $a(s) = \Gamma(s, v(s), v'(s)) \in M$, and the corresponding value of the game is $v(s) = v(s, a)$.

Example. This example is a two-person game. Let the state space and sets of admissible control values equal the unit interval. Let $d(s, a_1, a_2) = 1$, $b(s, a_1, a_2) = a_1 + a_2$, and $c_i(s, a_1, a_2) = C + a_i$, $i = 1, 2$, where C is a constant. Suppose the boundary condition at r_0 is reflection and that absorption occurs at r_1 with cost λ_1. Following the above procedure, we have for $i = 1, 2$ that $a_i(s, t, u) = 1$ for $u_i \leqq 1$ and $a_i(s, t, u) = 0$ otherwise. By symmetry we have that $v_1(s) = v_2(s)$ is the solution $v(s)$ to

$$v''(s) = -2a_1(s, v(s), v'(s))v'(s) + \lambda v(s) - K - a_1(s, v(s), v'(s)) \tag{32}$$

satisfying $v'(r_0) = 0$ and $v(r_1) = \lambda_1$. In some neighborhood of r_0 we have $v'(s) > -1$, so in this neighborhood $a(s, v(s), v'(s)) = (0, 0)$ and, for some constant q, $v(s) = q\,e^{\sqrt{\lambda}s} + q\,e^{-\sqrt{\lambda}s} + C/\lambda$. If

$$\lambda_1 \geqq \frac{e^{\sqrt{\lambda}} + e^{-\sqrt{\lambda}}}{\sqrt{\lambda}(e^{-\sqrt{\lambda}} - e^{\sqrt{\lambda}})} + \frac{C}{\lambda},$$

then q can be chosen so that $v(1) = \lambda_1$ and $v'(s) \geqq -1$ for all $s \in S$, in which case $a(s) = (0, 0)$ is optimal for all $s \in S$. If

$$\lambda_1 < \frac{e^{\sqrt{\lambda}} + e^{-\sqrt{\lambda}}}{\sqrt{\lambda}(e^{-\sqrt{\lambda}} - e^{\sqrt{\lambda}})} + \frac{C}{\lambda},$$

then for some $s_0 \in (0, 1)$ we have $v'(s_0) = -1$ and $a(s) = (0, 0)$ optimal for all $s \in [0, s_0)$. For $v(s)$ to exist, we must have $v''(s_0) < 0$ when $a_1(s_0, v(s_0), v'(s_0)) = 1$ in (32). But this is easily verified, so for all $s \geqq s_0$ in some neighborhood of s_0 we have $a(s) = (1, 1)$ optimal and $v(s) = t_1 e^{u_1 s} + t_2 e^{u_2 s} + (C + 1)/\lambda$, where $u_1 = -1 + \sqrt{1 + \lambda}$, $u_2 = -1 - \sqrt{1 + \lambda}$, and t_1 and t_2 are constants.

It remains to show that $a(s) = (1, 1)$ is optimal for all $s \geqq s_0$. Suppose not, but that $s_1 < 1$, say, is the smallest $s > s_0$ such that $v'(s_1) = -1$. Then $v''(s_1) < v''(s_0) < 0$, a contradiction. The unknown constants q, t_1, t_2, and s_0 can be solved from the boundary conditions and the fact that $v'(s)$ is continuous with $v'(s_0) = -1$.

7. The nonzero sum problem with undiscounted costs. The nonzero sum, N-person diffusion process game problem with undiscounted costs will be one of two types, depending on whether the boundary conditions are conservative or nonconservative. The results in this section parallel those of § 4 and § 6, and, consequently, they will be brief. The conservative case will be treated in the second half of this section.

For the purposes of this section, the boundary conditions are said to be nonconservative if at least one boundary is absorbing and neither boundary is purely adhesive, that is,

$$\kappa_0 + \kappa_1 > 0, \quad \kappa_j + \pi_j + \theta_j > 0, \qquad j = 0, 1.$$

Let $v(s, a) = v(s, a_1, \cdots, a_N) = v(s)$ denote the expected undiscounted cost of such a process corresponding to the admissible control $a \in M$. Then $v(s, a)$ will be the unique solution of (1), (2) with $\lambda = 0$. The control $a \in M$ is said to be optimal if it defines a Nash equilibrium point with respect to the expected cost functions, that is,

$$v_i(s, a) \leqq v_i(s, \hat{a}_1, \cdots, \hat{a}_{i-1}, a_i, \hat{a}_{i+1}, \cdots, \hat{a}_N)$$

for all $s \in S$, all $a_i \in M_i$, and each $i = 1, \cdots, N$. In this case $v(s, \hat{a})$ is said to be a value of the game.

THEOREM 22. *With nonconservative boundary conditions, a control $a \in M$ is optimal if and only if for each $s \in S$ which is a continuity point of $a(s)$,*

$$d(s, a(s))^{-1}[b(s, a(s))v'(s) + c(s, a(s))] \in \operatorname*{val}_{a \in A_s}\{d(s, a)^{-1}[b(s, a)v'(s) + c(s, a)]\},$$

where $v(s) = v(s, a)$, and

$$\sigma_j(c(r_j, a(r_j)) - \gamma_j) = 0, \qquad j = 0, 1,$$

where $\gamma_j \in \mathrm{val}_{a \in A_{r_j}} c(r_j, a)$, $j = 0, 1$.

The proof is essentially the same as that for Theorem 18, so it will be omitted. Moreover, the remarks and computational procedure that follow Theorem 18 apply to this case as well.

Example. This example is identical to that of the preceding section except that the costs are undiscounted. Proceeding in a similar manner, we have that $v_1(s) = v_2(s)$ is the solution $v(s)$ to

$$v''(s) = -2a_1(s, v(s), v'(s))v'(s))v'(s) - C - a_1(s, v(s), v'(s))$$

satisfying $v'(r_0) = 0$ and $v(r_1) = \lambda_1$, and that $a_1(s, v(s), v'(s)) = a_2(s, v(s), v'(s)) = 1\ (=0)$ if $v'(s) \leqq -1\ (\geqq -1)$. If $C \leqq 1$, then $a(s) = (0, 0)$ is optimal for all $s \in S$ and $v(s) = \lambda_1 + C(1 - s^2)/2$. If $C > 1$, then $a(s) = (0, 0)$ is optimal and $v(s) = \lambda_1 + C/2 + (C - 1)(\exp(-2 + 2/C) - 1)/4 - Cs^2/2$ on $[0, 1/C)$, and $a(s) = (1, 1)$ is optimal and $v(s) = \lambda_1 + (C + 1)(1 - s)/2 + \exp(-2 + 2/C)(1 - \exp(2 - 2s))(C - 1)/4$ on $(1/C, 1]$.

We now discuss the other type of undiscounted cost problem, the conservative case. For purposes of this section, the boundary conditions are said to be conservative if neither boundary is absorbing and at least one boundary is not purely adhesive, that is,

$$\kappa_0 + \kappa_1 = 0, \qquad \pi_0 + \theta_0 + \pi_1 + \theta_1 > 0.$$

Let $\Theta(a) = \Theta(a_1, \cdots, a_N) = \Theta$ denote the vector of mean costs per unit time of such a process corresponding to the control $a \in M$. Then $\Theta(a)$ is the unique vector to which there exists a solution $w(s, a)$ to (3) and (4). The control $\hat{a} \in M$ is said to be optimal if it defines a Nash equilibrium point with respect to the mean costs, that is,

$$\Theta_i(a) \leqq \Theta_i(\hat{a}_1, \cdots, \hat{a}_{i-1}, a_i, \hat{a}_{i+1}, \cdots, \hat{a}_N)$$

for all $a_i \in M_i$, $i = 1, \cdots, N$. In this case $\Theta(\hat{a})$ is said to be a value of the game.

THEOREM 23. *With conservative boundary conditions, a control $a \in M$ is optimal if and only if for each $s \in S$ which is a continuity point of $a(s)$,*

$$d(s, a(s))^{-1}[b(s, a(s))w(s) - \Theta + c(s, a(s))]$$
$$\in \operatorname*{val}_{a \in A_s} \{d(s, a)^{-1}[b(s, a)w(s) - \Theta + c(s, a)]\},$$

where $w(s) = w(s, a)$ and $\Theta = \Theta(a)$, and

$$\sigma_j(c(r_j, a(r_j)) - \gamma_j) = 0, \qquad j = 0, 1,$$

where $\gamma_j \in \mathrm{val}_{a \in A_{r_j}} c(r_j, a), j = 0, 1$.

The proof is essentially the same as that for Theorem 18, so it will be omitted. Moreover, the remarks and computational procedure that follow Theorem 18 apply to this case as well.

Example. This is an example of an N-person game. Let $S = A_s^i = [-1, 1]$ for $i = 1, \cdots\cdot, N$ and all $s \in S$, $d(s, a) = 1$, $b(s, a) = a_1 + \cdots + a_N$, and $c_i(s, a) = |s|$ for $i = 1, \cdots, N$. Suppose reflection occurs at each boundary. The ith control $a_i(s) = -1$ if $w_i(s) \geqq 0$ and $a_i(s) = 1$ otherwise. By symmetry, $w_1(s) = \cdots = w_N(s)$ so $w_1(s)$ is the solution to

$$w_1'(s) = \begin{cases} -Nw_1(s) - \Theta + |s|, & w_1(s) \geqq 0, \\ Nw_1(s) - \Theta + |s|, & w_1(s) \leqq 0, \end{cases}$$

satisfying $w_1(1) = w_1(-1) = 0$. By symmetry we must have $w_1(0) = 0$, so we verify that $\Theta = 1/(1 - e^{-N}) - 1/N$ and that for $i = 1, \cdots, N$,

$$a_i(s) = \begin{cases} 1, & s \in [-1, 0), \\ -1, & s \in (0, 1]. \end{cases}$$

Acknowledgments. The author is grateful to Martin L. Puterman and Robert W. Rosenthal for their helpful comments during the course of this research. The author is especially indebted to Arthur F. Veinott, Jr., for his guidance of the author's doctoral dissertation, on part of which this paper is based.

REFERENCES

[1] C. BERGE, *Topological Spaces*, Oliver & Boyd, London, 1963.

[2] J. CASE, *Toward a theory of many player differential games*, this Journal, 7 (1969), pp. 179–197.

[3] R. EDWARDS, *Functional Analysis, Theory and Applications*, Holt, Rinehart & Winston, New York, 1965.

[4] W. FELLER, *The parabolic differential equations and the associated semi-groups of transformations*, Ann. of Math., 55 (1952), pp. 468–519.

[5] ———, *Generalized second order differential operators and their lateral conditions*, Illinois J. Math., 1 (1957), pp. 459–504.

[6] I. GIRSANOV, *Minimax problems in the theory of diffusion processes*, Dokl. Akad. Nauk SSSR, 136 (1961), pp. 761–764.

[7] P. MANDL, *Analytical Treatment of One-dimensional Markov Processes*, Springer-Verlag, New York, 1968.

[8] S. PLISKA, *Single person controlled diffusions with discounted costs*, J. Optimization Theory Appl., 12 (1973).

[9] J. ROSEN, *Existence and uniqueness of equilibrium points for concave N-person games*, Econometrica, 33 (1965), pp. 520–534.

[10] M. SION, *On general minimax theorems*, Pacific J. Math., 8 (1958), pp. 171–176.

[11] A. STARR AND Y. HO, *Further properties of non-zero sum differential games*, J. Optimization Theory Appl., 3 (1969), pp. 207–219.

SIAM J. Control
Vol. 11, No. 4, November 1973

ON THE EXISTENCE OF OPTIMAL POLICIES IN STOCHASTIC CONTROL*

M. H. A. DAVIS†

Abstract. In this paper a sufficient condition is presented for the existence of an optimal control for a system described by stochastic differential equations, the solutions of which are defined for any nonanticipative control policy, by the Girsanov measure transformation technique. The result is that if a Hamiltonian function achieves its infimum pointwise then an optimal nonanticipative policy exists.

1. Introduction. In this paper conditions are given for the existence of an optimal control policy for a system described by stochastic functional differential equations of the form

$$dz_t = f(t, z, u_t)\,dt + \sigma(t, z)\,dB_t, \qquad t \in [0, 1], \tag{1.1}$$

where $\{z_t\}$ is the state and $\{B_t\}$ is a vector process of independent Brownian motions. The drift f depends at any time t on the past $\{z_s, s \leqq t\}$ of the state and on a control u_t which takes values in a metric space U and is itself allowed to depend on the complete past. The control policy is to be chosen so as to minimize the cost

$$J(u) = E\int_0^1 h(t, z, u_t)\,dt.$$

The solution to (1.1) is defined by the Girsanov measure transformation technique (see § 2). The result is that if f satisfies the technical conditions F1–F3 in § 2, then an optimal policy exists so long as a Hamiltonian function (2.4) achieves its absolute minimum at each time t and past trajectory up to t.

This problem has been considered in the same framework by Beneš [2] and Duncan and Varaiya [5]. The result here is obtained under significantly weaker hypotheses (the stipulation that $f(t, x, U)$ be convex, made in [2], [5], is not required here); and the approach is entirely different. References [2] and [5] are based on compactness arguments, whereas the present paper relies on the Hamilton–Jacobi theory developed in [4] to construct an optimal control policy (though no computational scheme is, even in principle, provided). The methods are in fact more akin to those of Beneš' earlier paper [1] and to those used in Markovian problems [6], [11]. See § 4 for further comments on these points.

2. The notation is as follows: C is the set of continuous functions from $[0, 1]$ to R^n, $\{z_t\}$ is the family of evaluation functionals on C, and, for each $t \in [0, 1]$, $\mathscr{F}_t$ is the σ-field of subsets of C generated by $\{z_s, s \leqq t\}$. $\mathscr{A}$ is the σ-field of subsets A of $[0, 1] \times C$ having the property that the section of A at t is in $\mathscr{F}_t$ for each $t \in [0, 1]$ and the section of A at x is Lebesgue measurable for each $x \in C$. A function g on $[0, 1] \times C$ is $\mathscr{A}$-measurable if and only if $g(t, \cdot)$ is $\mathscr{F}_t$-measurable for each t and $g(\cdot, x)$ is Lebesgue measurable for each x.

Let U be a separable metric space which is a countable union of its compact subsets and $\mathscr{Q}$ be the Borel sets of U. An *admissible control policy* is a measurable

* Received by the editors June 29, 1972, and in revised form November 7, 1972.

† Department of Computing and Control, Imperial College, London SW7 2BT, England.

function $u:([0,1] \times C, \mathscr{A}) \to (U, \mathscr{Q})$. The set of such functions is denoted by $\mathscr{U}$.

The function f is assumed to satisfy the following conditions ($\mathscr{R}^n$ is the Borel σ-field of R^n):

(F1) $f:[0,1] \times C \times U \to R^n$ is measurable with respect to the σ-field $\mathscr{A} * \mathscr{Q}$.

(F2) For each $(t,x) \in [0,1] \times C$, $f(t,x,\cdot)$ is continuous on U.

(F3) There exists a constant K such that

$$|f(t,x,u)| \leqq K(1 + \|x\|) \quad \text{for all } (t,x,u) \in [0,1] \times C \times U,$$

where $\|\cdot\|$ is the uniform norm in C.

$\sigma(t,x)$ is an $n \times n$ matrix for each $(t,x) \in [0,1] \times C$ whose elements $\sigma^{ij}(t,x)$ are $\mathscr{A}$-measurable functions and satisfy a uniform Lipschitz condition in x. The inverse matrix $\sigma^{-1}(t,x)$ is assumed to exist and be bounded for $(t,x) \in [0,1] \times C$. The *cost rate* h is a bounded function from $[0,1] \times C \times U \to R^+$ which also satisfies F1–F2.

Let $(\Omega, \mathscr{B}, \mu)$ be a probability space carrying an n-dimensional separable Brownian motion process $\{w_t\}$. With the above conditions the following equation has a unique solution (with fixed initial value $z_0 \in R^n$):

$$d\xi_t = \sigma(t,\xi)\,dw_t.$$

Since $\{\xi_t\}$ has continuous paths it induces a measure P_0 on its sample space $(C, \mathscr{F}_1)$ according to the formula

$$P_0 A = \mu\{\omega : \xi_{.}(\omega) \in A\}, \qquad A \in \mathscr{F}_1.$$

The probability space $(C, \mathscr{F}_1, P_0)$ will be taken as the "basic space" from now on.

Let Φ denote the set of $\mathscr{A}$-measurable functions $\phi:[0,1] \times C \to R^n$ such that $|\phi(t,x)| \leqq K(1 + \|x\|)$. For $\phi \in \Phi$ define

$$\zeta(\phi) = \int_0^1 \phi_t \cdot a_t^{-1}\,dz_t - \frac{1}{2}\int_0^1 \phi_t \cdot a_t^{-1}\phi_t\,dt,$$ [1]

where

$$\phi_t = \phi(t,z)$$

and

$$a_t = \{a^{ij}(t,z)\} = \sigma(t,z)\sigma'(t,z).$$

Let

$$\mathscr{D}(\Phi) = \{\exp[\zeta(\phi)] : \phi \in \Phi\}.$$

LEMMA 1 (Beneš [2], Girsanov [7]). *Let the measure P_ϕ on $(C, \mathscr{F}_1)$ be defined by*

$$P_\phi A = \int_A \exp[\zeta(\phi)]\,dP_0, \qquad A \in \mathscr{F}_1.$$

[1] The dot denotes the R^n inner product.

Then

(i) P_ϕ *is a probability measure*;

(ii) P_ϕ *is mutually absolutely continuous with respect to* P_0.

(iii) $\{B_t, t \in [0, 1]\}$ *is a Brownian motion under* P_ϕ, *where*

$$dB_t = \sigma^{-1}(t, z)[dz_t - \phi(t, z)\,dt].$$

Lemma 1 allows the solution of (1.1) to be defined in the following way: for $u \in \mathscr{U}$ and $(t, x) \in [0, 1] \times C$, define

$$\begin{aligned} f^u(t, x) &= f(t, x, u(t, x)), \\ h^u(t, x) &= h(t, x, u(t, x)). \end{aligned} \tag{2.1}$$

Now $f^u \in \Phi$; denoting $P_u = P_{f^u}$, P_u is the measure (it can be shown to be unique; see [5]) corresponding to the solution of (1.1) in the sense that under P_u,

$$dz_t = f(t, z, u(t, z))\,dt + \sigma(t, z)\,dB_t, \tag{2.2}$$

where $\{B_t\}$ is a Brownian motion.

The purpose of condition (F3) is to ensure that (i) holds in Lemma 1, i.e., $P_\phi C = 1$; (F3) could be replaced by any other sufficient condition for this. Let E_u denote integration with respect to measure P_u; then the cost of a control policy $u \in \mathscr{U}$ is

$$J(u) = E_u \int_0^1 h(t, z, u(t, z))\,dt. \tag{2.3}$$

It is the objective of the controller to minimize the cost. The main result of this paper is the following. For $p \in R^n, (t, x, u) \in [0, 1] \times C \times U$, the Hamiltonian function $\mathscr{H}'$ is defined by

$$\mathscr{H}'(t, x, u, p) = p \cdot f(t, x, u) + h(t, x, u). \tag{2.4}$$

The following condition will be required:

(H) For each $(t, x, p) \in [0, 1] \times C \times R^n$, $\mathscr{H}'(t, x, p, \cdot)$ achieves its minimum; i.e., there exists $u_0 \in U$ such that

$$\mathscr{H}(t, x, p) \triangleq \min_{u \in U} \mathscr{H}'(t, x, p, u) = \mathscr{H}'(t, x, p, u_0).$$

Notice that (H) is satisfied if, for example, U is compact.

THEOREM. *Suppose condition* (H) *is satisfied. Then an optimal policy exists.*

The proof of the theorem, and the form of the optimal control policy, are given in the next section. It will be seen that the result follows easily from the results of [4], encapsulated in Lemma 2 below, together with Lemma 1 of [1]. The idea in [4] is to notice that the value function $\{W_t\}$ is a semimartingale [9], and to show, using martingale-theoretic methods, that its bounded variation and local martingale components are as in (3.3). Thus one might say that the "basic existence results" at work here are really Meyer's decomposition of supermartingales [8] and Beneš' implicit function lemma.

3. Full details of the following will be found in [4] and [10].

For $u \in \mathscr{U}$ and $t \in [0, 1]$, the expected remaining cost from time t on is

$$\psi_u(t) = E_u\left[\int_t^1 h^u(s, z)\, ds | \mathscr{F}_t\right].$$

It can be shown that $\psi_u(t)$ only depends on the control policy used on $[t, 1]$. Since $\{\psi_u(t) : u \in \mathscr{U}\}$ is a subset of $L_\infty(C, \mathscr{F}_t, P_0)$, which is a complete lattice, the following infimum exists in L_∞:

$$W_t = \bigwedge_{u \in \mathscr{U}} \psi_u(t). \tag{3.1}$$

This is the "value function" for the control problem. Notice in particular that

$$\begin{aligned} W_1 &= 0 \quad \text{a.s.}, \\ W_0 &= \inf_{u \in \mathscr{U}} J(u) \triangleq J^*. \end{aligned} \tag{3.2}$$

It is shown in [4] that W_t has an Ito process representation. Using this, conditions for optimality analogous to the Hamilton–Jacobi equation can be obtained, involving the drift and diffusion terms in the representation of W_t. These facts may be summarized as follows.

LEMMA 2. (i) *There exist processes* $\{\Lambda W_t\}, \{\nabla W_t\}$ *taking values in* R, R^n *respectively, and adapted to* $\mathscr{F}_t$,[2] *such that*

$$\begin{gathered} \int_0^1 |\nabla W_t|^2\, dt < \infty \quad a.s., \qquad E\int_0^1 |\Lambda W_t|\, dt < \infty, \\ W_t = J^* + \int_0^t \Lambda W_s\, ds + \int_0^t \nabla W_s\, dz_s \end{gathered} \tag{3.3}$$

almost surely under measure P_0.

(ii) *For any* $u \in \mathscr{U}$,

$$\Lambda W_t + \nabla W_t \cdot f(t, z, u_t) + h(t, z, u_t) \geqq 0 \tag{3.4}$$

for almost all (t, z). $u^0 \in \mathscr{U}$ *is optimal if and only if* (3.4) *holds with equality a.e. for* $u = u^0$.

The next result about the set of densities $\mathscr{D}(\Phi)$ will also be required.

LEMMA 3. $\mathscr{D}(\Phi)$ *is compact in the* L_∞ *topology of* L_1.

This fact is not explicitly stated in the literature but is implicit in [2], [5]. Lemma 1 of [2] shows that $\mathscr{D}(\Phi)$ is uniformly integrable, and hence relatively compact in $\sigma(L_1, L_\infty)$. The argument of Theorem 2 of [5] shows that $\mathscr{D}(\Phi)$ is convex and strongly closed, and hence weakly closed.

Proof of Theorem. The optimal policy u^* is constructed in the following way: for each (t, x, p), $y^*(t, x, p)$ is a control value which minimizes $\mathscr{H}'$, i.e.,

$$\mathscr{H}(t, x, p) = p \cdot f(t, x, y^*[t, x, p]) + h(t, x, y^*[t, x, p]). \tag{3.5}$$

[2] I.e., the r.v.'s at each time t are $\mathscr{F}_t$-measurable.

Then the function $u^*: [0, 1] \times C \to U$ is defined by

$$u^*(t, x) = y^*(t, x, \nabla W[t, x]), \tag{3.6}$$

where ∇W_t is the function appearing in Lemma 2. Thus the proof consists of showing, first that $u^* \in \mathscr{U}$, i.e., that the function y^* can be selected in a suitably measurable way, then that u^* achieves the minimal cost.

For fixed (t, x, p), $\mathscr{H}'$ is continuous in u, while for fixed $u \in U$, it is evident that $\mathscr{H}'$ is measurable with respect to $\mathscr{M} = \mathscr{A} * \mathscr{R}^n$ in (t, x, p). Let S be a countable dense subset of U. Since $\mathscr{H}'$ is continuous in u,

$$\mathscr{H}(t, x, p) = \inf_{u \in S} \mathscr{H}'(t, x, p, u).$$

Thus

$$\{t, x, p : \mathscr{H}(t, x, p) < a\} = \bigcup_{u \in S} \{t, x, p : \mathscr{H}'(t, x, p, u) < a\}$$

so that $\mathscr{H}$ is $\mathscr{M}$-measurable. In view of (H),

$$\mathscr{H}(t, x, p) \in \mathscr{H}'(t, x, p, U) \quad \text{for each } (t, x, p).$$

According to Lemma 1 of [1] these facts are sufficient to guarantee the existence of an $\mathscr{M}$-measurable function $y^*: [0, 1] \times C \times R^n \to U$ satisfying (3.5). Define the measurable function $\psi: ([0, 1] \times C, \mathscr{A}) \to ([0, 1] \times C \times R^n, \mathscr{M})$ by:

$$\psi(t, x) = (t, x, \nabla W(t, x)).$$

Then, from (3.6), $u^* = y^* \circ \psi$, so that u^* is $\mathscr{A}$-measurable, i.e., $u^* \in \mathscr{U}$.

To prove that u^* is optimal one has to show that

$$\Lambda W(t, z) = -[\nabla W_t \cdot f(t, z, u^*(t, z)) + h(t, z, u^*(t, z))] \quad \text{a.e.} \tag{3.7}$$

For denote the right side of (3.7) by $\alpha(t, z)$ and let $u \in \mathscr{U}$. Using (2.2), the representation (3.3) of W_t can be written

$$W_t = J^* + \int_0^t (\Lambda W_s + \nabla W_s \cdot f_s^u)\, ds + \int_0^t \nabla W_s \cdot \sigma_s\, dB_s,$$

where $(B_t, \mathscr{F}_t, P_u)$ is a Brownian motion (see [7, Lemma 6]). Taking expectations at $t = 1$ and using (3.2) gives

$$J^* + E_u \int_0^1 (\Lambda W + \nabla W \cdot f^u)\, dt = 0. \tag{3.8}$$

In view of (H),

$$\nabla W \cdot f^u + h^u \geqq -\alpha \quad \text{a.e.} \tag{3.9}$$

and hence

$$J^* \leqq E_u \int_0^1 h^u\, dt - E_u \left[\int_0^1 (\Lambda W - \alpha)\, dt \right]. \tag{3.10}$$

Equality holds a.e. in (3.9) for $u = u^*$ so that

$$J^* = E_{u^*}\int_0^1 h^{u^*}\,dt - E_{u^*}\left[\int_0^1 (\Lambda W - \alpha)\,dt\right]. \tag{3.11}$$

Now (3.10)–(3.11) say that u^* is optimal, as long as (3.7) holds.

Let $X = \int_0^1 (\Lambda W - \alpha)\,dt$; then X is an a.s. positive random variable, from (3.4). Define for each integer N,

$$X^N(z) = \min(N, X(z)).$$

In view of (3.2), given any $\varepsilon > 0$ there is a $u \in \mathscr{U}$ such that

$$E_u\int_0^1 h^u\,dt < J^* + \varepsilon.$$

Then, from (3.10), $E_uX < \varepsilon$. Thus there exists a sequence $\{u_n\} \subset \mathscr{U}$ such that

$$E_{u_n}X \to 0.$$

Since $0 \leqq X^N \leqq X$, $E_{u_n}X^N \to 0$, $n \to \infty$. Let $\phi_n = \exp[\zeta(f^{u_n})]$. Then $\phi_n \in \mathscr{D}(\Phi)$ for each n and $E_0\phi_nX^N \to 0$. From Lemma 3 there is a subsequence, also denoted by ϕ_n, and an element $\phi \in \Phi$ such that $\phi_n \to \phi$ in $\sigma(L_1, L_\infty)$. Thus

$$\lim_{n\to\infty} E_0\phi_nX^N = 0 = E_0\phi X^N.$$

Now $\phi > 0$ a.s. in view of (ii) of Lemma 1, so that $X^N = 0$ a.s. Since this applies for every N, $X = 0$ a.s., so that (3.7) holds and the proof is complete.

4. Additional remarks.

(a) In Markovian problems, as studied by W. H. Fleming [6], R. W. Rishel [11] and others, the system is described by

$$dz_t = f(t, z_t, u_t)\,dt + \sigma(t, z_t)\,dB_t, \qquad z(0) = x_0 \in R^n,$$

where $u_t = u(t, z_t)$ is a function of the current state only, and the process $\{z_t\}$ is stopped at the first exit time from a cylinder $Q = [0, 1] \times B$, $B \subset R^n$. The value function $W(t, x)$ is obtained formally as the solution of the semilinear parabolic equation

$$\begin{aligned}&\frac{\partial W}{\partial t} + \frac{1}{2}\sum_{i,j} a^{ij}(t,x)\frac{\partial^2 W}{\partial x_i\,\partial x_j} + \min_{u\in U}\{\nabla_x W\cdot f(t,x,u) + h(t,x,u)\} = 0 \quad \text{in } Q,\\ &W(t,x) = 0, \qquad (t,x) \in \partial Q - \{0\} \times B.\end{aligned} \tag{4.1}$$

Results of the following type are obtained: If (i) there is a function $u^0(t, x, p)$ such that $[p\cdot f + h]$ is minimum on U when $u = u^0(t, x, p)$, and

(ii) (4.1) has a suitably smooth solution $W(t, x)$, then $u^0(t, x, \nabla_x W)$ is an optimal control, with cost $W(0, x_0)$. Attention is then devoted to finding conditions under which (i) and (ii) are true. This gives a constructive procedure for finding u^0: solve (4.1) for $W(t, x)$; then $u_t^0 = u^0(t, x, \nabla_x W)$. With the present approach, however, it is known in advance that a function ∇W (or, more generally, a process ∇W_t in the non-Markov case) exists which plays the role of $\nabla_x W$ above; thus nothing corresponding to (ii) is required. This generality is achieved at the expense of the constructive procedure: there is no algorithm for computing the function ∇W. However, from (3.7) one always has

$$-\Lambda W(t, z) = \mathscr{H}(t, z, \nabla W_t). \tag{4.2}$$

In Markovian problems it is known [4, § 6] that the value function as defined by (3.1) just depends on the current state, i.e., there is a function $\overline{W}:[0, 1] \times R^n \to R$ such that $W(t, z) = \overline{W}(t, z_t)$. If $\overline{W}$ is smooth enough, then Λ and ∇ are differential operators:

$$\begin{aligned} \Lambda\overline{W} &= \frac{\partial \overline{W}}{\partial t} + \frac{1}{2}\sum a^{ij}\frac{\partial^2 \overline{W}}{\partial x_i \partial x_j}, \\ \nabla\overline{W} &= \frac{\partial \overline{W}}{\partial x}. \end{aligned} \tag{4.3}$$

Combining (4.2) and (4.3), one recovers the Hamilton–Jacobi equation (4.1).

(b) The system considered in [11] is the following:

$$dx_t = f(t, x_t, y_t)\,dt, \tag{4.4a}$$

$$dy_t = g(t, x_t, y_t, u_t)\,dt + \sigma(t, y_t)\,dB_t. \tag{4.4b}$$

Here f is Lipschitz in (x_t, y_t), so that (4.4a) has a unique nonanticipative solution $\{x_t, t \in [0, 1]\}$ for any trajectory $y \in C$; denote this solution symbolically by

$$x_t = \chi_t(y). \tag{4.5}$$

Then (4.4) is equivalent to

$$dy_t = g(t, \chi_t(y), y_t, u_t)\,dt + \sigma(t, y_t)\,dB_t,$$

which is of the form (1.1). So there is an optimal policy $u^0 \in \mathscr{U}$ for this problem. Notice from (4.5) that $\sigma(x_s, s \leqq t) \subset \sigma(y_s, s \leqq t)$, so that $\mathscr{U}$ consists of nonanticipative functions of y. It is intuitively clear that u_t^0 must be just a function of (x_t, y_t), i.e., there is a function $\tilde{u}^0$ such that

$$u^0(t, y) = \tilde{u}^0(t, \chi_t(y), y_t).$$

This is indeed the case, but more work—along the lines of [4, Lemma 6.2]—is required to prove it; see [3].

(c) If the function f of (1.1) does not depend on the control, so that the controller does not affect the system trajectory but only (through h) the payoff structure, then the present result is a special case of Beneš' "small investor" problem [1]. That problem has the property that an optimal policy exists for a wide class of information patterns. It would be very desirable to extend the present result also to the case of partial observations, but this does not appear to be possible, for the following reason. Suppose $l < n$ and the observation σ-fields are $\mathscr{Y}_t = \sigma\{z_s^1, z_s^2, \cdots, z_s^l, s \leqq t\}$, i.e., the first l components of $\{z_t\}$ are observed.

The "remaining cost" function $\psi_u(t)$ is now

$$\psi_u(t) = E_u\left[\int_t^1 h(s, z, u_s)\,ds \middle| \mathscr{Y}_t\right]$$

and this depends on the control policy used over the entire interval [0, 1]. Processes $W_u(t)$, $\nabla W_u(t)$, $\Lambda W_u(t)$ are obtained (in some cases) as before, but these are now indexed by u, the policy used up to time t, reflecting the fact that past controls affect the expectation of future performance. Inequality (3.4) becomes

$$\Lambda W_u(t) + E_u[\nabla W_u(t) \cdot f(t, z, u_t) + h(t, z, u_t)|\mathscr{Y}_t] \geqq 0 \quad \text{a.e.} \tag{4.6}$$

with equality a.e. if u is optimal; however it is still far from clear how u is to be chosen to minimize (4.6).

Acknowledgment. I am indebted to E. Wong and to J. M. C. Clark for helpful discussions.

REFERENCES

[1] V. E. Beneš, *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*, this Journal, 9 (1971), pp. 446–475.

[3] M. H. A. Davis, *Optimal control of a degenerate Markovian system*, Res. Rep. 72/29, Department of Computing and Control, Imperial College, London. To appear in Proc. I.M.A. Conference on Recent Mathematical Developments in Control, University of Bath, 1972.

[4] M. H. A. Davis and P. P. Varaiya, *Dynamic programming conditions for partially observable stochastic systems*, this Journal, 11 (1973), pp. 226–261.

[5] T. E. Duncan and P. P. Varaiya, *On the solutions of a stochastic control system*, this Journal, 9 (1971), pp. 354–371.

[6] W. H. Fleming, *Optimal continuous-parameter stochastic control*, SIAM Rev., 11 (1969), pp. 470–509.

[7] 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.

[8] P. A. Meyer, *Probability and Potentials*, Blaisdell, Waltham, Mass., 1966.

[9] ———, *Non-square integrable martingales*, Martingales, H. Dinges, ed., Lecture Notes in Mathematics, vol. 190, Springer-Verlag, Berlin, 1971.

[10] R. W. Rishel, *Necessary and sufficient dynamic programming conditions for continuous-time optimal control*, this Journal, 8 (1970), pp. 559–571.

[11] ———, *Weak solutions of a partial differential equation of dynamic programming*, this Journal, 9 (1971), pp. 519–528.

SIAM J. CONTROL
Vol. 11, No. 4, November 1973

AN EXISTENCE THEOREM FOR OPTIMIZATION PROBLEMS INVOLVING INTEGRAL EQUATIONS*

DAVID E. COWLES†

Abstract. An existence theorem is proven for Lagrange problems of optimal control in which the state equation is an abstract integral equation. We formulate the existence theorem using properties of variable sets defined by Cesari. The proof involves the application of a lower closure theorem in conjunction with a growth condition. One- and multidimensional examples are given to illustrate possible applications of the existence theorem.

1. Introduction.

We consider one- and multidimensional problems of optimization of the Lagrange type in which the state variable satisfies an integral equation. Our state variable, x, will be a member of a topological space, S. Let G be a bounded, measurable subset of the t-space E^ν, $\nu \geqq 1$. Let $\mathscr{L}$ and $\mathscr{M}$ be two operators on S with values in $(L_1(G))^{r'}$ and $(L_1(G))^s$ respectively. For every $t = (t^1, \cdots, t^\nu)$ in the closure of G, let $A(t)$ be a nonempty closed subset of the y-space E^s, $y = (y^1, \cdots, y^s)$. Let A be the set of all (t, y) such that $t \in \mathrm{cl}(G)$ and $y \in A(t)$. For every (t, y) in A, let $U(t, y)$ be a nonempty subset of the u-space E^m, $u = (u^1, \cdots, u^m)$. Let M denote the set, $M = \{(t, y, u) | (t, y) \in A \text{ and } u \in U(t, y)\}$.

We investigate the problem of finding an element x of S and a measurable vector-valued control $u(t)$, $t \in G$, so as to minimize the cost functional

$$I(x, u) = \int_G f_0(t, (\mathscr{M}x)(t), u(t))\, dt \tag{1.1}$$

subject to the state equation

$$(\mathscr{L}x)(t) = \phi\left(t, (\mathscr{M}x)(t), \int_G K(t, \tau, (\mathscr{M}x)(\tau)) f(\tau, (\mathscr{M}x)(\tau), u(\tau))\, d\tau\right) \tag{1.2}$$

for almost every $t \in G$, and the constraints

$$(\mathscr{M}x)(t) \in A(t), \qquad u(t) \in U(t, (\mathscr{M}x)(t)) \qquad \text{a.e. } t \in G. \tag{1.3}$$

Here $K(t, \tau, y)$ is a $q \times r$ matrix defined on $G \times A$ and f is a vector-valued function defined on M with values in E^r. Also, $\phi(t, y, d)$ is a vector-valued function defined on $A \times E^q$ with values in $E^{r'}$.

2. Preliminaries.

In stating our theorems, we will use a property of set-valued functions called property (Q). In defining property (Q), we shall use the notations of paragraph one. Also, given a point $(t_0, y_0) \in A$ and a number $\delta > 0$, $N_\delta(t_0, y_0)$ will denote the set of all points $(t, y) \in A$ at a distance less than or equal to δ from (t_0, y_0). We have the following definition of property (Q).

* Received by the editors May 31, 1972.

[†] Department of Mathematics, Eastern Michigan University, Ypsilanti, Michigan 48197. This paper was written in the frame of AFOSR Research Project 71-2122 at the University of Michigan.

DEFINITION 1. For every $(t, y) \in A$, let $Q(t, y)$ be a subset of the z-space E^{r+1}, $z = (z^0, \cdots, z^r)$. We say that the sets $Q(t, y)$ have *Cesari's upper semicontinuity or property* (Q) *at* (t_0, y_0) *in* A *provided that*

$$Q(t_0, y_0) = \bigcap_{\varepsilon > 0} \operatorname{cl} \operatorname{co} Q(t_0, y_0, \varepsilon),$$

where

$$Q(t_0, y_0, \varepsilon) = \bigcup_{(t,y) \in N_\varepsilon(t_0, y_0)} Q(t, y).$$

In our proof of the existence theorem we will use the following.

LEMMA 2.1. *Let G be a subset of the t-space E^ν which has finite measure. For a fixed real number p, $1 \leqq p < \infty$, let f, $\bar{f}$, and f_n be functions in $L_p(G)$, $n = 1, 2, \cdots$. Suppose that $\lim_{n \to \infty} f_n = f$ weakly in $L_p(G)$ and $\lim_{n \to \infty} f_n = \bar{f}$ pointwise a.e. on G. Then $f(t) = \bar{f}(t)$ for a.e. $t \in G$. If $p > 1$, there is a subsequence, say still $n = 1, 2, \cdots$, such that $\lim_{n \to \infty} f_n = f$ strongly in $L_1(G)$. If $p = 1$ and the functions f_n are equiabsolutely integrable, then there is a subsequence, say still $n = 1, 2, \cdots$, such that $\lim_{n \to \infty} f_n = f$ strongly in $L_1(G)$.*

Proof. Given $\varepsilon > 0$, we may apply Egoroff's theorem to find a subsequence, say still $n = 1, 2, \cdots$, and a set K, $K \subseteq G$, such that $|G| - \varepsilon < |K|$ and $f_n \to \bar{f}$ uniformly on K, and hence in $L_p(K)$. Since we also have $\lim_{n \to \infty} f_n = f$ weakly in $L_p(K)$, $f = \bar{f}$ a.e. on K. From the fact that $\varepsilon > 0$ is arbitrary, we infer that $f = \bar{f}$ a.e. on G.

If $p > 1$,

$$\begin{aligned} \int_{G-K} |f_n - f| \, dt &\leqq \left(\int_{G-K} |f_n - f|^p \, dt \right)^{1/p} \left(\int_{G-K} 1^q \, dt \right)^{1/q} \\ &\leqq \left(\int_{G-K} |f_n - f|^p \, dt \right)^{1/p} \varepsilon^{1/q}. \end{aligned}$$

Since $\lim_{n \to \infty} f_n = f$ weakly in $L_p(G)$, the sequences $\|f_n\|_{L_p(G)}$ and $\|f_n - f\|_{L_p(G)}$, $n = 1, 2, \cdots$, are bounded. Because of these facts and the fact that $\varepsilon > 0$ is arbitrary, f_n approaches f strongly in $L_1(G)$. If $p = 1$, the functions $|f_n - f|$ are assumed to be equiabsolutely integrable. Hence, given $\varepsilon' > 0$, we may assume that we have chosen ε above small enough that

$$\int_{G-K} |f_n - f| \, dt \leqq \varepsilon'.$$

Since $\varepsilon' > 0$ is arbitrary and $\lim_{n \to \infty} f_n = f$ strongly in $L_1(K)$, $\lim_{n \to \infty} f_n = f$ strongly in $L_1(G)$.

3. A lower closure theorem. In this section we present a lower closure theorem on which our existence theorem is based. The concept of lower closure, introduced by Cesari [3] in connection with his existence theorems for optimal solutions, has the same role for Lagrange problems that Tonelli's lower semicontinuity has for free problems.

We will use the notations of paragraph one for the sets G, A, M. In addition, let $\mathbf{f}(t, y, u) = (f_0, f_1, \cdots, f_r)$ be a continuous $(r + 1)$-vector function on M, and for every $(t, y) \in A$, let $\mathbf{Q}(t, y)$ denote the set

$$\mathbf{Q}(t, y) = \{\mathbf{z} = (z^0, z) = (z^0, z^1, \cdots, z^r) | z^0 \geqq f_0(t, y, u),$$
$$z = f(t, y, u), u \in U(t, y)\}.$$

We consider the functional

$$I(y, u) = \int_G f_0(t, y(t), u(t))\, dt.$$

In the lower closure theorem below we shall deal with sequences of functions all defined on G:

$$\begin{aligned} z(t) &= (z^1, \cdots, z^r), & z_k(t) &= (z_k^1, \cdots, z_k^r), \\ y(t) &= (y^1, \cdots, y^s), & y_k(t) &= (y_k^1, \cdots, y_k^s), \\ u_k(t) &= (u_k^1, \cdots, u_k^m), & t &\in G, \quad k = 1, 2, \cdots. \end{aligned}$$

THEOREM 3.1 (A lower closure theorem). *Let G be bounded and measurable, A, M be closed, and assume that the sets $\mathbf{Q}(t, y)$ have property* (Q) *on A. Let us assume that there is a function $\psi(t) \geqq 0$, $t \in G$, $\psi \in L_1(G)$, such that $f_0(t, y, u) \geqq -\psi(t)$ for all (t, y, u) in M. Let us assume that the functions $z^i(t), z_k^i(t), y^j(t), y_k^j(t), i = 1, \cdots, r$, $j = 1, \cdots, s$, are in $L_1(G)$, that the functions $u_k(t)$ are measurable on G, that $f_0(t, y_k(t), u_k(t))$ are in $L_1(G)$, and that*

$$y_k(t) \in A(t), \quad u_k(t) \in U(t, y_k(t)),$$
$$z_k^i(t) = f_i(t, y_k(t), u_k(t)) \quad \text{a.e. on } G, \qquad k = 1, 2, \cdots.$$

Finally, let us assume that as $k \to \infty$ we have

(3.1) $$z_k^i(t) \to z^i(t) \quad \text{weakly in } L_1(G), \qquad i = 1, \cdots, r,$$

(3.2) $$y_k^j(t) \to y^j(t) \quad \text{pointwise a.e. in } G, \qquad j = 1, \cdots, s,$$

and

$$\liminf_{k \to \infty} I(y_k, u_k) \leqq a_0 < +\infty.$$

Then $y(t) \in A(t)$ a.e. on G and there is a measurable function $u(t) = (u^1, \cdots, u^m)$, $t \in G$, such that $f_0(t, y(t), u(t))$ is in $L_1(G)$, $u(t) \in U(t, y(t))$, and

$$z^i(t) = f_i(t, y(t), u(t)), \quad \text{a.e. on } G, \qquad i = 1, \cdots, r,$$

with $I(y, u) \leqq a_0$.

Lower closure theorem 3.1 is a simplified statement of lower closure theorem 3.1 of [7] for the case in which $\rho = r$. The reader will note that the hypothesis (2.2) in [7]:

$$y_k^i(t) \to y^i(t) \quad \text{strongly in } L_1(G), \qquad i = 1, 2, \cdots, s,$$

has been changed to the present hypothesis (3.2) of pointwise convergence. This change is possible since in the proof of Theorem 3.1 of [7], hypothesis (2.2) was only used to obtain the pointwise convergence (3.2) stated above.

4. An existence theorem. In this section we state an existence theorem for optimal control of processes described by integral equations. For this discussion we will let p, $1 \leqq p < \infty$, be a fixed real number and let p' be the conjugate real number which satisfies $p^{-1} + (p')^{-1} = 1$, with $p' = \infty$ if $p = 1$.

We will say that the pair (x, u) is an admissible pair for the problem described in paragraph one provided that $x \in S$, $u(t)$ is measurable on G, $f_0(t, (\mathscr{M}x)(t), u(t)) \in L_1(G)$, $f(t, (\mathscr{M}x)(t), u(t)) \in (L_p(G))^r$, the components of the matrix $K(t, \tau, (\mathscr{M}x)(\tau))$ are in $L_{p'}(G)$ for almost every t in G, and the equations (1.2) and (1.3) hold. A class Ω of admissible pairs (x, u) is said to be closed if the following holds:

If $(x_k, u_k) \in \Omega$, $k = 1, 2, \cdots$, if x_k approaches x_0 in S with $\lim_{k \to \infty} \inf I(x_k, u_k) \leqq a_0 < \infty$, and if there are admissible pairs (x_0, u) with $I(x_0, u) \leqq a_0$, then there is an admissible pair $(x_0, u_0) \in \Omega$ with $I(x_0, u_0) \leqq a_0$. For a class Ω of admissible pairs, we denote by $\{x\}_\Omega$ the subset of S defined by

$$\{x\}_\Omega = \{x \in S | (x, u) \in \Omega \text{ for some measurable control } u\}.$$

In the present situation we shall require a suitable growth condition called condition (H):

(H) If $p = 1$, we assume that, given $\varepsilon > 0$, there are functions $\phi_\varepsilon \geqq 0$, $\phi_\varepsilon \in L_1(G)$, such that

$$|f(t, y, u)| \leqq \phi_\varepsilon(t) + \varepsilon f_0(t, y, u) \quad \text{for all } (t, y, u) \in M.$$

If $p > 1$, we assume that there is a function $\phi_0(t) \geqq 0$, $\phi_0 \in L_1(G)$, and a constant $a > 0$ such that

$$|f(t, y, u)|^p \leqq \phi_0(t) + a f_0(t, y, u) \quad \text{for all } (t, y, u) \in M.$$

For $p = 1$, this condition has been systematically used by Cesari [4], [5] as a suitable extension of previous more restrictive growth hypotheses used by Tonelli and McShane. We note that if f is continuous on M, if M is compact, and if there is a function $\psi(t) \geqq 0$, $t \in G$, with $\psi \in L_1(G)$ and $f_0(t, y, u) \geqq -\psi(t)$ for every $(t, y, u) \in M$, then growth condition (H) is satisfied for any p.

THEOREM 4.1 (An existence theorem). *Let G be a bounded, measurable subset of E^ν. Suppose that A and M are closed, that* $\mathbf{f}(t, y, u) = (f_0, f_1, \cdots, f_r)(t, y, u)$ *is continuous on M, and that the sets*

$$\mathbf{Q}(t, y) = \{(z^0, \cdots, z^r) | z^0 \geqq f_0(t, y, u),$$
$$(z^1, \cdots, z^r) = f(t, y, u),\ u \in U(t, y)\}$$

have property (Q) *on A. Let us assume that there is a function $\psi(t) \geqq 0$, $t \in G$, $\psi \in L_1(G)$, such that $f_0(t, y, u) \geqq -\psi(t)$ for all (t, y, u) in M. Also assume that for every sequence (x_k, u_k), $k = 1, 2, \cdots$, of admissible pairs for which $x_k \to x$ in S, there is some subsequence, say still (x_k, u_k), $k = 1, 2, \cdots$, such that*

(4.1) $(\mathscr{M}x_k)(t) \to (\mathscr{M}x)(t)$ *pointwise a.e. in G,*

(4.2) $(\mathscr{L}x_k)(t) \to (\mathscr{L}x)(t)$ *weakly in* $(L_1(G))^{r'}$, *and*

(4.3) *$K(t, \tau, (\mathscr{M}x_k)(\tau)) \to K(t, \tau, (\mathscr{M}x)(\tau))$ componentwise strongly in $L_{p'}(G)$ for almost every t in G, as k approaches infinity. Let $\phi(t, y, d)$ be a continuous r'-vector-valued function defined on $A \times E^q$. Finally, assume that growth condition* (H) *holds.*

Let Ω be a nonempty closed class of admissible pairs (x, u) such that the set $\{x\}_\Omega$ is sequentially relatively compact in the topology on S. Then the cost functional (1.1) *has an absolute minimum in Ω.*

Proof. Let i denote the infimum of $I(x, u)$ in the class Ω. Since $f_0 \geqq -\psi$ and Ω is nonempty, i is finite. Let (x_k, u_k), $k = 1, 2, \cdots$, be a sequence for which $I(x_k, u_k) \to i$ as $k \to \infty$. Since we have assumed that the set $\{x\}_\Omega$ is sequentially relatively compact, there is a subsequence, say still $k = 1, 2, \cdots$, and an element $x \in S$ such that $x_k \to x$ in S as $k \to \infty$. We may even assume that the subsequence has been chosen so that the limit relations (4.1), (4.2) and (4.3) hold. Let $z_k(t) = f(t, (\mathscr{M}x_k)(t), u_k(t)), t \in G$, $k = 1, 2, \cdots$. By growth condition (H) and the boundedness of $I(x_k, u_k)$ for all k, we see that if $p > 1$, the functions $z_k(t)$, $k = 1, 2, \cdots$, are equibounded in the norm of $(L_p(G))^r$. If $p = 1$, it follows from an argument of Cesari [4] that the functions $z_k(t)$ are equiabsolutely integrable in G. In any case, there exists a subsequence, say still $k = 1, 2, \cdots$, and a function $z(t)$, $t \in G$, $z \in (L_p(G))^r$, such that $\lim_{k \to \infty} z_k = z$ weakly in $(L_p(G))^r$. In other words,

$$f(t, (\mathscr{M}x_k)(t), u_k(t)) \to z(t) \quad \text{weakly in } (L_p(G))^r \tag{4.4}$$

as $k \to \infty$, while

$$\lim_{k \to \infty} I(x_k, u_k) = i \tag{4.5}$$

and

$$(\mathscr{M}x_k)(t) \in A(t), \qquad u_k(t) \in U(t, (\mathscr{M}x_k)(t)) \quad \text{a.e. in } G \tag{4.6}$$

for all $k = 1, 2, \cdots$. The limit relations (4.3) and (4.4) imply that

$$\begin{aligned} &\lim_{k \to \infty} K(t, \tau, (\mathscr{M}x_k)(\tau)) \cdot f(\tau, (\mathscr{M}x_k)(\tau), u_k(\tau)) \\ &\quad = K(t, \tau, (\mathscr{M}x)(\tau)) \cdot z(\tau) \quad \text{weakly in } L_1(G) \quad \text{for a.e. } t \text{ in } G. \end{aligned}$$

Therefore,

$$\begin{aligned} &\lim_{k \to \infty} \int_G K(t, \tau, (\mathscr{M}x_k)(\tau)) \cdot f(\tau, (\mathscr{M}x_k)(\tau), u_k(\tau))\, d\tau \\ &\quad = \int_G K(t, \tau, (\mathscr{M}x)(\tau)) \cdot z(\tau)\, d\tau \end{aligned} \tag{4.7}$$

for almost every t in G. The convergences (4.1), (4.7), and the continuity of ϕ imply that

$$\begin{aligned} &\lim_{k \to \infty} \phi(t, (\mathscr{M}x_k)(t), \int_G K(t, \tau, (\mathscr{M}x_k)(\tau)) \cdot f(\tau, (\mathscr{M}x_k)(\tau), u_k(\tau))\, d\tau) \\ &\quad = \phi(t, (\mathscr{M}x)(t), \int_G K(t, \tau, (\mathscr{M}x)(\tau)) \cdot z(\tau)\, d\tau) \end{aligned} \tag{4.8}$$

pointwise for almost every $t \in G$. Lemma 2.1 and the convergences (4.2) and (4.8) imply that

$$(4.9) \qquad (\mathscr{L}x)(t) = \phi(t, (\mathscr{M}x)(t), \int_G K(t, \tau, (\mathscr{M}x)(\tau)) \cdot z(\tau)\, d\tau)$$

for almost every $t \in G$.

Relations (4.1), (4.4) and (4.5) show that we may apply lower closure theorem 3.1. To apply Theorem 3.1 we have assumed that the sets $\mathbf{Q}(t, y)$ have property (Q), that M is closed, that $\mathbf{f}$ is continuous, and that $f_0(t, y, u) \geqq -\psi(t)$, $\psi(t) \in L_1(G)$, for all $(t, y, u) \in M$. From Theorem 3.1 we conclude that $(\mathscr{M}x)(t) \in A(t)$ a.e. on G and that there is a measurable control $u(t), t \in G$, such that

$$(4.10) \qquad u(t) \in U(t, (\mathscr{M}x)(t)), \qquad z(t) = f(t, (\mathscr{M}x)(t), u(t))$$

a.e. on G with $f_0(t, (\mathscr{M}x)(t), u(t)) \in L_1(G)$ and $I(x, u) \leqq i$. Relations (4.9) and (4.10) imply that

$$(\mathscr{L}x)(t) = \phi(t, (\mathscr{M}x)(t), \int_G K(t, \tau, (\mathscr{M}x)(\tau)) f(\tau, (\mathscr{M}x)(\tau), u(\tau))\, d\tau).$$

Thus, the pair (x, u) is admissible, and since Ω is closed, $(x, u) \in \Omega$ and $I(x, u) \geqq i$. Therefore, $I(x, u) = i$ and existence theorem 4.1 is proven.

Remark 1. A sequence $(x_k, u_k) \in \Omega$, $k = 1, 2, \cdots$, chosen so that $I(x_k, u_k)$ approaches the infimum of $I(x, u)$ in Ω, is called a minimizing sequence in Ω. Actually, since we apply convergence properties (4.1)–(4.3) to a minimizing sequence, it is enough to verify these convergence properties for a minimizing sequence in Ω. Since $I(x_k, u_k)$ is bounded for a minimizing sequence, we may, alternatively, verify (4.1)–(4.3) for any sequence of admissible pairs $(x_k, u_k) \in \Omega$ with $I(x_k, u_k)$ bounded, $k = 1, 2, \cdots$.

5. Examples and applications.

Example 1. We consider the one-dimensional case; that is, $t \in E^1$. We seek the minimum of the cost functional

$$I(x, u) = \int_0^1 f_0(t, x(t), u(t))\, dt,$$

where $x(t)$ is an n-dimensional continuous column vector which satisfies the equation

$$(5.1) \qquad x(t) = g(t) + \int_0^t A(t, \tau)x(\tau)\, d\tau + \int_0^t B(t, \tau)u(\tau)\, d\tau$$

for t in $[0, 1]$. We wish to have the m-dimensional column vector $u(t)$ take its values in a fixed, compact, convex subset U of E^m for every t in $[0, 1]$. Here $g(t)$, $t \in [0, 1]$, is assumed to be an n-dimensional continuous column vector, and $B(t, \tau)$ and $A(t, \tau)$ are continuous $n \times n$ and $n \times m$ matrices on $\{(t, \tau) | 0 \leqq \tau \leqq t \leqq 1\}$.

The above problem may be placed in the frame of the existence theorem by using the following designations of the objects S, $\mathscr{L}x$, $\mathscr{M}x$, K, f, and ϕ:

$S = (L_1[0, 1])^n$, $(\mathscr{L}x)(t) = x(t) - g(t)$, $(\mathscr{M}x)(t) = y(t) = (y^1(t), y^2(t))$ with $y^1(t) = \int_0^t A(t, \tau)x(\tau)\,d\tau$, $y^2(t) = x(t)$, $f(t, y, u) = u$, $\phi(t, y, d) = y^1 + d$, and $K(t, \tau) = B(t, \tau)\chi_t(\tau)$, where $\chi_t(\tau)$ is the characteristic function of the interval $[0, t]$. With these designations $r' = n$, $r = m$, $s = 2n$, $q = n$, and we set $p = 1$ with $p' = \infty$.

In order to apply existence theorem 4.1, we now show that the set Ω of all admissible pairs (x, u) has the property that $\{x\}_\Omega$ is sequentially relatively compact in the weak topology on S. Also, we verify convergence properties (4.1)–(4.3). For this purpose, let $(x_k, u_k), k = 1, 2, \cdots$, be a sequence of admissible pairs. Since U is compact, the functions $u_k(t)$ are equibounded and we may choose a subsequence, say still $[k]$, so that $u_k(t)$ approaches a function $u^*(t)$ weakly in $(L_1[0, 1])^m$ as k approaches infinity. We have $x_k(0) = g(0)$ and for any solution $x(t)$ of (5.1),

$$\begin{aligned}|x(t)| &\leqq |g(t)| + \left|\int_0^t A(t, \tau)x(\tau)\,d\tau\right| + \left|\int_0^t B(t, \tau)u(\tau)\,d\tau\right| \\ &\leqq |g(t)| + \int_0^t |A(t, \tau)||x(\tau)|\,d\tau + \int_0^t |B(t, \tau)||u(\tau)|\,d\tau.\end{aligned}$$

By the boundedness of g, A, B, and U, there are positive constants m_0 and m_1 such that

$$|x(t)| \leqq m_0 + m_1 \int_0^t |x(\tau)|\,d\tau.$$

Hence, by Gronwall's lemma, $|x_k(t)| \leqq L = m_0 \exp(m_1)$, and we may take for A and M the bounded sets $A = [0, 1] \times ([-L, L])^n$ and $M = A \times U$. By the boundedness of the set A, we infer that there is a further subsequence, say $[k]$ for the sake of simplicity, for which $x_k(t)$ converges weakly in $(L_1[0, 1])^n$ to a function $x^*(t)$. This assures that $\{x\}_\Omega$ is weakly sequentially relatively compact in $(L_1[0, 1])^n$ and that property (4.2) holds. From here it follows that $\int_0^t A(t, \tau)x_k(\tau)\,d\tau$ and $\int_0^t B(t, \tau)u_k(\tau)\,d\tau$ converge pointwise to $\int_0^t A(t, \tau)x^*(\tau)\,d\tau$ and $\int_0^t B(t, \tau)u^*(\tau)\,d\tau$ for every $t \in [0, 1]$. Thus by (5.1) for the pairs (x_k, u_k), we conclude that $x_k(t)$ converges pointwise in $[0, 1]$ to $x^*(t)$. This assures that property (4.1) is satisfied. Finally, $K(t, \tau, y) = B(t, \tau)\chi_t(\tau)$ and property (4.3) is trivial.

To show that $I(x, u)$ attains an absolute minimum in any nonempty closed class Ω of admissible pairs, we must verify that

(i) there is a function $\psi(t), t \in [0, 1]$, $\psi(t) \geqq 0$, $\psi(t) \in L_1[0, 1]$, for which $f_0(t, y, u) \geqq -\psi(t)$ for all (t, y, u) in M;

(ii) $f_0(t, y, u)$ is continuous on M;

(iii) the sets $\mathbf{Q}(t, y) = \{(z^0, z) | z^0 \geqq f_0(t, y, u), z = u, u \in U\}$ satisfy property (Q).

Since M is bounded, we need not verify that growth condition (H) holds. For, as we remarked earlier, if M is compact and (i) is verified, we may infer that growth condition (H) holds.

We note a particular case. Here a square matrix A will be called nonnegative, written $A \geqq 0$, if for any vector z we have $(Az, z) \geqq 0$. Let $C(x, t), x \in E^r$ and $t \in [0, 1]$, be a continuous $m \times m$ matrix for which there is a fixed $\theta > 0$ such that

$C(x, t) - \theta I \geqq 0$, where I is the identity matrix, for all (t, x) in A. Let $I(x, u)$ be the following cost functional:

$$I(x, u) = \int_0^1 [(C(x(t), t)u(t), u(t)) + (a(x(t), t), u(t)) + b(x(t), t)]\, dt,$$

where $a(x, t)$ is a continuous m-vector function and $b(x, t)$ is a continuous real-valued function on A. We take $f_0(t, y, u) = (C(y, t)u, u) + (a(y, t), u) + b(y, t)$ for $(t, y, u) \in M$. We may verify (i), (ii) and (iii) as follows.

By the assumed continuity of C, a, and b, we verify the continuity of $f_0(t, y, u)$. By the boundedness of the sets A and U and the continuity of f_0, we may verify (i) for a sufficiently large constant function $\psi(t)$. Finally, by applying a lemma of Cesari [4, p. 521] and using the fact that $C(x, t) - \theta I \geqq 0$, we may verify that the sets $\mathbf{Q}(t, y)$ satisfy property (Q). Hence, the problem has an optimal solution in any nonempty closed class of admissible pairs. By using other methods, V. R. Vinokurov obtained this result in [8].

Example 2. We consider a problem of optimization involving an evolution equation. For this purpose, let G' be a compact subset of the $\tau = (\tau^1, \cdots, \tau^\nu)$-space E^ν, let T be a positive real number, and let G be the set $G = [0, T] \times G'$. Our state variable x will depend on the time variable t and the space variable τ and will be an element of the set $W = \{x(t, \tau) | x(t, \tau) \in L_2(G) \text{ and } \partial x/\partial t \in L_2(G)\}$. We give W the norm

$$\|x(t, \tau)\|_W = \|x(t, \tau)\|_{L_2(G)} + \left\| \frac{\partial x}{\partial t} \right\| L_2(G).$$

We consider the problem of minimizing the cost functional

$$I(x, u) = \int_G \left[(x(t, \tau))^2 + \left(\frac{\partial x}{\partial t} \right)^2 + (u(t, \tau))^2 \right] dt\, d\tau,$$

where the state variable $x(t, \tau)$ satisfies the equation

$$\begin{aligned} \frac{\partial x}{\partial t} &= g(t, \tau) + \int_{[0,t] \times G'} A(t, \tau, s, \sigma) x(s, \sigma)\, ds\, d\sigma \\ &\quad + \int_{[0,t] \times G'} B(t, \tau, s, \sigma) u(s, \sigma)\, ds\, d\sigma \end{aligned} \tag{5.2}$$

for almost every (t, τ) in G, with $x(0, \tau) = g(\tau)$ for almost every τ in G'. In the integral equation (5.2), the variable s is in $[0, T]$, $\sigma = (\sigma^1, \cdots, \sigma^\nu)$ is in G', and $u(s, \sigma)$ may take any values in E^m. Also, $A(t, \tau, s, \sigma)$ is a real-valued square integrable function on $G \times G$, $B(t, \tau, s, \sigma)$ is a square integrable m-vector function on $G \times G$, $g(t, \tau)$ is a square integrable function on G, and $g(\tau)$ is continuous on G'.

The above problem may be placed in the frame of the existence theorem by making the following designations: S is the set of functions W with the weak topology,

$$(\mathscr{L}x)(t, \tau) = \frac{\partial x}{\partial t} - g(t, \tau),$$

$$(\mathscr{M}x)(t, \tau) = y(t, \tau) = (y^1, y^2, y^3)(t, \tau)$$

with

$$y_1(t,\tau) = x(t,\tau), \qquad y_2(t,\tau) = \frac{\partial x}{\partial t},$$

$$y_3(t,\tau) = \int_{[0,t]\times G'} A(t,\tau,s,\sigma)x(s,\sigma)\,ds\,d\sigma,$$

$K(t,\tau,s,\sigma) = B(t,\tau,s,\sigma)\chi_t(s)$, $f_0(t,\tau,y,u) = y_1^2 + y_2^2 + u^2$, $f(t,\tau,y,u) = u$, $A = G \times E^1$, $U(t,\tau,y) = E^m$, and $\phi(t,\tau,y,d) = y_3 + d$. Here $\chi_t(s)$ is the characteristic function of the interval $[0,t]$. We also take $p = p' = 2$.

Let Ω be any nonempty closed class of pairs (x,u) admissible for this problem. Since $f_0 = x^2 + (\partial x/\partial t)^2 + u^2$ and we are minimizing $I(x,u)$, we may assume that the sets $\{x\}_\Omega$ and $\{u\}_\Omega$ are norm bounded in W and $[L_2(G)]^m$ respectively. Hence, $\{x\}_\Omega$ is sequentially relatively compact in the weak topology on $S = W$.

Before applying Theorem 4.1 to the set Ω, we must verify convergence properties (4.1), (4.2), and (4.3). For this purpose, let (x_k, u_k) be a sequence of admissible pairs in Ω, $k = 1, 2, \cdots$. Since the functions x_k are uniformly bounded, we may, by a theorem of Aubin [1], find a function $x^*(t,\tau), (t,\tau) \in G$, in W such that x_k approaches x^* weakly in W as k approaches infinity.

Similarly, since the functions u_k are uniformly bounded in $[L_2(G)]^m$, we may find a function $u^*(t,\tau), (t,\tau) \in G$, in $[L_2(G)]^m$ and a subsequence, say still $[k]$, such that u_k approaches u^* weakly in $[L_2(G)]^m$ as k approaches infinity.

Hence, $x_k(t,\tau)$ approaches $x^*(t,\tau)$ weakly in $L_2(G)$, $(\mathscr{L}x_k)(t,\tau)$ approaches $(\mathscr{L}x^*)(t,\tau)$ weakly in $L_2(G)$ as k approaches infinity and (4.2) is verified. By the weak convergence in L_2 of $x_k(t,\tau)$ to $x^*(t,\tau)$ and $u_k(t,\tau)$ to $u^*(t,\tau)$,

$$\lim \int_{[0,t]\times G'} A(t,\tau,s,\sigma)x_k(s,\sigma)\,ds\,d\sigma = \int_{[0,t]\times G'} A(t,\tau,s,\sigma)x^*(s,\sigma)\,ds\,d\sigma$$

and

$$\lim \int_{[0,t]\times G'} B(t,\tau,s,\sigma)u_k(s,\sigma)\,ds\,d\sigma = \int_{[0,t]\times G'} B(t,\tau,s,\sigma)u^*(s,\sigma)\,ds\,d\sigma$$

pointwise for almost every (t,τ) in G as k approaches infinity. Since the functions $\partial x_k/\partial t$ satisfy (5.2), $\partial x_k/\partial t$ approaches $\partial x^*/\partial t$ pointwise on G as k approaches infinity. In addition, because

$$x_k(t,\tau) = g(\tau) + \int_0^t \frac{\partial x_k}{\partial t}\,ds$$

for almost every (t,τ) in G, $x_k(t,\tau)$ approaches $x^*(t,\tau)$ pointwise in G as k approaches infinity. From the above statements we conclude that (4.1) is verified. Finally, (4.3) is trivial.

We may now easily verify that the other conditions of Theorem 4.1 are satisfied, for instance, that growth condition (H) holds with $p = 2$, and that the sets

$$\mathbf{Q}(t,y) = \{(z^0, \cdots, z^m) | z^0 \geqq y_1^2 + y_2^2 + u^2, z = u, u \in E^m\}$$

satisfy property (Q). By applying Theorem 4.1 to Ω, we conclude that $I(x,u)$ attains its infimum in Ω.

Example 3. We consider the multidimensional case. For this purpose, let G be a compact subset of E^2 with $t = (t^1, t^2) \in G$. We seek to minimize the cost functional

$$I(x, u) = \int_G (|x(t)| - \sin(u(t)))\, dt,$$

where $x(t), t \in G$, is an L_1 integrable function on G which satisfies the integral equation

$$x(t) = g(t) + \int_G B(t, \tau)u(\tau)\, d\tau \tag{5.3}$$

and $u(t), t \in G$, is a measurable function on G which is constrained to take its values in the set $U = [0, \pi]$. Here $g(t), t \in G$, is an integrable function on G, and $B(t, \tau)$ is integrable on $G \times G$.

This problem is placed in the framework of the above analysis by making the following designations: $S = L_1(G)$ with the weak topology, $(\mathscr{L}x)(t) = x(t) - g(t)$, $(\mathscr{M}x)(t) = x(t), f_0(t, x, u) = |x| - \sin(u), f(t, x, u) = u, A = G \times E^1, M = A \times U$, and $K(t, \tau, x) = B(t, \tau)$.

We now show that if Ω is any nonempty closed class of admissible pairs (x, u), then $\{x\}_\Omega$ is weakly sequentially relatively compact in $L_1(G)$ and we may obtain convergence properties (4.1)–(4.3). Let (x_k, u_k), $k = 1, 2, \cdots$, be a sequence of admissible pairs in Ω. Since $u_k(t) \in [0, \pi]$, we may extract a subsequence, say still $[k]$, and find a function $u^*(t), t \in G$, so that $u_k(t)$ approaches $u^*(t)$ weakly in $L_\infty(G)$ as k approaches infinity. Let $x^*(t), t \in G$, denote the function

$$x^*(t) = g(t) + \int_G B(t, \tau)u^*(\tau)\, d\tau.$$

Since the functions $x_k(t)$ satisfy (5.3), for any function $\theta(t)$ in $L_\infty(G)$,

$$\begin{aligned}\lim_{k\to\infty} &\int_G (x_k(t) - x^*(t))\theta(t)\, dt \\ &= \lim_{k\to\infty} \int_G \int_G B(t, \tau)(u_k(\tau) - u^*(\tau))\theta(t)\, d\tau\, dt = 0.\end{aligned}$$

Hence, $x_k(t)$ approaches $x^*(t)$ weakly in $L_1(G)$ as $k \to \infty$, and we have obtained (4.2) and the result that $\{x\}_\Omega$ is weakly sequentially relatively compact. Since all pairs (x_k, u_k), $k = 1, 2, \cdots$, and (x^*, u^*) satisfy (5.3) with $\lim_{k\to\infty} u_k(t) = u^*(t)$ weakly in $L_\infty(G)$,

$$\lim_{k\to\infty} x_k(t) = x^*(t) \quad \text{pointwise on } G,$$

and (4.1) is verified. Finally, (4.3) is trivial.

Now that we have established (4.1)–(4.3), the other conditions of Theorem 4.1 follow directly. For example, property (H) follows with $p = 1$ since $|u| \leqq \pi + f_0(t, x, u)$ for every (t, x, u) in M. Also, it is easy to see that the set

$$\mathbf{Q}(t, x) = \{(z^0, z^1) | z^0 \geqq |x| - \sin(u), z^1 = u, u \in [0, \pi]\}$$

is convex and has property (Q). Hence, Theorem 4.1 assures that for the above problem $I(x, u)$ attains its infimum in any nonempty closed class of admissible pairs (x, u).

Example 4. In this example, G is a compact subset of E^2 with $t = (t^1, t^2) \in G$. We seek to minimize the cost functional

$$I(x, u) = \int_G (x(t))^2 + (u(t))^2 \, dt,$$

where $x(t)$ satisfies the integral equation

$$x(t) = \int_G [A(t, \tau)x(\tau) + B(t, \tau)(u(\tau) + 2^{-1}|u(\tau)|)] \, d\tau \tag{5.4}$$

and $u(t), t \in G$, may take any values in E^1. Here $A(t, \tau)$ and $B(t, \tau)$ are L_2 integrable functions on $G \times G$.

In order to apply existence theorem 4.1 we make the following designations: $S = L_2(G)$ with the weak topology, $(\mathscr{L}x)(t) = x(t)$, $(\mathscr{M}x)(t) = (y^1(t), y^2(t)) = y(t)$ where $y^1(t) = x(t)$ and $y^2(t) = \int_G A(t, \tau)x(\tau) \, d\tau$, $f(t, y, u) = u + 2^{-1}|u|$, $f_0(t, y, u) = (y^1)^2 + (u)^2$, $\phi(t, y, d) = y^2 + d$, $r = r' = s = q = 1$, and $p = p' = 2$.

Let Ω be any nonempty closed class of pairs (x, u) admissible for this problem. Since $f_0 = x^2 + u^2 \geqq 0$ and we are minimizing $I(x, u)$, we may assume that the sets $\{x\}_\Omega$ and $\{u\}_\Omega$ are norm bounded in $L_2(G)$. Hence, $\{x\}_\Omega$ is sequentially relatively compact in the weak topology on $S = L_2(G)$.

We now verify convergence properties (4.1) and (4.2) of Theorem 4.1. For this example, (4.3) is trivial. Let (x_k, u_k), $k = 1, 2, \cdots$, be a sequence of pairs in Ω. Since $\{x\}_\Omega$ and $\{u\}_\Omega$ are norm bounded in $L_2(G)$, we may find functions $x^*(t)$, $u^*(t)$, and $v^*(t)$, each in $L_2(G)$, and extract a subsequence, say still $[k]$, such that

$$\begin{aligned} &\lim x_k(t) = x^*(t), \\ &\lim u_k(t) = u^*(t), \\ &\lim [u_k(t) + 2^{-1}|u_k(t)|] = v^*(t), \end{aligned}$$

weakly in $L_2(G)$ as k approaches infinity. Hence, (4.2) is verified. We also have

$$\lim \int_G A(t, \tau)x_k(\tau) \, d\tau = \int_G A(t, \tau)x^*(\tau) \, d\tau,$$

$$\lim \int_G B(t, \tau)(u_k(\tau) + 2^{-1}|u_k(\tau)|) \, d\tau = \int_G B(t, \tau)v^*(\tau) \, d\tau,$$

pointwise for every t in G as k approaches infinity. Since the pairs (x_k, u_k), $k = 1, 2, \cdots$, satisfy (5.4), we have

$$\lim x_k(t) = x^*(t) = \int_G A(t, \tau)x^*(\tau) \, d\tau + \int_G B(t, \tau)v^*(\tau) \, d\tau$$

pointwise for every t in G as k approaches infinity. This shows that for the chosen subsequence $(\mathscr{M}x_k)(t)$ approaches $(\mathscr{M}x^*)(t)$ pointwise as k approaches infinity. Convergence property (4.1) is thus verified. We note that although u_k approaches u^* and $[u_k + 2^{-1}|u_k|]$ approaches v^* weakly in $L_2(G)$, it may well be that v^* is not equal to $u^* + 2^{-1}|u^*|$.

The reader may easily verify that the other conditions of Theorem 4.1 are satisfied, for instance, that property (H) holds with $p = 2$, $\phi_0 = 0$, $a = 4$, and that the sets

$$\mathbf{Q}(t, x) = \{(z^0, z) | z^0 \geqq x^2 + u^2, z = u + 2^{-1}|u|, u \in E^1\}$$

satisfy property (Q). Hence, Theorem 4.1 guarantees that $I(x, u)$ attains its infimum in any nonempty closed class Ω of admissible pairs (x, u).

Acknowledgment. The author is indebted to Professor Lamberto Cesari for his valuable encouragement.

REFERENCES

[1] J. P. AUBIN, *Un théorème de compacité*, C. R. Acad. Sci. Paris, 256 (1963), pp. 5042–5044.

[2] A. G. BUTKOVSKY, A. I. EGOROV AND K. A. LURIE, *Optimal control of distributed systems* (*a survey of Soviet publications*), this Journal, 6 (1968), pp. 437–476.

[3] L. CESARI, *Existence theorems for weak and usual optimal solutions in Lagrange problems with unilateral constraints, I and II*, Trans. Amer. Math. Soc., 124 (1966), pp. 369–412, 413–430.

[4] ———, *Existence theorems for optimal controls of the Mayer type*, this Journal, 6 (1968), pp. 517–552.

[5] ———, *Existence theorems for abstract multidimensional control problems*, J. Optimization Theory and Appl., 6 (1970), pp. 210–236.

[6] D. E. COWLES, *Upper semicontinuity properties of variable sets in optimal control*, Ibid., 10 (1972), pp. 222–236.

[7] ———, *Lower closure theorems for Lagrange problems of optimization with distributed and boundary controls*, Ibid., 10 (1972), pp. 300–320.

[8] V. R. VINOKUROV, *Optimal control of processes described by integral equations, I and II*, this Journal, 7 (1969), pp. 324–336, 337–345.

SIAM J. CONTROL
Vol. 11, No. 4, November 1973

GLOBAL CONTROLLABILITY AND BANG-BANG STEERING OF CERTAIN NONLINEAR SYSTEMS*

GUNNAR ARONSSON†

Abstract. Controllability is proved for certain nonlinear systems obtained by perturbing controllable linear systems. A simple covering lemma which treats the change of the reachable set caused by the nonlinear term is essential for Theorems 1–3. If instead a "bang-bang covering lemma" is used, we obtain theorems on "bang-bang controllability" (Theorems 4–7).

1. Introduction. In this paper we shall establish the controllability of certain control systems

$$\dot{x} = A(t)x + B(t)u + g(t, x, u),$$

by assuming that $\dot{x} = A(t)x + B(t)u$ is controllable and imposing various restrictions on the magnitude of $g(t, x, u)$.

All the results concern *global controllability on a fixed interval* $0 \leqq t \leqq T$. The theorems of § 4 treat "usual" controllability, whereas those of § 6 concern "bang-bang" controllability. Our method of proof is somewhat similar to one used by Markus in [6]. An estimate for the difference between a solution of the linear and a solution of the nonlinear system is combined with a topological covering argument.

A paper with similar results is Lukes [5]. Most controllability results of [5] are considerably generalized in this paper, but, unlike Lukes, we shall not consider partially controllable systems, continuous controls, etc.. Lukes' method of proof is quite different from ours.

Linear perturbations of linear control systems were studied by Dauer [2].

A survey of controllability by special controls is given in Strauss [7, Chap. 5].

2. Preliminaries on the control systems. We introduce here some notation and basic facts concerning the control systems.

2.1. The linear control system. Consider the linear control system $\dot{x} = A(t)x + B(t)u$ on the interval $0 \leqq t \leqq T$. Here, $x \in R^n$, $u \in R^m$, and the matrices $A(t)$ and $B(t)$ have elements in $L^1(0, T)$. The admissible controls are all m-vector functions with elements in $L^\infty(0, T)$. We assume that the system is *controllable on the interval* $[0, T]$, that is, it can be steered from any initial point x_0 at $t = 0$ to any final point x_T at $t = T$ by means of an admissible control. For a control function $u(t)$, we shall use the norm

$$\|u\| = \max_{1 \leqq i \leqq m} \|u_i(t)\|_{L^\infty}.$$

* Received by the editors July 13, 1972, and in revised form November 29, 1972.

† Department of Mathematics, Chalmers Institute of Technology and the University of Göteborg, Göteborg 5, Sweden.

Introduce the class of control functions $\Omega_M = \{u| \|u\| \leqq M\}$. The solution of the control system, starting at x_0, satisfies

$$x(T) = X(T,0)x_0 + \int_0^T X(T,s)B(s)u(s)\,ds,$$

where $X(t,s)$ is the transition matrix of the homogeneous system $\dot{x} = A(t)x$. Write

$$X_T = X(T,0)x_0, \qquad \varphi(u) = \int_0^T X(T,s)B(s)u(s)\,ds,$$

and

$$\Phi_0(u) = X_T + \varphi(u).$$

The image set $\varphi(\Omega_1)$ is compact and convex [4, p. 69], and $0 \in \varphi(\Omega_1)$. Now 0 must be an interior point of $\varphi(\Omega_1)$, since otherwise $\varphi(\Omega_1)$ and each $\varphi(\Omega_N) = N\varphi(\Omega_1)$ would lie in a certain half-space and this contradicts the controllability, which requires that

$$\bigcup_{N=1}^{\infty} \Phi_0(\Omega_N) = R^n.$$

Put

$$S(r) = \{x|x \in R^n, \|x\| \leqq r\}.$$

(Norms of vectors are Euclidean, unless otherwise stated.) We have

$$d = \max\{r|S(r) \subset \varphi(\Omega_1)\} > 0.$$

If $M \cdot d > \|X_T\|$, it follows that

$$\Phi_0(\Omega_M) \supset S(M \cdot d - \|X_T\|).$$

Observe that $\Phi_0(\Omega_M)$ is a set of attainability for the linear system.

We shall also consider the class of *bang-bang control functions* $\tilde{\Omega}_M$ which consists of all control functions $u(t)$ such that each $u_j(t)$ is measurable on $(0, T)$ and $|u_j(t)| = M$ a.e., for $j = 1, 2, \cdots, m$. According to the famous bang-bang principle of LaSalle, we have $\varphi(\Omega_M) = \varphi(\tilde{\Omega}_M)$. (Concerning this, see [3, pp. 25, 46], or [4, pp. 79–80].)

2.2. The nonlinear control system. Consider a system $\dot{x} = A(t)x + B(t)u + g(t,x,u)$, where the matrices $A(t)$ and $B(t)$ are the same as before. We need some assumptions concerning regularity and magnitude of $g(t,x,u)$. For the control vector u we use the norm $\|u\| = \max_{1 \leqq j \leqq m} |u_j|$. We assume that:

(i) $g(t,x,u)$ is measurable in t for each fixed $(x,u) \in R^{n+m}$, $g(t,x,u)$ is continuous in (x,u) for each fixed $t \in [0, T]$;

(ii) $\|g(t,0,0)\| \in L^1(0,T)$;

(iii) for each $M > 0$, there is $a(t) \in L^1(0,T)$ such that

$$\|g(t,x_1,u) - g(t,x_2,u)\| \leqq a(t)\|x_1 - x_2\|$$

provided that $\|u\| \leqq M$;

(iv) for each $M > 0$, there is $b(t) \in L^1(0, T)$ and a continuous nondecreasing function $\mu(s)$, satisfying $\mu(0) = 0$, such that

$$\|g(t, x, u_1) - g(t, x, u_2)\| \leqq b(t)(\|x\| + 1)\mu(\|u_1 - u_2\|)$$

provided that $\|u_1\| \leqq M$ and $\|u_2\| \leqq M$.

Naturally, the functions $a(t)$, $b(t)$ and $\mu(s)$ will in general depend on M. Put

$$f(t, x, u) \equiv A(t)x + B(t)u + g(t, x, u).$$

If $\|u\| \leqq M$, then we have

$$\begin{aligned}\|g(t, x, u)\| &\leqq \|g(t, x, u) - g(t, 0, u)\| + \|g(t, 0, u) - g(t, 0, 0)\| + \|g(t, 0, 0)\| \\ &\leqq a(t)\|x\| + b(t)\mu(M) + \|g(t, 0, 0)\| \leqq c(t)(\|x\| + 1)\end{aligned}$$

for some $c(t) \in L^1(0, T)$. Take a control function $\bar{u}(t) \in \Omega_M$, and consider the ordinary differential system $\dot{x} = f(t, x, \bar{u}(t))$. It is now clear that

$$\|f(t, x, \bar{u}(t))\| \leqq d(t)(\|x\| + 1) \tag{1}$$

for some $d(t) \in L^1(0, T)$. It follows from this and condition (i) that Carathéodory's existence theorem can be applied to the system and hence a solution exists through any given point. Clearly,

$$\|f(t, x_1, \bar{u}(t)) - f(t, x_2, \bar{u}(t))\| \leqq e(t)\|x_1 - x_2\|$$

for some $e(t) \in L^1(0, T)$, and uniqueness of the solution follows easily from this, as is well known. Furthermore, it follows from (1) by standard arguments that the solution can be continued over the whole interval $[0, T]$ and that there is a uniform bound: $\|x(t)\| \leqq L = L(x_0, M)$ for the solution to $\dot{x} = f(t, x, \bar{u}(t))$, $x(0) = x_0$, still provided that $\bar{u}(t) \in \Omega_M$.

Put $\nu(s) = s + \mu(s)$. If $\|u_1\| \leqq M$ and $\|u_2\| \leqq M$, then we have

$$\begin{aligned}\|f(t, x_1, u_1) - f(t, x_2, u_2)\| &\leqq \|A(t)(x_1 - x_2)\| + \|B(t)(u_1 - u_2)\| + \|g(t, x_1, u_1) \\ &\qquad - g(t, x_1, u_2)\| + \|g(t, x_1, u_2) - g(t, x_2, u_2)\| \\ &\leqq \|A(t)\| \cdot \|x_1 - x_2\| + \|B(t)\| \cdot \|u_1 - u_2\| \\ &\qquad + b(t) \cdot (L + 1) \cdot \mu(\|u_1 - u_2\|) + a(t)\|x_1 - x_2\| \\ &\leqq h(t) \cdot \nu(\|u_1 - u_2\|) + k(t) \cdot \|x_1 - x_2\|,\end{aligned}$$

for certain nonnegative functions $h(t)$ and $k(t)$ in $L^1(0, T)$. Take two control functions $u_1(t)$, $u_2(t)$ in Ω_M and put

$$\Theta(t) = \|u_1(t) - u_2(t)\|.$$

Consider solutions $x_1(t)$, $x_2(t)$ of $\dot{x}_i = f(t, x_i, u_i(t))$ such that $x_1(0) = x_2(0) = x_0$, and put $v(t) = \|x_1(t) - x_2(t)\|$. Then $v(t)$ is absolutely continuous, $v(0) = 0$, and

$$\dot{v}(t) \leqq h(t) \cdot \nu(\Theta(t)) + k(t) \cdot v(t) \quad \text{a.e.}$$

Integration gives

$$\|x_1(T) - x_2(T)\| \leqq \exp\left(\int_0^T k(s)\, ds\right) \cdot \int_0^T h(t)\nu(\Theta(t))\, dt.$$

We need some more notation. Consider a fixed initial point $x(0) = x_0$ and a variable admissible control $u(t)$. For each $u(t)$ we get a solution $x(t)$ of $\dot{x} = f(t, x, u)$, starting at x_0. It satisfies

$$x(T) = X(T, 0)x_0 + \int_0^T X(T, s)B(s)u(s)\,ds + \int_0^T X(T, s)g(s, x(s), u(s))\,ds.$$

Denote this quantity by $\Phi(u)$. We thus have

$$\Phi(u) = X_T + \varphi(u) + \int_0^T X(T, s)g(s, x(s), u(s))\,ds.$$

Put

$$K = \max_{0 \leqq s \leqq t \leqq T} \|X(t, s)\|,$$

where

$$\|X(t, s)\| = \sup\{\|X(t, s)x\| \,|\, x \in R^n, \|x\| \leqq 1\}.$$

Thus

$$\|\Phi(u) - X_T - \varphi(u)\| \leqq K \int_0^T \|g(s, x(s), u(s))\|\,ds,$$

or

$$\|\Phi(u) - \Phi_0(u)\| \leqq K \int_0^T \|g(s, x(s), u(s))\|\,ds.$$

It is clear that

$$K \leqq \exp\left(\int_0^T \|A(s)\|\,ds\right),$$

where

$$\|A(s)\| = \sup\{\|A(s)x\| \,|\, x \in R^n, \|x\| \leqq 1\}.$$

Note finally that $\Phi(\Omega_M)$ is a set of attainability for the nonlinear system.

3. A basic covering lemma.

LEMMA 1. *Consider the control systems* $\dot{x} = A(t)x + B(t)u + g(t, x, u)$ *and* $\dot{x} = A(t)x + B(t)u$. *Here,* $A(t), B(t)$ *and* $g(t, x, u)$ *satisfy all requirements of* § 2. *We also retain the notation of* § 2. *Assume that* $\Phi_0(\Omega_M) \supset S(r_1)$ *and that* $\|\Phi(u) - \Phi_0(u)\| \leqq \rho < r_1$ *for* $u \in \Omega_M$. *Let* $0 < r_2 < r_1 - \rho$.

Then $\Phi(\Omega_M) \supset S(r_2)$.

Proof. The method will be as follows: we shall consider a suitable sphere $S(r) \subset \Phi_0(\Omega_M)$ and for each $x \in S(r)$ we construct a certain control $u_x(\cdot) \in \Omega_M$ which steers the linear system to x, i.e., $\Phi_0(u_x) = x$. Then we consider the continuous mapping $x \to \Phi(u_x)$. The result will follow by considering the image of $S(r)$ under this mapping. The control u_x will be constructed as a convex combination of a finite number of "basis control functions" $\{u_j\}_{j=1}^N$.

Choose r such that $r_2 + \rho < r < r_1$. We divide R^n into cubes defined by

$$\frac{k_i}{2^p} \leqq x_i \leqq \frac{k_i + 1}{2^p}, \qquad i = 1, 2, \cdots, n.$$

Here, x_i is a component of $x \in R^n$, $\{k_i\}_1^n$ are arbitrary integers, and p is a natural number. Let E be the union of all such closed cubes that are contained in $S(r_1)$. Fix p so large that $E \supset S(r)$. Let $X_1, X_2, \cdots, X_N$ be the vertices of the cubes in E. We shall need functions $\mu_1(x), \mu_2(x), \cdots, \mu_N(x)$ such that:

(a) $\mu_j(x) \in C(E)$,

(b) $0 \leqq \mu_j(x) \leqq 1$; $\sum_{j=1}^N \mu_j(x) = 1$ if $x \in E$,

(c) $x = \sum_{j=1}^N \mu_j(x) X_j$ if $x \in E$.

Consider the function of $s \in R^1$:

$$\mu(s) = \begin{cases} 1 - 2^p \cdot s & \text{for } 0 \leqq s \leqq 2^{-p}, \\ 1 + 2^p \cdot s & \text{for } -2^{-p} \leqq s \leqq 0, \\ 0 & \text{for } |s| \geqq 2^{-p}. \end{cases}$$

One easily verifies the identities

$$\sum_{k=-\infty}^{\infty} \mu\left(s - \frac{k}{2^p}\right) \equiv 1 \quad \text{and} \quad \sum_{k=-\infty}^{\infty} \mu\left(s - \frac{k}{2^p}\right) \cdot \frac{k}{2^p} \equiv s.$$

Now, if

$$X_j = \left(\frac{k_1}{2^p}, \frac{k_2}{2^p}, \cdots, \frac{k_n}{2^p}\right)$$

is one of the vertices in question we put

$$\mu_j(x) = \mu\left(x_1 - \frac{k_1}{2^p}\right) \cdot \mu\left(x_2 - \frac{k_2}{2^p}\right) \cdots \mu\left(x_n - \frac{k_n}{2^p}\right).$$

The continuity of $\mu_j(x)$ is clear and the properties (b) and (c) follow easily from the identities above. Now $S(r) \subset E \subset S(r_1) \subset \Phi_0(\Omega_M)$, and $X_j \in E$, for $j = 1, 2, \cdots, N$. Hence there exist $u_j \in \Omega_M$ such that $\Phi_0(u_j) = X_j, j = 1, 2, \cdots, N$. Let x be a variable point in $S(r)$ and put

$$u_x(t) = \sum_{j=1}^N \mu_j(x) u_j(t).$$

It is clear that $u_x \in \Omega_M$ and that u_x depends continuously on x, even in the L^∞-norm. Further,

$$\begin{aligned} \Phi_0(u_x) &= X_T + \varphi(u_x) = X_T + \sum_{j=1}^N \mu_j(x)\varphi(u_j) \\ &= \sum_{j=1}^N \mu_j(x)(X_T + \varphi(u_j)) = \sum_{j=1}^N \mu_j(x)\Phi_0(u_j) = \sum_{j=1}^N \mu_j(x) X_j = x. \end{aligned}$$

Consider the mapping $S(r) \ni x \xrightarrow{T} \Phi(u_x) \in R^n$. It is seen from above and § 2.2 that it is continuous, and we further have $\|x - T(x)\| = \|\Phi_0(u_x) - \Phi(u_x)\| \leqq \rho$. From a basic topological result [4, pp. 251–252] we have then $T(S(r)) \supset S(r - \rho) \supset S(r_2)$. Hence $\Phi(\Omega_M) \supset T(S(r)) \supset S(r_2)$, which proves the lemma.

COROLLARY. *There are controls $\{u_j\}_1^N$, all in Ω_M, such that the system $\dot{x} = A(t)x + B(t)u + g(t, x, u)$ can be steered to an arbitrary point in $S(r_2)$ by means of some convex combination of $\{u_j\}_1^N$.*

4. Controllability of some nonlinear systems. We will establish controllability of certain nonlinear systems

$$(2) \qquad \dot{x} = A(t)x + B(t)u + g(t, x, u)$$

by means of our covering lemma. We know (§ 2.1) that we can choose $r_1 = r_1(M) = M \cdot d - \|X_T\|$ and we need an estimate of ρ, $\rho = \rho(M)$, such that $\overline{\lim}_{M \to \infty} (r_1(M) - \rho(M)) = \infty$. If this can be achieved for any x_0, then the system (2) is controllable on $[0, T]$. We shall perform this under varying conditions on $g(t, x, u)$. *Throughout this section, $A(t)$, $B(t)$ and $g(t, x, u)$ are assumed to satisfy the conditions of* § 2. Our first theorem is a trivial consequence of the lemma, and may serve as a starting point for a brief discussion of the restrictions on $\|g(t, x, u)\|$.

THEOREM 1. *Consider the control system*

$$(3) \qquad \dot{x} = A(t)x + B(t)u + g(t, x, u),$$

where

$$\|g(t, x, u)\| \leqq \alpha(t) \cdot \psi(\|u\|).$$

Here, α and ψ are nonnegative functions, $\alpha(t) \in L^1(0, T)$ and ψ is nondecreasing. Assume that $\dot{x} = A(t)x + B(t)u$ is controllable on $[0, T]$. Put $A = \int_0^T \alpha(t)\, dt$.

If $\overline{\lim}_{M \to \infty} (M \cdot d - K \cdot A \cdot \psi(M)) = +\infty$, then the system (3) *is controllable on* $[0, T]$. *In particular, this is true if*

$$\underline{\lim}_{M \to \infty} \frac{\psi(M)}{M} < \frac{d}{K \cdot A}.$$

Proof. We apply Lemma 1, choosing $r_1(M) = M \cdot d - \|X_T\|$ and $\rho(M) = K \cdot A \cdot \psi(M)$ (see end of § 2). Clearly,

$$\overline{\lim}_{M \to \infty} (r_1(M) - \rho(M)) = \overline{\lim}_{M \to \infty} \left[M \left(d - K \cdot A \cdot \frac{\psi(M)}{M} - \frac{\|X_T\|}{M} \right) \right] = +\infty,$$

which proves the theorem.

Remark. This theorem is a generalization of Theorem 2.1 in Lukes [5] in that it allows $\|g(t, x, u)\|$ to be unbounded in u although in the bounded case Lukes' proof requires only continuity of g and avoids conditions (iii) and (iv).

It is thus seen that the nonlinear system is controllable, if the function ψ in Theorem 1 grows more slowly than a linear function. The same conclusion holds if ψ grows like a linear function, and A is small enough. The last condition is essential, as follows from the trivial example $g(t, x, u) \equiv -B(t)u$. Consider a system (for $n = m = 1$)

$$(4) \qquad \dot{x} = u + \varepsilon \cdot h(u),$$

where $h(u)$ is continuous and $\lim_{|u| \to \infty} (h(u)/|u|) = +\infty$. It is not controllable on $[0, T]$ for any $\varepsilon > 0\,(<0)$, since the right member of (4) is bounded from below (above). Hence, we *cannot* allow $\|g(t, x, u)\|$ to grow faster than a linear function in $\|u\|$ (without imposing conditions of some other type). Further, we assumed in § 2.2 that $\|g(t, x, u)\|$ (for $\|u\| \leqq M$) grows no faster than a linear function in $\|x\|$.

It therefore seems reasonable to consider perturbed systems $\dot{x} = A(t)x + B(t)u + g(t, x, u)$ under the restriction

$$\|g(t, x, u)\| \leqq \alpha_1(t)\|x\| + \alpha_2(t)\|u\| + \beta(t).$$

As is seen from above, $\int_0^T \alpha_2(t)\, dt$ must be assumed small enough. But this is also the case for $\int_0^T \alpha_1(t)\, dt$, as follows from the example $g \equiv -A(t)x$. Since we are not aiming at the best constants, we simply assume

$$\|g(t, x, u)\| \leqq \alpha(t)(\|x\| + \|u\|) + \beta(t).$$

THEOREM 2. *Consider the control system*

$$\dot{x} = A(t)x + B(t)u + g(t, x, u), \tag{5}$$

where

$$\|g(t, x, u)\| \leqq \alpha(t)(\|x\| + \|u\|) + \beta(t).$$

Here, $\alpha(t) \geqq 0$ *and* $\beta(t) \geqq 0$ *both belong to* $L^1(0, T)$. *Assume that* $\dot{x} = A(t)x + B(t)u$ *is controllable on* $[0, T]$.

Then there exists an $A_0 > 0$, *which only depends on the matrix functions* $A(t)$ *and* $B(t)$, *such that* (5) *is controllable on* $[0, T]$, *provided that* $\int_0^T \alpha(t)\, dt \leqq A_0$.

Proof. (a) Let $\dot{x}_1 = A(t)x_1 + B(t)u(t) + g(t, x_1, u(t))$ and $\dot{x}_2 = A(t)x_2 + B(t)u(t)$, with the same control $u(t)$. Further, let $x_1(0) = x_2(0)$. We then know from § 2.2 that

$$\|x_1(T) - x_2(T)\| \leqq \exp\left(\int_0^T \|A(s)\|\, ds\right) \cdot \int_0^T \|g(t, x_1(t), u(t))\|\, dt.$$

(b) We now use the assumption

$$\|g(t, x, u)\| \leqq \alpha(t)(\|x\| + \|u\|) + \beta(t),$$

and also assume that $u(t) \in \Omega_M$. Further, put

$$C_1 = \exp\left(\int_0^T \|A(s)\|\, ds\right) \quad \text{and} \quad A = \int_0^T \alpha(t)\, dt.$$

We find that

$$\|x_1(T) - x_2(T)\| \leqq C_1 \cdot \int_0^T [\alpha(t)(\max_{0 \leqq s \leqq T} \|x_1(s)\| + M) + \beta(t)]\, dt,$$

or

$$\|x_1(T) - x_2(T)\| \leqq C_1 \cdot A \cdot (\max_s \|x_1(s)\| + M) + C_1 \int_0^T \beta(t)\, dt.$$

(c) We need an estimate for $\max_s \|x_1(s)\|$. We have

$$x_1(t) = X(t, 0)x_0 + \int_0^t X(t, s)[B(s)u(s) + g(s, x_1(s), u(s))]\, ds.$$

Put

$$\|B(s)\| = \sup\{\|B(s)u\|_{R^n} | \max_{1 \leqq j \leqq m} |u_j| \leqq 1\}.$$

Assuming that $u(t) \in \Omega_M$, we thus have

$$\|x_1(t)\| \leqq K \cdot \|x_0\| + K \cdot \int_0^t [\|B(s)u(s)\| + \alpha(s)(\|x_1(s)\| + M) + \beta(s)]\, ds,$$

or

$$\|x_1(t)\| \leqq K \cdot \|x_0\| + K \cdot \int_0^T \|B(s)\|\, ds \cdot M + K \cdot M \cdot A + K \cdot \int_0^T \beta(s)\, ds$$
$$+ K \cdot \int_0^t \alpha(s)\|x_1(s)\|\, ds.$$

Put

$$C_2 = K \cdot \|x_0\| + K \cdot \int_0^T \beta(s)\, ds \quad \text{and} \quad C_3 = K \cdot \int_0^T \|B(s)\|\, ds + K \cdot A.$$

We can then write

$$\|x_1(t)\| \leqq (C_2 + C_3 \cdot M) + K \cdot \int_0^t \alpha(s)\|x_1(s)\|\, ds.$$

Gronwall's inequality gives

$$\|x_1(t)\| \leqq (C_2 + C_3 \cdot M) \cdot \exp\left\{\int_0^t K \cdot \alpha(s)\, ds\right\}.$$

Hence

$$\max_{0 \leqq t \leqq T} \|x_1(t)\| \leqq (C_2 + C_3 M) \cdot e^{KA}.$$

(d) Combining our results from (b) and (c) we find

$$\|x_1(T) - x_2(T)\| \leqq C_1 \cdot A[(C_2 + C_3 \cdot M)e^{KA} + M] + C_1 \int_0^T \beta(t)\, dt.$$

We thus have

$$\|x_1(T) - x_2(T)\| \leqq \rho(M) = C_4 + C_5 \cdot M,$$

where

$$C_5 = C_1 A(C_3 e^{KA} + 1) = \exp\left(\int_0^T \|A(s)\|\, ds\right) \cdot A$$
$$\cdot \left(\left[K \cdot \int_0^T \|B(s)\|\, ds + KA\right] e^{KA} + 1\right).$$

This expression does not depend on x_0 or $\beta(t)$ and tends to zero with A. Put $r_1(M) = d \cdot M - \|X_T\|$. Now, there exists $A_0 > 0$, *independent of* x_0 *and* $\beta(t)$, such that $C_5 < d$ if $A \leqq A_0$. For such values of A we thus have $\lim_{M \to \infty} (r_1(M) - \rho(M)) = \infty$, and it follows from the covering lemma that the nonlinear system is controllable on $[0, T]$.

This completes the proof of Theorem 2.

Remarks. If $\dot{x} = A(t)x + B(t)u$ is controllable on $[0, T]$, then it follows from the above theorem that $\dot{x} = (A(t) + A_1(t))x + (B(t) + B_1(t))u$ is also controllable on $[0, T]$, provided that $\int_0^T (\|A_1(t)\| + \|B_1(t)\|)\,dt$ is small enough. This result is identical with the case $p = 1$, $q = \infty$, of Theorem 4 in Dauer [2].

Also, the above theorem generalizes Theorem 2.2 in Lukes [5], as far as the controllability is concerned. (Lukes assumes $g(t, x, u)$ continuous and imposes a uniform Lipschitz condition in (x, u).)

The condition that $\int_0^T \alpha(t)\,dt$ is small enough can be dropped if we replace $\|x\| + \|u\|$ by $\varphi(\|x\|) + \psi(\|u\|)$, where φ and ψ grow sufficiently slowly.

THEOREM 3. *Consider the control system*

$$\dot{x} = A(t)x + B(t)u + g(t, x, u), \tag{6}$$

where

$$\|g(t, x, u)\| \leqq \alpha(t)(\varphi(\|x\|) + \psi(\|u\|)) + \beta(t).$$

Here, $\alpha(t) \geqq 0$ *and* $\beta(t) \geqq 0$ *both belong to* $L^1(0, T)$. *Further,* φ *and* ψ *are nonnegative, continuous, nondecreasing and satisfy* $\lim_{r\to\infty}(\varphi(r)/r) = 0$ *and* $\underline{\lim}_{M\to\infty}(\psi(M)/M) = 0$. *Assume that* $\dot{x} = A(t)x + B(t)u$ *is controllable on* $[0, T]$.

Then the system (6) *is controllable on* $[0, T]$.

Proof. We have $\varphi(r) \leqq r$ for $r \geqq r_0$, and since we are free to add a multiple of $\alpha(t)$ to $\beta(t)$, it is clear that we may assume that $\varphi(r) \leqq r$ for all $r \geqq 0$. Let $\dot{x}_1 = A(t)x_1 + B(t)u(t) + g(t, x_1, u(t))$, $\dot{x}_2 = A(t)x_2 + B(t)u(t)$, $u \in \Omega_M$, and $x_1(0) = x_2(0)$. As before, we have

$$\|x_1(T) - x_2(T)\| \leqq \exp\left(\int_0^T \|A(s)\|\,ds\right) \cdot \int_0^T \|g(t, x_1(t), u(t))\|\,dt,$$

or

$$\|x_1(T) - x_2(T)\| \leqq C_1 \cdot \int_0^T \|g(t, x_1(t), u(t))\|\,dt.$$

Using the condition on $g(t, x, u)$, we find that

$$\|x_1(T) - x_2(T)\| \leqq C_1\left[A \cdot \varphi(\max_s \|x_1(s)\|) + A \cdot \psi(M) + \int_0^T \beta(t)\,dt\right].$$

In order to estimate $\max_s \|x_1(s)\|$, we write

$$x_1(t) = X(t, 0)x_0 + \int_0^t X(t, s)[B(s)u(s) + g(s, x_1(s), u(s))]\,ds.$$

Thus

$$\|x_1(t)\| \leqq K \cdot \|x_0\| + K \cdot \int_0^t \|B(s)\|\,ds \cdot M + K$$
$$\cdot \int_0^t [\alpha(s)(\varphi(\|x_1(s)\|) + \psi(M)) + \beta(s)]\,ds.$$

Write $B = \int_0^T \|B(s)\|\,ds$ and $B_1 = \int_0^T \beta(s)\,ds$. Since $\varphi(r) \leqq r$, we then have

$$\|x_1(t)\| \leqq K(\|x_0\| + B\cdot M + A\cdot\psi(M) + B_1) + K\cdot\int_0^t \alpha(s)\|x_1(s)\|\,ds.$$

Gronwall's inequality gives

$$\|x_1(t)\| \leqq K(\|x_0\| + B\cdot M + A\cdot\psi(M) + B_1)\cdot e^{KA}.$$

Thus $\max_{0\leqq t\leqq T}\|x_1(t)\| \leqq C_2 + C_3\cdot M + C_4\cdot\psi(M)$. Finally, we obtain

$$\|x_1(T) - x_2(T)\| \leqq C_1[A\varphi(C_2 + C_3M + C_4\psi(M)) + A\psi(M) + B_1] \equiv \rho(M).$$

It follows easily from our conditions on φ and ψ that $\underline{\lim}_{M\to\infty}(\rho(M)/M) = 0$. Write $r_1(M) = d\cdot M - \|X_T\|$. We thus have $\overline{\lim}_{M\to\infty}(r_1(M) - \rho(M)) = \infty$, and the controllability of the nonlinear system follows from the covering lemma.

This completes the proof.

Remark. It is seen from the proof that the condition $\underline{\lim}_{M\to\infty}(\psi(M)/M) = 0$ can be replaced by the weaker assumption $\underline{\lim}_{M\to\infty}(\psi(M)/M) < d/(C_1A)$.

5. A bang-bang covering lemma. We first agree to say that a system $\dot{x} = f(t, x, u)$ is *bang-bang controllable on* $[0, T]$ if it can be steered from any initial point x_0 at $t = 0$ to any final point x_T at $t = T$ by means of a control in $\tilde{\Omega}_M$, for some $M(=M(x_0, x_T))$. It follows from the bang-bang principle that a controllable linear system is also bang-bang controllable.

In order to establish bang-bang controllability for certain nonlinear systems, we need a "bang-bang covering lemma."

LEMMA 2. *Consider the control systems* $\dot{x} = A(t)x + B(t)u + g(t, x, u)$ *and* $\dot{x} = A(t)x + B(t)u$. *Here,* $A(t)$, $B(t)$ *and* $g(t, x, u)$ *satisfy all requirements of* § 2. *Assume that* $\Phi_0(\Omega_M) \supset S(r_1)$ *and that* $\|\Phi(u) - \Phi_0(u)\| \leqq \rho < r_1$ *for* $u \in \tilde{\Omega}_M$. *Let* $0 < r_2 < r_1 - \rho$.

Then $\Phi(\tilde{\Omega}_M) \supset S(r_2)$.

Proof. Choose r such that $r_2 + \rho < r < r_1$. As in Lemma 1, we divide R^n into cubes: $k_i/2^p \leqq x_i \leqq (k_i + 1)/2^p$, $i = 1, 2, \cdots, n$. Let E be the union of all such closed cubes that are contained in $S(r_1)$. Fix p so large that $E \supset S(r)$. Let $X_1, X_2, \cdots, X_N$ be the vertices of the cubes in E. Exactly as in Lemma 1, we construct functions $\mu_1(x), \mu_2(x), \cdots, \mu_N(x)$ such that:

(a) $\mu_j(x) \in C(E)$;

(b) $0 \leqq \mu_j(x) \leqq 1$, $\sum_{j=1}^N \mu_j(x) = 1$ if $x \in E$;

(c) $x = \sum_{j=1}^N \mu_j(x)X_j$ if $x \in E$.

Now $S(r) \subset E \subset S(r_1) \subset \Phi_0(\Omega_M) = \Phi_0(\tilde{\Omega}_M)$ (see § 2.1), and all X_j lie in E. Hence there are bang-bang controls $u_j \in \tilde{\Omega}_M$ such that $\Phi_0(u_j) = X_j$, $j = 1, 2, \cdots, N$. Let $x \in S(r)$. We want a control $u_x \in \tilde{\Omega}_M$, such that $\Phi_0(u_x) = x$, but, unlike the case in Lemma 1, we *cannot* use a convex combination of $\{u_j\}_1^N$, since a convex combination of bang-bang controls is in general not a bang-bang control. Instead, we shall use a partition lemma of Markus [6, pp. 81–82], [4, pp. 372–373] which goes back to a theorem of Blackwell [1, p. 392, Theorem 2]. Let $\mu = (\mu_1, \cdots, \mu_N)$ be a variable vector in R^N satisfying $\sum_{j=1}^N \mu_j = 1$, $\mu_j \geqq 0$ for $j = 1, 2, \cdots, N$. According to Markus' lemma there exists a continuous family of N-partitions of $[0, T]$, $\mu \to (A_1(\mu), A_2(\mu), \cdots, A_N(\mu))$, such that the function in $L^\infty(0, T)$:

$$u(t, \mu) = \begin{cases} u_1(t) & \text{for } t \in A_1(\mu), \\ u_2(t) & \text{for } t \in A_2(\mu), \\ \text{------------} & \\ u_N(t) & \text{for } t \in A_N(\mu) \end{cases}$$

satisfies

$$\int_0^T X(T, s)B(s)u(s, \mu)\,ds = \sum_{j=1}^N \mu_j \int_0^T X(T, s)B(s)u_j(s)\,ds$$

for all μ in question. The sets $A_j(\mu)$ are Lebesgue measurable. Clearly, $u(t, \mu) \in \tilde{\Omega}_M$. Now consider the bang-bang control $u(t, \mu(x))$. It satisfies

$$\int_0^T X(T, s)B(s)u(s, \mu(x))\,ds = \sum_{j=1}^N \mu_j(x) \int_0^T X(T, s)B(s)u_j(s)\,ds = \sum_{j=1}^N \mu_j(x)\varphi(u_j)$$
$$= \sum_{j=1}^N \mu_j(x)(X_j - X_T) = x - X_T.$$

Hence $\Phi_0(u(\cdot, \mu(x))) = x$ and $u(\cdot, \mu(x)) \in \tilde{\Omega}_M$. Now let $x, x_0 \in S(r)$, and consider x_0 as fixed. Let $E(x)$ be the set where $u(t, \mu(x)) \neq u(t, \mu(x_0))$. Then $\lim_{x \to x_0} mE(x) = 0$ (see [6, p. 81]). Further, $u(t, \mu(x)) \in \Omega_M$ and $u(t, \mu(x_0)) \in \Omega_M$. It follows easily from this and the inequality

$$\|x_1(T) - x_2(T)\| \leqq \exp\left(\int_0^T k(s)\,ds\right) \cdot \int_0^T h(t)v(\Theta(t))\,dt$$

of § 2.2 that the mapping $S(r) \ni x \xrightarrow{T} \Phi(u(\cdot, \mu(x)))$ is continuous. From our assumptions we have

$$\|x - T(x)\| = \|\Phi_0[u(\cdot, \mu(x))] - \Phi[u(\cdot, \mu(x))]\| \leqq \rho.$$

Now the rest of the proof follows as in Lemma 1.

6. Bang-bang controllability of some nonlinear systems. By using Lemma 2 instead of Lemma 1 in the proofs, we can change Theorems 1, 2 and 3 into analogous theorems on the global *bang-bang* controllability of the system $\dot{x} = A(t)x + B(t)u + g(t, x, u)$. We also write $\tilde{\Omega}_M$ instead of Ω_M throughout the proofs.

Since these are the only changes in the proofs, we simply list the new theorems without further discussion. This type of theorem seems to be new in the literature.

As before, $A(t)$, $B(t)$ and $g(t, x, u)$ are assumed to satisfy the conditions of § 2.

THEOREM 4. *Consider the control system*

$$\dot{x} = A(t)x + B(t)u + g(t, x, u), \tag{7}$$

where $\|g(t, x, u)\| \leqq \alpha(t) \cdot \psi(\|u\|)$. Here, α are ψ are nonnegative functions, $\alpha(t) \in L^1(0, T)$ and $\psi(M)$ is nondecreasing. Assume that $\dot{x} = A(t)x + B(t)u$ is controllable on $[0, T]$. *Put $A = \int_0^T \alpha(t)\,dt$.*

If $\overline{\lim}_{M \to \infty} (M \cdot d - K \cdot A \cdot \psi(M)) = +\infty$, *then the system* (7) *is bang-bang controllable on* $[0, T]$.

In particular, this is true if $\underline{\lim}_{M\to\infty}(\psi(M)/M) < d/(K\cdot A)$.

THEOREM 5. *Consider the control system*

$$\dot{x} = A(t)x + B(t)u + g(t,x,u), \tag{8}$$

where

$$\|g(t,x,u)\| \leqq \alpha(t)(\|x\| + \|u\|) + \beta(t).$$

Here, $\alpha(t) \geqq 0$ *and* $\beta(t) \geqq 0$ *both belong to* $L^1(0,T)$. *Assume that* $\dot{x} = A(t)x + B(t)u$ *is controllable on* $[0,T]$.

Then there exists an $A_0 > 0$, *which only depends on the matrix functions* $A(t)$ *and* $B(t)$, *such that* (8) *is bang-bang controllable on* $[0,T]$, *provided that* $\int_0^T \alpha(t)\,dt \leqq A_0$.

THEOREM 6. *Consider the control system*

$$\dot{x} = A(t)x + B(t)u + g(t,x,u), \tag{9}$$

where

$$\|g(t,x,u)\| \leqq \alpha(t)(\varphi(\|x\|) + \psi(\|u\|)) + \beta(t).$$

Here, $\alpha(t) \geqq 0$ *and* $\beta(t) \geqq 0$ *both belong to* $L^1(0,T)$. *Further,* φ *and* ψ *are nonnegative, continuous, nondecreasing and satisfy* $\lim_{r\to\infty}(\varphi(r)/r) = 0$ *and* $\underline{\lim}_{M\to\infty}(\psi(M)/M) = 0$. *Assume that* $\dot{x} = A(t)x + B(t)u$ *is controllable on* $[0,T]$.

Then the system (9) *is bang-bang controllable on* $[0,T]$.

An important detail in Lemma 2 is that the estimate $\|\Phi(u) - \Phi_0(u)\| \leqq \rho$ is only assumed to hold for $u \in \tilde{\Omega}_M$ and not necessarily for all $u \in \Omega_M$. This will be used in the next theorem.

THEOREM 7. *Consider the control system*

$$\dot{x} = A(t)x + B(t)u + g(t,x,u), \tag{10}$$

where

$$\|g(t,x,u)\| \leqq \alpha(t)(\|x\| + \|u\|)\psi(\|u\|) + \beta(t).$$

Here, $\alpha(t) \geqq 0$ *and* $\beta(t) \geqq 0$ *both belong to* $L^1(0,T)$. *Further,* ψ *is nonnegative, continuous and* $\underline{\lim}_{M\to\infty}\psi(M) = 0$. *Assume that* $\dot{x} = A(t)x + B(t)u$ *is controllable on* $[0,T]$.

Then the system (10) *is bang-bang controllable on* $[0,T]$.

Proof. As in the proof of Theorem 2, let $x_1(t)$ be a solution of the perturbed system, and $x_2(t)$ a solution of the unperturbed system, corresponding to the same control $u(t)$ and satisfying $x_1(0) = x_2(0)$. Then

$$\|x_1(T) - x_2(T)\| \leqq C_1 \cdot \int_0^T \|g(t, x_1(t), u(t))\|\,dt.$$

Let $u(t) \in \tilde{\Omega}_M$. We then have

$$\|x_1(T) - x_2(T)\| \leqq C_1 \cdot \int_0^T [\alpha(t)(\|x_1(t)\| + M)\psi(M) + \beta(t)]\,dt,$$

or

$$\|x_1(T) - x_2(T)\| \leqq C_1(A\|x_1(\cdot)\|_{L^\infty}\psi(M) + A \cdot M \cdot \psi(M) + B_1),$$

using the notation of § 4.

In order to estimate $\max_t \|x_1(t)\|$, we write

$$x_1(t) = X(t,0)x_0 + \int_0^t X(t,s)[B(s)u(s) + g(s,x_1(s),u(s))]\,ds.$$

Thus

$$\|x_1(t)\| \leqq K\|x_0\| + K \cdot \int_0^T \|B(s)\|\,ds \cdot M$$
$$+ K \cdot \int_0^t [\alpha(s)(\|x_1(s)\| + M)\psi(M) + \beta(s)]\,ds,$$

or

$$\|x_1(t)\| \leqq K(\|x_0\| + B_1 + B \cdot M + A \cdot M \cdot \psi(M)) + K \cdot \psi(M) \cdot \int_0^t \alpha(s)\|x_1(s)\|\,ds.$$

Gronwall's lemma gives

$$\max_t \|x_1(t)\| \leqq K(\|x_0\| + B_1 + BM + AM\psi(M)) \exp(AK\psi(M)).$$

Now choose a sequence $M_i \to \infty$, such that $\lim_{i\to\infty} \psi(M_i) = 0$. We may assume that $\psi(M_i) \leqq 1$, and if $u \in \tilde{\Omega}_{M_i}$, we thus have $\max_t \|x_1(t)\| \leqq D_1 + D_2 \cdot M_i$, where the constants D_1, D_2 are easily specified. For $u \in \tilde{\Omega}_{M_i}$ we thus have

$$\|x_1(T) - x_2(T)\| \leqq C_1[A(D_1 + D_2M_i)\psi(M_i) + AM_i\psi(M_i) + B_1] \equiv \rho_i.$$

We can now put $r_{1,i} = M_id - \|X_T\|$, and apply Lemma 2. Since $\lim_{i\to\infty}(r_{1,i} - \rho_i) = +\infty$, it follows that $\bigcup_{i=1}^{\infty} \Phi(\tilde{\Omega}_{M_i}) = R^n$, which proves the theorem.

Remarks. It is seen from the proof that the assumption $\underline{\lim}_{M\to\infty} \psi(M) = 0$ can be replaced by the weaker condition that $\underline{\lim}_{M\to\infty} \psi(M)$ is small enough.

It is interesting that the bang-bang principle for linear systems via Lemma 2 thus offers a method to prove controllability for certain nonlinear systems. Apparently, many variations of this theme are possible, but we hope that the above theorem gives a sufficient illustration of the method. See also [8].

REFERENCES

[1] D. Blackwell, *The range of certain vector integrals*, Proc. Amer. Math. Soc., 2 (1951), pp. 390–395.

[2] J. P. Dauer, *Perturbations of linear control systems*, this Journal, 9 (1971), pp. 393–400.

[3] H. Hermes and J. P. LaSalle, *Functional Analysis and Time Optimal Control*, Academic Press, New York, 1969.

[4] E. B. Lee and L. Markus, *Foundations of Optimal Control Theory*, John Wiley, New York, 1967.

[5] D. L. Lukes, *Global controllability of nonlinear systems*, this Journal, 10 (1972), pp. 112–126.

[6] L. Markus, *Controllability of nonlinear processes*, this Journal, 3 (1965), pp. 78–90.

[7] A. Strauss, *An Introduction to Optimal Control Theory*, Lecture Notes in Operations Research and Mathematical Economics, Springer-Verlag, Berlin, 1968.

[8] G. Aronsson, *A new approach to nonlinear controllability*, J. Math. Anal. Appl., to appear.

SIAM J. CONTROL
Vol. 11, No. 4, November 1973

SUFFICIENT CONDITIONS FOR A STRONG MINIMUM IN SINGULAR CONTROL PROBLEMS*

H. GARDNER MOYER†

Abstract. This paper derives the conditions guaranteeing that a singular extremal that joins fixed endpoints provides a strong minimum for the independent time variable. For the nonsingular case, Weierstrass has shown that the extremal must be embedded in a field. The principal conditions that imply the existence of a field for a nonsingular problem with an n-dimensional state vector $\mathbf{x}$ and a scalar control variable u are that $\partial^2 H/\partial u^2$ is unequal to zero and that the $n \times (n + 1)$ matrix $[\partial \mathbf{x}(t)/\partial \boldsymbol{\lambda}(t_0), \dot{\mathbf{x}}(t)]$ has rank n. Here t is the time, H is the generalized Hamiltonian, and $\boldsymbol{\lambda}$ is the adjoint vector. This paper shows that under proper assumptions the field concept can be extended to the singular case. The condition on $\partial^2 H/\partial u^2$ is replaced by

$$\frac{\partial}{\partial u}\frac{d^2}{dt^2}\frac{\partial H}{\partial u} \neq 0.$$

The above matrix, whose first n columns are obtained in a uniform manner, is replaced by a matrix whose vectors are obtained by diverse procedures. Weirstrass' analysis is carried out in detail, using a singular nominal extremal and its field.

1. Introduction. The Pontryagin maximum principle states that the control that is optimal is the one for which the generalized Hamiltonian (see § 2) takes on its absolute maximum. If the graph of the Hamiltonian versus the control is a horizontal line, then all controls qualify, and if the circumstances permit the control to be chosen so that this line remains horizontal during a nonvanishing interval, a singular extremal is generated [2], [8], [11]. If, on the other hand, the maximum principle determines a unique control at almost all times, the extremal is called nonsingular. Of course some extremals are composed of both singular and nonsingular subarcs.

When an extremal is compared to curves that have neighboring values of both slope and position, it is a candidate to provide a weak local minimum. When the restriction on slope is removed, the extremal is a candidate to provide a strong local minimum. When the slope and position of the comparison curves are unrestricted, the extremal is a candidate to provide a global minimum.

Conditions guaranteeing that a singular extremal provides a weak local minimum can be derived from an analysis of the second variation [4], [7], [9], [13], [16]. Of course, there is but little justification for studying the weak minimum problem unless the strong minimum problem is considered to be intractable. Sufficient conditions for a strong minimum have been found by methods outside the calculus of variations, but this work is restricted to systems that are autonomous and whose control variables are all singular and number one less than the state variables [5], [12], [15].

The problem studied in the present paper is much more likely to occur in practice, although it is still far from being completely general. A system of differential equations is presented in Mayer format for a state vector $\mathbf{x}$, whose initial and final values are all specified. The single control variable u appears linearly

* Received by the editors May 31, 1972, and in revised form November 16, 1972.

† Research Department, Grumman Aerospace Corporation, Bethpage, New York 11714.

and is constrained to lie within finite upper and lower limits. The nominal extremal has no nonsingular subarcs. The objective is to find sufficient conditions for a strong, local minimum for the final value of the independent time variable t.

Weierstrass found that conditions that guarantee a strong minimum for nonsingular problems could be obtained by studying the extremal field rather than the second variation [3, pp. 143–149]. An extremal plotted in $\mathbf{x}$-space is said to be embedded in a *central Mayer field* if its neighborhood (excluding $\mathbf{x}_0$) is covered simply by the extremals that begin at $\mathbf{x}_0, t_0$ (Fig. 1). The members of the field define time contours $\bar{t}$, called *wavefronts* (Fig. 1). The overbar appearing in $\bar{t}$ distinguishes the time of the field from that defined in the manner described below.

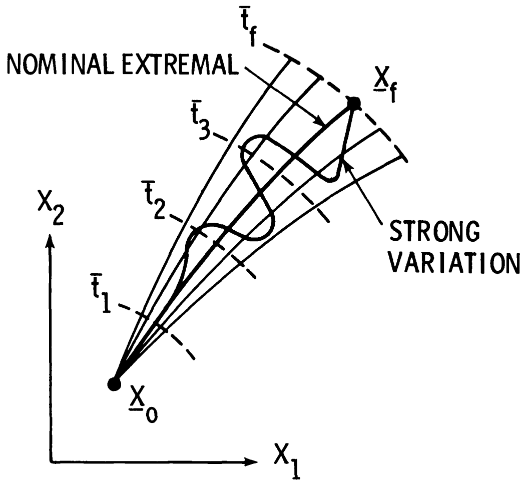


FIG. 1. *A family of adjacent extremals for a two state variable problem. Superimposed time contours (wavefronts) and a typical strong variation are shown*

It is possible to derive an expression for the gradient $\partial\bar{t}/\partial x_i$ at each point of the field. When this expression is substituted into the identity $\bar{t} = \bar{t}_0 + \int_{t_0}^{t} (\partial\bar{t}/\partial x_i)\dot{x}_i \, dt$, the Hilbert integral is obtained. Here $\dot{\mathbf{x}}$ and t are defined along the path of a comparison arc (strong variation) that is required only to obey the system equations for an admissible piecewise continuous u, and to have values of $\mathbf{x}$ that are in some sense close to those of the nominal extremal. The value of $\partial\bar{t}/\partial x_i$ to be used is that defined by the particular extremal the comparison arc currently intersects. Since $t - \bar{t} = \int_{t_0}^{t} [1 - (\partial\bar{t}/\partial x_i)\dot{x}_i] \, dt$, we see that in order to prove that $\bar{t}$ is less than, or equal to, the t of a strong variation governed by autonomous equations, the integrand (known as the Weierstrass excess function) must be shown to be greater than, or equal to, zero. In the nonautonomous case, the excess function is required to be nonnegative only when t equals $\bar{t}$.

The state and adjoint vectors of a nonsingular extremal with a scalar control obey

$$\text{(1a)} \qquad \dot{x}_i = \frac{\partial H}{\partial \lambda_i}, \quad \dot{\lambda}_i = -\frac{\partial H}{\partial x_i}, \qquad i = 1, \cdots, n,$$

$$\frac{\partial H}{\partial u} = 0, \tag{1b}$$

with $H \equiv \boldsymbol{\lambda} \cdot \dot{\mathbf{x}}$. If the Weierstrass condition is satisfied in strong form, so that corners are excluded, the derivatives of $\delta\mathbf{x}$ and $\delta\boldsymbol{\lambda}$ are given by the equations of variation of (1a). Provided

$$\frac{\partial^2 H}{\partial u^2} \neq 0, \tag{2}$$

δu may be obtained from the equation of variation of (1b). The $\delta\mathbf{x}(t)$ defined by a neighboring extremal that begins at $\mathbf{x}_0$ may then be obtained by employing the initial conditions $\delta\mathbf{x}(t_0) = \mathbf{0}$ with $\delta\boldsymbol{\lambda}(t_0)$ linearly independent of $\boldsymbol{\lambda}(t_0)$. The nominal extremal will be embedded in a central Mayer field if an arbitrary $\Delta\mathbf{x}$ determines a unique adjacent extremal and wavefront. Thus if a field exists, there must be a set of (nonunique) parameters $\delta\boldsymbol{\lambda}^i(t_0)$, $i = 1, \cdots, n-1$, such that the matrix

$$[\delta\mathbf{x}^1, \cdots, \delta\mathbf{x}^{n-1}, \dot{\mathbf{x}}] \tag{3}$$

has rank n, i.e., its vectors span n-dimensions, during $t_0 < t \leqq t_f$.

The present paper will find the conditions under which the field concept can be extended to singular extremals and it thus has the same general motivation as Caratheodory's thesis on broken extremals. A singular extremal for a two state variable problem is shown in Fig. 2. The branches are generated by jumping u from its singular value to either limit. We see that, although this family of arcs has a very different structure from that of Fig. 1, the property of simple covering appears here also, so that the general features of Weierstrass' argument can be applied. Note, however, that since the arcs of Fig. 2 are not generated by varying $\delta\boldsymbol{\lambda}(t_0)$, a new parameter must be found. Although the field approach could easily handle singular problems with $n = 2$ and would extend the earlier work to the

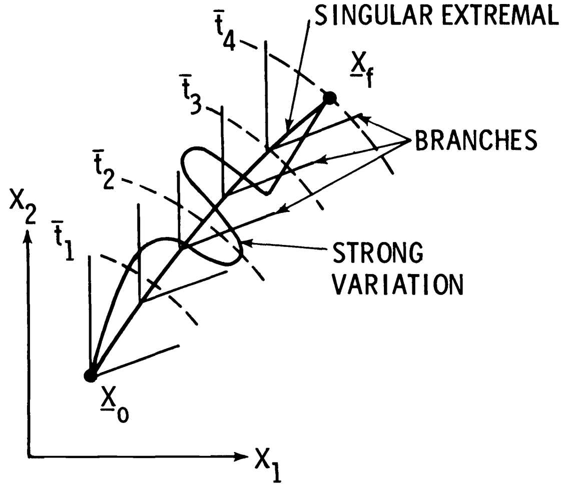


FIG. 2. *Singular extremal with branches for a two state variable problem. Superimposed time contours (wavefronts) and a typical strong variation are shown*

nonautonomous case, a detailed discussion will not be presented because it differs in certain respects from that for problems with $n > 2$.

Therefore, we proceed directly to three state variable problems in §§ 2–7. Sections 2–4 develop four conditions and show that they ensure that a unique singular extremal begins at $\mathbf{x}_0$ and that one, and only one, adjacent extremal passes through each point in its neighborhood. Sections 5 and 6 show that this "singular field," with no additional assumptions, guarantees that the nominal extremal provides a strong minimum. Section 7 investigates whether the conditions are satisfied for a sample problem. Section 8 outlines the modifications required for problems with $n > 3$. Section 9 touches briefly on the two state variable problem and indicates directions for further research. In view of the length and complexity of this paper, a complete listing of symbol definitions is given in § 10.

2. Singular extremals. We consider a Mayer control problem whose three dimensional state vector $\mathbf{x}$ obeys the following differential equations,

$$\dot{\mathbf{x}} = \mathbf{A}(\mathbf{x}, t) + \mathbf{B}(\mathbf{x}, t)u \tag{4}$$

and terminal constraints

$$\mathbf{x}(t_0) = \mathbf{x}_0, \quad \mathbf{x}(t_f) = \mathbf{x}_f. \tag{5}$$

The initial time t_0 is given while the final time t_f is to be minimized. The scalar control variable u is constrained by the finite constants $u_{\min}$ and $u_{\max}$ according to

$$u_{\min} \leqq u \leqq u_{\max}. \tag{6}$$

Assumption 2.1. All elements of the vectors $\mathbf{A}$ and $\mathbf{B}$ are of class C^3 in all their arguments.

We have seen that for nonsingular problems the expressions for $\dot{\mathbf{x}}$ are required merely to be of class C^2. The vector $\boldsymbol{\lambda}$ of adjoint variables obeys

$$\dot{\lambda}_i = -\frac{\partial H}{\partial x_i}, \qquad i = 1, 2, 3, \tag{7}$$

where the generalized Hamiltonian H is defined as the scalar product of $\dot{\mathbf{x}}$ and $\boldsymbol{\lambda}$,

$$H(\mathbf{x}, t, u, \boldsymbol{\lambda}) = \dot{\mathbf{x}} \cdot \boldsymbol{\lambda} = \mathbf{A} \cdot \boldsymbol{\lambda} + \mathbf{B} \cdot \boldsymbol{\lambda} u. \tag{8}$$

The switch function $H_u = \mathbf{B} \cdot \boldsymbol{\lambda}$ and its time derivative are formally independent of u and both vanish along a singular subarc,

$$H_u = \mathbf{B} \cdot \boldsymbol{\lambda} = 0, \tag{9}$$

$$\begin{aligned} \dot{H}_u &= \left[\frac{\partial B_i}{\partial x_j}(A_j + B_j u) + \frac{\partial B_i}{\partial t}\right]\lambda_i - B_j\left[\frac{\partial A_i}{\partial x_j} + \frac{\partial B_i}{\partial x_j}\mathrm{u}\right]\lambda_i \\ &= \left(\frac{\partial B_i}{\partial x_j}A_j + \frac{\partial B_i}{\partial t} - \frac{\partial A_i}{\partial x_j}B_j\right)\lambda_i \equiv \mathbf{C} \cdot \boldsymbol{\lambda} = 0. \end{aligned} \tag{10}$$

Assumption 2.2. The vectors $\mathbf{B}(t_0)$ and $\mathbf{C}(t_0)$ are linearly independent.

The special case with $\mathbf{B}$ and $\mathbf{C}$ collinear has been excluded merely to conserve space. Although this assumption can be retained for problems with $n > 3$, it has to be replaced for two state variable problems. For three dimensional problems, Assumption 2.2 implies that (9) and (10) determine two diametrically opposite directions for $\lambda(t_0)$. Setting the next derivative of the switch function to zero yields equations of the form

$$(11) \qquad \ddot{H}_u = f + gu_s = 0, \quad u_s = -f/g.$$

Assumption 2.3. The control variable appears explicitly in $\ddot{H}_u$ (i.e., $g \neq 0$) and the singular control is interior to the admissible region (i.e., $u_{\min} < u_s < u_{\max}$) for $t_0 \leqq t \leqq t_f$.

Note that u_s has the same value for both of the directions for λ mentioned above. Thus, if the complete family of extremals that issue from $\mathbf{x}_0$ were to be generated by varying λ_0, the singular curve would be found to be associated with two diametrically opposite λ directions. This phenomenon cannot occur for nonsingular extremals since it requires the graph of H versus u to be horizontal at all times.

Remark. Of the two possible directions for λ, the one that will be used throughout the rest of the paper makes

$$(12) \qquad g \equiv \frac{\partial}{\partial u}\frac{d^2}{dt^2}\frac{\partial}{\partial u}H > 0.$$

It will be shown that this λ, together with the maximum principle, yields a singular field. Had the minimum principle convention been adopted, it would have been necessary to select the other λ. Assumption 2.3 and inequality (12) require the plot of $\ddot{H}_u$ versus u to have the form of Fig. 3 so that $\ddot{H}_u(u_{\min}) < 0$ and $\ddot{H}_u(u_{\max}) > 0$.

Assumptions 2.1, 2.2, and 2.3 establish the existence of a unique singular extremal with initial point $\mathbf{x}_0$. Given that this curve passes through $\mathbf{x}_f$, we seek the conditions that guarantee that it provides a strong local minimum for t_f.

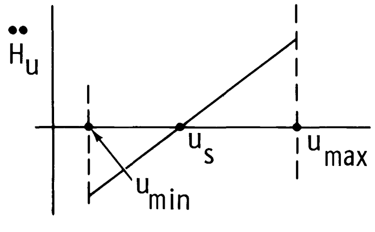


FIG. 3. *Second derivative of switch function versus control variable*

3. Branch manifolds and the singular manifold. At each point of a singular subarc, the control u can be jumped to either its upper or lower limit without violating the maximum principle. Subarcs that branch off from the singular subarc are then generated, and if they cover simply a two dimensional surface, that surface will be called the *branch manifold.* The branch manifold associated with the nominal extremal will be called the *nominal branch manifold* (Fig. 4). It can be seen that the nominal branch manifold will exist, provided $u_{\min} < u_s < u_{\max}$ and $[\mathbf{B}, \dot{\mathbf{x}}_s]$ has rank two. The former condition was included in Assumption 2.3 and the latter condition will be included in Assumption 4.1 of the next section.

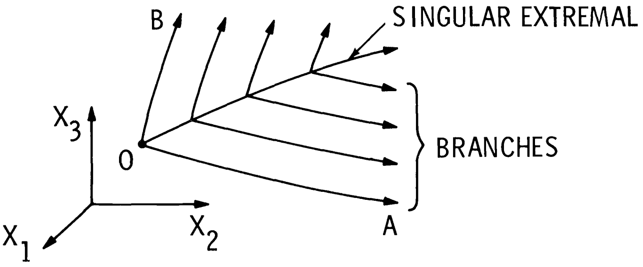


FIG. 4. *Extremal flow on the nominal branch manifold*

If $\lambda(t_0)$ does not satisfy both (9) and (10), the optimal u will be uniquely determined by the maximum principle at almost all times, and the extremal initially proceeds along either OA or OB (Fig. 4 or 5). If (9) and (10) should subsequently happen to be simultaneously satisfied, it would become possible to initiate a singular regime.

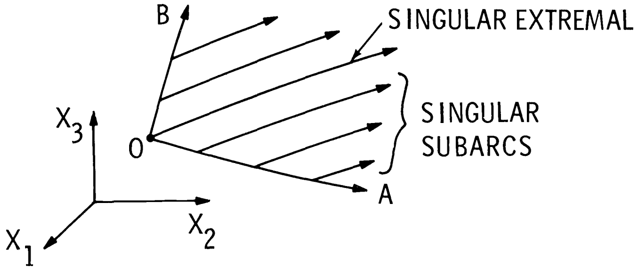


FIG. 5. *Extremal flow on the singular manifold*

THEOREM 3.1. *If a singular extremal satisfies Assumptions* 2.1, 2.2, *and* 2.3, *it is possible to increment its* $\lambda(t_0)$ *so that a unique singular subarc can be entered at any given point of* OA *or* OB *in the neighborhood of* O (*Fig.* 5).

Proof. When the $\lambda(t_0)$ of the singular extremal is incremented, (9) and (10) indicate that

$$H_u(t_0, \lambda + \delta\lambda) = \mathbf{B} \cdot \delta\lambda, \tag{13a}$$

$$\dot{H}_u(t_0, \lambda + \delta\lambda) = \mathbf{C} \cdot \delta\lambda. \tag{13b}$$

These equations are exact because λ appears linearly in (9) and (10). If the latter equations hold at time $t_0 + dt$, then

$$H_u(t_0 + dt, \lambda + \delta\lambda) = \mathbf{B} \cdot \delta\lambda + \dot{H}_u dt \simeq \mathbf{B} \cdot \delta\lambda = 0, \tag{14a}$$

$$\dot{H}_u(t_0 + dt, \lambda + \delta\lambda) = \mathbf{C} \cdot \delta\lambda + \ddot{H}_u dt = 0 \tag{14b}$$

if higher order terms such as $\dot{H}_u dt = \mathbf{C} \cdot \delta\lambda\, dt$ are (temporarily) neglected. The argument of $\ddot{H}_u$ is either $u_{\min}$ or $u_{\max}$ as appropriate. It is well known that all adjoint vectors in a given direction are associated with the same extremal. Thus, there is no loss in generality if the $\lambda + \delta\lambda$ vector is fixed by requiring $\delta\lambda$ to be normal to $\boldsymbol{\lambda}$,

$$\boldsymbol{\lambda} \cdot \delta\boldsymbol{\lambda} = 0. \tag{14c}$$

Equations (14) determine a unique $\delta\boldsymbol{\lambda}(t_0)$ for any choice of dt and a branch OA or OB (Fig. 5), provided that the matrix $[\mathbf{B}, \mathbf{C}, \boldsymbol{\lambda}]$ is nonsingular. But this property is implied by Assumption 2.2 and the discussion in the paragraph that follows it, so that the theorem has been proved.

Since the sign of the switch function on the initial subarc is not determined by the first order approximation (14a), before proceeding we should establish that it is in accord with the value of u used in (14b). Inserting (13) and (14b) into a second order expansion for the incremented switch function shows that the following relation holds at the corner point:

$$\begin{aligned} H_u(t_0 + dt, \boldsymbol{\lambda} + \delta\boldsymbol{\lambda}) &= H_u(t_0, \boldsymbol{\lambda} + \delta\boldsymbol{\lambda}) + \dot{H}_u(t_0, \boldsymbol{\lambda} + \delta\boldsymbol{\lambda})\, dt + \tfrac{1}{2}\ddot{H}_u(t_0, \boldsymbol{\lambda} + \delta\boldsymbol{\lambda})\, dt^2 \\ &= \mathbf{B} \cdot \delta\boldsymbol{\lambda} + \mathbf{C} \cdot \delta\boldsymbol{\lambda}\, dt + \tfrac{1}{2}\ddot{H}_u\, dt^2 \\ &= \mathbf{B} \cdot \delta\boldsymbol{\lambda} - \ddot{H}_u\, dt^2 + \tfrac{1}{2}\ddot{H}_u\, dt^2 \\ &= \mathbf{B} \cdot \delta\boldsymbol{\lambda} - \tfrac{1}{2}\ddot{H}_u\, dt^2 = 0. \end{aligned} \tag{15}$$

Thus, if, for example, $u_{\max}$ is being used, $\ddot{H}_u$ is positive (Fig. 3) and (15) makes the initial value of the switch function positive, so that there is no conflict. If, however, the diametrically opposite $\boldsymbol{\lambda}$ (which § 2 demonstrated to be also associated with the singular extremal) were to be used in $\ddot{H}_u$ together with $u_{\max}$, $\ddot{H}_u$ and the switch function would be negative and the maximum principle would be contradicted.

The time history of the $\delta\mathbf{x}(t)$ determined by the nominal singular extremal and the second subarc of one of the broken extremals of Fig. 5 is determined by the equations of variation of (4), (7), and (11).

$$\delta\dot{x}_i = \left(\frac{\partial A_i}{\partial x_j} + u\frac{\partial B_i}{\partial x_j}\right)\delta x_j + B_i\,\delta u, \qquad i = 1, 2, 3, \tag{16}$$

$$\delta\dot{\lambda}_i = -\frac{\partial^2 H}{\partial x_i \partial x_j}\delta x_j - \frac{\partial^2 H}{\partial x_i \partial \lambda_j}\delta\lambda_j - \frac{\partial^2 H}{\partial x_i \partial u}\delta u, \qquad i = 1, 2, 3, \tag{17}$$

$$\left(\frac{\partial f}{\partial x_i} + \frac{\partial g}{\partial x_i}u\right)\delta x_i + \left(\frac{\partial f}{\partial \lambda_i} + \frac{\partial g}{\partial \lambda_i}u\right)\delta\lambda_i + g\,\delta u = 0. \tag{18}$$

To solve (18) for δu, g (see (12)) must be nonzero. But this has already been stipulated by Assumption 2.3. Recall that at the corresponding point of the nonsingular analysis, it was necessary to assume $\partial^2 H/\partial u^2 \neq 0$.

If $u = u_{\max}$ on the initial infinitesimal subarc, then the values of $\delta\mathbf{x}$ and $\delta\boldsymbol{\lambda}$ at the start of the singular subarc are

$$\delta\mathbf{x}(t_0 + dt_0) = \mathbf{B}(t_0)(u_{\max} - u_s)\,dt_0, \tag{19}$$

$$\delta\lambda_i(t_0 + dt_0) = \delta\lambda_i(t_0) - \frac{\partial^2 H}{\partial x_i \partial u}(u_{\max} - u_s)\,dt_0, \qquad i = 1, 2, 3. \tag{20}$$

Here $\delta\boldsymbol{\lambda}(t_0)$ is determined by (14) with $u = u_{\max}$. If the singular subarc succeeds an infinitesimal subarc with $u = u_{\min}$, $u_{\min}$ replaces $u_{\max}$ in (19), (20), and (14).

The time histories of $\delta\mathbf{x}$ and $\delta\boldsymbol{\lambda}$ have the form shown in Fig. 6. Consequently, when $\delta\mathbf{x}(t_0 + dt_0)$ and $\delta\boldsymbol{\lambda}(t_0 + dt_0)$ as given by (19) and (20) are used as initial conditions for the integration of (16)–(18) with initial time t_0 rather than $t_0 + dt_0$, the resulting errors are infinitesimals of higher order than $\delta\mathbf{x}(t)$ and $\delta\boldsymbol{\lambda}(t)$ that can, therefore, be neglected.

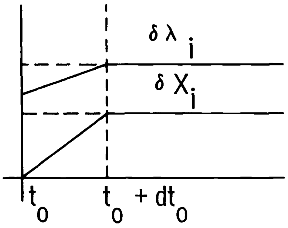


FIG. 6. *Time histories for δx_i and $\delta\lambda_i$ determined by an adjacent extremal that enters a singular regime at time $t_0 + dt_0$*

We have, thus, constructed a one-parameter family of neighboring broken extremals whose second subarc is singular. The two dimensional surface swept out by these subarcs will be called the *singular manifold* (Fig. 5). In view of Assumption 2.1 and the strong form of the inequalities appearing in Assumption 2.3, these inequalities and the linear independence of $\mathbf{B}$ and $\dot{\mathbf{x}}_s$ (see the beginning of this section) apply to the adjacent singular subarcs as well as to the nominal extremal. Therefore, each adjacent subarc will possess its own branch manifold.

The next theorem will determine whether the maximum principle or the minimum principle is satisfied on the branches. This information will be required when we decide whether the nominal extremal provides a maximum or a minimum.

THEOREM 3.2. *Let a singular subarc obey inequality* (12) *and Assumptions* 2.1 *and* 2.3, *and let a branch be generated at any point of the subarc by holding the control at either limit for a small interval; then the strong form of the maximum principle is satisfied on the branch.*

Proof. Expanding the switch function about the junction (as in (15) but with $\delta\boldsymbol{\lambda} = \mathbf{0}$) yields

$$(21)\qquad H_u(t + dt) = H_u(t) + \dot{H}_u(t)\,dt + \tfrac{1}{2}\ddot{H}_u(t)\,dt^2 + \cdots = \tfrac{1}{2}\ddot{H}_u\,dt^2 + \cdots.$$

Since $\ddot{H}_u(u_{\min}) < 0$ and $\ddot{H}_u(u_{\max}) > 0$ (Fig. 3), (21) shows that the switch function is positive on a branch with $u = u_{\max}$ and negative on a branch with $u = u_{\min}$ as required by the maximum principle.

4. Singular analog of the Mayer field. The parameter δp_0 will now be introduced into (19) and (20) so that these equations become

$$(22)\qquad \delta\mathbf{x}(t_0 + dt_0) = \mathbf{B}\delta p_0,$$

$$(23)\qquad \delta\lambda_i(t_0 + dt_0) = \delta\lambda_i(t_0) - \frac{\partial^2 H}{\partial x_i \partial u}\delta p_0, \qquad i = 1, 2, 3.$$

Thus, for δp_0 negative, $u_{\min}$ is used, and for δp_0 positive, $u_{\max}$ is used in (19) and (20). It is not difficult to show that the $\delta\mathbf{x}(t)$ that lie in the singular manifold vary directly with δp_0. That is, the ratio $\delta\mathbf{x}(t)/\delta p_0$ is independent of the magnitude of δp_0 and can be written $\partial\mathbf{x}/\partial p_0 \equiv \mathbf{x}_{p_0}$. To illustrate, two $\delta\mathbf{x}$'s determined by two equal and opposite δp_0's will be compared. If

$$\delta p_0^{(1)} = -\delta p_0^{(2)}, \tag{24}$$

then

$$(u_{\max} - u_s)\, dt_0^{(1)} = -(u_{\min} - u_s)\, dt_0^{(2)}. \tag{25}$$

The latter equation is multiplied by g; $(f + gu_s)dt_0^{(1)} = 0$ (see (11)) is added to the left; and $(f + gu_s)\, dt_0^{(2)} = 0$ is subtracted from the right.

$$(f + gu_{\max})\, dt_0^{(1)} = -(f + gu_{\min})\, dt_0^{(2)}. \tag{26}$$

Thus, (11) implies that

$$\ddot{H}_u(u_{\max})\, dt_0^{(1)} = -\ddot{H}_u(u_{\min})\, dt_0^{(2)}. \tag{27}$$

Therefore, the $\delta\boldsymbol{\lambda}^{(1)}(t_0)$ and $\delta\boldsymbol{\lambda}^{(2)}(t_0)$ obtained from (14) are equal and opposite. That $\delta\mathbf{x}^{(1)}(t) = -\delta\mathbf{x}^{(2)}(t)$ now follows easily from (22), (23), (18), (16), and (17).

Extremals that branch away from a singular subarc at the time $t > t_0$ define in the neighborhood of the branch point

$$\delta\mathbf{x}(t + dt) = \mathbf{B}(t)(u_{\min} - u_s)\, dt \tag{28}$$

and

$$\delta\mathbf{x}(t + dt) = \mathbf{B}(t)(u_{\max} - u_s)\, dt. \tag{29}$$

The parameter δp_t will be introduced in the same way as δp_0. Equations (28) and (29) can then be combined, and if only first order terms are retained, we have

$$\delta\mathbf{x}(t) = \mathbf{B}(t)\delta p_t = \frac{\partial\mathbf{x}(t)}{\partial p_t}\delta p_t \equiv \mathbf{x}_{p_t}\,\delta p_t. \tag{30}$$

DEFINITION 4.1. A singular extremal will be said to be embedded in a *singular field* if for $t_0 < t \leqq t_f$ it possesses a branch manifold on each side of which open neighborhoods are simply covered by singular subarcs and their branches (see Fig. 7). For the purpose of this definition, a singular subarc is counted only once despite its channel character.

Extremal families can fail to provide a field for various reasons. Should the state variable differential equations include a holonomic constraint, the extremals will fill only a subspace. Should the nominal extremal contact an envelope, the adjacent extremals cover one side doubly and the other side not at all. Envelope

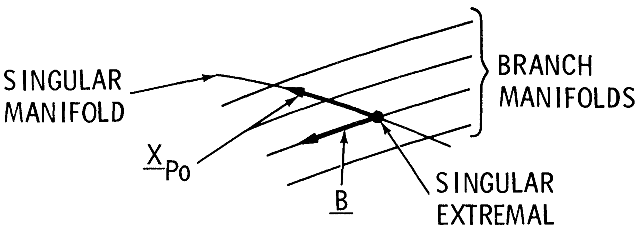


FIG. 7. *Trace of singular field in wavefront*

contacts are associated either with the Jacobi condition or with H passing through zero. In the former case, one covering is supplied by saddle point arcs and the other by either minimizing or maximizing arcs. In the latter case, the double covering consists of minimum and maximum time arcs, and the difficulty is not intrinsic in that it is (usually) possible to project the extremal family onto (x_1, x_2, t)-space and obtain simple covering by extremals that either minimize or maximize x_3. The present discussion will be limited to deriving the conditions that ensure simple covering in (x_1, x_2, x_3)-space of the neighborhood of the nominal extremal.

Assumption 4.1. The determinant

$$\begin{aligned}\det[\mathbf{P}(t)] &\equiv \det[\mathbf{x}_{p_0}, \mathbf{x}_{p_t}, \dot{\mathbf{x}}_s] \\ &= \det[\mathbf{x}_{p_0}, \mathbf{B}, \mathbf{A} + \mathbf{B}u_s] = \det[\mathbf{x}_{p_0}, \mathbf{B}, \mathbf{A}]\end{aligned} \tag{31}$$

has well-behaved elements and is nonzero for $t_0 < t \leqq t_f$.

We will show that this provision excludes all the pathological situations mentioned above. The relationship between det $\mathbf{P}$ and the value of H will be taken up explicitly in Theorem 5.2. The matrix of (31) is the singular counterpart of the matrix (3).

THEOREM 4.1. *A singular extremal is embedded in a singular field if Assumptions* 2.1, 2.2, 2.3, *and* 4.1 *are satisfied.*

Proof. Assumption 4.1 implies that $\mathbf{B}$ and $\dot{\mathbf{x}}_s$ are nonnull vectors that are linearly independent. These properties, together with the inequalities mentioned in Assumption 2.3, imply the existence of a two-dimensional nominal branch manifold (Fig. 4). Theorem 3.1 states that there are adjacent singular subarcs branching away from the subarcs OA and OB of this manifold (Fig. 5). In accordance with the discussion preceding Theorem 3.2, each of these subarcs possesses its own branch manifold. Assumption 4.1 indicates that the vectors $\mathbf{x}_{p_0}$ and $\mathbf{B}$ are linearly independent as shown in Fig. 7 so that the singular subarcs straddle the nominal branch manifold. The implicit function theorem permits the matrix equation

$$\Delta\mathbf{x} = [\mathbf{x}_{p_0}, \mathbf{B}, \dot{\mathbf{x}}_s][\delta p_0, \delta p_t, dt]^T \tag{32}$$

to be inverted to

$$[\delta p_0, \delta p_t, dt]^T = [\mathbf{x}_{p_0}, \mathbf{B}, \dot{\mathbf{x}}_s]^{-1}\Delta\mathbf{x}. \tag{33}$$

Thus an arbitrary $\Delta\mathbf{x}$ is associated with a unique adjacent singular subarc, branch, and wavefront.

Figure 7 indicates that if $\mathbf{x}_{p_0}$ and $\mathbf{B}$ should become collinear (with $\mathbf{B}\cdot\mathbf{B} > 0$), an adjacent singular subarc would cross the nominal branch manifold. Assumption 4.1 has ruled out the possibility of such an intersection. It would be clearly analogous to a Jacobi conjugate point, which is defined for nonsingular problems as the point at which neighboring extremals (with $\delta\mathbf{x}(t_0) = \mathbf{0}$) meet at a common time [3, pp. 105–124].

5. The Hilbert invariant integral. Before the existence of an invariant integral can be established, the properties of $\boldsymbol{\lambda}$ and H must be investigated. In the nonsingular case it is well known that $\boldsymbol{\lambda}$ is normal to the wavefront [3, p. 216]. This property is easily extended to the singular case.

THEOREM 5.1. *The vector* $\boldsymbol{\lambda}$ *associated with a singular extremal that obeys Assumptions* 2.1, 2.2, 2.3, *and* 4.1 *is normal to the wavefront.*

Proof. If $\delta\mathbf{x}$ is in the nominal branch manifold, (9) indicates that

$$\boldsymbol{\lambda}\cdot\delta\mathbf{x} = \boldsymbol{\lambda}\cdot\mathbf{B}\delta p_t = H_u\delta p_t = 0. \tag{34}$$

Recall that $\mathbf{B}$ is not a null vector by virtue of Assumption 4.1. If $\delta\mathbf{x}$ is in the singular manifold,

$$\frac{d}{dt}(\boldsymbol{\lambda}\cdot\delta\mathbf{x}) = \dot{\boldsymbol{\lambda}}\cdot\delta\mathbf{x} + \boldsymbol{\lambda}\cdot\delta\dot{\mathbf{x}} = 0 \tag{35}$$

by virtue of (7), (16), and (9). Thus, $\boldsymbol{\lambda}\cdot\delta\mathbf{x} =$ const. and the constant is zero because at $t = t_0 + dt_0$, $\boldsymbol{\lambda}\cdot\delta\mathbf{x} = \boldsymbol{\lambda}\cdot\mathbf{B}\,\delta p_0 = 0$. In the general case, $\delta\mathbf{x}$ can be decomposed into a linear combination of $\mathbf{B}\,\delta p_t$ and $\mathbf{x}_{p_0}\,\delta p_0$.

THEOREM 5.2. *The Hamiltonian* $H = \boldsymbol{\lambda}\cdot\dot{\mathbf{x}}_s$ *never vanishes during* $t_0 \leqq t \leqq t_f$ *on a singular extremal that obeys Assumptions* 2.1, 2.2, 2.3, and 4.1.

Proof. The coefficients $\partial(A_i + B_iu)/\partial x_j$, $i,j = 1,2,3$, of λ_i in (7) (after substitution from (8)) are required by Assumption 2.1 to be bounded. It follows from this and the homogeneity of $\boldsymbol{\lambda}$ in (7), that if $\boldsymbol{\lambda}$ were null at one point of the extremal, it would be null at all points. If $\dot{\mathbf{x}}_s$ (the other vector appearing in the definition of H) were to be null, Assumption 4.1 would be violated. This proviso also implies that the vectors $\mathbf{x}_{p_0}$, $\mathbf{B}$, and $\dot{\mathbf{x}}_s$ span three dimensions. Since Theorem 5.1 demonstrated that $\boldsymbol{\lambda}$ is normal to $\mathbf{x}_{p_0}$ and $\mathbf{B}$, this vector must have a component in the $\dot{\mathbf{x}}_s$ direction. Thus $H \neq 0$ on the nominal singular extremal. In view of the strong form of the latter inequality and the continuity properties assumed for $\mathbf{A}$ and $\mathbf{B}$, there must be an open neighborhood about the nominal extremal in which the singular field defines H to be nonzero.

DEFINITION 5.1. Let the path of the nominal singular extremal be denoted by $\bar{\mathbf{x}}_s(t)$ and that of a comparison arc that obeys (4) and (6) by $\mathbf{x}(t)$. Assume that at each point $\mathbf{x}(t)$, the quantity $(\mathbf{x}(t) - \bar{\mathbf{x}}_s(\bar{t}))\cdot(\mathbf{x}(t) - \bar{\mathbf{x}}_s(\bar{t}))$ is a second order infinitesimal or smaller for at least one value of $\bar{\mathbf{x}}_s(\bar{t})$. Then the comparison arc is called a *strong variation.* The adjective "strong" connotes that u and $\dot{\mathbf{x}}$ on the varied arc are not required to be close to the corresponding quantities on the nominal extremal and may in fact have a finite number of jump discontinuities.

THEOREM 5.3. *Let Assumptions* 2.1, 2.2, 2.3, *and* 4.1 *be satisfied. Let* $\boldsymbol{\lambda}$, $\bar{t}$, $\bar{u}$, *and* $\bar{H} \equiv \boldsymbol{\lambda}\cdot\dot{\mathbf{x}}(\mathbf{x}, \bar{t}, \bar{u})$ *be defined in the neighborhood of the nominal singular extremal by the singular field. Let* t, u, *and* $H \equiv \boldsymbol{\lambda}\cdot\dot{\mathbf{x}}(\mathbf{x}, t, u)$ *be defined by a strong variation whose initial and final points coincide with those of the nominal singular extremal. Then the integral*

$$\int_{t_0}^{t_f} (H/\bar{H})\,dt \tag{36}$$

taken along the path of the variation is independent of the path and the interval $t_f - t_0$.

Proof. Theorem 5.1 established

$$\boldsymbol{\lambda}\cdot\delta\mathbf{x} = 0 \tag{37}$$

for any $\delta\mathbf{x}$ that lies in the wavefront tangent plane. Taking the scalar product of $\boldsymbol{\lambda}$ and

$$\Delta\mathbf{x} \equiv \delta\mathbf{x} + \dot{\mathbf{x}}(\mathbf{x},\bar{t},\bar{u})\,d\bar{t} \tag{38}$$

and using (37) and (8), we obtain

$$\boldsymbol{\lambda}\cdot\Delta\mathbf{x} = \boldsymbol{\lambda}\cdot\dot{\mathbf{x}}\,d\bar{t} = \bar{H}\,d\bar{t}. \tag{39}$$

Note that although $\bar{u}$ in (38) is triple-valued whenever the comparison arc is on the singular manifold, $\bar{H}$ in (39) is always unique. The latter equation implies

$$\frac{\partial\bar{t}}{\partial x_i} = \frac{\lambda_i}{\bar{H}}, \qquad i = 1,2,3. \tag{40}$$

Recall that $\bar{H}$ was shown to be nonzero at the conclusion of the proof of Theorem 5.2. Therefore, along the line integral of the variation,

$$\bar{t}_f - \bar{t}_0 = \int_{\mathbf{x}_0}^{\mathbf{x}_f} \frac{\partial\bar{t}}{\partial x_i}\Delta x_i = \int_{t_0}^{t_f} \frac{\partial\bar{t}}{\partial x_i}\dot{x}_i\,dt = \int_{t_0}^{t_f} \frac{\lambda_i}{\bar{H}}\dot{x}_i\,dt = \int_{t_0}^{t_f} \frac{H}{\bar{H}}\,dt. \tag{41}$$

The last member is (36) while the first is a function of the field only.

Although it has been helpful to think of a strong variation when proving this theorem, it can be seen that t plays a dummy role and that the varied arc is required to obey neither (6) nor even (4). Note that Hilbert's invariant integral for nonsingular Mayer control problems is formally identical to (36).

6. Sufficient conditions for a strong local minimum. An adequate background is now available for the proof of the following fundamental sufficiency theorem.

THEOREM 6.1. *Let $\bar{t}_0$ and $\bar{t}_f$ be defined by a nominal singular extremal that obeys inequality* (12) *and Assumptions* 2.1, 2.2, 2.3, *and* 4.1; *let $\bar{H}$ be defined in its neighborhood by the singular field implied by Theorem* 4.1; *let H be defined between $\bar{t}_0$ and t_f by a strong variation whose initial and final points coincide with those of the nominal extremal. Then $\bar{t}_f \leqq t_f$ if $\bar{H} > 0$ and $\bar{t}_f \geqq t_f$ if $\bar{H} < 0$.*

Proof. Since the assumptions of Theorem 4.1 have been postulated, the singular extremal is embedded in a singular field and (41) can be applied.

$$\int_{t_0}^{t_f} \frac{\bar{H} - H}{\bar{H}}\,dt = (t_f - \bar{t}_0) - (\bar{t}_f - \bar{t}_0) = t_f - \bar{t}_f. \tag{42}$$

The integrand is the Weierstrass excess function for a Mayer problem. To be definite, $\bar{H}$ (the denominator of the integrand) will be assumed positive. By Theorem 3.2 the numerator is positive or zero whenever $t = \bar{t}$. Since these times are equal initially, t cannot be less than $\bar{t}$ when t is small. Depending upon the definition of H, it may be possible to have $H > \bar{H}$ when $t > \bar{t}$ so that $d\bar{t}/dt = H/\bar{H} > 1$ (see (41)) and the comparison arc reduces the gap with the wavefront. The difference $t - \bar{t}$ could never become zero, however, since this would imply $H > \bar{H}$ at $t = \bar{t}$.

7. Example. The system governed by the following equations has been previously studied for a variable final point problem in [6], [13], and [14].

$$\dot{\mathbf{x}} = \mathbf{A} + \mathbf{B}u, \tag{43}$$

$$\mathbf{A}^T \equiv [x_2, 0, \tfrac{1}{2}(x_2^2 - x_1^2)], \tag{44}$$

$$\mathbf{B}^T \equiv [0, 1, 0], \tag{45}$$

$$-2 \leqq u \leqq 2. \tag{46}$$

The initial conditions are

$$t_0 = 0, \quad \mathbf{x}_0^T = [0, 1, 0]. \tag{47}$$

The problem is to minimize the time to a fixed final point $\mathbf{x}_f$. The location of $\mathbf{x}_f$ will be left unspecified except that it lies on the singular extremal described below. The adjoint variables obey

$$\dot{\boldsymbol{\lambda}}^T = [x_1\lambda_3, -\lambda_1 - x_2\lambda_3, 0]. \tag{48}$$

The switch function and its derivative both vanish on a singular extremal (see (9) and (10)).

$$H_u = \mathbf{B} \cdot \boldsymbol{\lambda} = \lambda_2 = 0, \qquad \dot{H}_u = \mathbf{C} \cdot \boldsymbol{\lambda} = -\lambda_1 - x_2\lambda_3 = 0. \tag{49}$$

The vector $\mathbf{B}$ was given by (45) and $\mathbf{C}$ is

$$\mathbf{C}^T = [-1, 0, -x_2]. \tag{50}$$

The singular control is obtained from

$$\ddot{H}_u = -(x_1 + u)\lambda_3 = 0, \quad u_s = -x_1. \tag{51}$$

A vector $\boldsymbol{\lambda}(t_0)$ perpendicular to the linearly independent vectors $\mathbf{B}(t_0)$ and $\mathbf{C}(t_0)$ is

$$\boldsymbol{\lambda}_s^T(0) = [x_2(0), 0, -1] = [1, 0, -1]. \tag{52}$$

This vector, rather than its opposite, will be employed because it makes $\partial\ddot{H}_u/\partial u > 0$ and is therefore associated with a singular field when the maximum principle convention is adopted. The equations for the singular extremal are

$$\mathbf{x}_s^T(t) = [\sin t, \cos t, \tfrac{1}{4}\sin 2t]. \tag{53}$$

The equations of variation of (43) and (51) are

$$\delta\dot{\mathbf{x}}^T = [\delta x_2, \delta u, x_2\,\delta x_2 - x_1\,\delta x_1], \tag{54}$$

$$\delta u = -\delta x_1. \tag{55}$$

The solution to (54) and (55) with $\delta\mathbf{x}(0) = \mathbf{B}(0)$ (see the discussion following (20)) is

$$\mathbf{x}_{p_0}^T = [\sin t, \cos t, \tfrac{1}{2}\sin 2t]. \tag{56}$$

Using (44), (45), (56), and (53) to evaluate (31) yields

$$\det \mathbf{P} = \det[\mathbf{x}_{p_0}, \mathbf{B}, \mathbf{A}] = -\tfrac{1}{2}\sin t. \tag{57}$$

Therefore the sufficiency conditions 2.1, 2.2, 2.3, and 4.1 are satisfied for $0 < t < \pi$, and during this interval the singular extremal provides a strong local minimum because $H(t) = \frac{1}{2} > 0$. At $t = \pi$ the vectors $\mathbf{x}_{p_0}$ and $\mathbf{B}$ (Fig. 7) are collinear, and the singular subarc of the adjacent broken extremal (Fig. 5) intersects the nominal branch manifold. The character of the nominal arc after this time is beyond the scope of the present paper.

Note that the sufficiency established above does not contradict the assertion that the extremal is nonoptimal for $t_f < \pi$ when the final point is variable [6], [13]. Both conclusions can be correct if the point conjugate to the final point always precedes the focal point. This has been proved for the nonsingular case [1, p. 175] and also presumably holds for singular problems.

8. More than three state variables. In this section $\mathbf{x}$ will be defined as a general n-vector ($n \geqq 3$) although u will still be scalar. We have seen that when $n = 3$, Assumption 2.2 imposes two conditions on $\boldsymbol{\lambda}$ ((9) and (10)) so that (provided Assumption 2.3 is satisfied) a unique singular extremal originates at a given $\mathbf{x}_0$. When $n > 3$ an $(n - 3)$-parameter family of singular extremals begins at the initial point. The members of this family satisfy (12) and Assumption 2.3 because of the strong form of these conditions. Each member adjacent to a given nominal member $\mathbf{x}_s(t)$ defines a $\delta\mathbf{x}(t)$ that obeys (16)–(18) with $\delta\mathbf{x}(t_0) = \mathbf{0}$. The initial values of the $\boldsymbol{\lambda}_s + \delta\boldsymbol{\lambda}^i$, $i = 1, \cdots, n - 3$, must of course satisfy (9) and (10),

$$(58) \qquad \mathbf{B}\cdot(\boldsymbol{\lambda}_s + \delta\boldsymbol{\lambda}^i) = \mathbf{B}\cdot\delta\boldsymbol{\lambda}^i = 0, \qquad \mathbf{C}\cdot(\boldsymbol{\lambda}_s + \delta\boldsymbol{\lambda}^i) = \mathbf{C}\cdot\delta\boldsymbol{\lambda}^i = 0$$
$$i = 1, \cdots, n - 3,$$

and if the $\delta\mathbf{x}^i$ are to be linearly independent, the rank of $[\boldsymbol{\lambda}, \delta\boldsymbol{\lambda}^i, \cdots, \delta\boldsymbol{\lambda}^{n-3}]$ must equal $n - 2$. These relations are unchanged if each $\delta\boldsymbol{\lambda}^i(t_0)$ is scaled to have unit magnitude. We can then regard the $\delta\mathbf{x}^i(t)$ as partial derivatives and designate them as $\mathbf{x}_{p_i}$, $i = 1, \cdots, n - 3$.

A $\delta\boldsymbol{\lambda}(t_0)$ to be used to generate an extremal that enters a singular regime at a given point of OA or OB (Fig. 5) must satisfy only (14). Thus an $(n - 3)$-parameter family of singular subarcs begins at each of these points. When the latter point is regarded as an additional parameter, there is altogether an $(n - 2)$-parameter family of neighboring broken extremals. To obtain an $\mathbf{x}_{p_0}(t)$ in the simplest way, integrate (16)–(18) with $\delta\mathbf{x}(t_0) = \mathbf{B}(t_0)$, $\delta\lambda_i(t_0) = 0$, $i = 4, \cdots, n$, and $\delta\lambda_1$, $\delta\lambda_2$, and $\delta\lambda_3$ chosen to satisfy (14).

Thus Theorem 6.1 holds for $n > 3$ provided Assumption 4.1 is replaced by the following.

Assumption 8.1. The determinant

$$(59) \qquad \det \mathbf{P} \equiv \det [\mathbf{x}_{p_0}, \mathbf{x}_{p_1}, \cdots, \mathbf{x}_{p_{n-3}}, \mathbf{B}, \mathbf{A}]$$

has well-behaved elements and is nonzero during $t_0 < t \leqq t_f$. As discussed above, the $\mathbf{x}_{p_i}$, $i = 1, \cdots, n - 3$, are defined by adjacent members of the $(n - 3)$-parameter family of curves that are singular for $t \geqq t_0$, and $\mathbf{x}_{p_0}$ is defined by any member of the $(n - 2)$-parameter family of adjacent broken extremals whose second subarc is singular.

9. Concluding remarks. When Assumptions 2.1, 2.2, and 2.3 are satisfied for a problem with a linear, scalar control variable, an $(n - 3)$-parameter family of singular extremals issues from the initial point. After introducing the three additional parameters δp_0, δp_t, and dt, the $n \times n$ matrix $\mathbf{P}$ can be formed. If Assumption 8.1 is also satisfied, the neighborhood in $\mathbf{x}$-space of the nominal arc is covered simply by a "singular extremal field." That the nominal arc takes less time than any strong variation joining $\mathbf{x}_0$ and $\mathbf{x}_f$ can then be proved by an analysis identical to that used by Weierstrass for the nonsingular case.

It can be shown that the set of four assumptions mentioned above still guarantees optimality when the problem is changed to allow x_{n_f} to vary while t_f is fixed. If λ_{n_f} is positive, x_{n_f} is a maximum and if λ_{n_f} is negative, x_{n_f} is a minimum. As mentioned at the end of § 7, these assumptions are no longer sufficient if two or more of the $n + 1$ variables $\mathbf{x}_f$, t_f are not fixed or if any of the initial values can vary.

On comparing the assumptions mentioned above with those for the nonsingular case, we see that the derivatives of the expressions for $\dot{\mathbf{x}}$ that are now required are of higher order, the condition $\partial \ddot{H}_u/\partial u \neq 0$ replaces (2), and the matrix of (59) replaces the matrix (3). The remaining condition—that excluding the special case with **B** and **C** collinear—was inserted merely to save space and thus corresponds to the exclusion of broken and singular extremals from Weierstrass' original theory.

As in the studies of a weak minimum in singular problems, one of these conditions contains a matrix. However, the elements of the matrices found in [4], [7], [9], [13], and [16] were obtained in a uniform manner from a single set of either linear or Riccati differential equations. The matrix of (59), on the contrary, reflects the nonuniform character of the singular field. The vector $\mathbf{x}_{p_0}$ is obtained from the differential equations (16)–(18) with $\delta\mathbf{x}(t_0) = \mathbf{B}(t_0)$. The vectors $\mathbf{x}_{p_i}$, $i = 1, \cdots, n - 3$, are obtained from the same equations but with $\delta\mathbf{x}^i(t_0) = \mathbf{0}$ and with constraints on the $\delta\boldsymbol{\lambda}^i(t_0)$ that are of a different form. Finally, the vectors **B** and **A** of **P** are taken directly from the nominal singular extremal. On the other hand, in the nonsingular case identical matrices can be used to analyze the second variation and to establish the existence of a field.

The existence and optimality of a singular extremal with two state variables can also be established by four conditions. Assumptions 2.1 and 2.3 can be taken over without change, but the stipulation that **B** and **C** are collinear should replace Assumption 2.2 and $\mathbf{x}_{p_0}$ should be deleted from the matrix appearing in Assumption 4.1. These conditions can be applied more generally than those of [5], [12], and [15] because the system differential equations are not required to be autonomous. It can be shown that for the latter special case, the present analysis is equivalent to the earlier work.

Opportunities for further research are obtained when the assumptions of the present paper are weakened. Thus, investigations can be conducted into the conditions that arise when $\mathbf{B}(t_0)$ is collinear with $\mathbf{C}(t_0)$, when the first explicit appearance of u occurs in a time derivative of H_u higher than the second, when $\mathbf{B}(t)$ becomes null at an intermediate point, when $H(t)$ has a varying sign, and when the nominal extremal has both singular and nonsingular subarcs. More general problem formulations can also be explored. These include problems with more than one singular control variable and variable endpoint problems.

The ultimate goal of a set of conditions for a strong minimum that are both necessary and sufficient remains to be achieved. The necessity of the condition $\partial \ddot{H}_u, \partial u \geqq 0$ (the weak form of inequality (12)) in conjunction with the maximum principle has already been established [10], [11], but not the necessity of the new analog of the Jacobi condition obtained for the first time in this paper. The latter requires the matrix obtained by deleting the last column of the matrix appearing in (59),

$$[\mathbf{x}_{p_0}, \cdots, \mathbf{x}_{p_{n-3}}, \mathbf{B}], \tag{60}$$

(thereby retaining only the partial derivatives with time fixed) to have rank $n - 1$ (whenever $\mathbf{B} \cdot \mathbf{B} > 0$ during $t_0 < t < t_f$) irrespective of the boundary conditions and performance index. Note that when $n = 2$, all the columns of this matrix are undefined except the last. It follows that the concept of conjugate points is then inapplicable.

10. Notation.

$\mathbf{A}, \mathbf{B}$ = n-vectors appearing in (4).

$\mathbf{C}$ = n-vector appearing in (10).

f, g = see (11) and (12).

H = generalized Hamiltonian; see (8).

H_u = switch function; see (9).

n = number of state variables, equals 3 in §§ 2–7 and is >3 in §§ 8 and 9.

$\mathbf{P}$ = matrix defined by (31) for $n = 3$ and (59) for $n > 3$.

δp_0 = either $[u_{\max} - u_s(t_0)]\,dt_0$ or $[u_{\min} - u_s(t_0)]\,dt_0$. The $\delta\mathbf{x}(t_0 + dt_0)$ defined by the nominal extremal and an extremal that lies in the singular manifold equals $\mathbf{B}(t_0)\delta p_0$ at the time the singular subarc of the neighboring extremal begins (Fig. 5). The singular manifold can be swept out by varying δp_0 and t.

δp_t = either $[u_{\max} - u_s(t)]\,dt$ or $[u_{\min} - u_s(t)]\,dt$. The $\delta\mathbf{x}(t + dt)$ defined by the nominal extremal and an extremal that branched away from it at time t (Fig. 4) is equal to $\mathbf{B}(t)\delta p_t$. The nominal branch manifold can be swept out by varying δp_t and t.

t = time, the independent variable.

u = scalar control variable.

$u_{\max}, u_{\min}$ = upper and lower limits, respectively, of the admissible values of u.

$\mathbf{x}$ = n-vector of state variables.

$\mathbf{x}_{p_0}$ = $\partial\mathbf{x}(t)/\partial p_0$ (Fig. 7). See the definition of δp_0.

$\mathbf{x}_{p_t}$ = $\partial\mathbf{x}(t)/\partial p_t = \mathbf{B}(t)$ (Fig. 7). See the definition of δp_t.

$\mathbf{x}_{p_i}$ = $\partial\mathbf{x}(t)/\partial p_i$, where the p_i, $i = 1, \cdots, n - 3$, are any set of parameters that can be used to describe the $(n - 3)$-parameter family of singular extremals that issue from $\mathbf{x}_0$.

$\boldsymbol{\lambda}$ = n-vector of adjoint variables associated with $\mathbf{x}$.

$(\ )_0$ = initial value.

$(\ )_f$ = final value.

$(\ )_s$ = value on the nominal singular extremal.

$(\ )^T$ = transpose of a matrix.

$\delta(\ )$ = variation with time fixed.

$\Delta(\ )$ = variation with time not fixed.

$(\dot{\ })$ = derivative with respect to time.

$(\bar{\ })$ = when a variable is defined both by a comparison arc and the extremals of the field, the latter variable has an overbar.

$(\ )\cdot(\ )$ = scalar product of two vectors.

$(\ )_i(\ )_i$ = summation over the appropriate range with i as the running variable.

REFERENCES

[1] G. A. Bliss, *Lectures on the Calculus of Variations*, Univ. of Chicago, Chicago, 1946.

[2] R. Gabasov and F. M. Kirillova, *High order necessary conditions for optimality*, this Journal, 10 (1972), pp. 127–168.

[3] I. M. Gelfand and S. V. Fomin, *Calculus of Variations*, Prentice-Hall, Englewood Cliffs, N.J., 1963.

[4] B. S. Goh, *A theory of the second variation in optimal control*, unpublished report, Division of Applied Mechanics, University of California, Berkeley, 1970.

[5] G. W. Haynes, *On the optimality of a totally singular vector control: An extension of the Green's theorem approach to higher dimensions*, this Journal, 4 (1966), pp. 662–685.

[6] D. H. Jacobson, *On Conditions of Optimality for Singular Control Problems*, IEEE Trans. Automatic Control, AC-15 (1970), pp. 109–110.

[7] ———, *Sufficient conditions for non-negativity of the second variation in singular and nonsingular control problems*, this Journal, 8 (1970), pp. 403–423.

[8] ———, *Totally singular quadratic minimization problems*, IEEE Trans. Automatic Control, AC-16 (1971), pp. 651–658.

[9] D. H. Jacobson and J. L. Speyer, *Necessary and sufficient conditions for optimality for singular control problems: A limit approach*, J. Math. Anal. Appl., 34 (1971), pp. 239–266.

[10] H. J. Kelley, *A second variation test for singular extremals*, AIAA J., 2 (1964), pp. 1380–1382.

[11] H. J. Kelley, R. E. Kopp and H. G. Moyer, *Singular extremals*, Topics in Optimization, Academic Press, New York, 1967, pp. 63–101.

[12] J. D. Mancill, *Identically non-regular problems in the calculus of variations*, Matematica Y Fisica Teorica, ser. A, 7 (1950), pp. 131–139.

[13] J. P. McDanell and W. F. Powers, *New Jacobi-type necessary and sufficient conditions for singular optimization problems*, AIAA J., 8 (1970), pp. 1416–1420.

[14] ———, *Necessary conditions for joining optimal singular and nonsingular subarcs*, this Journal, 9 (1971), pp. 161–173.

[15] A. Mieli, *Extremization of linear integrals by Green's theorem*, Optimization Techniques with Application to Aerospace Systems, Academic Press, New York, 1962, pp. 69–98.

[16] J. L. Speyer and D. H. Jacobson, *Necessary and sufficient conditions for optimality for singular control problems: A transformation approach*, J. Math. Anal. Appl., 33 (1971), pp. 163–187.

SIAM J. Control
Vol. 11, No. 4, November 1973

A DESCENT NUMERICAL METHOD FOR OPTIMIZATION PROBLEMS WITH NONDIFFERENTIABLE COST FUNCTIONALS*

DIMITRI P. BERTSEKAS† and SANJOY K. MITTER‡

Abstract. In this paper we consider the numerical solution of convex optimization problems with nondifferentiable cost functionals. We propose a new algorithm, the ε-subradient method, a large step, double iterative algorithm which converges rapidly under very general assumptions. We discuss the application of the algorithm in some problems of nonlinear programming and optimal control and we show that the ε-subgradient method contains as a special case a minimax algorithm due to Pshenichnyi [5].

1. General remarks. One of the most common approaches toward the numerical solution of optimization problems with or without constraints is the use of descent algorithms such as the steepest descent, conjugate gradient, quasi-Newton methods, methods of feasible directions, etc. These decent methods have enjoyed a great deal of popularity due to their reliability, simplicity, and good convergence properties. In their usual form all these algorithms require the existence of the gradient of the function to be minimized both for explicit use in the calculations and as a guarantee of their convergence to a local minimum. In many optimization problems, however, often arising in an economics framework, the natural cost functional of the problem turns out to be nondifferentiable. Such problems have received considerable attention recently and are the subject of this paper.

Early work on optimization problems with nondifferentiable cost functionals can be traced to the early sixties with the research of Dubovitskii and Milyutin [1], [2] which apparently served as a starting point for subsequent work of Soviet scientists [3]–[6]. At about the same time the theory of subdifferentiability of convex functions was developed by Moreau [7], [8], Rockafellar [9], [10], and Brøndsted and Rockafellar [11]. The notion of the subdifferential of a convex function (set of all supporting hyperplanes to the graph of the function) provided an efficient generalization of the notion of the ordinary gradient and formed the basis for the development of generalized necessary and sufficient conditions for optimality (see e.g. [10]). Necessary conditions which generalize the Pontryagin maximum principle of optimal control in very elegant form have been given by Neustadt [12], Heins and Mitter [13], and Rockafellar [14]. The latter reference contains also some generalizations of known results in the calculus of variations.

* Received by the editors March 27, 1972, and in revised form November 12, 1972.

† Department of Engineering-Economic Systems, Stanford University, Stanford, California 94305. The work of this author was supported by the National Science Foundation under Grant NSF-GR-29237.

‡ Decision and Control Sciences Group, Electronic Systems Laboratory, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139. The work of this author was supported by the Air Force Office of Scientific Research under Grant AFOSR 70-1941 and by NASA NGL-22-009-124.

Further necessary conditions for optimal control problems with nondifferentiable cost functionals were given by Luenberger [15]. Some additional results along the same lines can be found in the thesis by Ghanem [16]. Luenberger's results were somewhat generalized for the case of discrete-time systems using subdifferential theory by the authors [17]. Questions related to stochastic optimization problems with nondifferentiable cost functionals have been examined in [35], [36]. Such problems occur often in stochastic programming. A method for approximating a nondifferentiable convex function by a smooth function was also given in reference [35]. Necessary conditions for optimality for nonlinear, nonconvex programming problems without differentiability were obtained by Bazaraa, Goode and Shetty [18], [19] and for minimax problems by Danskin, Dem'yanov and Pschenichyni [20], [21], [5]. Among existing nonlinear programming algorithms, the convex cutting plane algorithm [25], [37] can be used for the solution of convex nondifferentiable optimization problems.

In the area of descent numerical methods a minimization algorithm has been reported by Ermol'ev [22], [23] and credited to Shor [24]. This algorithm is applicable to unconstrained convex programming problems with nondifferentiable cost. It reportedly has slow convergence properties [33] although computational examples using the algorithm are not available in the English literature. A similar algorithm has been proposed by Polyak [33]. Decent algorithms for the solution of minimax problems have been given by Dem'yanov [21], Pshenichnyi [5], Birzak and Pshenichnyi [26], and Levitin [34]. It should be noted that many optimization problems with nondifferentiable cost functionals can be converted into minimax problems. The generalization of the steepest descent method for the numerical solution of optimization problems with nondifferentiable cost functions was given by Luenberger [15]; however, a proof of convergence of this algorithm is not presently available. The problem appears to be that the algorithmic map in this algorithm is not closed (using Zangwill's terminology [25]). The ε-subgradient method, first presented in [17], circumvents this closure problem as will be seen in what follows. Other papers related to optimization problems with nondifferentiable cost functionals include those of Polyak [38], Minch [39], Auslender [40], [41], and Butz [42].

In this paper we present a new descent algorithm for constrained or unconstrained minimization problems where the cost function is convex but not necessarily differentiable. This algorithm, the ε-subgradient method, is a large step, double iterative algorithm that converges rapidly under very general assumptions. The algorithm was first presented in [17] and is based on the notion of the ε-subgradient of a convex function. In § 2 we describe the algorithm and we prove its convergence. In § 3 we consider some practical aspects of the algorithm and we demonstrate by means of examples its application. Finally, in § 4 we delineate some classes of problems for which the ε-subgradient method compares favorably with existing algorithms. In addition we show that the ε-subgradient method contains as a special case a minimax algorithm due to Pshenichnyi [5].

2. The ε-subgradient method. In this section we describe a descent algorithm for the minimization of a convex function subject to convex constraints. Rather than considering explicitly the constraints, however, we shall allow the function to be minimized to take the value $+\infty$. Thus the problem of finding the minimum of a function $g(\cdot)$ over a set X is equivalent to finding the minimum of the extended real-valued function $f(x) = g(x) + \delta(x|X)$, where $\delta(\cdot|X)$ is the indicator function of X, i.e., $\delta(x|X) = 0$ for $x \in X$, $\delta(x|X) = \infty$ for $x \notin X$. Stating the problem formally:

Find $\inf_x f(x)$ where $f: R^n \to (-\infty, +\infty]$ is a convex function which is lower semicontinuous with $\inf_x f(x) > -\infty$ and $f(x) < +\infty$ for at least one $x \in R^n$.

With the above assumptions, the function f is a closed proper convex function as defined in [10]. A detailed discussion of closed proper convex functions can be found in the same reference. A basic concept for the algorithm that we shall present is the notion of ε-subgradient. This notion was introduced in [9], [11] in connection with investigations related to the existence and characterization of subgradients of convex functions.

Let x be a point such that $f(x) < \infty$ and $\varepsilon > 0$ any positive scalar. A vector $x^* \in R^n$ is said to be an *ε-subgradient* of f at x if

$$(1) \qquad f(z) \geqq f(x) - \varepsilon + \langle z - x, x^* \rangle \quad \text{for all } z \in R^n,$$

where $\langle \cdot, \cdot \rangle$ denotes the usual inner product in R^n. The set $\partial_\varepsilon f(x)$ of all ε-subgradients at x will be called the *ε-subdifferential* of f at x. This set is nonempty, closed and convex. It is evident that for $0 < \varepsilon_1 < \varepsilon_2$ we have

$$\partial f(x) \subset \partial_{\varepsilon_1} f(x) \subset \partial_{\varepsilon_2} f(x).$$

A useful characterization of the set $\partial_\varepsilon f(x)$ is given by the equation [10, p. 220]

$$(2) \qquad \partial_\varepsilon f(x) = \{x^* | f^*(x^*) + f(x) - \langle x, x^* \rangle \leqq \varepsilon\},$$

where

$$(3) \qquad f^*(x^*) = \sup_x \{\langle x, x^* \rangle - f(x)\}$$

is the conjugate convex function of f [10]. The support function of $\partial_\varepsilon f(x)$ is given by the following useful equation [10, p. 220]:

$$(4) \qquad \sigma[y|\partial_\varepsilon f(x)] = \sup_{x^* \in \partial_\varepsilon f(x)} \langle y, x^* \rangle = \inf_{\lambda > 0} \frac{f(x + \lambda y) - f(x) + \varepsilon}{\lambda}.$$

The set $\partial_\varepsilon f(x)$ has some interesting properties from the algorithmic point of view as shown by the following two propositions.

PROPOSITION 1. *Let x be a vector such that $f(x) < \infty$. Then*

$$0 \leqq f(x) - \inf_z f(z) \leqq \varepsilon \leftrightarrow 0 \in \partial_\varepsilon f(x).$$

Proof. By the definition (1),

$$0 \in \partial_\varepsilon f(x) \leftrightarrow f(z) \geqq f(x) - \varepsilon \quad \text{for all } z \in R^n,$$

which is equivalent to the desired relations. Q.E.D.

PROPOSITION 2. *Let x be a point such that $f(x) < \infty$ and $0 \notin \partial_\varepsilon f(x)$. Let y be any vector such that*

$$(5) \qquad \sup_{x^* \in \partial_\varepsilon f(x)} \langle y, x^* \rangle < 0.$$

Then we have

$$f(x) - \inf_{\lambda \geqq 0} f(x + \lambda y) > \varepsilon. \tag{6}$$

Proof. Assume the contrary, i.e., $\inf_{\lambda \geqq 0} f(x + \lambda y) - f(x) + \varepsilon \geqq 0$. Then we have

$$\frac{f(x + \lambda y) - f(x) + \varepsilon}{\lambda} \geqq 0 \quad \text{for all } \lambda > 0.$$

This implies by using (4)

$$\sup_{x^* \in \partial_\varepsilon f(x)} \langle x^*, y\rangle = \inf_{\lambda > 0} \frac{f(x + \lambda y) - f(x) + \varepsilon}{\lambda} \geqq 0.$$

Since $\partial_\varepsilon f(x)$ is closed this implies that $0 \in \partial_\varepsilon f(x)$ which contradicts the hypothesis. Q.E.D.

In the case where $0 \notin \partial_\varepsilon f(x)$, a possible method for finding a vector $\bar{y}(x) \in R^n$ such that $\sup_{x^* \in \partial_\varepsilon f(x)} \langle x^*, \bar{y}(x)\rangle < 0$ is the following. Let $\|\cdot\|$ be the usual Euclidean norm in R^n and let $\bar{x}^*$ be the unique vector of minimum norm in $\partial_\varepsilon f(x)$. Then the vector

$$\bar{y}(x) = -(\bar{x}^*/\|\bar{x}^*\|) \tag{7}$$

satisfies $\sup_{x^* \in \partial_\varepsilon f(x)} \langle \bar{y}(x), x^*\rangle = -\|\bar{x}^*\| < 0$.

Propositions 1 and 2 form the basis for the algorithm that we shall present. The former provides a termination criterion for the algorithm. The latter states that whenever the value $f(x)$ exceeds the optimal value by more than ε, then by a descent along a vector y satisfying (5) we can decrease the value of the cost by at least ε. Consider the following algorithm.

ε-SUBGRADIENT METHOD.

Step 1. Select a vector x_0 such that $f(x_0) < \infty$, a scalar $\varepsilon_0 > 0$ and a scalar a, $0 < a < 1$.

Step 2. Given x_n and $\varepsilon_n > 0$, set $\varepsilon_{n+1} = a^k \varepsilon_n$, where k is the smallest nonnegative integer such that $0 \notin \partial_{\varepsilon_{n+1}} f(x_n)$.

Step 3. Find a vector y_n such that

$$\sup_{x^* \in \partial_{\varepsilon_{n+1}} f(x_n)} \langle y_n, x^*\rangle < 0. \tag{8}$$

Step 4. Set $x_{n+1} = x_n + \lambda_n y_n$, where $\lambda_n > 0$ is such that

$$f(x_n) - f(x_{n+1}) > \varepsilon_{n+1}.$$

Return to Step 2.

It should be mentioned that if x_n is not a minimizing point of f there always exists a nonnegative integer k such that $0 \notin \partial_{a^k \varepsilon_n} f(x_n)$, since by Proposition 1 we have

$$0 \notin \partial_{a^k \varepsilon_n} f(x_n) \leftrightarrow f(x_n) - \inf_x f(x) > a^k \varepsilon_n .$$

Also by Proposition 2 there exists a scalar λ_n such that

$$f(x_n) - f(x_n + \lambda_n x_n) > \varepsilon_{n+1}, \tag{9}$$

thus showing that Step 4 can always be carried out. In fact, one can show that the set of all scalars λ_n satisfying (9) is an open bounded interval or an open half-line. One way of finding a scalar λ_n satisfying (9) is by means of the one-dimensional minimization

$$f(x_n + \lambda_n y_n) = \min_{\lambda > 0} f(x_n + \lambda y_n),$$

assuming the minimum is attained. This in turn can be guaranteed whenever the set of minimizing points of f is nonempty and compact, since in this case all the level sets of f are compact [10, Cor. 8.7.1]. We note also that Steps 2 and 3 can be carried out by means of an auxiliary minimization problem as will be discussed in detail in the next section.

We now prove the convergence of the ε-subgradient method.

PROPOSITION 3. *Consider the vectors x_n generated by the ε-subgradient method. Then either $f(x_m) = \min_x f(x)$ for some $m \geqq 0$ or the generated infinite sequence $\{x_n\}$ satisfies*

(a) $\lim_{n \to \infty} f(x_n) = \inf_x f(x)$.

If, in addition, the set $M = \{\bar{x} | f(\bar{x}) = \min_x f(x)\}$ is nonempty and bounded, then:

(b) *Every convergent subsequence of $\{x_n\}$ has its limit in M, and at least one such subsequence exists.*

(c) *For every $\varepsilon > 0$ there exists an $m \geqq 0$ such that $x_n \in M + \varepsilon B$ for all $n \geqq m$, where $B = \{x | \|x\| \leqq 1\}$ is the unit ball in R^n.*

(d) *If the minimum of f is attained at a single point $\bar{x}$ then $\{x_n\} \to \bar{x}$.*

Proof. By Proposition 2 we have

$$f(x_n) - f(x_{n+1}) > \varepsilon_{n+1} \quad \text{for all } n \geqq 0$$

and hence,

$$f(x_0) - \sum_{i=1}^{n} \varepsilon_i > f(x_n) > \inf_x f(x) \quad \text{for all } n \geqq 1 .$$

Since $\varepsilon_i > 0$ the above inequality implies $\{\varepsilon_i\} \to 0$. This implies that $\varepsilon_{i+1} < \varepsilon_i$ for an infinite number of integers i. In view of Step 2 of the algorithm we have for those integers: $0 < f(x_i) - \inf_x f(x) \leqq \varepsilon_i$. Since $\{f(x_n)\}$ is a decreasing sequence, it follows that $\lim_{n \to \infty} f(x_n) = \inf_x f(x)$, and (a) is proved. To prove (b) notice that $x_n \in F_0$, where $F_0 = \{x | f(x) \leqq f(x_0)\}$ and since M is nonempty and bounded, F_0 is compact (see [10, Cor. 8.7.1]). Therefore the sequence $\{x_n\}$ has at least one convergent subsequence. The fact that the limits of all convergent subsequences belong to M follows from (a) and Cor. 27.2.1 in [10]. Part (c) follows from (a) and Thm. 27.2 in [10]. Part (d) follows from (a) and Cor. 27.2.2 in [10]. Q.E.D.

The above proposition establishes that the ε-subgradient method has attractive convergence properties. In fact, it converges to the optimal value even if an optimal solution does not exist. A further attractive feature of the method is that it guarantees substantial progress at every iteration (Step 4) and that the progress of the computation is monitored constantly via the parameter ε (Step 2). The price for this substantial progress is the computations necessary to find the direction of descent in Steps 2 and 3. In the next section we shall describe some practical aspects of the algorithm and demonstrate by means of examples its application.

3. Practical aspects of the ε-subgradient method. A cursory examination of the ε-subgradient method reveals that in fact the most difficult step in a single iteration is finding the direction of descent y_n. However, contrary to most descent algorithms, the chosen direction of descent in the ε-subgradient method can lead to guaranteed substantial reduction of the value of the cost functional in a single iteration. To demonstrate this fact consider the following lemma.

LEMMA. *Assume that the scalars ε_0 and a in the ε-subgradient method are such that*

$$f(x_0) - \inf_x f(x) \leqq \varepsilon_0, \qquad 1/2 \leqq a < 1. \tag{10}$$

Then for all $n \geqq 1$,

$$f(x_n) - \inf_x f(x) < ((1 - a)/a)\varepsilon_n \leqq (1 - a)\varepsilon_{n-1}. \tag{11}$$

Proof. We have $f(x_0) - \inf_x f(x) \leqq \varepsilon_0$ implying that $0 \in \partial_{\varepsilon_0}(x_0)$. Hence in Step 2 we have $\varepsilon_1 \neq \varepsilon_0$. This in turn implies that $0 \in \partial_{\varepsilon_1/a} f(x_0)$, or equivalently,

$$f(x_0) - \inf_x f(x) \leqq \varepsilon_1/a.$$

On the other hand,

$$f(x_0) - f(x_1) > \varepsilon_1.$$

Combining the last two inequalities we have

$$f(x_1) - \inf_x f(x) < ((1 - a)/a)\varepsilon_1,$$

proving (11) for $n = 1$. Since $1/2 \leqq a < 1$, the last inequality implies that $f(x_1) - \inf_x f(x) < \varepsilon_1$ and the same argument as above can be used to prove (11) for $n = 2$ and every n. Q.E.D.

It is evident now from (11) that a substantial reduction of the value of the cost functional is possible by choosing the value of the parameter a high enough. On the other hand, a value of the parameter a close to unity leads to an increased number of iterations in order to find the scalar ε_{n+1} from ε_n in Step 2 of the algorithm. Thus, in practice, one must settle on a compromise value for the parameter a depending on how difficult it is to carry out a single check $0 \in \partial_{a^k\varepsilon_n} f(x_n)$ in Step 2. Another possibility is to modify the algorithm so that the value of the parameter a is adjusted during the iterations in Step 2 on the basis of information already obtained. A number of convergent schemes are possible. We do not discuss these schemes since they are not theoretically interesting but rather relate to the intelligent programming of the method.

We now turn to the important question of how the calculation of the direction of descent is to be carried out once the value of the parameter a is selected. As mentioned in the previous section it is possible to carry out Steps 2 and 3 of the algorithm by solving the following minimization problem:

$$\min_{x^* \in \partial_{a^k \varepsilon_n} f(x_n)} \|x^*\| . \tag{12}$$

Now clearly we have $0 \in \partial_{a^k \varepsilon_n} f(x_n)$ if and only if problem (12) has a zero optimal value and therefore Step 2 of the algorithm can be carried out by solving problem (12) successively for $k = 0, 1, \cdots$. There exists an integer k for which problem (12) has a nonzero optimal value. Let $\bar{x}^*$ be the optimal solution of problem (12) for the first such integer k. Then a suitable direction of descent y_n satisfying (8) in Step 3 of the algorithm is given by

$$y_n = - \bar{x}^*/\|\bar{x}^*\| . \tag{13}$$

One efficient method for solving the minimization problem (12) is to solve successively the unconstrained problem

$$\min_{x^*} \{ \|x^*\|^2 + P_k(x^*) \} , \tag{14}$$

where $P_k(\cdot)$ is a (moderate) penalty function

$$\begin{gathered} P_k(x^*) \geqq 0 \quad \text{for all } x^*, \\ P_k(x^*) = 0 \quad \text{if and only if } x^* \in \partial_{a^k \varepsilon_n} f(x_n). \end{gathered} \tag{15}$$

It is clear that problem (14) has a zero optimal value if and only if problem (12) has a zero optimal value. Furthermore, when k is such that problem (12) has a nonzero value, problem (14) yields an approximate solution $\tilde{x}^*$ to problem (12). In this case one can either increase the penalty and obtain a more accurate solution or obtain an approximate direction of descent $\tilde{y}_n$ from

$$\tilde{y}_n = - \tilde{x}^*/\|\tilde{x}^*\| .$$

The approximate direction $\tilde{y}_n$ is considered acceptable if it yields a point x_{n+1} satisfying $f(x_n) - f(x_{n+1}) > \varepsilon_{n+1}$ in Step 4. If $\tilde{y}_n$ is not acceptable we increase the penalty in problem (14) and resolve the problem in order to obtain a more accurate direction of descent.

The preceding discussion clearly demonstrates that the application of the ε-subgradient method to a specific problem requires the solution of minimization problems of the form

$$\min_{x^* \in \partial_\varepsilon f(x)} \|x^*\| . \tag{16}$$

At first sight it would therefore appear that the ε-subgradient method can be applied only to the limited class of functions for which the ε-subdifferential $\partial_\varepsilon f(x)$ has a convenient characterization. We shall demonstrate in what follows in this section that this is not the case and, in fact, the method can be applied to most functions likely to be encountered in practice. This is due to the fact that problem (16) can be cast into the usual nonlinear programming framework even if a convenient closed form characterization of the set $\partial_\varepsilon f(x)$ is not available.

By making use of the characterization (2) of the ε-subdifferential $\partial_\varepsilon f(x)$ in terms of the conjugate convex function f^*, problem (16) can be written as

$$\text{minimize } \|x^*\| \tag{17}$$

subject to

$$f^*(x^*) + f(x) - \langle x, x^* \rangle \leqq \varepsilon.$$

Now there is a class of simple functions f for which the conjugate

$$f^*(x^*) = \sup_x \{\langle x, x^* \rangle - f(x)\}$$

has a convenient closed form. Such functions include:

(a) Positively homogeneous closed convex functions, i.e., support functions of given sets [10, § 13]. Thus if

$$f(x) = \sigma(x|X) = \sup_{x^* \in X} \langle x, x^* \rangle,$$

then

$$f^*(x^*) = \delta(x^*|\bar{X}) = \begin{cases} 0 & \text{if } x^* \in \bar{X}, \\ \infty & \text{if } x^* \notin \bar{X}, \end{cases}$$

where $\bar{X}$ is the closure of the convex hull of X. This class includes all norms and seminorms in R^n as well as linear functions. In addition, the conjugates of powers greater than one of norms and seminorms in R^n (including quadratic forms) are given in [10, § 15].

(b) Exponentials and logarithms of coordinates of x (see [10, § 12]).

(c) Indicator functions of affine sets (linear manifolds), convex cones and unit balls with respect to a norm or a seminorm [10, § 13].

(d) Indicator functions of sets with known support functions, [10, § 13]. If X is a closed convex set and

$$f(x) = \delta(x|X),$$

then

$$f^*(x^*) = \sigma(x^*|X) = \sup_{x \in X} \langle x, x^* \rangle.$$

We note that constraint sets which are characterized by their support function are encountered, for example, in some optimal control problems as will be discussed in some detail in § 4.

Now from this class of simple functions one can build more complicated functions by means of various operations such as summation, affine transformation, maximization, etc. The conjugates of such functions are characterized by the following well-known relations:

$$(f_1 + f_2 + \cdots + f_m)^*(x^*) = \min_{\sum_{i=1}^m x_i^* = x^*} \left\{ \sum_{i=1}^m f_i^*(x_i^*) \right\} \quad ([10, \text{Thm. } 16.4]), \tag{18}$$

where f_i, $i = 1, \cdots, m$, are closed proper convex functions with a common point in the relative interior of their effective domain, and the function $f_1 + \cdots + f_m$ is defined by

$$(f_1 + f_2 + \cdots + f_m)(x) = f_1(x) + f_2(x) + \cdots + f_m(x),$$

$$(f \cdot A)^*(x^*) = \min_{A^*y^* = x^*} f^*(y^*), \tag{19}$$

where $f : R^m \to R_e$ is a closed proper convex function, A is a linear transformation from R^n to R^m, A^* denotes its adjoint, the function $f \cdot A$ is the composition of f and A, and, in addition, the range of A contains a point in the relative interior of the effective domain of f.

$$(\max\{f_1, \cdots, f_m\})^*(x^*) = \min_{\substack{x^* = \sum_{i=1}^m \lambda_i x_i^* \\ \lambda_i \geqq 0 \\ \sum_{i=1}^m \lambda_i = 1}} \left\{ \sum_{i=1}^m \lambda_i f_i^*(x_i^*) \right\} \quad \text{([10, Thm. 16.5])}, \tag{20}$$

where f_i, $i = 1, \cdots, m$, are convex real-valued functions and the function $\max\{f_1, \cdots, f_m\}$ is defined by

$$(\max\{f_1, \cdots, f_m\})(x) = \max\{f_1(x), \cdots, f_m(x)\},$$

$$g^*(x^*) = f^*(x^*) + \langle c, x^* \rangle, \tag{21}$$

where $g(x) = f(x - c), f = R^n \to (-\infty, +\infty]$ is a closed proper convex function and $c \in R^n$ is a given vector.

The equations (18)–(21) can be used in order to put the minimization problem (17) in the standard nonlinear programming framework for a wide variety of functions. As an illustration, consider the case where the function f to be minimized by means of the ε-subgradient method has the form

$$f(x) = f_1(x) + f_2(x) + \cdots + f_m(x).$$

By using (18) the optimization problem (17) can be written as

$$\text{minimize } \|x^*\|$$

subject to

$$\min_{\sum_{i=1}^m x_i^* = x^*} \left\{ \sum_{i=1}^m f_i^*(x_i^*) \right\} + f(x) - \langle x^*, x \rangle \leqq \varepsilon.$$

It can be easily seen that the above problem is equivalent to

$$\text{minimize } \left\| \sum_{i=1}^m x_i^* \right\|$$

subject to

$$\sum_{i=1}^m f_i^*(x_i^*) + f(x) - \sum_{i=1}^m \langle x_i^*, x \rangle \leqq \varepsilon.$$

This latter problem is in the standard nonlinear programming framework whenever the functions f_i belong to the class of simple functions mentioned earlier. As another example consider the case where the function f has the form

$$f(x) = \max\{f_1(A_1x), \cdots, f_m(A_mx)\},$$

where $A_1, \cdots, A_m$ are linear transformations and $f_1, \cdots, f_m$ are real-valued convex functions. By using (19), (20), the optimization problem (17) for this function can be written as

$$\text{minimize } \|x^*\|$$

subject to

$$\min_{\substack{x^* = \sum_{i=1}^{m} \lambda_i x_i^* \\ \lambda_i \geqq 0 \\ \sum_{i=1}^{m} \lambda_i = 1}} \left\{ \sum_{i=1}^{m} \lambda_i \min_{A_i^* y_i^* = x_i^*} f_i^*(y_i^*) \right\} + f(x) - \langle x^*, x \rangle \leqq \varepsilon,$$

or equivalently,

$$\text{minimize } \left\| \sum_{i=1}^{m} \lambda_i A_i^* y_i^* \right\|$$

subject to

$$\sum_{i=1}^{m} \lambda_i f_i^*(y_i^*) + f(x) - \sum_{i=1}^{m} \lambda_i \langle A_i^* y_i^*, x \rangle \leqq \varepsilon,$$

$$\lambda_i \geqq 0, \quad \sum_{i=1}^{m} \lambda_i = 1.$$

Similarly, one can write the optimization problem (17) in standard form whenever the function to be minimized involves simultaneously sums, compositions with linear transformations and maxima of the basic simple functions referred to earlier. Thus the ε-subgradient method can be applied for the minimization of a wide class of functions. This class of functions can be further enlarged by making use of the following technique to eliminate some of the constraints of the minimization problem.

Consider the convex programming problem

$$\text{minimize } f_0(x) \tag{22}$$

subject to

$$x \in X, \quad f_i(x) \leqq 0, \qquad i = 1, \cdots, m,$$

where $f_0, f_1, \cdots, f_m$ are real-valued convex functions and X is a closed convex set. Let $\bar{x}$ be an optimal solution of this problem and assume that there exists a point $\tilde{x} \in X$ such that $f_i(\tilde{x}) < 0$, $i = 1, \cdots, m$. Then there exist nonnegative Lagrange multipliers, $\lambda_1, \cdots, \lambda_m$, corresponding to $\bar{x}$ [25], [37] such that $\bar{x}$ minimizes

$$f_0(x) + \sum_{i=1}^{m} \lambda_i f_i(x)$$

subject to $x \in X$. Furthermore, it is known [15] that if k is a scalar such that

$$k > \max\{\lambda_1, \cdots, \lambda_m\} \tag{23}$$

then $\bar{x}$ is an optimal solution to the problem

$$\text{minimize } f_0(x) + k \sum_{i=1}^{m} \max\,[0, f_i(x)] \tag{24}$$

subject to $x \in X$.

Conversely, every optimal solution of problem (24) is an optimal solution of problem (22) so that the two problems are equivalent and either one of the two can be solved in place of the other. Concerning the selection of the scalar k, it can be easily proved that if μ is a strict lower bound for the optimal value of problem (22), then

$$k = \max\left\{\frac{f_0(\tilde{x}) - \mu}{-f_1(\tilde{x})}, \cdots, \frac{f_0(\tilde{x}) - \mu}{-f_m(\tilde{x})}\right\}$$

satisfies (23), where $\tilde{x}$ is a vector such that $\tilde{x} \in X$ and $f_i(\tilde{x}) < 0$, $i = 1, 2, \cdots, m$.

We shall close this section by showing explicitly the form of the auxiliary minimization problem (17) for a specific problem.

Example. Consider the problem

$$\text{minimize}\left\{\max_{\substack{y_i \geqq 0 \\ \|y\|^- \leqq 1}} \langle x, y\rangle + \max\,[0, \tfrac{1}{2}x'Qx + \langle c, x\rangle]\right\}$$

subject to $x \in X = \{x | x_i \geqq 0, i = 1, \cdots, m\}$.

In the above problem, x, y are vectors in R^n, $\|\cdot\|$ denotes the Euclidean norm in R^n, Q is a positive definite matrix and c is a given vector. By defining

$$f_1(x) = \max_{\substack{y_i \geqq 0 \\ \|y\|^- \leqq 1}} \langle x, y\rangle,$$

$$f_2(x) = (1/2)x'Qx + \langle c, x\rangle,$$

$$f_3(x) = \delta(x|X) = \begin{cases} 0 & \text{if } x \in X, \\ \infty & \text{if } x \notin X, \end{cases}$$

the problem is written

$$\text{minimize } f(x) = f_1(x) + \max\,[0, f_2(x)] + f_3(x).$$

The auxiliary optimization problem to be solved in Steps 2 and 3 of the ε-subgradient method is

$$\text{minimize } \|x^*\|$$

subject to $f^*(x^*) + f(x) - \langle x^*, x\rangle \leqq a^k\varepsilon$. By using (18) and (20) and the fact that the conjugate of the zero function is the function

$$(0)^*(x^*) = \begin{cases} 0 & \text{if } x^* = 0, \\ \infty & \text{if } x^* \neq 0, \end{cases}$$

the above problem is equivalent to

$$\text{minimize } \|x_1^* + \lambda x_2^* + x_3^*\|^2 \tag{25}$$

subject to

$$f_1^*(x_1^*) + \lambda f_2^*(x_2^*) + f_3^*(x_3^*) + f(x) - \langle x_1^* + \lambda x_2^* + x_3^*, x\rangle \leqq a^k\varepsilon, \qquad 0 \leqq \lambda \leqq 1.$$

We have

$$f_1^*(x_1^*) = \begin{cases} 0 & \text{if } x_1^* \geqq 0, \quad \|x_1^*\| \leqq 1, \\ \infty & \text{otherwise,} \end{cases}$$

$$f_2^*(x_2^*) = \tfrac{1}{2}(x_2^* - c)'Q^{-1}(x_2^* - c),$$

$$f_3^*(x_3^*) = \begin{cases} 0 & \text{if } x_3^* \leqq 0, \\ \infty & \text{otherwise,} \end{cases}$$

where the inequalities $x_1^* \geqq 0$, $x_3^* \leqq 0$ are interpreted to be componentwise. Hence problem (25) takes the form

$$\text{minimize } \|x_1^* + \lambda x_2^* + x_3^*\|^2$$

subject to

$$(\lambda/2)(x_2^* - c)'Q^{-1}(x_2^* - c) + f(x) - \langle x_1^* + \lambda x_2^* + x_3^*, x\rangle \leqq a^k\varepsilon,$$

$$0 \leqq x_1^*, \quad \|x_1^*\| \leqq 1, \quad x_3^* \leqq 0, \quad 0 \leqq \lambda \leqq 1,$$

a nonlinear program with linear and quadratic constraints. If $(\bar{x}_1^*, \bar{x}_2^*, \bar{x}_3^*, \bar{\lambda})$ is an optimal solution of the above problem then $\bar{x}^* = \bar{x}_1^* + \bar{\lambda}\bar{x}_2^* + \bar{x}_3^*$ is an optimal solution of the auxiliary optimization problem of Steps 2 and 3 of the ε-subgradient method.

4. Applications. In this section we attempt to delineate some classes of problems for which the ε-subgradient method compares favorably with existing methods. It is well known that many optimization problems with nondifferentiable cost functionals can be converted into nonlinear programming problems where all functions involved are differentiable. For example consider the problem

$$\text{(26)} \qquad \text{minimize } \max\{f_1(x), \cdots, f_m(x)\},$$

where the functions f_i are convex and differentiable. This problem is equivalent to the problem

$$\text{(27)} \qquad \text{minimize } y$$

subject to

$$f_i(x) \leqq y, \qquad i = 1, \cdots, m,$$

where y is a scalar auxiliary variable. This latter problem can be solved by any of the existing algorithms for differentiable functions such as, for instance, the ε-perturbation feasible direction method [25]. Also problem (26) can be solved by using Dem'yanov's minimax algorithm [21] which is closely related to the feasible direction method mentioned above. It appears that either one of the two algorithms is preferable to the ε-subgradient method for the solution of problem (26). This is due to the considerable computation necessary in order to find the direction of descent in the ε-subgradient method. More generally, one can say that if the optimization problem can be converted to a nonlinear program where all functions involved are differentiable, standard methods should, in most cases, be preferable over the ε-subgradient method.

The ε-subgradient method, however, should be considered advantageous when applied to problems which cannot be converted to nonlinear programming problems involving differentiable functions since it has the advantage of fast convergence. One class of such problems is characterized by the presence of terms of the form $\max_{y \in Y} \langle x, y \rangle$ either in the cost function or the constraints. The first known algorithm involving functions of the form $\max_{y \in Y} \langle x, y \rangle$ is the one of Pshenichnyi [5] who considered the problem

$$\text{minimize} \max_{y \in Y} \langle x, y \rangle \tag{28}$$

subject to

$$x \in A,$$

where Y is a convex compact set and A is a given hyperplane. When the ε-subgradient method is applied to problem (26), the direction of descent is determined by solving the auxiliary optimization problem

$$\text{minimize } \|x_1^* + x_2^*\|$$

subject to

$$x_1^* \in Y, \quad \max_{y \in Y} \langle x, y \rangle - \langle x_1^*, x \rangle \leqq \varepsilon,$$

$$x_2^* \in A^{\perp},$$

where $A^{\perp}$ is the one-dimensional subspace orthogonal to the hyperplane A. This is exactly the same optimization problem by means of which the direction of descent is determined in Pshenichnyi's method and thus the ε-subgradient method and Pshenichnyi's method are identical when applied to problem (28). The ε-subgradient method, however, can be applied to much more general problems involving terms of the form $\max_{y \in Y} \langle x, y \rangle$. One such example was given in the previous section. For such problems the ε-subgradient method compares favorably with, for example, Dem'yanov's minimax algorithm which involves comparable computations for finding the direction of descent but does not converge as fast as the ε-subgradient method.

The ε-subgradient method can also be used effectively for problems where some of the constraint sets are not given explicitly but instead can be specified from their support function. For such problems methods of feasible directions, for example, are not applicable. As an example, consider the following optimal control problem where some of the constraint sets are characterized as reachable sets of a differential system.

Consider the linear system

$$\dot{x}(t) = A(t)x(t) + B(t)u(t) \tag{29}$$

over the time interval $[t_0, T]$ which is controllable from t_0 to T and where $A(t)$ is a Lebesgue integrable $n \times n$ matrix, and $B(t)$ is a continuous $n \times m$ matrix function on $[t_0, T]$. The m-vector-valued function $u(t)$ is assumed to be measurable in $[t_0, T]$ and such that

$$u(t) \in U \quad \text{almost everywhere in } [t_0, T], \tag{30}$$

where U is a nonempty compact subset of R^n. Assume further that the initial condition is constrained to lie in X_0, a convex compact subset of R^n:

$$x(t_0) \in X_0. \tag{31}$$

Consider the problem of minimizing

$$J[x(t_0), u] = F[x(T)], \tag{32}$$

where F is a closed proper convex function in R^n subject to the constraints (29)–(31).

Then under our assumptions, for every pair $(x(t_0), u)$ satisfying (30) and (31), there exists a unique absolutely continuous solution of (29). The set $X(T)$ of reachable states $x(T)$ at time T corresponding to the constraints (30), (31) is convex and compact by a theorem of Neustadt [30], and its support function is given by ([31], [32])

$$\sigma[x^*|X(T)] = \sigma[\Phi'(t_0, T)x^*|X_0] + \int_{t_0}^{T} \sigma[B'(t)\Phi'(t, T)x^*|U]\, dt,$$

where $\Phi(t, \tau)$ is the unique absolutely continuous transition matrix corresponding to the matrix $A(t)$.

The problem can now be recast as one of minimizing the extended real-valued convex function

$$f[x(T)] = F[x(T)] + \delta[x(T)|X(T)]$$

and the ε-subgradient method can be used for its solution. The direction of descent is determined by solving the optimization problem

$$\text{minimize } \|x_1^* + x_2^*\|$$

subject to

$$F^*(x_1^*) + \sigma[x_2^*|X(T)] + F[x(T)] - \langle x_1^* + x_2^*, x(T)\rangle \leqq \varepsilon.$$

For the problem that we consider there is some difficulty associated with the one-dimensional line search in Step 4 of the ε-subgradient method since it is not easy to check feasibility of any given terminal state. This difficulty can be circumvented by finding a point along the direction of descent such that the value of the function F has decreased by ε or a little less. It can be easily seen that such a point is feasible and that the algorithm will still be convergent.

5. Conclusions. The ε-subgradient method is a descent algorithm which can solve efficiently some convex minimization problems with nondifferentiable cost functionals which cannot be solved by standard nonlinear programming methods. It converges fast under very general assumptions but requires the solution of an auxiliary optimization problem in order to determine the direction of descent at each iteration. Presently, we do not have any computational experience with the method. It is hoped that such computational experience will be gained in the near future.

REFERENCES

[1] A. Y. Dubovitskii and A. A. Milyutin, *Extremum problems in the presence of constraints*, Dokl. Akad. Nauk SSSR, 149 (1963), pp. 452–455; English transl., Soviet Math Dokl., 4 (1963), pp. 452–455.

[2] ———, *Extremum problems in the presence of constraints*, Zh. Vychisl. Mat. i Mat. Fiz., 5 (1965), pp. 395–453.

[3] V. F. Dem'yanov and A. M. Rubinov, *Minimization of functionals in normed spaces*, this Journal, 6 (1968), pp. 73–89.

[4] ———, *Approximate Methods in Optimization Problems*, American Elsevier, New York, 1970.

[5] B. N. Pshenichnyi, *Dual methods in extremum problems*, Kibernetika, 1 (1965), no. 3, pp. 89–95.

[6] ———, *Convex programming in a normed space*, Ibid., 1 (1965), no. 5, pp. 46–54.

[7] J. J. Moreau, *Fonctionelles sous-differentiables*, C. R. Acad. Sci. Paris, 257 (1963), pp. 4117–4119.

[8] ———, *Semi-continuité de sous-gradient d'une fonctionelle*, Ibid., 360 (1965), pp. 1057–1070.

[9] R. T. Rockafellar, *Characterization of the subdifferentials of convex functions*, Pacific J. Math., 17 (1966), pp. 497–510.

[10] ———, *Convex Analysis*, Princeton University Press, Princeton, N.J., 1970.

[11] A. Brøndsted and R. T. Rockafellar, *On the subdifferentiability of convex functions*, Proc. Amer. Math. Soc., 16 (1965), pp. 605–611.

[12] L. W. Neustadt, *A general theory of extremals*, J. Comput. System Sci., 3 (1969), pp. 57–92.

[13] W. Heins and S. K. Mitter, *Conjugate convex functions, duality, and optimal control problems, I. Systems governed by ordinary differential equations*, Information Sci., 2 (1970), pp. 211–243.

[14] R. T. Rockafellar, *Conjugate convex functions in optimal control and the calculus of variations*, J. Math. Anal. Appl., 32 (1970), pp. 174–222.

[15] D. G. Luenberger, *Control problems with kinks*, IEEE Trans. Automatic Control, AC-15 (1970), pp. 570–575.

[16] M. Z. E. Ghanem, *Optimal control problems with nondifferentiable cost functionals*, Ph.D. dissertation, Dept. of Engineering–Economic Systems, Stanford University, Stanford, Calif., 1970.

[17] D. P. Bertsekas and S. K. Mitter, *Steepest descent for optimization problems with nondifferentiable cost functionals*, Proc. 5th Annual Princeton Conference on Information Sciences and Systems, Princeton, N.J., 1971.

[18] M. S. Bazaraa, J. J. Goode and C. M. Shetty, *Optimality criteria in nonlinear programming without differentiability*, Operations Res., 19 (1971), pp. 77–86.

[19] M. S. Bazaraa, *Nonlinear programming: nondifferentiable functions*, Ph.D. dissertation, Georgia Institute of Technology, Atlanta, Ga., 1969.

[20] J. M. Danskin, *The theory of max-min with applications*, SIAM J. Appl. Math., 14 (1966), pp. 641–664.

[21] V. F. Dem'yanov, *The solution of several minimax problems*, Kibernetika, 2 (1966), no. 6, pp. 58–66.

[22] Y. M. Ermol'ev, *Methods of solution of nonlinear extremal problems*, Ibid., 2 (1966), no. 4, pp. 1–17.

[23] Y. M. Ermol'ev and N. Z. Shor, *On the minimization of nondifferentiable cost functions*, Ibid., 3 (1967), no. 1, pp. 101–102.

[24] N. Z. Shor, *On the structure of algorithms for the numerical solution of problems of optimal programming and design*, Dissertation, Kiev, 1964.

[25] W. I. Zangwill, *Nonlinear Programming: A Unified Approach*, Prentice-Hall, Englewood Cliffs, N.J., 1969.

[26] B. Birzak and B. N. Pshenichnyi, *Some problems of the minimization of unsmooth functions*, Kibernetika, 2 (1966), no. 6, pp. 43–46.

[27] L. V. Kantorovich and K. P. Akilov, *Functional Analysis in Normed Spaces*, Pergamon Press, New York, 1965, Chap. 15.

[28] D. G. Luenberger, *Optimization by Vector Space Methods*, John Wiley, New York, 1969.

[29] D. P. Bertsekas, *Control of uncertain systems with a set-membership description of the uncertainty*, Ph.D. thesis, Dept. of Electrical Engineering, Mass. Inst. of Technology, Cambridge, Mass., 1971.

[30] L. W. NEUSTADT, *The existence of optimal controls in the absence of convexity conditions*, J. Math. Anal. Appl., 7 (1963), pp. 110–117.

[31] H. S. WITSENHAUSEN, *Minimax control of uncertain systems*, M.I.T. Electronics Systems Lab. Rep. ESL-R-269, Cambridge, Mass., 1966.

[32] ———, *A minimax control problem for sampled linear systems*, IEEE Trans. Automatic Control, AC-13 (1968), pp. 5–21.

[33] B. T. POLYAK, *Minimization of unsmooth functionals*, Zh. Vychisl. Mat. i Mat. Fiz., 9 (1969), no. 3, pp. 509–521.

[34] E. S. LEVITIN, *A general minimization method for unsmooth extremal problems*, Ibid., 9 (1969), no. 4, pp. 783–806.

[35] D. P. BERTSEKAS, *Stochastic optimization problems with nondifferentiable cost functionals*, J. Optimization Theory Appl., Aug., 1973.

[36] ———, *Stochastic optimization problems with nondifferentiable cost functionals with an application in stochastic programming*, Proc. 1972 Conference on Decision and Control, New Orleans, La., 1972.

[37] D. G. LUENBERGER, *Introduction to Linear and Nonlinear Programming*, Addison-Wesley, Reading, Mass., 1973.

[38] B. T. POLYAK, *A general method for solving extremal problems*, Dokl. Akad. Nauk SSSR, 174 (1967), no. 1, pp. 33–36.

[39] R. A. MINCH, *Applications of symmetric derivatives in mathematical programming*, Math. Programming, 1 (1971), pp. 307–321.

[40] A. AUSLENDER, *Méthodes numeriques pour la décomposition et la minimization de fonctions non différentiables*, Numer. Math., 18(1971), pp. 213–223.

[41] ———, *Recherche de points de selle d'une fonction*, Cahiers Centre Etudes Recherche Opér., 12 (1970), no. 2.

[42] A. BUTZ, *Iterative saddle point techniques*, SIAM J. Appl. Math., 15 (1967), pp. 719–726.

SIAM J. Control
Vol. 11, No. 4, November 1973

NECESSARY CONDITIONS FOR MULTIPLE CONSTRAINT OPTIMIZATION PROBLEMS*

M. L. J. HAUTUS†

Abstract. In this paper a necessary condition is given for a real-valued function f to attain a maximum at a point b subject to the condition $x \in S$, where S is given as an intersection of a finite number of sets in an n-dimensional Euclidean space. It is shown that well-known necessary conditions in mathematical programming, like the Lagrange multipliers theorem and results of F. John, Mangasarian and Fromovitz, are immediate consequences of this general condition. The result is also used to derive a general necessary condition for discrete-time optimal control problems, which contains the results of Halkin (discrete maximum principle), Jordan and Polak, and Canon, Cullum and Polak as special cases. As a final application of the necessary condition a simple proof of the Pontryagin maximum principle for continuous-time control problems is given.

Introduction. In optimization problems the constraint set S often is an intersection of a finite number of sets $S_1, \cdots, S_k$, each of which is of a fairly simple nature, whereas the intersection is very complicated. It is therefore desirable in those cases to give necessary conditions for optimality in terms of $S_1, \cdots, S_k$.

In this paper the concept of derived cone will be used to express such a necessary condition. A derived cone is a kind of first order approximation of the constraint set in the neighborhood of an optimal point. It is a simple matter to give a necessary condition for optimality if a derived cone of the set S is given. We are, however, interested in the situation, where only derived cones for the sets $S_1, \cdots, S_k$ are available. This will often be the case if the sets $S_1, \cdots, S_k$ have a simple character and S is of a complicated nature.

In § 1 we introduce some notation, we give the definition of a derived cone and we state our main result, Theorem 1.5. This result will be proved in § 2. In order to utilize the theorem, one must know derived cones for certain sets. For convex sets and sets given by a single equality or inequality relation, such a derived cone is easily found. This will be shown in § 3. In the same section we introduce substitution rules, which facilitate the use of the main theorem, in particular to situations where constraint sets occur given by multiple equalities and inequalities. In § 4 the theory is applied to discrete-time optimal control problems. A general type of necessary condition is shown to be an easy consequence of Theorem 1.5. This necessary condition in its turn contains well-known necessary conditions (discrete maximum principle [4], [3, § 4.3], conditions by Jordan and Polak [10], Canon, Cullum and Polak [2]) as special cases. In § 5 we give a simple proof of Pontryagin's maximum principle with a general type of initial and final constraint. In Appendix B a generalization of Theorem 1.5 is given.

1. Statement of the main result.

We introduce some notations. The n-dimensional Euclidean space of column (row) vectors is denoted R^n (R_n). The set of vectors $\tau \in R^k$, satisfying $\tau_i \geqq 0, i = 1, \cdots, k$ (notation $\tau \geqq 0$), will be denoted by R^k_+. The interior of a set S is denoted int S, the relative interior of a convex set S by ri S. Various properties of the relative interior given in [18, pp. 43–50] will be

* Received by the editors May 31, 1972.

† Department of Mathematics, Technological University Eindhoven, Eindhoven, The Netherlands.

used without further mentioning. The transpose of a vector $b \in R^n$ is denoted by b' (hence $b' \in R_n$). If $b_1, \cdots, b_k$ are vectors in R^n, then $(b_1; \cdots; b_k) := (b'_1, \cdots, b'_k)'$. Note that this is an element of R^{nk}. Furthermore, if S, $T \subseteq R^n$, $b \in R^n$, and A is a linear map, we define $S + T := \{x + y | x \in S,\ y \in T\}$, $S \pm b := \{x \pm b | x \in S\}$, $AS := \{Ax | x \in S\}$. If $S \subseteq R^n$, the closed convex cone generated by S is denoted by $\mathrm{cc}(S)$. Also, we write $\mathrm{cc}(y_1, \cdots, y_k)$ for $\mathrm{cc}(\{y_1, \cdots, y_k\})$, where $y_1, \cdots, y_k$ are vectors in R^n. If $S \subseteq R^n$, then $S^0 := \{\eta \in R_n | \eta x \leqq 0 \text{ for all } x \in S\}$ is called the *polar cone* of S. Obviously, S^0 is a closed convex cone in R_n. If $C_1, \cdots, C_k$ are convex closed cones, then

$$(1.1) \qquad (C_1 + \cdots + C_k)^0 = C_1^0 \cap \cdots \cap C_k^0.$$

If, in addition, $(\mathrm{ri}\, C_1) \cap \cdots \cap (\mathrm{ri}\, C_k) \neq \emptyset$, then (see [18, p. 146])

$$(1.2) \qquad (C_1 \cap \cdots \cap C_k)^0 = C_1^0 + \cdots + C_k^0.$$

We say that $f : S \to R^m$ (where $S \subseteq R^n$) is $\mathscr{C}'$ (or a $\mathscr{C}'$ map) in S if f is continuously differentiable in S. The function f will be called $\mathscr{C}'$ at a point $b \in R^n$ if f is $\mathscr{C}'$ in a neighborhood of b. The notation $\mathscr{N}$ (sometimes provided with an index $\mathscr{N}_1$, $\mathscr{N}_2$) will be used for a (usually unspecified) neighborhood of the origin in some space R^n. A statement like: "There exists a $\mathscr{C}'$ map $f : \mathscr{N} \to R^1$" should be read: "There exists a neighborhood $\mathscr{N}$ of the origin and a $\mathscr{C}'$ map $f : \mathscr{N} \to R^1$". We often will consider $\mathscr{C}'$ maps defined on a set of the form $\mathscr{N} \cap R_+^k$. Differentiability and derivatives on the boundary of this set (particularly in 0) are to be understood in the obvious sense.

We are now in the position to define the concept derived cone.

Definition 1.3. If $b \in S \subseteq R^n$, a closed convex cone C is called a *derived cone* of S at b if for any collection of vectors $p_1, \cdots, p_k$ in $\mathrm{ri}\, C$, there exists a $\mathscr{C}'$ map $\xi : \mathscr{N} \cap R_+^k \to S$ satisfying

$$(1.4) \qquad \xi(\tau) = b + \sum_{i=1}^{k} \tau_i p_i + o(\tau), \qquad \tau \to 0.$$

We will say that ξ is a *map corresponding to* the vectors $p_1, \cdots, p_k$. Note that the linearization $\hat{\xi}(\tau) := b + \sum_{i=1}^k \tau_i p_i$ maps R_+^k onto $b + \mathrm{cc}(p_1, \cdots, p_k)$. In that sense a derived cone is a linearization of the set S in the neighborhood of b.

The following theorem is the main result of this paper.

Theorem 1.5. *Let $S_1, \cdots, S_k$ be sets in R^n, $b \in S := S_1 \cap \cdots \cap S_k$ and let C_i be a derived cone of S_i at b for $i = 1, \cdots, k$. If a function $f : R^n \to R^1$ is $\mathscr{C}'$ at b and if f attains a maximum at b subject to the condition $x \in S$ (that is, $f(x) \leqq f(b)$ for all $x \in S$), then there exist $\rho \geqq 0$, $\psi_i \in C_i^0$, $i = 1, \cdots, k$, such that*

$$(1.6) \qquad (\rho, \psi_1, \cdots, \psi_k) \neq 0,$$

$$(1.7) \qquad \rho \nabla f(b) = \psi_1 + \cdots + \psi_k.$$

Note that we consider $\nabla f(b)$ a row vector. The condition (1.6) will be called the *nontriviality condition.* The vectors $\psi_1, \cdots, \psi_k$ are called *polar vectors* and formula (1.7) will be referred to as the *polar rule.*

2. Proof of Theorem 1.5. The major step in the proof is the following lemma, the proof of which is postponed to Appendix A.

LEMMA 2.1. *Let S_1, S_2 be sets in R^n, $b \in S_1 \cap S_2$ and C_1, C_2 derived cones of S_1, S_2 respectively, at b. If C_1 and C_2 are not separated (that is, if there does not exist a nonzero vector $\psi \in R_n$ such that $\psi x \leqq 0 \leqq \psi y$ for all $x \in C_1, y \in C_2$), then $C_1 \cap C_2$ is a derived cone of $S_1 \cap S_2$ at b.*

DEFINITION 2.2. A system $(C_1, \cdots, C_k)$ of closed convex cones is called *separated* if there exist vectors $\psi_1, \cdots, \psi_k$ satisfying $\psi_i \in C_i^0$, $i = 1, \cdots, k$, $\psi_1 + \cdots + \psi_k = 0, (\psi_1, \cdots, \psi_k) \neq 0$.

It is easily seen that this definition coincides with the well-known concept of separation in the case $k = 2$.

We have the following extension of Lemma 2.1.

LEMMA 2.3. *If $(C_1, \cdots, C_k)$ is a nonseparated system of derived cones of the sets $S_1, \cdots, S_k$ at b, then* $\mathrm{ri}\, C_1 \cap \cdots \cap \mathrm{ri}\, C_k \neq \emptyset$ *and $C_1 \cap \cdots \cap C_k$ is a derived cone of $S_1 \cap \cdots \cap S_k$.*

Proof. We use induction. Since $(C_1, \cdots, C_{k-1})$ is not separated we have: $\mathrm{ri}\, C_1 \cap \cdots \cap \mathrm{ri}\, C_{k-1} \neq \emptyset$ and $\hat{C} := C_1 \cap \cdots \cap C_{k-1}$ is a derived cone of $\hat{S} := S_1 \cap \cdots \cap S_{k-1}$. Also $\hat{C}$ and C_k are not separated, since $\hat{\psi} \in \hat{C}^0$ implies $\hat{\psi} = \psi_1 + \cdots + \psi_{k-1}$ with $\psi_i \in C_i^0$, $i = 1, \cdots, k - 1$ (see (1.2)). It follows that $\mathrm{ri}\, \hat{C} \cap \mathrm{ri}\, C_k = \mathrm{ri}\, C_1 \cap \cdots \cap \mathrm{ri}\, C_k \neq \emptyset$ and by Lemma 2.1, $\hat{C} \cap C_k = C_1 \cap \cdots \cap C_k$ is a derived cone of $\hat{S} \cap S_k = S_1 \cap \cdots \cap S_k$. Q.E.D.

LEMMA 2.4. *If C is a derived cone of S at b and if $f: R^n \to R^1$ is $\mathscr{C}'$ at b and attains a maximum at b subject to the condition $x \in S$, then $\nabla f(b) \in C^0$.*

Proof. If $p \in \mathrm{ri}\, C$, there exists a function $\xi : [0, \varepsilon) \to S$ for some $\varepsilon > 0$, satisfying $\xi(\tau) = b + p\tau + o(\tau)$, $\tau \to 0$. Then $f(b) \geqq f(\xi(\tau)) = f(b) + \nabla f(b)p\tau + o(\tau)$, $\tau \to 0$, implies $\nabla f(b)p \leqq 0$. By the continuity of the inner product we have $\nabla f(b)p \leqq 0$ for all $p \in C$, hence $\nabla f(b) \in C^0$. Q.E.D.

Proof of Theorem 1.5. If $(C_1, \cdots, C_k)$ is not separated, then $C_1 \cap \cdots \cap C_k$ is a derived cone of $S_1 \cap \cdots \cap S_k$ and $\mathrm{ri}\, C_1 \cap \cdots \cap \mathrm{ri}\, C_k \neq \emptyset$. By (1.2) and Lemma 2.4 we have $\nabla f(b) \in (C_1 \cap \cdots \cap C_k)^0 = C_1^0 + \cdots + C_k^0$, say $\nabla f(b) = \psi_1 + \cdots + \psi_k$, with $\psi_i \in C_i^0$. If $(C_1, \cdots, C_k)$ is separated, then there exist vectors $\psi_i \in C_i^0$, not all zero, such that $\psi_1 + \cdots + \psi_k = 0$. Then (1.7) is satisfied with $\rho = 0$. Q.E.D.

Whenever one of the constraint sets S_i is itself given as an intersection $S_i = S_{i1} \cap \cdots \cap S_{ir}$, with derived cones $C_{i1}, \cdots, C_{ir}$, then the polar vector ψ_i corresponding to the constraint set S_i is to be replaced with $\varphi_{i1} + \cdots + \varphi_{ir}$ in the polar rule and with $\varphi_{i1}, \cdots, \varphi_{ir}$ in the nontriviality condition. Here $\varphi_{ij} \in C_{ij}^0$, $j = 1, \cdots, r$. We call this operation a *substitution rule*. The validity of the substitution rule is immediately clear from Theorem 1.5. We give another argument, however, which is also applicable in more general situations: If $C_{i1}, \cdots, C_{ir}$ are not separated, then $C_{i1} \cap \cdots \cap C_{ir}$ is a derived cone of S_i and $\mathrm{ri}\, C_{i1} \cap \cdots \cap \mathrm{ri}\, C_{ir} \neq \emptyset$. Hence $\psi_i \in (C_{i1} \cap \cdots \cap C_{ir})^0 = C_{i1}^0 + \cdots + C_{ir}^0$. If $C_{i1}, \cdots, C_{ir}$ are separated, there exist $\varphi_{i1}, \cdots, \varphi_{ir}$, with $\varphi_{ij} \in C_{ij}^0$, $\varphi_{i1} + \cdots + \varphi_{ir} = 0$. Then we may set ρ and every other polar vector zero, and the nontriviality condition and the polar rule will be satisfied.

3. Derived cones of some simple sets and applications to mathematical programming. In order to apply Theorem 1.5, we have to find derived cones for simple sets. We give three examples of sets where derived cones are easily given.

LEMMA 3.1. (i) *If* $S := \{x \in R^n | g(x) \leqq 0\}$, *where* $g: R^n \to R^1$ *is* $\mathscr{C}'$ *at b, then* $C = R^n$ *is a derived cone of S at b if* $g(b) < 0$ *and* $C = \{p | \nabla g(b)p \leqq 0\}$ *is a derived cone if* $g(b) = 0$, $\nabla g(b) \neq 0$.

The corresponding polar cones are $\{0\}$ *and* $\{\mu \nabla g(b) | \mu \geqq 0\}$ *respectively. In either case,* $\{\mu \nabla g(b) | \mu \geqq 0, \mu g(b) = 0\}$ *is the polar of a derived cone.*

(ii) *If* $S := \{x \in R^n | r(x) = 0\}$, *where* $r: R^n \to R^1$ *is* $\mathscr{C}'$ *at b,* $r(b) = 0$ *and* $\nabla r(b) \neq 0$, *then* $C := \{p | \nabla r(b)p = 0\}$ *is a derived cone. The corresponding polar cone is* $\{\lambda \nabla r(b) | \lambda \in R^1\}$.

(iii). *If S is convex,* $b \in S$, *then* $C = \operatorname{cc}(S - b)$ *is a derived cone of S at b.*

Proof. (i) If $g(b) < 0$, then the map $\xi(\tau) := b + \sum \tau_i p_i$ corresponds to $p_1, \cdots, p_k$, since $b \in \operatorname{int} S$. If $g(b) = 0$, and $p_1, \cdots, p_k \in \operatorname{ri} C$, that is $\nabla g(b)p_i < 0$, then again $\xi(\tau) := b + \sum \tau_i p_i$ corresponds to $p_1, \cdots, p_k$, since $g(\xi(\tau)) = \sum \tau_i \nabla g(b) p_i + o(\tau) \leqq 0$ for sufficiently small $\tau \in R^k_+$.

(ii) Without loss of generality we assume $b = 0$. Consider the function $h(p, \omega) := r(p + \omega(\nabla r(0))')$ for $p \in R^n$, $\omega \in R^1$. Since $h(0, 0) = 0$, $(\partial h/\partial \omega)(0, 0) = \|\nabla r(0)\|^2 \neq 0$, there exists (by the implicit function theorem) a $\mathscr{C}'$ function $\omega = \varphi(p)$ such that $r(p + \varphi(p)(\nabla r(0))') = 0$ for $p \in \mathscr{N}$. Furthermore, $(\partial h/\partial p)(0, 0)p = \nabla r(0)p = 0$ and hence, $\nabla \varphi(0)p = 0$ for $p \in C$. Therefore the map η defined by

$$\eta(\tau) := \sum \tau_i p_i + \varphi(\sum \tau_i p_i)(\nabla r(0))'$$

corresponds to $p_1, \cdots, p_k$.

(iii) Again we assume $b = 0$. Furthermore, it is no loss of generality to assume that $\dim S = n$, so that $\operatorname{ri} S = \operatorname{int} S$. If $p \in \operatorname{ri} \operatorname{cc}(S)$, then there exists $\varepsilon > 0$, such that $\varepsilon p \in S$. Otherwise, the ray $\{\lambda p | \lambda \geqq 0\}$ would be separated from S and hence from $\operatorname{cc}(S)$ in contradiction to $p \in \operatorname{ri} \operatorname{cc}(S) = \operatorname{int} \operatorname{cc}(S)$. It follows, that if $p_1, \cdots, p_k$ are in $\operatorname{ri} \operatorname{cc}(S)$, then $\mathscr{N} \cap \operatorname{cc}(p_1, \cdots, p_k) \subseteq S$ for some $\mathscr{N}$. Therefore $\xi(\tau) = \sum \tau_i p_i$ is a map corresponding to $p_1, \cdots, p_k$. Q.E.D.

Lemma 3.1 gives rise to some other substitution rules: (i) If one of the constraint sets S_i in Theorem 1.5 is given by $S_i := \{x \in R^n | g(x) \leqq 0\}$ with $g: R^n \to R^1$, $\mathscr{C}'$ at b, then we may replace the corresponding polar vector ψ_i in the polar rule by $\mu \nabla g(b)$, with the conditions $\mu \geqq 0$, $\mu g(b) = 0$, and in the nontriviality condition by μ.

Indeed, if $g(b) < 0$, then $\{0\}$ is the polar cone of a derived cone and we must have $\psi_i = 0$. If $g(b) = 0$, $\nabla g(b) \neq 0$, then the polar vectors are of the form $\psi_i = \mu \nabla g(b)$, according to Lemma 3.1.

Finally, if $g(b) = 0$, $\nabla g(b) = 0$, then we may simply set $\mu = 1$, and ρ and the remaining polar vectors equal to zero. Then the polar rule is trivially satisfied. Similarly we have:

(ii) If one of the constraint sets S_i is equal to $\{x \in R^n | r(x) = 0\}$, where $r: R^n \to R^1$ is $\mathscr{C}'$ at b, then we may replace ψ_i with $\lambda \nabla r(b)$ in the polar rule and with λ in the nontriviality condition.

DEFINITION 3.2. A constraint set S is called of the type *EI* if it is of the form

$$S = \{x \in R^n | r_i(x) = 0, i = 1, \cdots, l, g_j(x) \leqq 0, j = 1, \cdots, m\}, \tag{3.3}$$

where $r_i: R^n \to R^1$, $g_j: R^n \to R^1$ are $\mathscr{C}'$. If $l = 0$ we say that S is of the type *I*, if $m = 0$, S is called of the type *E*.

A set of the type *EI* is an intersection of sets of the types given in (i) and (ii).

Therefore, combining the substitution rules of this section and § 2, we obtain the *EI-substitution rule*: If some S_ν is of the form (3.3), then we may replace the corresponding polar vector ψ_ν in the polar rule by

$$\sum \lambda_i \nabla r_i(b) + \sum \mu_j \nabla g_j(b) \tag{3.4}$$

with λ_i real, $\mu_j \geqq 0$, $\mu_j g_j(b) = 0$, $i = 1, \cdots, l; j = 1, \cdots, m$, and in the nontriviality condition by $\lambda_1, \cdots, \lambda_l, \mu_1, \cdots, \mu_m$.

A general type of mathematical programming problem is to determine the maximum of a function $f(x)$ subject to $x \in S$, where S is of the type *EI*. Theorem 1.5 and the *EI*-substitution rule yield immediately the following necessary condition.

THEOREM 3.5 (Mangasarian and Fromovitz [13]). *Let $f: R^n \to R$ be $\mathscr{C}'$ at b and maximal at b subject to the condition $x \in S$, where S is given by* (3.3). *Then there exist numbers $\rho \geqq 0$, $\lambda_i \in R^1$, $i = 1, \cdots, l$, $\mu_j \geqq 0$, $j = 1, \cdots, m$, with $\mu_j g_j(b) = 0$, such that*

$$(\rho, \lambda_1, \cdots, \lambda_l, \mu_1, \cdots, \mu_m) \neq 0, \tag{3.6}$$

$$\rho \nabla f(b) = \sum \lambda_i \nabla r_i(b) + \sum \mu_j \nabla g_j(b). \tag{3.7}$$

In [12, p. 168] the problem is considered of maximizing $f(x)$ subject to $x \in S_1 \cap S_2$, where S_1 is of the type *EI* and S_2 is convex. Using Lemma 3.1 (iii) and the *EI*-substitution rule we find immediately from Theorem 1.5, the following.

THEOREM 3.8. *Let $f: R^n \to R^1$ be $\mathscr{C}'$ at b and maximal at b subject to $x \in S_1 \cap S_2$, where S_1 is given by* (3.3) *and S_2 is convex. Then there exist numbers $\rho \geqq 0$, $\lambda_i \in R$, $\mu_j \geqq 0$ with $\mu_j g_j(b) = 0$, satisfying* (3.6) *and $(\rho \nabla f(b) - \sum \lambda_i \nabla r_i(b) - \sum \mu_j \nabla g_j(b))(x - b) \leqq 0$ for all $x \in S_2$.*

Proof. Note that $\eta \in (\mathrm{cc}\,(S_2 - b))$ implies $\eta(x - b) \leqq 0$ for all $x \in S_2$. Q.E.D.

If in Theorem 3.5, $m = 0$, then we have the Lagrange multiplier theorem. If $l = 0$, the result is F. John's necessary condition.

4. Discrete-time optimal control problems.

We give a parameter-free formulation of the discrete-time optimal control problem.

Problem 4.1. Given a positive integer N, sets $X_k \subseteq R^n$, $k = 0, \cdots, N$, set-valued maps $V_k(\cdot)$, $k = 0, \cdots, N - 1$, (that is, $V_k(x) \subseteq R^n$ for every $x \in R^n$) and a function $h: R^n \to R^1$, determine a sequence x_k, $k = 0, \cdots, N$, such that $h(x_N)$ is maximized subject to the constraints $x_k \in X_k$, $k = 0, \cdots, N$, $x_{k+1} \in V_k(x_k)$, $k = 0, \cdots, N - 1$.

In order to give necessary conditions for the solution of Problem 4.1 we need an extension of the concept derived cone.

DEFINITION 4.2. If $V(\cdot)$ is a set-valued function for $x \in S \subseteq R^{n_1}$, such that $V(x) \subseteq R^{n_2}$, $x \in S$, and if s is a *selection* of V, that is, $s(x) \in V(x)$ for $x \in S$, then a set-valued function $D(\cdot)$ is called a *family of derived cones of* V at $b \in S$ (with respect to the selection s), if $D(x)$ is a closed convex cone for $x \in S$ and if for every collection $p_1, \cdots, p_k \in \operatorname{ri} D(b)$ there exists a $\mathscr{C}'$ function $\xi: (b + \mathscr{N}_1) \times (\mathscr{N}_2 \cap R^k_+) \to R^{n_2}$ satisfying $\xi(x, \tau) \in V(x)$ and

$$\xi(x, \tau) = s(x) + \sum_{i=1}^{k} \tau_i p_i + o(\tau, x - b), \qquad \tau \to 0, \quad x \to b, \tag{4.3}$$

where $o(\tau, x - b)$, $\tau \to 0$, $x \to b$, stands for a function of the form $(\|\tau\| + \|x - b\|) \cdot \varepsilon(\tau, x - b)$, with $\varepsilon(\tau, y) \to 0$, $\tau \to 0$, $y \to 0$.

We say that ξ is a *family of maps corresponding to* $p_1, \cdots, p_k$. It follows from Definition 4.2, that $s(x)$ is $\mathscr{C}'$ if there exists a family of derived cones.

The following lemma will be used in the proof of Theorem 4.5 (i and ii) and in the next section, (iii).

LEMMA 4.4. (i) *If* $b_i \in S_i \subseteq R^{n_i}$ *and* C_i *is a derived cone of* S_i *at* b_i, $i = 1, \cdots, k$, *then* $C_1 \times \cdots \times C_k$ *is a derived cone of* $S_1 \times \cdots \times S_k$ *at* $(b_1; \cdots; b_k)$. *The corresponding polar cone is* $C_1^0 \times \cdots \times C_k^0$.

(ii) *If* $b \in S \subseteq R^n$, C *is a derived cone of* S *at* b, V *is a set-valued function with selection* s *and a family of derived cones* D *at* b, *then*

$$E := \{(p; q) | p \in C, q - \nabla s(b)p \in D(b)\}$$

is a derived cone of $T := \{(x; y) | x \in S, y \in V(x)\}$ *at* $(b; s(b))$. *The polar cone is given by*

$$E^0 = \{(\varphi - \psi \nabla s(b), \psi) | \varphi \in C^0, \psi \in D^0(b)\}.$$

(iii) *If* b, S, C, V, s, *and* D *are as in* (ii), *then* $F := D(b) + \nabla s(b)C$ *is a derived cone of* $\bigcup_{x \in S} V(x)$ *at* b.

Here $\nabla s(b)$ *denotes the functional matrix* $(\partial s_i / \partial x_j)$ *at the point* b.

Proof. (i) If $\mathbf{p}_j = (p_{1j}; \cdots; p_{kj})$, $j = 1, \cdots, m$, are vectors in $\operatorname{ri}(C_1 \times \cdots \times C_k) = \operatorname{ri} C_1 \times \cdots \times \operatorname{ri} C_k$ and ξ_i corresponds to $p_{i1}, \cdots, p_{im}$ for $i = 1, \cdots, k$, then $\boldsymbol{\xi}(\tau) = (\xi_1(\tau); \cdots; \xi_k(\tau))$ corresponds to $\mathbf{p}_1, \cdots, \mathbf{p}_m$.

(ii) Let $(p_i; q_i)$ be in $\operatorname{ri} E$ for $i = 1, \cdots, k$. Then $p_1, \cdots, p_k \in \operatorname{ri} C$, $r_i := q_i - \nabla s(b)p_i \in \operatorname{ri} D(b)$. If $\xi(\tau)$ corresponds to $p_1, \cdots, p_k$ and $\eta(x, \tau)$ is a family of maps corresponding to $r_1, \cdots, r_k$, then $\zeta(\tau) := (\xi(\tau); \eta(\xi(\tau), \tau))$ is a map corresponding to $(p_1; q_1), \cdots, (p_k; q_k)$.

(iii) If $p_1, \cdots, p_k$ are in $\operatorname{ri}(D(b) + \nabla s(b)C)$, say $p_i = r_i + \nabla s(b)q_i$ with $r_i \in \operatorname{ri} D(b)$, $q_i \in \operatorname{ri} C$, then there exists a map $\xi(\tau)$ corresponding to $q_1, \cdots, q_k$ and a family $\eta(x, \tau)$ corresponding to $r_1, \cdots, r_k$. It is easily seen that $\zeta(\tau) := \eta(\xi(\tau), \tau)$ corresponds to $p_1, \cdots, p_k$. Q.E.D.

The following theorem gives general necessary conditions for the solution of Problem 4.1.

THEOREM 4.5. *Let* $\bar{x}_k$ *be a solution of Problem* 4.1. *Let* $s_k(x)$ *be a selection of* $V_k(x)$, *satisfying* $\bar{x}_{k+1} = s_k(\bar{x}_k)$ *for* $k = 0, \cdots, N - 1$, *and* $D_k(x)$ *a family of derived cones of* $V_k(x)$ *at* $\bar{x}_k$. *Finally, let* C_k *be a derived cone of* X_k. *Then there exist a number* $\rho \geqq 0$ *and row vectors* ψ_k, φ_k, $k = 0, \cdots, N$, *satisfying*

$$(\rho, \psi_0, \cdots, \psi_N, \varphi_0, \cdots, \varphi_N) \neq 0, \tag{4.6}$$

$$\begin{aligned} &\varphi_k \in C_k^0, && k = 0, \cdots, N, \\ &\psi_{k+1} \in D_k^0(\bar{x}_k), && k = 0, \cdots, N - 1, \end{aligned} \tag{4.7}$$

$$\psi_k = \psi_{k+1} \nabla s_k(\bar{x}_k) - \varphi_k, \qquad k = 0, \cdots, N - 1, \tag{4.8}$$

$$\psi_0 = 0, \qquad \psi_N = \rho \nabla h(\bar{x}_N) - \varphi_N. \tag{4.9}$$

Proof. We introduce the $(N + 1)$ n-dimensional vector $\mathbf{x} := (x_0; \cdots; x_N)$. The optimal control problem is to maximize $\mathbf{f}(\mathbf{x}) := h(x_N)$ subject to the constraints

$\mathbf{x} \in \mathbf{T}_0 \cap \cdots \cap \mathbf{T}_N$, where $\mathbf{T}_k := \{\mathbf{x} | x_k \in X_k, x_{k+1} \in V(x_k)\}$, $k = 0, \cdots, N-1$, and $\mathbf{T}_N := \{\mathbf{x} | x_N \in X_N\}$. Hence,

$$\mathbf{T}_k = R^n \times \cdots \times T_k \times \cdots \times R^n,$$

where $T_k := \{(x; y) | x \in X_k, y \in V_k(x)\}$, $k = 0, \cdots, N-1$, and $T_N = X_N$. In order to apply Theorem 1.5, we compute derived cones of $\mathbf{T}_k$ at $\bar{\mathbf{x}} := (\bar{x}_0; \cdots; \bar{x}_N)$. According to Lemma 4.4 (ii), E_k is a derived cone of T_k at $\bar{x}_k$, where

$$E_k := \{(p; q) | p \in C_k, q - \nabla s_k(\bar{x}_k)p \in D_k(\bar{x}_k)\}, \qquad k = 0, \cdots, N-1,$$

and $E_N := C_N$. The polar cones are:

$$E_k^0 = \{(\varphi - \psi \nabla s_k(\bar{x}_k), \psi) | \varphi \in C_k^0, \psi \in D_k^0(\bar{x}_k)\}, \qquad k = 0, \cdots, N-1,$$

and $E_N^0 = C_N^0$. Because of Lemma 4.4 (i),

$$\mathbf{E}_k := R^n \times \cdots \times E_k \times \cdots \times R^n$$

is a derived cone of $\mathbf{T}_k$ at $\bar{\mathbf{x}}$ and the corresponding polar cone is $\mathbf{E}_k^0 = \{0\} \times \cdots \times E_k^0 \times \cdots \times \{0\}$.

Now we are in the position to apply Theorem 1.5. We find that there exist $\rho \geqq 0$, $\boldsymbol{\varphi}_k \in \mathbf{E}_k^0$, $k = 0, \cdots, N$, not all zero, such that

$$\rho \nabla \mathbf{f}(\bar{\mathbf{x}}) = \boldsymbol{\varphi}_0 + \cdots + \boldsymbol{\varphi}_N. \tag{4.10}$$

We have

$$\nabla \mathbf{f}(\bar{\mathbf{x}}) = (0, \cdots, 0, \nabla h(\bar{x}_N)),$$

$$\boldsymbol{\varphi}_k = (0, \cdots, 0, \varphi_k - \psi_{k+1} \nabla s_k(\bar{x}_k), \psi_{k+1}, 0, \cdots, 0), \qquad k = 0, \cdots, N-1,$$

$$\boldsymbol{\varphi}_N = (0, \cdots, 0, \varphi_N),$$

with $\varphi_k \in C_k^0$, $\psi_{k+1} \in D_k^0(\bar{x}_k)$. Now (4.10) implies (4.7), (4.8), (4.9). Q.E.D.

Remark 4.11. The nontriviality condition (4.6) may be replaced with $(\rho, \psi_1, \cdots, \psi_N) \neq 0$, since it follows from (4.8) and (4.9) that $(\rho, \psi_1, \cdots, \psi_N) = 0$ implies $(\rho, \psi_0, \cdots, \psi_N, \varphi_0, \cdots, \varphi_N) = 0$. We have given the seemingly weaker nontriviality condition (4.6) in order to make the application of the substitution rule easier. Indeed, very often the constraint sets X_k are of the type *EI*, *E* or *I*. In those cases the *EI*-substitution rule for the polar vectors φ_k is easily seen to apply.

If it is known that $\nabla h(\bar{x}_N) \notin C_N^0$, we may replace the nontriviality condition with $(\psi_1, \cdots, \psi_N) \neq 0$ (see (4.9)).

Remark 4.12. We could of course have reduced Problem 4.1 to a problem without state constraints by defining $\hat{V}_k(x) := V_k(x) \cap X_{k+1}$, $k = 0, \cdots, N-1$. However, it might be difficult to find a family of derived cones for $\hat{V}_k$.

Usually the sets $V_k(x)$ in (4.1) are given in parameter form, that is $V_k(x) = f_k(x, U_k) := \{f_k(x, u) | u \in U_k\}$, where U_k is a set and $f_k : R^n \times U_k \to R^n$. Then selections are easily given. For example, for every $u_0 \in U_k$, $s(x) := f_k(x, u_0)$ is a selection of $V_k(x)$. We give a parameter formulation of the discrete-time control problem.

Problem 4.13. Given a positive integer N, sets $X_k \subseteq R^n$, $k = 0, \cdots, N$, set U_k, $k = 0, \cdots, N-1$, functions $f_k : R^n \times U_k \to R^n$, such that $f_k(\cdot, u)$ is $\mathscr{C}'$

for every $u \in U_k$, and a $\mathscr{C}'$ function $h: R^n \to R^1$, determine sequences u_k, $k = 0, \cdots, N-1$; x_k, $k = 0, \cdots, N$, such that $h(x_N)$ is maximal subject to the constraints

$$x_{k+1} = f_k(x_k, u_k), \qquad k = 0, \cdots, N-1, \tag{4.14}$$

$$u_k \in U_k, \qquad k = 0, \cdots, N-1, \tag{4.15}$$

$$x_k \in X_k, \qquad k = 0, \cdots, N. \tag{4.16}$$

Necessary conditions for the solution of this problem can be derived from Theorem 4.5 if one knows derived cones C_k for the sets X_k and families of derived cones $D_k(x)$ for the sets $V_k(x) := f_k(x, U_k)$. In Theorem 4.18 and in Theorem 4.25 we postulate the existence of C_k (see also Remark 4.11, but we construct families $D_k(x)$. First, we give a lemma.

LEMMA 4.17. *If $f: R^n \times U \to R^n$ and $f(\cdot, u)$ is $\mathscr{C}'$ for $u \in U$, $V(x) := f(x, U)$ is convex for every $x \in R^n$, and $s(x) := f(x, u_0)$ for some $u_0 \in U$, then $D(x) :=$* cc $(V(x) - s(x))$ *is a family of derived cones of $V(x)$ with respect to $s(x)$ at each point $b \in R^n$.*

Proof. Let $p_1, \cdots, p_k \in \operatorname{ri} D(b)$. Then there exist positive numbers $\lambda_1, \cdots, \lambda_k$, such that $p_i \in \lambda_i(V(b) - s(b))$, say $p_i = \lambda_i(f(b, v_i) - s(b))$ with $v_i \in U$ (compare Lemma 3.1 (iii)). It follows that

$$\xi(x, \tau) := s(x) + \sum \lambda_i \tau_i (f(x, v_i) - s(x))$$

is a family of maps corresponding to $p_1, \cdots, p_k$. Indeed, since $V(x)$ is convex and $f(x, v_i) \in V(x)$, $s(x) \in V(x)$ we have that $\xi(x, \tau) \in V(x)$ for $0 \leqq \lambda_i \tau_i \leqq 1$, $i = 1, \cdots, k$. Furthermore, ξ is $\mathscr{C}'$ and

$$\xi(x, \tau) = s(x) + \sum p_i \tau_i + o(\tau, x - b), \qquad \tau \to 0, \quad x \to b. \qquad \text{Q.E.D.}$$

THEOREM 4.18 (Discrete maximum principle). *In Problem 4.13 let $V_k(x) := f_k(x, U_k)$ be convex for every $x \in R^n$, $k = 0, \cdots, N-1$. Let $\bar{u}_k$, $\bar{x}_k$ be a solution of the problem and C_k a derived cone of X_k at $\bar{x}_k$, $k = 0, \cdots, N$. Then there exist a number $\rho \geqq 0$, and row vectors ψ_k, φ_k, $k = 0, \cdots, N$, satisfying*

$$\psi_k = \psi_{k+1} \nabla_x f_k(\bar{x}_k, \bar{u}_k) - \varphi_k, \qquad k = 0, \cdots, N-1, \tag{4.19}$$

$$\varphi_k \in C_k^0, \tag{4.20}$$

$$\psi_{k+1} f_k(\bar{x}_k, \bar{u}_k) = \max_{v \in U_k} \psi_{k+1} f_k(\bar{x}_k, v), \tag{4.21}$$

$$\psi_0 = 0, \qquad \psi_N = \rho \nabla h(\bar{x}_N) - \varphi_N, \tag{4.22}$$

$$(\rho, \psi_0, \cdots, \psi_N, \varphi_0, \cdots, \varphi_N) \neq 0. \tag{4.23}$$

Proof. This result follows from Lemma 4.17 and Theorem 4.5, where we can take $s_k(x) = f_k(x, \bar{u}_k)$. Note that $\psi_{k+1} \in \{\text{cc}(V_k(\bar{x}_k) - s_k(\bar{x}_k))\}^0$ implies (4.21). Q.E.D.

Theorem 4.18 was given in [4] by Halkin for the case where X_0, X_N are of the type E and $X_k = R^n$ for $k = 1, \cdots, N-1$. In [3, (4.2)] the more general situation is treated where X_0, X_N are of the type EI and X_k of the type I for $k = 1, \cdots, N-1$.

Remark 4.24. In the case where $h(x) := -x_1$ (where x_1 is the first coordinate of x) and $f_k(x, u)$ does not depend on x_1, the convexity condition can be weakened as observed in [9]. Indeed, introducing the vector $e := (1, 0, \cdots, 0)'$, one can show that the necessary condition given in Theorem 4.18 remains valid if only the set

$W_k(x) := \{f_k(x, u) + u_0 e | u \in U,\ u_0 \geqq 0\}$ is convex. (In this case, the set $V_k(x) := f_k(x, U)$ is called *directionally convex*.) This is most easily seen by making the following substitutions in control problem (4.13): $\mathbf{U} := \{\mathbf{u} = (u_0; u) | u_0 \geqq 0, u \in U\}$, $\mathbf{f}_k(x, \mathbf{u}) := f_k(x, u) + u_0 e$, where $\mathbf{u} = (u_0; u)$, so that $\mathbf{f}_k : R^n \times \mathbf{U} \to R^n$, and observing that $\bar{\mathbf{u}}_k := (0; \bar{u}_k)$, $\bar{x}_k$ is an optimal control of Problem 4.13 with U, f replaced by $\mathbf{U}, \mathbf{f}$ (see also [7, III.1]). The situation mentioned above often occurs if Problem 4.13 is obtained by adding one state coordinate in an optimal control problem with a sum of the form $\sum_{k=1}^{N-1} f_{0k}(x_k, u_k)$ as performance index.

As a second application of Theorem 4.5 we mention the following result.

THEOREM 4.25. *Let $\bar{u}_k$, $k = 0, \cdots, N-1$, $\bar{x}_k$, $k = 0, \cdots, N$, be a solution of Problem 4.13. Suppose that $U_k \subseteq R^m$ and $f_k : R^n \times U_k \to R^n$ is a $\mathscr{C}'$ map. If E_k is a derived cone of U_k at $\bar{u}_k$ and C_k is a derived cone of X_k at $\bar{x}_k$, then there exist $\rho \geqq 0$, $\psi_k, \varphi_k \in R_n$, $k = 0, \cdots, N$, such that*

$$(4.26)\qquad (\rho, \psi_0, \cdots, \psi_N, \varphi_0, \cdots, \varphi_N) \neq 0,$$

$$(4.27)\qquad \psi_k = \psi_{k+1} \nabla_x f_k(\bar{x}_k, \bar{u}_k) - \varphi_k,$$

$$(4.28)\qquad \psi_{k+1} \nabla_u f_k(\bar{x}_k, \bar{u}_k) \in E_k^0,$$

$$(4.29)\qquad \varphi_k \in C_k^0,$$

$$(4.30)\qquad \psi_0 = 0, \qquad \psi_N = \rho \nabla h(\bar{x}_N) - \varphi_N.$$

The proof of this result is an immediate application of Theorem 4.5 and the following lemma.

LEMMA 4.31. *If $u_0 \in U \subseteq R^m$, E is a derived cone of U at u_0, $f : R^n \times R^m \to R^n$ is $\mathscr{C}'$, then $D(x) := \nabla_x f(x, u_0) E$ is a family of derived cones of $V(x) := f(x, U)$ at every $b \in R^n$ with respect to the selection $s(x) = f(x, u_0)$.*

Proof. Let $p_1, \cdots, p_k$ be in $\operatorname{ri} D(b)$, say $p_i = \nabla_x f(b, u_0) v_i$ with $v_i \in \operatorname{ri} E$. Let ω be a map corresponding to $v_1, \cdots, v_m$. Then $\xi(x, \tau) := f(x, \omega(\tau))$ is a family of maps corresponding to $p_1, \cdots, p_k$. Q.E.D.

Theorem 4.25 was given in a less general form in a number of papers. In all cases the state constraint sets were of the type *EI* in a more or less restricted form. In [3, § 4.1], [2], the situation was considered where the sets U_k are of abstract character and necessary conditions, in particular of the form (4.28), were also expressed in terms of approximating cones. In [2] it was assumed that the cone (the radial cone) was a linearization of the first kind. This condition is rather restrictive, since it excludes the case where U_k is given by nonlinear equalities. In [3] a less restrictive situation is considered. However, the proof is not complete, since a result similar to Lemma 4.4(i) is used without proof and unlike Lemma 4.4(i) the result is not obvious, because of an independence condition occurring in the definition of the approximating cone.

5. Continuous-time optimal control problems. We consider the following continuous-time optimal control problem.

Problem 5.1. Given a number $T > 0$, a set $U \subseteq R^n$, a continuous function $f : R^n \times U \to R^n$, such that $f(\cdot, u)$ is $\mathscr{C}'$ for every $u \in U$, a $\mathscr{C}'$ function $h : R^n \to R^1$ and sets X_0, X_1 in R^n, determine a point $b \in X_0$ and a piecewise continuous function $u:[0, T] \to U$ such that the solution x of the differential equation

$$\mathring{x}(t) = f(x(t), u(t)), \tag{5.2}$$

with initial condition $x(0) = b$, exists on $[0, T]$, satisfies $x(T) \in X_1$ and such that $h(x(T))$ is maximal.

In Appendix B we discuss optimal control problems with measurable instead of piecewise continuous controls.

The results of §§ 2, 3 are not immediately applicable, since the control function is not an element of a finite-dimensional vector space. However, we can restate the problem, using the following notations: Let $\Omega := \{u | u:[0, T] \to U,$ piecewise continuous$\}$ be the set of *admissible* controls and $x_u(\cdot, b)$ the solution of (5.2) corresponding to the control $u \in \Omega$ and $b \in R^n$. For every $b \in R^n$ we define

$$W(b) := \{x_u(T, b) | u \in \Omega \text{ and } x_u(t, b) \text{ exists for } 0 \leqq t \leqq T\}, \tag{5.3}$$

$$W := \bigcup_{b \in X_0} W(b). \tag{5.4}$$

W is called the *reachable* set. With these notations Problem 5.1 is equivalent to the following problem.

Problem 5.5. Given T, U, f, h, X_0, X_1 as in (5.1), determine max $h(x)$ subject to the condition $x \in W \cap X_1$.

Thus we have a finite-dimensional optimization problem with two constraint sets. We can apply Theorem 1.5 if we are able to find derived cones for W and X_1. We postulate a derived cone of X_1 (and so we do for X_0) and we construct a derived cone of W. For that aim, we define the *perturbation cone* already present in [16, Chap. II].

Let $b \in R^n$, $\hat{u} \in \Omega$ and let $\hat{x}(t) := x_{\hat{u}}(t, b)$ exist on $[0, T]$. Then we define

$$A(t) := \nabla_x f(\hat{x}(t), \hat{u}(t)). \tag{5.6}$$

Furthermore, for every $s, t \mapsto \Phi(t, s)$ is defined to be the solution of the matrix differential equation $\mathring{Y}(t) = A(t)Y(t)$, $t \geqq s$, satisfying the initial condition $Y(s) = I$. First, we consider an *elementary perturbation* π of $\hat{u}$, given by $\pi := (t_0, v)$, where $t_0 \in (0, T)$ is a point of continuity of $\hat{u}$ and $v \in U$. Then, for sufficiently small $\varepsilon > 0$, the control function

$$u_\pi(t, \varepsilon) := \begin{cases} v, & t_0 - \varepsilon < t \leqq t_0, \\ \hat{u}(t) & \text{elsewhere in } [0, T], \end{cases} \tag{5.7}$$

will be in Ω and will yield a trajectory $x_{u_\pi}(t) =: x_\pi(t, \varepsilon)$, which exists on $[0, T]$. This is a result of the theory of ordinary differential equations. As in [10], [11, Chap. 4] it is easily seen that

$$x_\pi(T, \varepsilon) = \hat{x}(T) + \varepsilon p_\pi + o(\varepsilon), \qquad \varepsilon \to 0, \tag{5.8}$$

where

$$p_\pi := \Phi(T, t_0)\{f(\hat{x}(t_0), v) - f(\hat{x}(t_0), \hat{u}(t_0))\}. \tag{5.9}$$

This estimate is locally uniform with respect to b, that is, uniform with respect to b for b in a sufficiently small neighborhood of a fixed value $\hat{b}$.

A *combined perturbation* is a system $\pi := (\pi_1, \cdots, \pi_k)$, where $\pi_i = (t_i, v_i)$, with $v_i \in U$ and where the numbers $t_i \in (0, T)$ are distinct and such that $\hat{u}$ is continuous at t_i. The corresponding perturbed control is defined by

$$(5.10) \qquad u_\pi(t, \varepsilon) := \begin{cases} v_i, & t_i - \varepsilon_i < t \leqq t_i, i = 1, \cdots, k, \\ \hat{u}(t) & \text{elsewhere on } [0, T], \end{cases}$$

where $\varepsilon = (\varepsilon_1, \cdots, \varepsilon_k)'$. For small $\varepsilon \geqq 0$, the control u_π is well-defined in Ω and the corresponding trajectory x_π exists on $[0, T]$ and satisfies

$$(5.11) \qquad x_\pi(t, \varepsilon) = \hat{x}(T) + \sum_{i=1}^{k} \varepsilon_i p_{\pi_i} + o(\varepsilon), \qquad \varepsilon \to 0.$$

Again this result is locally uniform with respect to b. Furthermore, $x_\pi(T, \varepsilon)$ is a $\mathscr{C}'$ function of b and ε, and $x_\pi(T, \varepsilon) \in W(b)$ for small $\varepsilon \geqq 0$.

DEFINITION 5.12. The *perturbation cone* of Problem 5.1 corresponding to the control $\hat{u} \in \Omega$ and $b \in R^n$ is denoted $P_{\hat{u}}(b)$ and defined as the closed convex cone generated by the vectors p_π.

LEMMA 5.13. *Let $\hat{u} \in \Omega$, $\hat{b} \in R^n$ be given such that $x_{\hat{u}}(t, \hat{b})$ exists on $[0, T]$. Then $P_{\hat{u}}(b)$ is a family of derived cones of $W(b)$ at $\hat{b}$, with respect to the selection $x_{\hat{u}}(T, b)$.*

Proof. Let $q_1, \cdots, q_k \in \operatorname{ri} P_{\hat{u}}(\hat{b})$. Then we may assume that the vectors q_i are of the form

$$(5.14) \qquad q_i = \sum_{j=1}^{l} \alpha_{ij} p_{\pi_j},$$

with $\alpha_{ij} > 0$, $\pi_j = (t_j, v_j)$, the numbers $t_j \in (0, T)$ distinct and $v_j \in U$.

To see this, we remark that, according to Lemma A.1, there exist vectors $q_1^*, \cdots, q_r^* \in \operatorname{ri} P_{\hat{u}}(b)$ such that $\dim \operatorname{cc}(q_1^*, \cdots, q_r^*) = \dim P_{\hat{u}}(\hat{b})$ and $q_1, \cdots, q_k \in \operatorname{ri} \operatorname{cc}(q_1^*, \cdots, q_r^*)$. It follows that the vectors q_i^* are of the form

$$q_i^* = \sum_{j=1}^{l} \mu_{ij} p_{\pi_{ij}^*},$$

where $\pi_{ij}^* = (t_{ij}^*, v_{ij}^*)$, $\mu_{ij} > 0$. Changing, if necessary, the numbers t_{ij}^* slightly, we obtain numbers $\bar{t}_{ij}$, which are distinct and such that, with $\bar{\pi}_{ij} := (\bar{t}_{ij}, v_{ij}^*)$ and $\bar{q}_i := \sum_{j=1}^{l} \mu_{ij} p_{\bar{\pi}_{ij}}$, we still have $\bar{q}_1, \cdots, \bar{q}_r \in \operatorname{ri} P_{\hat{u}}(\hat{b})$ and $q_1, \cdots, q_k \in \operatorname{ri} \operatorname{cc}(\bar{q}_1, \cdots, \bar{q}_r)$. Hence, there exist $\lambda_{si} > 0$, $s = 1, \cdots, k$; $i = 1, \cdots, r$, such that

$$q_s = \sum_{i=1}^{r} \lambda_{si} \bar{q}_i = \sum_{i,j} \lambda_{si} \mu_{ij} p_{\pi_{ij}},$$

which is a sum of the form (5.14) written as a double sum.

We define $\xi : (\hat{b} + \mathscr{N}_1) \times (\mathscr{N}_2 \cap R_+^k) \to R^n$ by

$$\xi(b, \tau) = x_{u_\pi(\cdot, \varepsilon_\tau)}(T, b),$$

where $\pi := (\pi_1, \cdots, \pi_l)$, $\varepsilon_\tau := (\varepsilon_{1\tau}, \cdots, \varepsilon_{l\tau})$ and $\varepsilon_{j\tau} := \sum_{i=1}^{k} \alpha_{ij} \tau_i, j = 1, \cdots, l$. It follows from (5.11) that $\xi(b, \tau) = x_{\hat{u}}(T, b) + \sum q_i \tau_i + o(\tau, b - \hat{b})$, $\tau \to 0$, $b \to \hat{b}$, so that $\xi(b, \tau)$ is a family of maps corresponding to $q_1, \cdots, q_k$. Q.E.D.

Now we state and prove a version of Pontryagin's maximum principle with a general type of initial and final state constraint.

THEOREM 5.15 (Maximum principle). *Let $\bar{b}, \bar{u}$ be a solution of Problem* 5.1 *and let $\bar{x}(t) := x_{\bar{u}}(t, \bar{b})$ be the corresponding trajectory. If E_0 is a derived cone of X_0 at $b = \bar{x}(0)$ and E_1 is a derived cone of X_1 at $\bar{x}(T)$, then there exist a function $\psi : [0, T] \to R_n$, a vector $\varphi \in R_n$, and a number $\rho \geqq 0$ satisfying*

$$(\rho, \psi(T), \varphi) \neq 0, \tag{5.16}$$

$$\overset{\circ}{\psi}(t) = -\psi(t)\nabla_x f(\bar{x}(t), \bar{u}(t)), \tag{5.17}$$

$$\psi(t) f(\bar{x}(t), \bar{u}(t)) = \max_{v \in U} \psi(t) f(\bar{x}(t), v) \tag{5.18}$$

for all $t \in (0, T)$ at which $\bar{u}$ is continuous,

$$\psi(0) = E_0^0, \qquad \varphi \in E_1^0, \tag{5.19}$$

$$\psi(T) = \rho \nabla h(\bar{x}(T)) - \varphi. \tag{5.20}$$

Proof. It follows from Lemma 4.4(iii) and Lemma 5.12, that $P_{\bar{u}}(\bar{b}) + \nabla_b x_{\bar{u}}(T, \bar{b})E_0$ is a derived cone of W at $\bar{x}(T)$. According to a well-known result in the theory of ordinary differential equations, we have $\nabla_b x_{\bar{u}}(T, \bar{b}) = \Phi(T, 0)$, where Φ is as defined before, with $b = \bar{b}$, $\hat{u} = \bar{u}$. Applying Theorem 1.5 to Problem 5.5, we find that there exist a number $\rho \geqq 0$ and vectors φ_1, φ_2 in R_n such that

$$(\rho, \varphi_1, \varphi_2) \neq 0, \quad \rho\nabla h(\bar{x}(T)) = \varphi_1 + \varphi_2, \qquad \varphi_1 \in E_1^0,$$

$$\varphi_2 \in (P_{\bar{u}}(\bar{b}) + \Phi(T, 0)E)^0 = P_{\bar{u}}^0(\bar{b}) \cap (\Phi(T, 0)E_0)^0 \qquad (\text{see } (1.1)).$$

We define $\psi : [0, T] \to R_n$ to be the solution of (5.17) with $\psi(T) = \varphi_2$, hence $\psi(t) = \varphi_2 \Phi(T, t)$.

From $\varphi_2 \in (\Phi(T, 0)E_0)^0$ it follows that $\varphi_2 \Phi(T, 0)p \leqq 0$ for all $p \in E_0$, hence $\psi(0) = \varphi_2 \Phi(T, 0) \in E_0^0$. Furthermore, it follows from $\varphi_2 \in P_{\bar{u}}(\bar{b})$ that $\varphi_2 p_\pi \leqq 0$ for all π, that is, $\psi(t_0)\{f(\bar{x}(t_0), v) - f(\bar{x}(t_0), \bar{u}(t))\} \leqq 0$ for all $v \in U$ and all $t_0 \in (0, T)$ at which $\bar{u}$ is continuous. This yields (5.18). Finally we define $\varphi := \varphi_1$. Q.E.D.

Remark 5.21. It is easily seen that the nontriviality condition can be replaced with $(\rho, \psi(0)) \neq 0$ or $(\rho, \psi(T)) \neq 0$. If it is known that $\nabla h(\bar{x}(T)) \notin E_1^0$, then we may even write $\psi(T) \neq 0$ or $\psi(0) \neq 0$. Again we have written condition (5.16) in Theorem 5.15 in order to facilitate the use of the substitution rules. Usually, the sets X_0 and X_1 are of the type E. In [17] the situation is considered where X_0 and X_1 are of the type EI.

Remark 5.22. Naturally, other types of optimal control problems, for example, problems with integral criteria, integral constraints, nonautonomous systems, equations with parameters (that have to be chosen in an optimal way), coupled initial and final constraints and optimization criteria depending on initial and final value of $x(t)$, can be transformed into (5.1) by addition of state variables. Problems, where the final time is not fixed, can be transformed into fixed-time problems by the substitution $t = \tau T$, so that τ becomes the new time variable in the fixed interval $[0, 1]$ and T is a parameter. Another way to deal with free end time problems is to prove a slight generalization of Lemma 5.13, namely, that the set-valued function $(b; T) \mapsto P_{\bar{u}}(b; T)$ is a family of derived cones of $W(b; T) := \{x_u(T, b) | u \in \Omega\}$.

In this way one can also obtain necessary conditions if T is the constraint in some interval $[T_0, T_1]$. Free initial time problems can be treated similarly.

Appendix A. This section is devoted to the proof of the basic Lemma 2.1. We need some preliminary results.

LEMMA A.1. *If* $x_1, \cdots, x_k$ *are contained in* $\operatorname{ri} C$, *where* C *is a closed convex cone, then there exist vectors* $y_1, \cdots, y_m$ *in* $\operatorname{ri} C$, *such that* $\dim \operatorname{cc}(y_1, \cdots, y_m) = \dim C$ *and* $x_1, \cdots, x_k \in \operatorname{ri} \operatorname{cc}(y_1, \cdots, y_m)$.

Proof. Let $x_{k+1}, \cdots, x_m$ be vectors in $\operatorname{ri} C$, such that $\dim \operatorname{cc}(x_1, \cdots, x_m) = \dim C$. Define $b := (x_1 + \cdots + x_m)/m$ and $y_i := x_i + \varepsilon(x_i - b)$, $i = 1, \cdots, m$. For sufficiently small $\varepsilon > 0$, we have $y_i \in \operatorname{ri} C$. Also, $(y_1 + \cdots + y_m)/m = b$ and hence $x_i = \sum_{j=1}^m \lambda_{ij} y_j$, where $\lambda_{ij} = \varepsilon/(m(1 + \varepsilon))$ if $i \neq j$ and $\lambda_{ii} = (m + \varepsilon)/(m(1 + \varepsilon))$. Hence $x_i \in \operatorname{ri} \operatorname{cc}(y_1, \cdots, y_m)$. Q.E.D.

LEMMA A.2. *If* $f : \mathscr{N}_1 \cap R^n_+ \to R^m$ *is* $\mathscr{C}'$, *then there exists a* $\mathscr{C}'$ *extension* $F : \mathscr{N} \to R^m$ *of* f.

Proof. Let $\mathscr{N} := \{x \in R^n | x = (x_1, \cdots, x_n)'$ and $(|x_1|, \cdots, |x_n|) \in \mathscr{N}_1\}$. We extend f stepwise. Define $S_k := \{x \in \mathscr{N} | x_i \geqq 0$ for $i = k + 1, \cdots, n\}$. Then $\mathscr{N} \cap R^n_+ = S_0 \subset S_1 \subset \cdots \subset S_n = \mathscr{N}$. Furthermore, define $f_0 := f$, so that $f_0 : S_0 \to R^m$. Suppose that $k \in \{0, 1, \cdots, n - 1\}$ and that $f_k : S_k \to R^m$ has been defined. Then we construct $f_{k+1} : S_{k+1} \to R^m$ as follows:

$$f_{k+1}(x) := f_k(x), \qquad x \in S_k; \tag{A.3}$$

$$\begin{aligned} f_{k+1}(x) = f_{k+1}(x_1, \cdots, x_k, \cdots, x_n) := {} & 4f_k(x_1, \cdots, x_{k-1}, -\tfrac{1}{2}x_k, x_{k+1}, \cdots, x_n) \\ & -3f_k(x_1, \cdots, x_{k-1}, -x_k, x_{k+1}, \cdots, x_n) \end{aligned}$$

for $x \in S_{k+1} \setminus S_k$. It is easily seen that the partial derivatives of f_{k+1} exist and are continuous in S_{k+1}. Hence f_{k+1} is $\mathscr{C}'$. In this way $f_1, \cdots, f_n$ can be constructed successively. Finally we define $F := f_n$. Q.E.D.

Proof of Lemma 2.1. Without loss of generality we assume $b = 0$. Suppose that C_1 and C_2 are not separated. Let $x_1, \cdots, x_k$ be vectors in $\operatorname{ri}(C_1 \cap C_2) = \operatorname{ri} C_1 \cap \operatorname{ri} C_2$. According to Lemma A.1, there exist vectors $y_1, \cdots, y_m \in \operatorname{ri} C_1$ and vectors $z_1, \cdots, z_l \in \operatorname{ri} C_2$ such that, if $Y := [y_1, \cdots, y_m]$ (that is, Y is the matrix formed by the columns $y_1, \cdots, y_m$) and $Z := [z_1, \cdots, z_l]$, then $\operatorname{rank} Y = \dim C_1$, $\operatorname{rank} Z = \dim C_2$ and

$$X = YM = ZN, \tag{A.4}$$

where $X = [x_1, \cdots, x_k]$ and M and N are matrices all entries of which are positive. Formula (A.4) expresses the fact, that $x_1, \cdots, x_k \in \operatorname{ri} \operatorname{cc}(y_1, \cdots, y_m) \cap \operatorname{ri} \operatorname{cc} \cdot (z_1, \cdots, z_l)$. Because the cones C_1 and C_2 are not separated, they are not contained in a common hyperplane. Hence,

$$\operatorname{rank}[Y, Z] = n. \tag{A.5}$$

If necessary, we renumber the columns of Y and Z so as to obtain that Y and Z have the form

$$Y = [Y_1, Y_2], \quad Z = [Z_1, Z_2], \tag{A.6}$$

where Y_1 is an $n \times m_1$ matrix, Z_1 is an $n \times l_1$ matrix, with $m_1 + l_1 = n$ and such that the $n \times n$-matrix $[Y_1, Z_1]$ is nonsingular.

The matrices M and N are partitioned accordingly:

$$M = \begin{bmatrix} M_1 \\ M_2 \end{bmatrix}, \quad N = \begin{bmatrix} N_1 \\ N_2 \end{bmatrix} \tag{A.7}$$

and it follows from (A.4) that

$$Y_1M_1 - Z_1N_1 = Z_2N_2 - Y_2M_2. \tag{A.8}$$

There exist $\mathscr{C}'$ maps $\eta : \mathscr{N}_1 \cap R^m_+ \to S_1$ and $\zeta : \mathscr{N}_2 \cap R^l_+ \to S_2$ satisfying

$$\eta(\tau) = Y\tau + o(\tau), \quad \tau \to 0; \qquad \zeta(\sigma) = Z\sigma + o(\sigma), \quad \sigma \to 0. \tag{A.9}$$

It follows from Lemma A.2 that these maps can be extended to $\mathscr{C}'$ maps defined in a neighborhood of the origin. These extensions will also be denoted η and ζ and (A.9) is still valid. According to the partitioning of Y and Z we write

$$\tau = \begin{bmatrix} \tau_1 \\ \tau_2 \end{bmatrix}, \quad \sigma = \begin{bmatrix} \sigma_1 \\ \sigma_2 \end{bmatrix}.$$

Then (A.9) reads

$$\begin{aligned} \eta(\tau_1, \tau_2) &= Y_1\tau_1 + Y_2\tau_2 + o(\tau), & \tau \to 0, \\ \zeta(\sigma_1, \sigma_2) &= Z_1\sigma_1 + Z_2\sigma_2 + o(\sigma), & \sigma \to 0. \end{aligned} \tag{A.10}$$

Consider now the equation

$$f(\tau_1, \sigma_1, \tau_2, \sigma_2) := \eta(\tau) - \zeta(\sigma) = 0. \tag{A.11}$$

Since $f(\tau_1, \sigma_1, \tau_2, \sigma_2) = Y_1\tau_1 - Z_1\sigma_1 + Y_2\tau_2 - Z_2\sigma_2 + o(\tau, \sigma)$ and $[Y_1, -Z_1]$ is nonsingular, it follows from the implicit function theorem, that there exist $\mathscr{C}'$ functions φ and ψ of (τ_2, σ_2) in a neighborhood of $(0, 0)$ such that

$$f(\varphi(\tau_2, \sigma_2), \psi(\tau_2, \sigma_2), \tau_2, \sigma_2) = 0. \tag{A.12}$$

If $\varphi(\tau_2, \sigma_2) = A\tau_2 + B\sigma_2 + o(\tau_2, \sigma_2)$ and $\psi(\tau_2, \sigma_2) = C\tau_2 + D\sigma_2 + o(\tau_2, \sigma_2)$, then substitution in (A.12) yields

$$Y_1A - Z_1C + Y_2 = 0, \tag{A.13}$$

$$Y_1B - Z_1D - Z_2 = 0. \tag{A.14}$$

Multiplying (A.13) from the right with M_2 and (A.14) with N_2 and adding, we obtain $Y_1(AM_2 + BN_2) - Z_1(CM_2 + DN_2) = Z_2N_2 - Y_2M_2$. From (A.8) and the regularity of the matrix $[Y_1, -Z_1]$ it follows that

$$AM_2 + BN_2 = M_1, \qquad CM_2 + DN_2 = N_1. \tag{A.15}$$

We define $\bar{\xi}(\tau_2, \sigma_2) := \eta(\varphi(\tau_2, \sigma_2), \tau_2) = \zeta(\psi(\tau_2, \sigma_2), \sigma_2)$ and $\xi(\rho) := \bar{\xi}(M_2\rho, N_2\rho)$ for ρ in some $\mathscr{N} \subseteq R^k$. Then a short calculation yields $\xi(\rho) = X\rho + o(\rho)$, $\rho \to 0$.

We show that there exists a neighborhood $\mathscr{N}$ of 0, such that $\xi(\mathscr{N} \cap R^k_+) \subseteq S_1 \cap S_2$. It follows from (A.15), that $\varphi(M_2\rho, N_2\rho) = M_1\rho + o(\rho) \geqq 0$ for small $\rho \geqq 0$. Also, $M_2\rho \geqq 0$ for $\rho \geqq 0$. Hence $\xi(\rho) = \eta(\varphi(M_2\rho, N_2\rho), M_2\rho) \in S_1$ for small $\rho \geqq 0$. Similarly $\xi(\rho) \in S_2$ for small $\rho \geqq 0$. This completes the proof of Lemma 2.1. Q.E.D.

Appendix B. A generalization. If in § 5 we define Ω to be the set of measurable functions $u:[0, T] \to U$ and try to find the optimal control in this class, then the concept derived cone as defined thus far turns out to be too restrictive because of the differentiability condition occurring in the definition. Therefore we introduce two more general concepts.

DEFINITION B.1. If $b \in S \subseteq R^n$, a closed convex cone C is called a *generalized derived cone* of S at b, if for any collection of vectors $p_1, \cdots, p_k$ in ri C there exists a continuous map $\xi:\mathcal{N} \cap R^k_+ \to S$ satisfying

$$\xi(\tau) = b + \sum p_k\tau_k + o(\tau), \qquad \tau \to 0. \tag{B.2}$$

C is called a *weak derived cone* if corresponding to $p_1, \cdots, p_k$ in ri C a (not necessarily continuous) map ξ can be found satisfying (B.2). Instead of Lemma 2.1 we now have the following result.

LEMMA B.3. *Let S_1, S_2 be sets in R^n, $b \in S_1 \cap S_2$ and C_1, C_2 generalized derived cones of S_1, S_2 respectively. If C_1 and C_2 are not separated, then $C_1 \cap C_2$ is a weak derived cone of $S_1 \cap S_2$ at b.*

The proof of this lemma is completely the same as the proof of Lemma 2.1, except, that instead of the implicit function theorem, one uses the following property.

LEMMA B.4. *If $f:\mathcal{N}_1 \times \mathcal{N}_2 \to R^n$ is a continuous function, where $\mathcal{N}_1 \subseteq R^m$, $\mathcal{N}_2 \subseteq R^n$ are neighborhoods of the origin, and if*

$$f(x, y) = Ax + By + g(x, y), \tag{B.5}$$

where A and B are linear maps, B is not singular, and $g(x, y) = (|x| + |y|)\varepsilon(x, y)$, with $\varepsilon(x, y) \to 0$, $x, y \to 0$, then there exists a function $\eta:\mathcal{N} \to R^n$, where $\mathcal{N} \subseteq R^m$ is a neighborhood of the origin, such that

$$f(x, \eta(x)) = 0, \tag{B.6}$$

$$\eta(x) = -B^{-1}Ax + o(x), \qquad x \to 0. \tag{B.7}$$

Proof. We consider the map F_x defined by $F_x(y) := -B^{-1}Ax - B^{-1}g(x, y)$. For small $|x|$ and $|y|$, say $|x|, |y| \leqq \varepsilon$, we have $|B^{-1}\varepsilon(x, y)| \leqq \frac{1}{2}$. Then for $|x| < \delta := \varepsilon/(1 + 2\|B^{-1}A\|)$, we have $|F_x(y)| \leqq |B^{-1}Ax| + \frac{1}{2}(|x| + |y|) \leqq \varepsilon$. Since F_x is continuous, it follows by Brouwer's fixed-point theorem, that for $|x| < \delta$, there exists $y = \eta(x)$ such that $\eta(x) = F_x(\eta(x))$, hence $f(x, \eta(x)) = 0$. Let us show that (B.7) is satisfied: $|\eta(x)| = |B^{-1}Ax - B^{-1}g(x, \eta(x))| \leqq |B^{-1}Ax| + \frac{1}{2}(|x| + |\eta(x)|)$ and hence $|\eta(x)| \leqq M|x|$ for some $M > 0$. Consequently, $|\eta(x) + B^{-1}Ax| \leqq (M + 1)|x| \cdot |\varepsilon(x, \eta(x))| = o(x)$, $x \to 0$. Q.E.D.

Using Lemmas B.3 and 2.1 the following generalization of Theorem 1.5 is easily obtained.

THEOREM B.8. *Let $S_0, \cdots, S_k$ be sets in R^n, $b \in S := S_0 \cap \cdots \cap S_k$ and let C_0 be a generalized derived cone of S_0 at b and C_i derived cones of S_i at b for $i = 1, \cdots, k$. If a function $f:R^n \to R$ is $\mathscr{C}'$ at b and if f attains a maximum at b subject to $x \in S$,* *then there exist* $\rho \geqq 0, \psi_i \in C_i^0, i = 0, \cdots, k$, *such that*

$$(\rho, \psi_0, \cdots, \psi_k) \neq 0, \tag{B.9}$$

$$\rho \nabla f(b) = \psi_0 + \cdots + \psi_k. \tag{B.10}$$

Instead of one constraint set with a generalized derived cone and an arbitrary number of sets with derived cones, we can also derive the polar rule for the case that there are two constraint sets with generalized derived cones and no other constraints. However, contrary to Lemma 2.1, Lemma B.3 cannot be used repeatedly, so that we cannot derive necessary conditions in the case where more sets with generalized derived cones occur.

In [2], Canon, Cullum and Polak consider the following problem: Minimize $f(z)$ subject to the constraints $z \in \Omega$, $r(z) = 0$, where $f: R^n \to R$ and $r: R^n \to R^m$ are $\mathscr{C}'$ functions and $\Omega \subseteq R^n$ is an arbitrary set. Applying Theorem B.8 to this problem, we obtain the necessary condition given in [2]. In [2], this result is applied to a number of mathematical programming problems and discrete-time optimal control problems. In all these problems, Theorem 1.5 can be applied in a simpler and more direct way, as we have seen in §§ 3 and 4. The result of [2] is extended to infinite-dimensional spaces in order to derive the maximum principle (see [3, Appendix B]).

A slightly more general result is obtained in [8, Chap. 4, Thm. 3.2], where also inequality constraints are given. Both results are easily obtained from Theorem B.8. For example, if the problem is to minimize $f(z)$ subject to the constraints $r(z) = 0$, $z \in \Omega$, where Ω is an arbitrary set, and $r: \Omega \to R^m$, then the following equivalent finite-dimensional problem is obtained by considering the map $F: \Omega \to R^{m+1}$, defined by $F(z) := (f(z); r(z))$: maximize y_0, subject to $y = (y_0, \cdots, y_m)' \in F(\Omega)$, $y_1 = \cdots = y_m = 0$."

It should be remarked, that it is usually simpler to apply the finite-dimensional result directly to the problem essentially making the same substitution, as is done in § 5. A similar method is used in [1] to derive the continuous maximum principle from the results in [2].

The results of H. Halkin and L. W. Neustadt [5], [6], [14], [15] are of a more general scope than Theorem B.8.

Acknowledgment. The author wishes to thank J. L. de Jong for many stimulating discussions about the paper.

REFERENCES

[1] E. J. Balder, *A short proof of the Pontryagin maximum principle for fixed time continuous optimal control problems using a fundamental theorem of Canon, Cullum and Polak*, Tech. Note, Department of Mathematics, Technological University Eindhoven, Eindhoven, The Netherlands.

[2] M. Canon, C. Cullum and E. Polak, *Constrained minimization problems in finite dimensional spaces*, this Journal, 4 (1966), pp. 528–547.

[3] ———, *Theory of Optimal Control and Mathematical Programming*, McGraw-Hill, New York, 1970.

[4] H. Halkin, *A maximum principle of the Pontryagin type for systems described by nonlinear difference equations*, this Journal, 4 (1966), pp. 90–111.

[5] ———, *A satisfactory treatment of equality and operator constraints in the Dubovitskii–Milyutin optimization formalism*, J. Optimization Theory Appl., 6 (1970), pp. 138–149.

[6] H. Halkin and L. W. Neustadt, *General necessary conditions for optimization problems*, Proc. Nat. Acad. Sci. U.S.A., 56 (1966), pp. 1066–1071.

[7] M. L. J. Hautus, *Optimal control of differential systems with discontinuous right-hand side*, Doctoral thesis, Technological University Eindhoven, Eindhoven, The Netherlands, 1970.

[8] M. R. Hestenes, *Calculus of Variations and Optimal Control Theory*, John Wiley, New York, 1966.

[9] H. Holtzman, *On a maximum principle for nonlinear discrete-time systems*, IEEE Trans. Automatic Control, AC 11 (1966), pp. 30–35.

[10] B. W. Jordan and E. Polak, *Theory of a class of discrete optimal control systems*, J. Electron. Contr., 17 (1964), pp. 697–713.

[11] E. B. Lee and L. Markus, *Foundations of Optimal Control Theory*, John Wiley, New York, 1967.

[12] O. L. Mangasarian, *Nonlinear Programming*, McGraw-Hill, New.York, 1969.

[13] O. L. Mangasarian and S. Fromovitz, *The Fritz John necessary optimality conditions in the presence of equality and inequality constraints*, J. Math. Anal. Appl., 17 (1967), pp. 37–47.

[14] L. W. Neustadt, *An abstract variational theory with applications to a broad class of optimization problems, I General theory, II Applications*, this Journal, 4 (1966), pp. 505–527, 5 (1967), pp. 90–137.

[15] ———, *A general theory of extremals*, J. Comput. System Sci., 3 (1969), pp. 57–92.

[16] L. S. Pontryagin, V. G. Boltyanskii, R. V. Gamkrelidze and E. F. Mishchenko, *The Mathematical Theory of Optimal Processes*, Interscience, New York, 1962.

[17] R. Pallu de Barriere, *On the cost of constraints in dynamical optimization*, Mathematical Theory of Control, Academic Press, New York, 1967, pp. 246–250.

[18] R. T. Rockafellar, *Convex Analysis*, Princeton University Press, Princeton, N.J., 1970.

SIAM J. CONTROL
Vol. 11, No. 4, November 1973

ON THE EQUIVALENCE OF CONTROL SYSTEMS AND THE LINEARIZATION OF NONLINEAR SYSTEMS*

ARTHUR J. KRENER†

Abstract. Given two control systems where the control enters linearly, a necessary and sufficient condition is derived that these systems be locally diffeomorphic, i.e., that there exist a local diffeomorphism between the state spaces which carries a trajectory of the first system for each control into the trajectory of the second system for the same control. As a corollary we derive necessary and sufficient conditions for a system to be locally diffeomorphic to a linear system.

1. Introduction. Consider the two control systems

$$\begin{aligned} \dot{x} &= a_0(x) + \sum_{i=1}^{k} u_i(t)a_i(x), \\ x(0) &= x^0, \end{aligned} \tag{1}$$

and

$$\begin{aligned} \dot{y} &= b_0(y) + \sum_{i=1}^{k} u_i(t)b_i(y), \\ y(0) &= y^0, \end{aligned} \tag{2}$$

where $x = (x_1, \cdots, x_m)$, $y = (y_1, \cdots, y_n)$ are vectors, $a_0(x), \cdots, a_k(x), b_0(x), \cdots, b_k(x)$ are analytic vector-valued functions and $u(t) = (u_1(t), \cdots, u_k(t))$ is a bounded measurable control.

The purpose of this paper is to give necessary and sufficient conditions that these two systems be equivalent, i.e., that there exist a local diffeomorphism from x-space to y-space which takes the solution of (1) for each control into the solution of (2) for the same control. As a corollary we derive necessary and sufficient conditions that there exist a local diffeomorphism which carries a nonlinear system into a linear one.

2. Preliminaries. If $a_i(x), a_j(x)$ are as above we define the Lie bracket $[a_i, a_j](x)$, another analytic vector-valued function, by

$$[a_i, a_j](x) = \frac{\partial a_j}{\partial x}(x)a_i(x) - \frac{\partial a_i}{\partial x}(x)a_j(x),$$

where $(\partial a_j/\partial x)(x)$ is the matrix of partial derivatives at x. Suppose $t, x \mapsto \alpha_i(t)x$ is the family of integral curves of $a_i(x)$, that is, $(d/dt)\alpha_i(t)x = a_i(\alpha_i(t)x)$ and $\alpha_i(0)x = x$. Then for fixed t, the map $x \mapsto \alpha_i(-t)x$ is a diffeomorphism from a neighborhood of $\alpha_i(t)x^0$ onto a neighborhood of x^0 and hence has a tangent map which we denote by $\alpha_i(-t)_*$. The derivative of the vector-valued curve $t \mapsto \alpha_i(-t)_* a_j(\alpha_i(t)x^0)$ at $t = 0$ is $[a_i, a_j](x^0)$ (Bishop and Crittenden [1, p. 17]). Since a_i, a_j are analytic, we obtain the Taylor series expansion $\alpha_i(-t)_* a_j(\alpha_i(t)x^0) = \sum_{h=0}^{\infty} (t^h/h!) ad^h(a_i)a_j(x^0)$, where $ad^0(a_i)a_j(x^0) = a_j(x^0)$ and $ad^h(a_i)a_j(x^0) = [a_i, ad^{h-1}(a_i)a_j](x^0)$.

* Received by the editors October 31, 1972, and in revised form January 8, 1973.

† Department of Mathematics, University of California–Davis, Davis, California 95616.

Following Haynes and Hermes [2] we define $D^0(A)$ to be a set of functions $\{a_i : i = 0, \cdots, k\}$ and $D^j(A) = D^{j-1}(A) \cup \{[a_i, c] : i = 0, \cdots, k, c \in D^{j-1}(A)\}$, for $j \leq 1$. The *completed system* of A is $D(A) = \bigcup_{j \geqq 0} D^j(A)$, and we define $D(A)_x = \{c(x) : c \in D(A)\} \subseteq \mathbb{R}^m$. The *rank* r *of* $D(A)$ *at* x is just the dimension of the span $D(A)_x$.

THEOREM (Nagano [4]). *Let the completed system of* (1) *have rank* r *at* x^0. *Then there exists a submanifold* M *of dimension* r *through* x^0, *which carries* (1). *That is, if* $u(t)$ *is any bounded measurable control and* $x(t)$ *is the corresponding solution of* (1), *then for some* $\varepsilon > 0$, $x(t) \in M$ *for* $|t| < \varepsilon$.

For generalizations of this result see Krener [3].

3. Equivalent systems.

THEOREM 1. *Consider the systems* (1) *and* (2). *Let* M *and* N *be submanifolds which carry* (1) *and* (2) *at* x^0 *and* y^0 *respectively. There exists a linear map* $l : \operatorname{span} D(A)_{x^0} \to \operatorname{span} D(B)_{y^0}$ *such that* $l(a_i(x^0)) = b_i(y^0)$ *for* $i = 0, \cdots, k$ *and*

$$l([a_{i_1}, \cdots [a_{i_{h-1}}, a_{i_h}] \cdots](x^0)) = [b_{i_1}, \cdots [b_{i_{h-1}}, b_{i_h}] \cdots](y^0)$$

for $h \leq 2$ *and* $1 \leqq i_j \leqq k$ *if and only if there exist neighborhoods* U *and* V *of* x^0 *and* y^0 *in* M *and* N *and an analytic map* $\lambda : U \to V$ *such that* λ *carries* (1) *into* (2). *That is, if* $x(t)$ *and* $y(t)$ *are the solutions of* (1) *and* (2) *for the same control* $u(t)$ *and* $x(t) \in U$ *for* $|t| < \varepsilon$, *then* $y(t) = \lambda(x(t)) \in V$ *for* $|t| < \varepsilon$. *Furthermore* l *is a linear isomorphism if and only if* λ *is a local diffeomorphism.*

Proof. We start by assuming l exists and constructing λ. Since the theorem is local in nature, we can assume that $M = \mathbb{R}^m$ and $N = \mathbb{R}^n$, then $\operatorname{span} D(A)_{x^0} = \mathbb{R}^m$ and $\operatorname{span} D(B)_{y^0} = \mathbb{R}^n$. Let $c_1(x^0), \cdots, c_h(x^0)$ be a maximal linearly independent subset of $D^0(A)_{x^0}$. Let $d_1(y), \cdots, d_h(y)$ be the corresponding elements of $D^0(B)$, that is, if $c_i(x) = a_j(x)$ then $d_i(y) = b_j(y)$. We choose $c_{h+1}(x), \cdots, c_m(x)$ from $D(A)$ so that $c_1(x^0), \cdots, c_m(x^0)$ forms a basis for $\mathbb{R}^m$. Let $d_{h+1}(y), \cdots, d_m(y)$ be the corresponding elements of $D(B)$, that is, if $c_i(x) = [a_{j_1}, \cdots [a_{j_{l-1}}, a_{j_l}] \cdots](x)$, then $d_i(y) = [b_{j_1}, \cdots [b_{j_{l-1}}, b_{j_l}] \cdots](y)$. Let $t, x \mapsto \alpha_i(t)x$, be the family of integral curves of $c_i(x)$ for $i = 1, \cdots, m$. That is $(d/dt)\alpha_i(t)x = c_i(\alpha_i(t)x)$ and $\alpha_i(0)x = x$. Similarly $t, y \mapsto \beta_i(t)y$, is defined by $(d/dt)\beta_i(t)y = d_i(\beta_i(t)y)$ and $\beta_i(0)y = y$, for $i = 1, \cdots, m$. Let $s = (s_1, \cdots, s_m)$ and define maps $g_1 : s \mapsto x$ and $g_2 : s \mapsto y$ by $g_1(s) = \alpha_m(s_m) \cdots \alpha_2(s_2)\alpha_1(s_1)x^0$ and $g_2(s) = \beta_m(s_m) \cdots \beta_2(s_2)\beta_1(s_1)y^0$. Then $(\partial g_1/\partial s_i)(0) = c_i(x^0)$, so g_1 has an inverse $g_1^{-1} : x \mapsto s$ defined for x in some neighborhood U of x^0. Let $\lambda : x \mapsto y$ be defined on U by $\lambda = g_2 \circ g_1^{-1}$.

We must now show that if $x(t)$ and $y(t)$ are the solutions of (1) and (2) respectively for the same control $u(t)$, then $\lambda(x(t)) = y(t)$. Since $\lambda(x(0)) = \lambda(x^0) = y^0 = y(0)$ it suffices to show that $(d/dt)\lambda(x(t)) = (d/dt)y(t)$ or $\lambda_*(\dot{x}(t)) = \dot{y}(t)$, where λ_* is the tangent map to λ at $x(t)$. This is true if $\lambda_*(a_i(x)) = b_i(\lambda(x))$, $i = 1, \cdots, k$, for all $x \in U$, which in turn would follow if $\lambda_*(c_i(x)) = d_i(\lambda(x))$, $i = 1, \cdots, m$, for all $x \in U$.

To show this we let $x = g_1(s)$, $x^i = g_1(s_1, \cdots, s_i, 0, \cdots, 0)$, for $i = 1, \cdots, m$, $y = \lambda(x) = g_2(s)$ and $y^i = g_2(s_1, \cdots, s_i, 0, \cdots, 0)$, for $i = 1, \cdots, m$. Then $x^m = x$ and for $i = 1, \cdots, m$, the map $\alpha_i(-s_i)(\cdot)$ takes x^i into x^{i-1} and is a local diffeomorphism with tangent at x^i denoted by $\alpha_i(-s_i)_*$. Similarly $y^m = y$, and the map

$\beta_i(s_i)(\cdot)$ takes y^{i-1} into y^i and is a local diffeomorphism with tangent at y^{i-1} denoted by $\beta_i(s_i)_*$.

We now show that $\lambda_* = \beta_m(s_m)_* \cdots \beta_1(s_1)_* \, l \, \alpha_1(-s_1)_* \cdots \alpha_m(-s_m)_*$. Since $\partial g_1(s)/\partial s_i$ forms a basis for $\mathbb{R}^m$, it suffices to show that the right side applied to $\partial g_1(s)/\partial s_i$ yields $\lambda_*(\partial g_1(s)/\partial s_i)$ which equals $\partial g_2(s)/\partial s_i$. But $\partial g_1(s)/\partial s_i = \alpha_m(s_m)_* \cdots \alpha_i(s_i)_* c_i(x^{i-1})$ and $\partial g_2(s)/\partial s_i = \beta_m(s_m)_* \cdots \beta_i(s_i)_* d_i(y^{i-1})$ so

$$\begin{aligned}
&\beta_m(s_m)_* \cdots \beta_1(s_1)_* l \alpha_1(-s_1)_* \cdots \alpha_m(-s_m)_* \frac{\partial g_1(s)}{\partial s_i} \\
&\quad = \beta_m(s_m)_* \cdots \beta_1(s_1)_* l \alpha_1(-s_1)_* \cdots \alpha_{i-1}(-s_{i-1})_* c_i(x^{i-1}) \\
&\quad = \beta_m(s_m)_* \cdots \beta_1(s_1)_* l \sum \frac{(s_1)^{h_1}}{h_1!} ad^{h_1}(c_1) \left(\cdots \sum \frac{(s_{i-1})^{h_{i-1}}}{(h_{i-1})!} ad^{h_{i-1}}(c_{i-1}) c_i \cdots \right) (x^0) \\
&\quad = \beta_m(s_m)_* \cdots \beta_1(s_1)_* \sum \frac{(s_1)^{h_1}}{h_1!} ad^{h_1}(d_1) \left(\cdots \sum \frac{(s_{i-1})^{h_{i-1}}}{(h_{i-1})!} ad^{h_{i-1}}(d_{i-1}) d_i \cdots \right) (y^0) \\
&\quad = \beta_m(s_m)_* \cdots \beta_i(s_i)_* d_i(y^{i-1}) = \frac{\partial g_2(s)}{\partial s_i}.
\end{aligned}$$

This implies that

$$\begin{aligned}
\lambda_*(c_i(x^m)) &= \beta_m(s_m)_* \cdots \beta_1(s_1)_* l \alpha_1(-s_1)_* \cdots \alpha_m(-s_m)_* c_i(x^m) \\
&= \beta_m(s_m)_* \cdots \beta_1(s_1)_* l \sum \frac{(s_1)^{h_1}}{h_1!} ad^{h_1}(c_1) \left(\cdots \sum \frac{(s_m)^{h_m}}{h_m!} ad^{h_m}(c_m) c_i \cdots \right) (x^0) \\
&= \beta_m(s_m)_* \cdots \beta_1(s_1)_* \sum \frac{(s_1)^{h_1}}{h_1!} ad^{h_1}(d_1) \left(\cdots \sum \frac{(s_m)^{h_m}}{h_m!} ad^{h_m}(d_m) d_i \cdots \right) (y^0) \\
&= d_i(y^m).
\end{aligned}$$

Notice that $\lambda_*(c_i(x^0)) = d_i(y^0)$ so $l = \lambda_*$ at x^0. It follows by the inverse function theorem that if l is a linear isomorphism then λ is a local diffeomorphism.

As for the converse, if λ exists and $\lambda(x(t)) = y(t)$ where $x(t)$ and $y(t)$ are the solutions of (1) and (2) for the same control $u(t)$, then clearly $\lambda_*(a_i(x)) = b_i(\lambda(x))$. It is a standard result of differential geometry (Bishop and Crittenden [1, p. 14]) that if $\lambda_*(c_i(x)) = d_i(\lambda(x))$, $i = 1, 2$, then $\lambda_*([c_1, c_2](x)) = [d_1, d_2](\lambda(x))$, and so $l = \lambda_*$ at x^0 satisfies the required condition. Q.E.D.

Remark. Since $g_1(s)$ covers a neighborhood of x^0 in M, the map λ is uniquely determined in that neighborhood by the condition that it take system (1) into system (2). Furthermore if M is connected and simply connected, then λ can be extended uniquely to a map defined on all M by standard arguments. See Example 3 below.

Example 1. Consider the two systems

$$\begin{aligned}
\dot{x}_1 &= u, & \dot{y}_1 &= u, \\
\dot{x}_2 &= u \cdot t, & \dot{y}_2 &= y_1.
\end{aligned}$$

Since the right-hand side of the first system depends on t, we introduce a new variable $x_0 = t$.

$$\begin{aligned}\dot{x}_0 &= 1,\\ \dot{x}_1 &= u,\\ \dot{x}_2 &= u\cdot x_0,\end{aligned}\qquad a_0(x)=\begin{pmatrix}1\\0\\0\end{pmatrix},\quad a_1(x)=\begin{pmatrix}0\\1\\x_0\end{pmatrix},\quad [a_0,a_1](x)=\begin{pmatrix}0\\0\\1\end{pmatrix},$$

$$b_0(y)=\begin{pmatrix}0\\y_1\end{pmatrix},\quad b_1(y)=\begin{pmatrix}1\\0\end{pmatrix},\quad [b_0,b_1](y)=\begin{pmatrix}0\\-1\end{pmatrix},$$

and all other brackets are zero.

For initial points $x^0 = (x_0^0, x_1^0, x_2^0)$, $y^0 = (y_1^0, y_2^0)$, let $l:\mathbb{R}^3 \mapsto \mathbb{R}^2$ be given by the matrix

$$\begin{pmatrix}0 & 1 & 0\\ y_1^0 & x_0^0 & -1\end{pmatrix}.$$

The hypotheses of Theorem 1 are satisfied and λ can be constructed as in the theorem. Let α_1, α_2 and α_3 be the families of integral curves of a_0, a_1 and $[a_0, a_1]$ and β_1, β_2 and β_3 be the families of integral curves of b_0, b_1 and $[b_0, b_1]$. Then

$$g_1(s_1,s_2,s_3) = \alpha_3(s_3)\alpha_2(s_2)\alpha_1(s_1)x^0 = (x_0^0 + s_1, x_1^0 + s_2, x_2^0 + (x_0^0 + s_1)s_2 + s_3),$$

$$g_2(s_1,s_2,s_3) = \beta_3(s_3)\beta_2(s_2)\beta_1(s_1)y^0 = (y_1^0 + s_2, y_2^0 + s_1y_1^0 - s_3),$$

and

$$\lambda(x) = g_2(g_1^{-1}(x)) = (y_1^0 + x_1 - x_1^0, y_2^0 + (x_0 - x_0^0)y_1^0 - (x_2 - x_2^0) + x_0(x_1 - x_1^0)).$$

Notice that $M = \mathbb{R}^3$, $N = \mathbb{R}^2$ and λ is defined for all $x \in \mathbb{R}^3$ and is onto $\mathbb{R}^2$. In fact, if we introduce a time coordinate $y_0 = t$ into the second system, then $N = \mathbb{R}^3$ and λ becomes a diffeomorphism from $\mathbb{R}^3$ onto $\mathbb{R}^3$:

$$\begin{aligned}\lambda(x) = (y_0^0 + x_0 - x_0^0, y_1^0 + x_1 - x_1^0, y_2^0 + (x_0 - x_0^0)y_1^0 - (x_2 - x_2^0)\\ + x_0(x_1 - x_1^0)).\end{aligned}$$

Example 2. Suppose we replace the second system of Example 1 with one similar to that of Haynes and Hermes [2]·

$$\begin{aligned}\dot{y}_0 &= 1,\\ \dot{y}_1 &= u,\\ \dot{y}_2 &= u\cdot y_0\cdot y_2,\end{aligned}\qquad b_0(y)=\begin{pmatrix}1\\0\\0\end{pmatrix},\quad b_1(y)=\begin{pmatrix}0\\1\\y_0\cdot y_2\end{pmatrix}\quad\text{and}\quad [b_0,b_1](y)=\begin{pmatrix}0\\0\\y_2\end{pmatrix},$$

and all other brackets are identically zero. The rank of $D(B)$ is 3 except at points where $y_2 = 0$, where it is 2. The system splits $\mathbb{R}^3$ into three disjoint manifolds $N_+ = \{y : y_2 > 0\}$, $N_0 = \{y : y_2 = 0\}$ and $N_- = \{y : y_2 < 0\}$. A trajectory of this system must lie wholly within one of these manifolds.

For initial points x^0 and y^0 we define l by

$$\begin{pmatrix}1 & 0 & 0\\ 0 & 1 & 0\\ 0 & (y_0^0 - x_0^0)y_2^0 & y_2^0\end{pmatrix}$$

and construct λ as before:

$$\lambda(x) = (y_0^0 + x_0 - x_0^0, y_1^0 + x_1 - x_1^0, y_2^0 \exp((y_0^0 - x_0^0)(x_1 - x_1^0) + x_2 - x_2^0)).$$

Notice if $y^0 \in N_+(N_-)$, then λ is a diffeomorphism $\lambda:\mathbb{R}^3 \to N_+(N_-)$. If $y^0 \in N_0$, then $\lambda:\mathbb{R}^3 \to N_0$ is onto.

Example 3. Consider the systems

$$\begin{aligned} \dot{x}_1 &= -ux_2, & \dot{y}_1 &= u, \\ \dot{x}_2 &= ux_1, & & \\ a_1 &= \begin{pmatrix} -x_2 \\ x_1 \end{pmatrix}, & b_1 &= (1), \end{aligned}$$

and of course there are no nontrivial brackets. If $x^0 = (1, 0)$ and $y^0 = 0$, then $M = \{(x_1, x_2): x_1^2 + x_2^2 = 1\}$ and $N = \mathbb{R}$. So $l = \begin{pmatrix} 0 \\ 1 \end{pmatrix}$ satisfies Theorem 1 and λ is defined in a neighborhood of $(1, 0)$ on M by $\lambda(x_1, x_2) = \arctan(x_2/x_1)$. It is clear that λ cannot be extended to a map on all of M.

4. The linearization of nonlinear systems. Consider the linear control system

$$\dot{y} = F(t)y(t) + G(t)u(t) + h(t), \tag{3}$$

where F and G are matrices, y and h are vectors and u is the control vector. As before we introduce time as a coordinate, $y_0 = t$. It is well known that there exists a change of the y coordinates which carries (3) into

$$\dot{y} = b_0 + \sum_{i=1}^{k} u_i b_i(y_0), \tag{4}$$

where

$$b_0 = \begin{pmatrix} 1 \\ 0 \\ \cdot \\ \cdot \\ \cdot \\ 0 \end{pmatrix}, \quad b_i = \begin{pmatrix} 0 \\ * \\ \cdot \\ \cdot \\ \cdot \\ * \end{pmatrix}, \quad i = 1, \cdots, k \quad \text{and} \quad y_0^0 = 0,$$

and where * denotes some real-valued function of y_0 alone.

The question we now answer is when does there exist a transformation $\lambda: x \mapsto y$ which carries a nonlinear system (1) into a linear system (4).

THEOREM 2. *Consider the system* (1). *Let n = rank of $D(A)_{x^0}$ and let M be the n-dimensional manifold which carries* (1). *There exists a linear system* (4), *a neighborhood U of x^0 in M, a neighborhood V of $y^0 = 0$ in $\mathbb{R}^n$ and a diffeomorphism $\lambda: U \mapsto V$ carrying* (1) *into* (4) *if and only if for all $1 \leqq i, j \leqq k$ and for all $h \geqq 0$,* $[a_i, ad^h(a_0)a_j](x^0) = 0$.

Proof. Suppose the system (4) and λ exist. Then λ_*, the tangent to λ at 0, is one-to-one and

$$\lambda^*([a_i, ad^h(a_0)a_j](x^0)) = [b_i, ad^h(b_0)b_j](0).$$

Then by induction for $h \geqq 0$,

$$ad^h(b_0)b_j = \begin{pmatrix} 0 & \cdots & 0 \\ * & & \\ \vdots & 0 & \\ \vdots & & \\ * & & \end{pmatrix}\begin{pmatrix} 1 \\ 0 \\ \vdots \\ \vdots \\ 0 \end{pmatrix} - \begin{pmatrix} \\ \\ 0 \\ \\ \\ \end{pmatrix}\begin{pmatrix} 0 \\ * \\ \vdots \\ \vdots \\ * \end{pmatrix} = \begin{pmatrix} 0 \\ * \\ \vdots \\ \vdots \\ * \end{pmatrix}$$

and

$$[b_i, ad^h(b_0)b_j] = \begin{pmatrix} 0 & \cdots & 0 \\ * & & \\ \vdots & 0 & \\ \vdots & & \\ * & & \end{pmatrix}\begin{pmatrix} 0 \\ * \\ \vdots \\ \vdots \\ * \end{pmatrix} - \begin{pmatrix} 0 & \cdots & 0 \\ * & & \\ \vdots & 0 & \\ \vdots & & \\ * & & \end{pmatrix}\begin{pmatrix} 0 \\ * \\ \vdots \\ \vdots \\ * \end{pmatrix} = \begin{pmatrix} 0 \\ \vdots \\ \vdots \\ \vdots \\ 0 \end{pmatrix}$$

and it follows that $[a_i, ad^h(a_0)a_j] = 0$.

On the other hand if $[a_i, ad^h(a_0)a_j](x^0) = 0$, we construct (4) as follows. Let $s, x \to \alpha_0(s)x$ be the family of integral curves of a_0. Define the system (4) by setting

$$b_0 = \begin{pmatrix} 1 \\ 0 \\ \vdots \\ \vdots \\ 0 \end{pmatrix} \quad \text{and} \quad b_j(y_0) = \alpha_0(-y_0)_* a_j(\alpha_0(y_0)x^0), \qquad j = 1, \cdots, k.$$

The Taylor series expansion of $b_j(y_0) = \sum_{h=0}^{\infty} ((y_0)^h/h!) ad^h(a_0)a_j(x^0)$ and it follows that

$$ad^h(b_0)b_j(0) = \frac{d^h}{dy_0^h} b_j(0) = ad^h(a_0)a_i(x^0) \quad \text{for } j = 1, \cdots, k \text{ and } h \geqq 0.$$

Also by hypothesis $[a_i, ad^h(a_0)a_j](x^0) = 0$ and we showed above for systems of type (4), $[b_i, ad^h(b_0)b_j](0) = 0$. Therefore the hypotheses of Theorem 1 are satisfied with l = identity map, and so we can construct λ. Q.E.D.

Example 4. Consider the nonlinear system

$$\dot{x}_1 = 1 + u \cdot x_3,$$

$$\dot{x}_2 = x_1^2 x_2 + u,$$

$$\dot{x}_3 = x_3,$$

$$a_0 = \begin{pmatrix} 1 \\ x_1^2 x_2 \\ x_3 \end{pmatrix}, \quad a_1 = \begin{pmatrix} x_3 \\ 1 \\ 0 \end{pmatrix}, \quad [a_0, a_1] = \begin{pmatrix} x_3 \\ -2x_1x_2x_3 - x_1^2 \\ 0 \end{pmatrix}$$

$$\text{and} \quad [a_1[a_0, a_1]] = \begin{pmatrix} 0 \\ -2x_2x_3^2 - 4x_1x_3 \\ 0 \end{pmatrix}.$$

Therefore the system is not linearizable in general. However if $x_3^0 = 0$, then the system is carried by $M = \{x : x_3 = 0\}$ and on this submanifold

$$a_0 = \begin{pmatrix} 1 \\ x_1^2 x_2 \\ 0 \end{pmatrix}, \quad a_1 = \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}, \quad [a_0, a_1] = \begin{pmatrix} 0 \\ -x_1^2 \\ 0 \end{pmatrix}, \quad [a_1[a_0, a_1]] = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix},$$

$$ad^2(a_0)a_1 = \begin{pmatrix} 0 \\ -2x_1 \\ 0 \end{pmatrix}, \quad [a_1, ad^2(a_0)a_1] = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \quad ad^3(a_0)a_1 = \begin{pmatrix} 0 \\ -2 \\ 0 \end{pmatrix}.$$

All higher brackets are zero, so the system is linearizable. We do not have to compute λ to describe the equivalent linear system. For example if $x^0 = (0, 0, 0)$ we define

$$b_0 = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}, \quad b_1(y_0) = \sum_{h=0}^{\infty} \frac{y_0^h}{h!} ad^h(a_0)a_1(x^0) = \begin{pmatrix} 0 \\ 1 - \dfrac{y_0^3}{3} \\ 0 \end{pmatrix}.$$

Since $y_0 = t$ and the y_2-coordinate is superfluous, this becomes

$$\dot{y}_1 = u(1 - t^3/3), \quad y_1(0) = 0.$$

REFERENCES

[1] R. L. Bishop and R. J. Crittenden, *Geometry of Manifolds*, Academic Press, New York, 1964.

[2] G. W. Haynes and H. Hermes, *Nonlinear controllability via Lie theory*, this Journal, 8 (1970), pp. 450–460.

[3] A. J. Krener, *A generalization of Chow's theorem and the bang-bang theorem to nonlinear control problems*, this Journal, 11 (1973).

[4] T. Nagano, *Linear differential system with singularities and an application to transitive Lie algebras*, J. Math. Soc. Japan, 18 (1966), pp. 398–404.

SIAM J. CONTROL
Vol. 11, No. 4, November 1973

ERRATUM: AN EXISTENCE THEOREM FOR LAGRANGE PROBLEMS WITH UNBOUNDED CONTROLS AND A SLENDER SET OF EXCEPTIONAL POINTS*

L. CESARI,† J. R. LA PALM‡ AND D. A. SANCHEZ§

On page 601 we reported from other papers [1c, (7.i); 1d, (6.5)] the statement of criterion (5.3). This criterion is not correct as stated. For a correct version and proof see [1e]. This criterion (5.3) was used on page 602 (line 7 f.t.b), to prove that an existence theorem for free problems (Theorem 6.3) can be derived from our existence theorems for Lagrange problems. For this derivation we do not need (5.3). Indeed, for free problems, that is, when $m = n, f = u$, $U = E_n$, condition (α) is trivial, and property (X) reduces to seminormality. In the proof of (6.3) we proved f_0 to be normal at $(\bar{t}, \bar{x})$, hence seminormal at the same point, and the easier criterion (5.2) of page 600 can be applied, instead of (5.3).

* This Journal, 9 (1971), pp. 590–605. Received by the editors October 2, 1972.

† Department of Mathematics, University of Michigan, Ann Arbor, Michigan 48104.

‡ Department of Mathematics, Wayne State University, Detroit, Michigan 48202.

§ Department of Mathematics, University of California, Los Angeles, California 90024.