

DIRECT MINIMAL BALANCED REALIZATION FROM IMPULSE RESPONSE MATRICES USING I/O MAP DECOMPOSITION*

B. S. LEE† AND F. W. FAIRMAN†

Abstract. Fundamental properties of I/O maps for linear continuous-time systems are used to formulate a method for determining a minimal (balanced) realization directly from a given impulse response matrix.

Key words. finite dimensional systems, minimal realization, balanced realization

1. Introduction. Work on the nature and properties of balanced realizations of finite dimensional linear time-invariant systems has been ongoing since the initial introduction of this class of realizations by Moore [1], [2]. A minimal realization is said to be balanced if its controllability and observability Gramians are equal and diagonal. Recent work has generalized and extended the original concept to *LQG* feedback control systems [3] and to time-varying systems [4], [5].

However, the determination of a balanced realization for a given transfer function has not received much attention in the literature. Except for the method given in [6] for scalar transfer functions having simple poles, to date the general methods for obtaining a balanced realization given by Moore [1], [2] and Laub [7] require the determination of an initial minimal realization. Starting with the system's transfer function matrix, Moore's algorithm involves the minimal realization of the given transfer function matrix, the solution of two Lyapunov equations and the singular value decomposition of two positive definite matrices. A step-by-step description of Moore's algorithm is given as

- (i) Determine a minimal realization.
- (ii) Solve the Lyapunov equation for the controllability Gramian of the realization determined in (i).
- (iii) Singular value decompose the Gramian determined in (ii).
- (iv) Calculate the realization which results from changing the coordinates through the use of a coordinate transformation matrix made up from the matrices determined in (iii).
- (v) Solve the Lyapunov equation for the observability Gramian for the realization determined in (iv).
- (vi) Singular value decompose the Gramian determined in (v).
- (vii) Calculate the balanced realization which results from changing the coordinates through the use of a co-ordinate transformation matrix composed from matrices determined in (vi).

On the other hand, Laub's method [7] requires both the controllability and observability Gramians of the initial realization (i.e., the solution of two Lyapunov equations). Moreover, the requirement for two singular value decompositions in [2] is replaced by a Cholesky decomposition and an eigenvalue-eigenvector problem for a real symmetric matrix in [7].

Starting with the system's transfer function matrix, the algorithm developed here for achieving a minimal (balanced) realization involves the partial fraction expansion of the given transfer function matrix, the Cholesky decomposition of a positive definite

* Received by the editors November 22, 1983, and in revised form November 30, 1984. This work was supported by the Natural Sciences and Engineering Research Council of Canada.

† Department of Electrical Engineering, Queen's University, Kingston, Ontario, Canada K7L 3N6.

real symmetric matrix and the singular value decomposition of a real matrix. The steps involved in the execution of this algorithm are given as

(1) Determine the poles and partial fraction coefficient matrices of the transfer function matrix.

(2) Form three real matrices \bar{K} , \bar{V} and $\bar{\Lambda}$ directly from the poles and partial fraction coefficient matrices.

(3) Perform Cholesky decomposition on the real positive definite matrix \bar{V} , i.e. $\bar{V} = \bar{L}\bar{L}^T$.

(4) Singular value decompose

$$\bar{L}_q^T \bar{K}^T [I_p \otimes \bar{V}] \bar{K} \bar{L}_q = [U_1 \ U_2] \begin{bmatrix} \Sigma^2 & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} U_1^T \\ U_2^T \end{bmatrix}$$

where $\bar{L}_q = [I_q \otimes \bar{L}]$.

(5) Calculate the balanced realization (A, B, C^T) as

$$A = T_1^{-1} \bar{A} T_1, \quad B = T_1^{-1} \bar{B}, \quad C^T = \bar{C}^T T_1,$$

where

$$\begin{aligned} \bar{A} &= [I_q \otimes \bar{\Lambda}], & \bar{B} &= [I_q \otimes \gamma(0)], \\ \bar{C}^T &= [I_p \otimes \gamma(0)^T] \bar{K}, & T_1 &= \bar{L}_q U_1 \Sigma^{-1/2} \end{aligned}$$

and $\gamma(t)$ is a vector of time multiplied exponentials which are related to the poles of the transfer function matrix and $[\cdot \otimes \cdot]$ is the Kronecker product.

Notice that an arbitrary minimal realization is not required and that there is no need to solve any Lyapunov equations.

The development of the algorithm is carried out by considering single-input single-output (SISO) systems first. The algorithm developed for this class of systems is then extended to enable the balanced realization of multi-input, multi-output (MIMO) systems.

Notation peculiar to this paper which is employed in the sequel is stated now as follows: I_r is the r by r identity matrix; Γ^i denotes the i th diagonal block of some square block diagonal matrix Γ ; E is a parity matrix, i.e., E is a diagonal matrix with diagonal elements $\{e_i; i = 1, 2, \dots, n\}$ where $e_i \in \{1, -1\}$; j is context dependent being either a positive integer or $(-1)^{1/2}$.

2. General properties of I/O maps. Let the system be specified by its transfer function $h_L(s)$ as

$$(1) \quad h_L(s) = \frac{n(s)}{d(s)}$$

where $n(s)$, $d(s)$ are coprime polynomials with real coefficients and all zeros of $d(s)$ have nonzero negative real part (stability assumption). In the sequel interest will focus on the inverse transform of $h_L(s)$ (impulse response) which is given in the general form (for proper rational $h_L(s)$) as

$$(2) \quad h(t) = \sum_{i=1}^m \sum_{j=1}^{n_i} m_{ij} t^{j-1} e^{\lambda_i t} = \alpha^T \gamma(t)$$

where $\lambda_i, m_{ij} \in C^1$ and $\lambda_i \neq \lambda_j$ when $i \neq j$. Also $n = \sum_{i=1}^m n_i$ and

$$\alpha^T = [\alpha_1^T, \alpha_2^T, \dots, \alpha_m^T], \quad \alpha_i^T = [m_{i1}, m_{i2}, \dots, m_{in_i}],$$

$$\gamma^T(t) = [\gamma_1^T(t), \gamma_2^T(t), \dots, \gamma_m^T(t)], \quad \gamma_i^T(t) = [1, t, \dots, t^{n_i-1}] e^{\lambda_i t}.$$

Next recall that a minimal realization (A, b, c^T) of $h_L(s)$ must satisfy

$$(3) \quad h_L(s) = c^T(sI - A)^{-1}b, \quad h(t) = c^T e^{At}b,$$

where $A \in C^{n \times n}$, $c \in C^n$ and n is minimal.

Then the controllability and observability Gramians W_c, W_o over the infinite interval $[0, \infty)$ as well as the input and output maps $f_c(t), f_o(t)$ are defined as

$$(4) \quad W_c = \int_0^\infty f_c(t) f_c^T(t) dt, \quad f_c(t) = e^{At}b,$$

$$(5) \quad W_o = \int_0^\infty f_o(t) f_o^T(t) dt, \quad f_o(t) = e^{A^T t} c^T,$$

where $W_c, W_o \in R^{n \times n}$ are nonsingular.

The existence of W_c, W_o follows from stability of $h_L(s)$ and the nonsingularity of these matrices is a consequence of the minimality of the realization. When the realization is balanced W_c, W_o are equal and diagonal, i.e., $W_c = W_o = \Sigma$. The I/O maps, $f_c(t)$ and $f_o(t)$ are considered now in some detail.

Properties of a matrix F , relating the I/O maps, are given in the following theorem.

THEOREM 1. *For a minimal realization (A, b, c^T) of a stable SISO system the I/O maps are related by a constant square matrix F as*

$$(6) \quad f_o(t) = F f_c(t)$$

where

$$(7) \quad A^T = F A F^{-1},$$

$$(8) \quad c = F b,$$

$$(9) \quad F \text{ is nonsingular and symmetric.}$$

Proof. Recall that a minimal realization and its dual, (A^T, c, b^T) , are related by a unique symmetric co-ordinate transformation matrix [8, Thm. 2.4.7 and Ex. 2.4.16]. Therefore it follows directly that (6) holds with F satisfying (7)–(9).

It should be pointed out that the foregoing result is closely related to a result obtained in [9]. Notice that the Gramians can now be related by using (6) as

$$(10) \quad W_o = F W_c F.$$

It is interesting to notice from (2), (4) that every entry of e^{At} is a linear combination of exponential functions of the form $t^i \exp[\lambda_p t]$ where λ_p is a pole of the transfer function and t^i is due to the multiplicity of that pole. This observation leads to the conclusion that each entry of the I/O maps must be a linear combination of the above-mentioned exponential functions. Thus the input and output maps can be expressed as

$$(11) \quad f_c(t) = G_c \gamma(t), \quad f_o(t) = G_o \gamma(t),$$

where $G_c, G_0 \in C^{n \times n}$ and $\gamma(t)$ was defined in (2). Moreover, the nature of $\gamma(t)$ implies that there exists a constant matrix Λ such that

$$(12) \quad \frac{d}{dt} \gamma(t) = \Lambda \gamma(t)$$

where

$$\Lambda = \text{Diag} [\Lambda^1, \Lambda^2, \dots, \Lambda^m],$$

$$\Lambda^i = \begin{bmatrix} \lambda_i & 0 & \cdots & 0 & 0 \\ 1 & \lambda_i & \cdots & 0 & 0 \\ 0 & 2 & \cdots & 0 & 0 \\ 0 & 0 & & \vdots & \vdots \\ \vdots & \vdots & & \lambda_i & 0 \\ 0 & 0 & \cdots & n_i - 1 & \lambda_i \end{bmatrix}.$$

Next (11) enables the Gramians, (4), (5), to be rewritten as

$$(13) \quad W_c = G_c V G_c^T, \quad W_0 = G_0 V G_0^T,$$

where $V = \int_0^\infty \gamma(t) \gamma^T(t) dt$ and $\gamma(t)$ is defined in (2). Notice that since W_c, W_0 are nonsingular, therefore, $G_c G_0$ and V must be nonsingular.

Now it is readily shown by using (11), (12) and comparing the expressions obtained for the time derivative of $f_c(t), f_0(t)$ obtained using (4), (5) with those obtained using (11) that

$$(14) \quad \begin{aligned} A &= G_c \Lambda G_c^{-1}, & A^T &= G_0 \Lambda G_0^{-1}, \\ b &= G_c \gamma(0), & c &= G_0 \gamma(0). \end{aligned}$$

An expression for the impulse response involving $\gamma(t)$, F and G_c is developed now. This expression will be of central importance to the realization algorithm to be developed later in the paper. Notice that if A is scaled by a nonzero scalar k and b and c are unchanged then $h(t)$, $\gamma(t)$, and hence both maps have their time variable scaled by k . This observation leads to the following I/O map decomposition of the impulse response

$$(15) \quad h(t) = c^T e^{(1/2)At} e^{(1/2)At} b = f_0^T(t/2) f_c(t/2).$$

Moreover, using (6) and (11) this expression for the impulse response can be rewritten as

$$(16) \quad h(t) = \gamma^T(t/2) K \gamma(t/2)$$

where

$$(17) \quad K = G_c^T F G_c.$$

The basis for the method being developed here can now be indicated. As can be seen from (8), (14) the determination of a balanced realization is essentially complete once *appropriate* values for F and G_c have been determined. In the sequel it is shown that F can always be assumed to be a parity (sign, signature) matrix E . Therefore, the determination of E and G_c so that the realization is balanced requires the simultaneous satisfaction of (13), (17) with F assumed to be an unknown parity matrix and W_c an unknown positive definite diagonal matrix.

Before taking up the problem of determining E , G_c from (13), (17), it is shown that V can be determined directly from the poles, that K can be determined directly from the partial fraction expansion coefficients, and that (13), (17) can be rewritten in terms of real matrices.

Notice that the matrix V which is needed in (13) can readily be computed from a knowledge of the factors of the denominator polynomial as

$$(18) \quad (V_{ij})_{pq} = \frac{(p+q-2)!}{[-(\lambda_i + \lambda_j)]^{p+q-1}}$$

where $V_{ij} \in C^{n_i \times n_j}$ is the ij th block of V and

$$V_{ij} = \int_0^\infty \Gamma_{ij}(t) e^{(\lambda_i + \lambda_j)t} dt,$$

$$(\Gamma_{ij}(t))_{pq} = t^{p+q-2}.$$

Moreover it is shown in the Appendix that the matrix K which is needed in (17) can be evaluated directly from the partial fraction expansion coefficients as

$$(19) \quad K = \text{Diag}[K^1, K^2, \dots, K^m]$$

where

$$(K^i)_{pq} = \begin{cases} \frac{(p+q-2)!}{(p-1)!(q-1)!} m_{i,p+q-1}, & p+q < n_i+1, \\ 0, & \text{otherwise.} \end{cases}$$

Thus the K matrix is block diagonal with diagonal blocks being upper cross-diagonal with coefficients on rays parallel to the cross diagonal being related to the coefficients in the binomial expansion, e.g.,

$$(20) \quad K^i = \begin{bmatrix} m_{i,1} & m_{i,2} & m_{i,3} & m_{i,4} & m_{i,5} & \cdot & \cdot & m_{i,n_i} \\ m_{i,2} & 2m_{i,3} & 3m_{i,4} & 4m_{i,5} & \cdot & \cdot & \cdot & 0 \\ m_{i,3} & 3m_{i,4} & 6m_{i,5} & \cdot & \cdot & \cdot & \cdot & 0 \\ m_{i,4} & 4m_{i,5} & \cdot & \cdot & \cdot & \cdot & \cdot & 0 \\ m_{i,5} & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & 0 \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & 0 & \cdot & \cdot & \cdot & \cdot & 0 \\ m_{i,n_i} & 0 & 0 & \cdot & \cdot & \cdot & \cdot & 0 \end{bmatrix}.$$

Notice from the foregoing that the first row of K^i is α_i^T , (2). Also notice that $\gamma_i^T(0)K^i = \alpha_i^T$ and hence that $\gamma^T(0)K = \alpha^T$. Thus it follows from this observation and from (12) that $h(t) = c^T e^{At} b$ when $A = \Lambda$, $b = \gamma(0)$ and $c^T = \gamma^T(0)K$. This simple fact will play an important role in the development of the realization algorithm.

Thus from the foregoing it is seen that the matrices K and V can be readily determined directly from the partial fraction expansion of the transfer function. However, these matrices are complex in general. Considerable reduction in the computational burden would result if it were possible to modify the foregoing approach so that all matrices involved in (13), (17) were real. It turns out that this is possible due to the fact that the impulse response function is real. For $h(t)$ to be real, any complex λ_i 's must occur in conjugate pairs, i.e., if $\text{Im}[\lambda_i] \neq 0$ then $\text{Im}[\lambda_i] = -\text{Im}[\lambda_j]$ for i, j integers $\in [1, m]$ and $i = j$. Therefore, the λ_i 's (2), can be arranged so that conjugate

pairs are denoted by consecutive indices and the complex λ_i 's are denoted by the lowest indices viz.

$$(21) \quad \begin{aligned} \lambda_{2i-1} &= \lambda_{2i}^*, & i &= 1, 2, \dots, \nu, \\ \lambda_i &= \text{real scalar}, & i &= 2\nu+1, 2\nu+2, \dots, m. \end{aligned}$$

It also follows from the realness of the impulse response that the coefficients in the expansion of the impulse response, (2), must satisfy

$$(22) \quad \begin{aligned} m_{2i-1,j} &= m_{2i,j}^*, & i &= 1, 2, \dots, \nu, \\ & & j &= 1, 2, \dots, n_i, \\ m_{ij} &= \text{real scalar}, & i &= 2\nu+1, 2\nu+2, \dots, m, \\ & & j &= 1, 2, \dots, n_i, \end{aligned}$$

where n_i is the multiplicity of the corresponding pole. Moreover, the block diagonal matrices Λ and K (12), (19) become

$$(23) \quad \begin{aligned} \Lambda &= \text{Diag} [\Lambda^1, \Lambda^{1*}, \Lambda^2, \Lambda^{2*}, \dots, \Lambda^\nu, \Lambda^{2\nu+1}, \Lambda^{2\nu+2}, \dots, \Lambda^m], \\ K &= \text{Diag} [K^1, K^{1*}, K^2, K^{2*}, \dots, K^\nu, K^{2\nu+1}, K^{2\nu+2}, \dots, K^m]. \end{aligned}$$

Consider next the block diagonal complex matrix J where

$$(24) \quad J = \text{Diag} [J^1, J^2, \dots, J^\nu, I_\delta]$$

where

$$\begin{aligned} \delta &= n - 2 \sum_{i=1}^{\nu} n_i, \\ J^i &= \frac{1}{\sqrt{2}} \begin{bmatrix} I_{n_i} & I_{n_i} \\ -jI_{n_i} & jI_{n_i} \end{bmatrix}, \quad \begin{aligned} i &= 1, 2, \dots, \nu, \\ j &= \sqrt{-1}, \end{aligned} \end{aligned}$$

and n_i is the multiplicity of the complex poles of $h_L(s)$ at λ_{2i} or λ_{2i-1} .

Then it can be seen by inspection of the matrices resulting from the indicated matrix multiplications that

$$(25) \quad \begin{aligned} \bar{K} &= J^{-T} K J^{-1} = \text{real, symmetric}, \\ \bar{\Lambda} &= J \Lambda J^{-1} = \text{real}, \\ \bar{V} &= J V J^T = \text{real, symmetric, positive definite.} \end{aligned}$$

The real symmetry of \bar{V} follows from

$$(26) \quad \bar{V} = \int_0^\infty \beta(t) \beta^T(t) dt$$

where

$$\begin{aligned} \beta(t) &= J \gamma(t), \\ \gamma(t) &= [\gamma_1(t), \gamma_1^*(t), \dots, \gamma_\nu^*(t), \gamma_{2\nu+1}(t), \dots, \gamma_m(t)]. \end{aligned}$$

Moreover the block matrices in \bar{K} and $\bar{\Lambda}$ are readily calculated as

$$(27) \quad \begin{aligned} \bar{K} &= \text{Diag} [\bar{K}^1, \bar{K}^2, \dots, \bar{K}^\nu, \bar{K}^{2\nu+1}, \bar{K}^{2\nu+2}, \dots, \bar{K}^m], \\ \bar{\Lambda} &= \text{Diag} [\bar{\Lambda}^1, \bar{\Lambda}^2, \dots, \bar{\Lambda}^\nu, \bar{\Lambda}^{2\nu+1}, \bar{\Lambda}^{2\nu+2}, \dots, \bar{\Lambda}^m] \end{aligned}$$

where

$$\bar{K}^i = \begin{cases} \left[\begin{array}{c|c} R_e[K^i] & -I_m[K^i] \\ \hline -I_m[K^i] & -R_e[K^i] \end{array} \right], & i = 1, 2, \dots, \nu, \\ K^i, & i = 2\nu + 1, \dots, m, \end{cases}$$

$$\bar{\Lambda}^i = \begin{cases} \left[\begin{array}{c|c} R_e[\Lambda^i] & I_m[\Lambda^i] \\ \hline I_m[\Lambda^i] & R_e[\Lambda^i] \end{array} \right], & i = 1, 2, \dots, \nu, \\ \Lambda^i, & i = 2\nu + 1, \dots, m, \end{cases}$$

and K^i, Λ^i are specified in (23).

In the sequel the positive definiteness of \bar{V} will be important. This property is given now in the following theorem.

THEOREM 2. *The \bar{V} matrix is real, symmetric and positive definite.*

Proof. Suppose (A, b, c^T) is a balanced realization. Then the controllability Gramian, (4), is diagonal and satisfies the following Lyapunov equation [10]

$$(28) \quad A\Sigma + \Sigma A^T = -bb^T.$$

Next, pre and post multiplying this equation by G_c^{-1} and G_c^{-T} , respectively, and using (13), (14) yields

$$(29) \quad \Lambda V + V\Lambda^T = -\gamma(0)\gamma^T(0).$$

Again pre- and post-multiplying (29) by J and J^T , respectively, and using (25) yields

$$(30) \quad \bar{\Lambda}\bar{V} + \bar{V}\bar{\Lambda}^T = -\bar{b}\bar{b}^T$$

where $\bar{b} = J\gamma(0)$.

Recall that $\bar{\Lambda}$ is real. It can also be seen from the block structure of J and $\gamma(0)$ that \bar{b} must be real. Moreover, $\bar{\Lambda}, \bar{b}$ are the system "A" and "b" matrices which result from changing the co-ordinates of the balanced realization by the co-ordinate transformation matrix $G_c J^{-1}$. Therefore, it follows that the controllability of (A, b) and stability of A guarantee the controllability of $(\bar{\Lambda}, \bar{b})$ and stability of $\bar{\Lambda}$. Thus it follows from these conditions and a well-known theorem [10, p. 86], that \bar{V} satisfying (30) is unique, real, symmetric and positive definite. This completes the proof of the theorem.

Before taking up the use of (18), (19), (25) in the development of a method for solving (13), (17) as discussed following (17), it needs to be established that for any transfer function there always exist balanced realizations with F being a parity matrix. This task is accomplished in the next section.

3. Symmetry properties of I/O maps for balanced realizations. Recall [2] that a minimal realization is said to be balanced if the controllability and observability Gramians (4), (5) are diagonal and equal, i.e.

$$(31) \quad W_c = W_o = \Sigma$$

where $\Sigma = \text{Diag}[\sigma_1, \sigma_2, \dots, \sigma_n]$ with $\sigma_i > 0$ and real for all integers $i \in [1, n]$. The σ_i 's have been referred to as the second order modes of the system [2].

In order to show the properties of F which result from having the realization balanced, a general result concerning a class of complex symmetric matrices is needed. This result is given now in the following

LEMMA 1. *If $Q \in C^{n \times n}$ is symmetric and $Q^2 = I$ then Q is similar to a parity matrix with the similarity transformation matrix $V \in C^{n \times n}$ satisfying*

$$V^T V = I_n.$$

Proof. The fact that Q is symmetric implies that its singular value decomposition can be written as

$$Q = V\Sigma V^T \quad \text{with } V^T V = I_n.$$

Then squaring both sides and using the fact that Q is self inverse leads to the conclusion that Σ^2 must be the identity matrix. Thus Σ must be a parity matrix and the proof is complete.

The properties of F that result from having the realization balanced are now stated in the following theorem.

THEOREM 3. *If (A, b, c^T) is a minimal balanced realization of a stable scalar transfer function then F in Theorem 1 is*

- 1) *a parity (signature, sign) matrix if all the second-order modes are distinct,*
- 2) *similar to a parity matrix if they are not distinct with the transformation matrix being orthogonal, i.e.*

$$V^T F V = \text{parity matrix with } V^T V = I_n.$$

Proof. Since the realization is assumed to be balanced (31) holds and (10) becomes

$$(32) \quad \Sigma = F\Sigma F.$$

Now since Σ and F are each nonsingular it follows that

$$(33) \quad F^{-1} = \Sigma^{-1} F \Sigma.$$

Now let Σ be arranged so that any repeated second order modes appear in succession along the diagonal, i.e.

$$\Sigma = \text{Diag} [\sigma_1 I_{r_1}, \sigma_2 I_{r_2}, \dots, \sigma_{r_m} I_{r_m}]$$

where there are m distinct second-order modes and σ_i is repeated r_i times. Using this form for Σ in (33) and equating like positioned blocks on either side of (33) yields

$$(34) \quad (F^{-1})_{ij} = \frac{\sigma_j}{\sigma_i} F_{ij}$$

where F is partitioned so as to enable block multiplication in (33) and F_{ij} refers to the r_i by r_j block viz.

$$F = \begin{bmatrix} F_{11} & F_{12} & \cdots & F_{1m} \\ F_{21} & F_{22} & \cdots & F_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ F_{m1} & F_{m2} & \cdots & F_{mm} \end{bmatrix}.$$

Now since F is symmetric so is F^{-1} and (34) implies

$$(35) \quad \frac{\sigma_i}{\sigma_j} F_{ji}^T = \frac{\sigma_j}{\sigma_i} F_{ij}.$$

But $F_{ij} = F_{ji}^T$ (Theorem 1) and $\sigma_i = \sigma_j$, $\sigma_i, \sigma_j > 0$ (assumed). Therefore, the off-diagonal blocks in F must be null, i.e.

$$(36) \quad F_{ij} = 0, \quad i \neq j.$$

Thus F must be a block diagonal matrix. Moreover, it is seen from (32) that the diagonal blocks must satisfy

$$(37) \quad F_{ii}^2 = I_{r_i}.$$

Then it follows from Lemma 1 and the block diagonal nature of F that F must be similar to a parity matrix.

Finally if all the second order modes are distinct, i.e. diagonal elements of Σ all different, then the foregoing shows that F must be a parity matrix. This completes the proof of the theorem.

From the foregoing it is now possible to state the following important theorem.

THEOREM 4. *It is always possible to choose the co-ordinates for a balanced realization such that F is a parity matrix.*

Proof. Recall that if T is a co-ordinate transformation matrix which transforms the realization (A, b, c^T) to $(\tilde{A}, \tilde{b}, \tilde{c}^T)$ then the controllability and observability Gramians in the original co-ordinates, namely W_c and W_o , are related to the controllability and observability Gramians in the new co-ordinates, namely \tilde{W}_c and \tilde{W}_o , as

$$(38) \quad \tilde{W}_c = T^{-1} W_c T^{-T}, \quad \tilde{W}_o = T^T W_o T.$$

Moreover, the relation between W_o , W_c and F , (10) implies that the F -matrix for (A, b, c^T) , namely F , is related to the F -matrix for $(\tilde{A}, \tilde{b}, \tilde{c}^T)$ namely \tilde{F} as

$$(39) \quad \tilde{F} = T^T F T.$$

Next suppose F is not a parity matrix. Then from the proof of Theorem 3, F is a block diagonal matrix having diagonal blocks which are self-inverse, i.e. diagonal blocks satisfying (37). Hence from Lemma 1 it follows that there exists a block diagonal matrix T having diagonal blocks satisfying

$$(40) \quad T_{ii}^{-1} F_{ii} T_{ii} = E_{ii},$$

where

$$E_{ii} = \text{parity matrix} \quad \text{and} \quad T_{ii}^{-1} = T_{ii}^T.$$

It follows that $T^T = T^{-1}$ and the system obtained by transforming the co-ordinates by T has a F -matrix, namely \tilde{F} , (39), which is a parity matrix. This completes the proof of Theorem 4.

The foregoing result is of importance in that it guarantees the existence of balanced realizations having the E -symmetry property corresponding to any stable transfer function. This fact was shown recently in a quite different fashion [11]. It is used in the next section to develop an algorithm for determining a balanced realization for a given transfer function by using the properties of I/O maps where it is assumed that the co-ordinates are such that F is a parity matrix.

4. Balancing algorithm: SISO case. Recall that a balanced realization has Gramians which satisfy (31). Recall also from Theorem 4 that it is always possible for F in Theorem 1 to be a parity matrix E . Therefore, when the realization is balanced with $F = E$, (13), (17) becomes

$$(41) \quad \Sigma = G_c V G_c^T,$$

$$(42) \quad K = G_c^T V G_c.$$

Now the need for complex matrices in these equations can be eliminated by using the results developed in § 3. Solving (25) for K and V and substituting in (41), (42) yields,

$$(43) \quad \Sigma = G \bar{V} G^T,$$

$$(44) \quad \bar{K} = G^T E G,$$

where

$$(45) \quad G = G_c J^{-1}.$$

The foregoing real equations can be combined to enable G , E and Σ to be determined as an eigenvalue-eigenvector problem. This is done by solving (44) for E and using the result to multiply (43). This gives rise to the following equation.

$$(46) \quad E\Sigma = G\bar{V}\bar{K}G^{-1}.$$

Notice that G diagonalizes $\bar{V}\bar{K}$ and that $E\Sigma$ and $\bar{V}\bar{K}$ are real. Therefore, G must be real. Also notice that in order for the G matrix satisfying (46) to yield a balanced realization, G must also satisfy (43) or (44). Since \bar{V} has been shown to be positive definite, real and symmetric, (43) is used with (46) to determine G using the numerically stable approach given in [12, pp. 337-338].

Notice that the product of the symmetric matrices \bar{V} and \bar{K} is symmetric only if \bar{V} and \bar{K} commute. In general these matrices do not commute. Therefore, to simplify the determination of G let the positive definite matrix \bar{V} be decomposed by Cholesky decomposition as

$$(47) \quad \bar{V} = LL^T$$

where L is real, lower triangular and nonsingular since \bar{V} is positive definite (Theorem 2). Therefore, it follows from (46) that if (α_i, u_i) is an eigenvalue-eigenvector pair for $L^T\bar{K}L$ then (α_i, Lu_i) is an eigenvalue-eigenvector pair for $\bar{V}\bar{K}$. However, since $L^T\bar{K}L$ is real and symmetric, its eigenvalues are real and it can be diagonalized by a real orthogonal matrix U .

From the foregoing it is seen that

$$(48) \quad \Sigma = \text{Diag} [|\alpha_1|, |\alpha_2|, \dots, |\alpha_n|],$$

$$(49) \quad E = \text{Diag} [\text{Sgn} [\alpha_1] \text{Sgn} [\alpha_2], \dots, \text{Sgn} [\alpha_n]].$$

Moreover $G^{-1} = LUS$ satisfies (46) with S diagonal. However, (44) must also be satisfied. This requirement is met by setting $S = \Sigma^{-1/2}$. Therefore using (14), (25), (45), it is seen that a balanced realization can be composed as

$$(50) \quad \begin{aligned} A &= \Sigma^{1/2} U^T L^{-1} \bar{A} L U \Sigma^{-1/2}, \\ b &= \Sigma^{1/2} U^T L^{-1} J \gamma(0), \\ c &= E b. \end{aligned}$$

To show that this realization is balanced, notice from (44) that the realization $(\bar{A}, J\gamma(0), \gamma^T(0)J^T\bar{K})$ equivalent to (50) with co-ordinate transformation matrix $T = \Sigma^{1/2} U^T L^{-1}$. Moreover, assuming $W_c = \Sigma$ in (38) gives $\tilde{W}_c = \bar{V}$ which was shown, (30), to be the controllability Gramian for $(\bar{A}, J\gamma(0))$. The fact that $W_0 = \Sigma$ follows from (10) since F is a parity matrix. Thus the realization determined by (50) is indeed balanced.

Finally, the realization (50) has impulse (2). This is seen by using the equivalence just mentioned and (25) to show that

$$c^T e^{At} b = \gamma^T(0) K e^{At} \gamma(0).$$

Then from the structure of $\gamma(0)$ and K , (2), (19), (20) as well as from (12) it follows that

$$\gamma^T(0) K = \alpha^T, \quad e^{At} \gamma(0) = \gamma(t).$$

Thus $c^T e^{At} b$ is seen to have an impulse response given by (2).

5. Balancing algorithm: MIMO case. In this section, Kronecker products [13] are used to extend the algorithm given in the previous section to enable the balancing of MIMO systems. This extension is not straightforward however, since F (Theorem 1) can no longer be assumed in general, to be a parity matrix.

Let the system's transfer function matrix be expanded in partial fractions as

$$(51) \quad H_L(s) = \sum_{i=1}^m \sum_{j=1}^{n_i} (s - \lambda_i)^{-j} H_{ij}, \quad r = \sum_{i=1}^m n_i$$

where $H_{ij} \in C^{p \times q}$ and $H_L(s)$ is assumed rational. Then the impulse response matrix (weighting matrix or pattern) is given as

$$(52) \quad H(t) = \sum_{i=1}^m e^{\lambda_i t} \sum_{j=1}^{n_i} M_{ij} t^{j-1} = \alpha^T [\gamma(t) \otimes I_q]$$

where $M_{ij} = (H_{ij}/(j-1)!)$ and $[\cdot \otimes \cdot]$ indicates Kronecker product.

$$\begin{aligned} \alpha^T &= [\alpha_1^T, \alpha_2^T, \dots, \alpha_m^T], & \alpha_i^T &= [M_{i1}, M_{i2}, \dots, M_{i,n_i}], \\ \gamma^T(t) &= [\gamma_1^T(t), \gamma_2^T(t), \dots, \gamma_m^T(t)], & \gamma_i^T &= [1, t, \dots, t^{n_i-1}] e^{\lambda_i t}. \end{aligned}$$

Notice that r , the sum of the n_i 's, is the degree of the least common denominator of $H_L(s)$ and also that r is the degree of the minimal polynomial of A , where (A, B, C^T) is any minimal realization of $H_L(s)$, [14, p. 108]. Moreover, since the maximal dimension of any of the blocks of the Jordan form for A associated with λ_i is n_i , [15, p. 226], it follows that the largest power of t multiplying $\exp(\lambda_i t)$ in any element of $\exp(At)$ is $n_i - 1$. Therefore, the least common denominator of the transfer function matrix plays the same role as the denominator polynomial of $h_L(s)$ in the SISO case for the determination of the vector of time multiplied exponentials, $\gamma(t)$. However, since the degree of the least common denominator r is in general not the same as the dimension of the minimal realization, it is important to notice that $\gamma(t)$ is defined by (2) with n replaced by r .

Next, let $H(t)$ be written as a matrix of scalar time functions

$$(53) \quad H(t) = [H_{ij}(t)], \quad i = 1, 2, \dots, p, \quad j = 1, 2, \dots, q.$$

Now each $H_{ij}(t)$ can be considered as the impulse response of a SISO system. Therefore, from (16) it is seen that

$$(54) \quad H_{ij}(t) = \gamma^T(t/2) K^{ij} \gamma(t/2)$$

where $\gamma(t)$ is determined from the least common denominator of $H_L(s)$ for reasons given earlier and K^{ij} is determined using (19). This observation enables the impulse response matrix to be written in terms of Kronecker products as

$$(55) \quad H(t) = [I_p \otimes \gamma^T(t/2)] K [I_q \otimes \gamma(t/2)]$$

where

$$K = [K^{ij}], \quad K \in C^{pr \times qr}.$$

In the sequel, the following rules for Kronecker products will be used [13],

$$(56) \quad \begin{aligned} [A \otimes B][C \otimes D] &= [AC \otimes BD], \\ [A \otimes B]^T &= [A^T \otimes B^T]. \end{aligned}$$

Proceeding in a similar fashion, the relations (11) can be generalized so that maps of a minimal realization of $H_L(s)$, (A, B, C^T) , having dimension n can be written as

$$(57) \quad \begin{aligned} e^{At}B &= F_c(t) = G_c[I_q \otimes \gamma(t)], \\ e^{A^T t}C &= F_0(t) = G_0[I_p \otimes \gamma(t)] \end{aligned}$$

where

$$G_c \in C^{n \times qr}, \quad G_0 \in C^{n \times pr}.$$

Now it follows from these expressions that

$$(58) \quad B = G_c[I_q \otimes \gamma(0)], \quad C = G_0[I_p \otimes \gamma(0)].$$

Moreover, proceeding in the same manner as was used to show (14) it follows that

$$(59) \quad A = G_c[I_q \otimes \Lambda]G_c^\#, \quad A^T = G_0[I_p \otimes \Lambda]G_0^\#$$

where $G_c^\#$ and $G_0^\#$ are right inverses of G_c and G_0 respectively. The fact that G_c and G_0 are each full row rank will become evident in what follows.

Notice at this point that once appropriate values for G_c and G_0 are obtained a realization, (A, B, C^T) , can be calculated from (58), (59).

Next, using the factorization, (15), and the expressions, (57), for the I/O maps it is seen from (55) that

$$(60) \quad K = G_0^T G_c.$$

Next, following the same procedure used to obtain (13) enables the Gramians to be expressed as

$$(61) \quad W_c = G_c[I_q \otimes V]G_c^T,$$

$$(62) \quad W_0 = G_0[I_p \otimes V]G_0^T$$

where V is given in (13) with $\gamma \in C^r$.

Note that the nonsingularity of W_c and W_0 due to the fact that (A, B, C^T) is a minimal realization implies that G_0 and G_c are full (row) rank.

The method to be developed for determining $G_c(G_0)$ depends on (60), (61) ((60), (62)). However, the need for working with complex matrices in connection with these equations can be eliminated in a manner similar to that used earlier in the SISO case. Ordering the poles in the manner specified in (21) it follows that

$$(63) \quad \begin{aligned} \bar{K} &= [I_p \otimes J^{-T}]K[I_q \otimes J^{-1}], \\ \bar{\Lambda} &= J\Lambda J^{-1}, \quad \bar{V} = JVJ^T, \end{aligned}$$

where K is defined in (55) and J is defined in (24) with n replaced by r .

A minimal balanced realization can now be determined in the manner specified in the following

THEOREM 5. *A minimal balanced realization can be obtained as*

$$A = T_1^{-1}\bar{A}T_1, \quad B = T_1^{-1}\bar{B}, \quad C^T = \bar{C}^T T_1$$

where

$$\begin{aligned} \bar{A} &= [I_q \otimes \bar{\Lambda}], \quad \bar{B} = [I_q \otimes J\gamma(0)], \quad \bar{C}^T = [I_p \otimes \gamma^T(0)J^T]\bar{K}, \\ T &= [T_1^1 \ T_2], \quad T_1 = \bar{L}_q U_1 \Sigma^{-1/2}, \quad T_2 = \bar{L}_q U_2 \end{aligned}$$

and U_1, U_2, Σ are determined using singular value decomposition on a real symmetric matrix viz.

$$\bar{L}_q^T \bar{K}^T [I_p \otimes \bar{V}] \bar{K} \bar{L}_q = [U_1^T \mid U_2^T] \left[\begin{array}{c|c} \Sigma^2 & 0 \\ \hline 0 & 0 \end{array} \right] \begin{bmatrix} U_1^T \\ U_2^T \end{bmatrix}$$

and \bar{L}_q is determined using Cholesky decomposition as

$$[I_q \otimes \bar{V}] = \bar{L}_q \bar{L}_q^T, \quad \bar{L}_q = [I_q \otimes \bar{L}], \quad \bar{L} \bar{L}^T = \bar{V}.$$

Proof. Proceeding in a manner similar to the end of the previous section, it can be shown that $H(t) = \bar{C}^T e^{\bar{A}t} \bar{B}$. Moreover, using (63), (12) it can be shown that the controllability Gramian, (4), for this realization is $[I_q \otimes \bar{V}]$. Since this Gramian is full rank, the realization $(\bar{A}, \bar{B}, \bar{C}^T)$ must be controllable. Moreover, from (57) it is seen that G_c for this realization is a qr dimension identity matrix and from (60), $\bar{G}_0 = K^T$. Hence, from (62) it is seen that the observability Gramian for this realization is $\bar{K}^T [I_p \otimes \bar{V}] \bar{K}$.

Let T be a co-ordinate transformation matrix which transforms $(\bar{A}, \bar{B}, \bar{C}^T)$ to $(\tilde{A}, \tilde{B}, \tilde{C}^T)$ and \bar{W}_0, \bar{W}_c to \tilde{W}_0, \tilde{W}_c respectively viz.

$$\tilde{A} = T^{-1} \bar{A} T, \quad \tilde{B} = T^{-1} \bar{B}, \quad \tilde{C}^T = \bar{C}^T T,$$

$$\tilde{W}_c = T^{-1} \bar{W}_c T^{-T}, \quad \tilde{W}_0 = T^T \bar{W}_0 T.$$

Let T be chosen so that $\tilde{W}_c \tilde{W}_0$ is diagonal, i.e.,

$$T^T \bar{W}_0 \bar{W}_c T^{-T} = \left[\begin{array}{c|c} \Sigma^2 & 0 \\ \hline 0 & 0 \end{array} \right].$$

Now since \bar{W}_0 and \bar{W}_c are each symmetric and \bar{W}_c is positive definite, T can be determined using the numerically stable approach given in [12, pp. 337-338] by determining U_1, U_2 as indicated in the statement of the theorem. Notice the T matrix which results yield

$$\begin{aligned} \tilde{W}_0 &= \left[\begin{array}{c|c} \Sigma & 0 \\ \hline 0 & 0 \end{array} \right], & \tilde{W}_c &= \left[\begin{array}{c|c} \Sigma & 0 \\ \hline 0 & I \end{array} \right], \\ \tilde{A} &= \left[\begin{array}{c|c} \tilde{A}_{11} & \tilde{A}_{12} \\ \hline \tilde{A}_{21} & \tilde{A}_{22} \end{array} \right]_{n \times (n+r)}, & \tilde{B} &= \left[\begin{array}{c} \tilde{B}_1 \\ \tilde{B}_2 \end{array} \right], \\ \tilde{C} &= [\tilde{C}_1 \mid \tilde{C}_2]_{(n+r) \times n} \text{ with } \tilde{A}_{12} = 0, \tilde{C}_2 = 0. \end{aligned}$$

Thus it follows from \tilde{W}_0 , and \tilde{W}_c and the zero block structure of $(\tilde{A}, \tilde{B}, \tilde{C}^T)$ that $(\tilde{A}_{11}, \tilde{B}_1, \tilde{C}_1^T)$ is a minimal balanced realization of $H(t)$. Moreover, $A = \tilde{A}_{11}$, $B = \tilde{B}_1$, $C^T = \tilde{C}_1$ follows by direct calculation.

The fact that \tilde{A}_{12} and \tilde{C}_2 are each null can be shown as follows. From the singular value decomposition indicated in the theorem it follows that

$$U_2^T \bar{L}_q^T \bar{K}^T [I_p \otimes \bar{V}] \bar{K} \bar{L}_q U_2 = 0$$

and since $[I_p \otimes \bar{V}]$ is nonsingular it follows that

$$\text{Range} [\bar{L}_q U_2] \subset \text{Null} [\bar{K}]$$

which from the dependency of \bar{W}_0 on \bar{K} implies that

$$\text{Range} [\bar{L}_q U_2] \subset \text{Null} [\bar{W}_0].$$

Next it is easily shown directly from the Lyapunov equation governing \bar{W}_0 that

$$\text{Range} [\bar{L}_q U_2] \subset \text{Null} [\bar{C}^T], \quad \text{Range} [\bar{L}_q U_2] \subset \text{Null} [\bar{C}^T \bar{A}].$$

Thus it follows that $\text{Range} [\bar{L}_q U_2] \subset \text{Null} [\bar{\Omega}_0]$ which implies that \tilde{A}_{12} and \tilde{C}_2 must each be null.

An analysis of the foregoing theorem reveals that use of the co-ordinate transformation matrix T simultaneously diagonalizes both Gramians and puts the original controllable realization $(\bar{A}, \bar{B}, \bar{C}^T)$ in a form (observability decomposed form, [8, p. 362]) which enables the immediate extraction of the observable part.

Finally notice that U_2 need not be computed as it is not needed in the formulation of the balanced realization. In addition the singular value decomposition mentioned in Theorem 5 can be replaced by the computationally less demanding task of determining the nonzero eigenvalues and corresponding orthonormal eigenvectors of real symmetrix $\bar{L}_q^T \bar{K}^T [I_p \otimes \bar{V}] \bar{K} \bar{L}_q$.

This completes the proof of the theorem.

6. Conclusion. A method for determining a minimal balanced realization from a given transfer function matrix has been developed. The method requires the Laplace inversion of the transfer function matrix, the Cholesky decomposition of a real, symmetric, positive definite matrix and the determination of the nonzero eigenvalues and corresponding eigenvectors of a real, symmetric matrix. Unlike Moore's method [2], and Laub's modification [7], the determination of a minimal realization is not required at the outset and no Lyapunov equations need to be solved. Moreover, in [2], [7] the extent to which the controllability Gramian of the initial realization is ill conditioned, depends on that realization. Thus there is no guarantee in [2], [7] that a bad choice for the initial realization will not be made causing numerical instability in subsequent calculations. Any ill conditioning that occurs in the proposed method arises directly from the nature of the given transfer function matrix and not from any choice involved in the execution of the algorithm.

Finally, the problem of obtaining an arbitrary minimal realization of a transfer function matrix is often quite difficult and has engaged the attention of many researchers in the past. The apparent numerically beneficial features of the present algorithm for determining balanced realizations would seem to argue well for its use in determining a minimal realization even when balancing is not required.

Appendix. Determination of K from impulse response coefficients. Since F is symmetric and nonsingular (Theorem 1) and G_c is nonsingular, (13), it follows from (17) that K must be symmetric and nonsingular. Moreover, it can be seen from a comparison of the expression (2), (16) for the impulse response that K can always be assumed to be a block diagonal matrix, i.e.

$$(A1) \quad K = \text{Diag} [K^1, K^2, \dots, K^m], \quad K^i \in C^{n_i \times n_i}.$$

Now it turns out that K is related to Λ in the manner specified in the following theorem.

THEOREM A1. $K\Lambda = \Lambda^T K$ where Λ is defined by (12).

Proof. Differentiating the impulse response as given in (16) and using (12) yields

$$(A2) \quad \frac{d}{dt} h(t) = \frac{1}{2} \gamma^T(t/2) [\Lambda^T K + K \Lambda] \gamma(t/2).$$

Alternatively, differentiating the first expression for the impulse response given in (15) and using (6), (11) yields

$$(A3) \quad \frac{d}{dt} h(t) = \gamma^T(t/2) G_c^T F A G_c \gamma(t/2).$$

Then using (14) and the relation for K , F and G_c , (17) enables (A3) to be rewritten as

$$(A4) \quad \frac{d}{dt} h(t) = \gamma^T(t/2) K \Lambda \gamma(t/2).$$

Finally, the desired relation (A1) results from the comparison of the right-hand sides of (A2) and (A4). This completes the proof of the theorem.

The foregoing theorem, together with the structure of Λ , implies that the elements of the diagonal blocks of K must satisfy a number of rather specific properties. These properties are given now in the following corollary.

COROLLARY A1. *Each diagonal block of K , say $K^p \in C^{n_p \times n_p}$ for any integer $p \in [1, m]$ has elements $\{K_{ij}^p: i, j \in [1, n_p]\}$ satisfying*

$$(A5) \quad K_{ij}^p = \frac{j}{i-1} K_{i-1, j+1}^p \quad \begin{array}{l} i = 2, 3, \dots, n_p, \\ j = 1, 2, \dots, n_p - 1, \end{array}$$

$$(A6) \quad = \frac{i}{j-1} K_{i+1, j-1}^p \quad \begin{array}{l} i = 1, 2, \dots, n_p - 1, \\ j = 2, 3, \dots, n_p, \end{array}$$

and

$$(A7) \quad K_{n_p, j}^p = 0 \quad j = 2, 3, \dots, n_p,$$

$$(A8) \quad K_{i, n_p}^p = 0 \quad i = 2, 3, \dots, n_p.$$

Proof. From Theorem 1 it follows that

$$(A9) \quad K^p \Lambda^p = (\Lambda^p)^T K^p \quad p = 1, 2, \dots, m.$$

Then the relations given in the corollary are obtained by equating like positioned entries on either side of (A9).

The foregoing properties of K are now used to show that all entries below (to the right of) the cross diagonal in each diagonal block of K must be zero while the entries on and to the left of the cross diagonal are each a scalar multiple of a coefficient in the expression for the impulse response (2).

COROLLARY A2. *K^p is upper cross triangular. Furthermore, nonzero (ij) th elements of K^p are related to the first row element $K_{1, i+j-1}^p$ as*

$$(A10) \quad K_{ij}^p = \frac{(i+j-2)!}{(i-1)!(j-1)!} K_{1, i+j-1}^p \quad i+j < n_p + 1.$$

In addition, the sums S_r of elements along rays parallel to the cross diagonal are given as

$$(A11) \quad S_r = \sum_{i=1}^r K_{i, r-i+1}^p = 2^{r-1} K_{1r}^p \quad r = 1, 2, \dots, n_p.$$

Proof. From recursion on (A5) and using (A7), it can be seen easily that K^p is upper cross triangular. Furthermore, recursion on (A6) yields (A10). Now, substitution of (A10) with $j = r - i + 1$ into (A11) yields

$$S_r = K_{1r}^p \sum_{i=0}^{r-1} \frac{(r-1)!}{i!(r-i+1)!} = K_{1r}^p (1+1)^{r-1} = 2^{r-1} K_{1r}^p.$$

This completes the proof.

COROLLARY A3. *If $K^p \in C^{n_p \times n_p}$ is a typical diagonal block of K , then its entries K_{ij}^p must satisfy the following property*

$$(A12) \quad \begin{aligned} K_{ij}^p &= \frac{(i+j-2)!}{(i-1)!(j-1)!} m_{p,i+j-1} & i+j < n_p+1, \\ &= 0 & i+j > n_p+1, \end{aligned}$$

where $\{m_{p,j}; j = 1, 2, \dots, n_p-1\}$ are the coefficients in the impulse response $h(t)$, (2).

Proof. Recall that the portion of the impulse response arising from the transfer function pole at λ_p of multiplicity n_p is given as

$$(A13) \quad h_p(t) = e^{\lambda_p t} [m_{p1} + m_{p2}t + m_{p3}t^2 + \dots + m_{p,n_p}t^{n_p-1}].$$

Now from (16) with $\gamma(t)$ and K , partitioned as in (2) and (19) it is seen that

$$(A14) \quad h_p(t) = \gamma_p^T(t/2) K^p \gamma_p(t/2).$$

Then with K_{ij}^p indicating the ij th element of K_p it is seen comparing (A13) and (A14) that

$$(A15) \quad k_{pj} = \sum_{i=1}^j K_{i,j+1-i}^p \left(\frac{1}{2}\right)^{j-1} \quad j = 1, 2, \dots, n_p,$$

$$(A16) \quad 0 = \sum_{j=1}^{n_p-1-l} K_{j+1+l, n_p+1-j}^p \quad l = 0, 1, \dots, n_p-2.$$

Now, referring to Corollary A2 and the expression for S_r , (A11), it is seen that

$$(A17) \quad m_{pj} = K_{1j}^p \quad j = 1, 2, \dots, n_p.$$

This completes the proof of Corollary A3 and the formula, (19), for determining the elements of K directly from the coefficients of the impulse response has been validated.

Acknowledgment. The authors wish to thank the referees for suggesting several improvements for the presentation of this work.

REFERENCES

- [1] B. C. MOORE, *Singular value analysis of linear systems: Parts I and II*, Univ. Toronto Report 7801, also IEEE Conf. on Decision and Control, 1978.
- [2] ———, *Principal component analysis in linear systems: controllability, observability and model reduction*, IEEE Trans. Automat. Control, AC-26 (1981), pp. 17–32.
- [3] E. A. JONCKHEERE AND L. M. SILVERMAN, *A new set of invariants for linear systems—application to reduced order compensator design*, Internat. Symp. Math. Th. of Nets. and Sys., (1981), also IEEE Trans. Automat. Control, AC-28 (1983), pp. 953–964.
- [4] S. SHOKOHI, L. M. SILVERMAN AND P. M. VANDOOREN, *Linear time variable systems: balancing and model reduction*, IEEE Trans. Automat. Control, AC-28 (1983), pp. 810–822.
- [5] E. I. VERRIEST AND T. KAILATH, *On generalized balanced realizations*, IEEE Trans. Automat. Control, AC-28 (1983), pp. 833–844.
- [6] F. W. FAIRMAN, S. S. MAHIL AND J. A. DEABREU, *Balanced realization algorithm for scalar continuous-time systems having simple poles*, Internat. J. Systems Sci., 15 (1984), pp. 685–694.
- [7] A. J. LAUB, *Computation of “balancing” transformations*, Joint Automatic Control Conference, San Francisco, FA8-E, 1980.
- [8] T. KAILATH, *Linear Systems*, Prentice-Hall, Englewood Cliffs, NJ, 1980.
- [9] B. D. O. ANDERSON, *On the computation of Cauchy index*, Quart. Appl. Math. (1972), pp. 577–582.
- [10] S. BARNETT, *Matrices in Control Theory*, Van Nostrand Reinhold, London, 1971.

- [11] D. A. WILSON AND A. KUMAR, *Symmetry properties of balanced systems*, IEEE Trans. Automat. Control, AC-28 (1983), pp. 927–929.
- [12] J. H. WILKINSON, *The Algebraic Eigenvalue Problem*, Oxford University Press, Oxford, 1965.
- [13] A. GRAHAM, *Kronecker Products and Matrix Calculus with Applications*, Ellis Horwood, Chichester, 1981.
- [14] R. W. BROCKETT, *Finite Dimensional Linear Systems*, John Wiley, New York, 1970.
- [15] S. LIPSCHUTZ, *Linear Algebra*, McGraw-Hill, New York, 1968.

ESTIMATION OF DISCONTINUOUS COEFFICIENTS IN PARABOLIC SYSTEMS: APPLICATIONS TO RESERVOIR SIMULATION*

PATRICIA K. LAMM†

Abstract. We present spline-based techniques for estimating spatially varying parameters that appear in parabolic distributed systems (typical of those found in reservoir simulation problems). In particular, we discuss the problem of determining discontinuous coefficients, estimating both the functional shape and points of discontinuity for such parameters. In addition, our ideas may also be applied to problems with unknown initial conditions and unknown parameters appearing in terms representing external forces. Convergence results and a summary of numerical performance of the resulting algorithms are given.

Key words. parameter estimation, discontinuous coefficients, parabolic distributed systems, spline approximations

AMS(MOS) subject classifications. 65K, 49

1. Introduction. We present here our efforts related to the estimation of discontinuous spatially varying coefficients in parabolic distributed systems. Although our ideas are applicable to a wide class of problems in which the determination of discontinuous coefficients is of importance (e.g., the propagation of waves through layered media; the dynamics of beams with “discontinuous” elastic properties), our work here is motivated by an inverse problem in reservoir simulation commonly referred to as “history matching”. The problem in this case is to determine unknown parameters (such as permeability, porosity) that appear as coefficients in model reservoir equations. “Optimal” choices of these parameters should provide the best match between the observed and simulated production history at one or more wells. Information about these coefficients (functional shape and location of discontinuities) provides insight into physical properties of the reservoir and can indicate the location of abrupt structural changes; in addition, precise determination of these parameters is essential to the process of accurately simulating and predicting reservoir behavior.

The governing reservoir equations describe mathematically the physical and chemical processes occurring during primary hydrocarbon recovery or during enhanced recovery efforts (secondary or tertiary forms of recovery). Mathematical models vary widely depending on the physical process being described (miscible or immiscible fluid flow, thermal or fluid injection, etc.) and the types of observations available. Common to each model however is a system of rate equations (derived from Darcy’s law, which relates flow rate to fluid pressure gradients) as well as appropriate conservation laws and equations of state. The resulting dynamical system is typically distributed in nature and of parabolic type [22], [23]; unknown parameters quite often include the porosity of surrounding rock, or the ratio of pore volume to total volume, and (relative) permeability, which is the ability of the rock to transmit fluid [23]. Due to spatial changes in underground structure, it is highly likely that these parameters will vary spatially and contain numerous discontinuities.

* Received by the editors February 7, 1984, and in revised form June 6, 1985. This research was supported in part by National Science Foundation grant MCS-8200883 and National Aeronautics and Space Administration grant NAG-1-258. Part of the research was carried out while the author was a visitor at the Institute for Computer Applications in Science and Engineering (ICASE), NASA Langley Research Center, Hampton, Virginia, which is operated under NASA contract nos. NAS1-17130 and NAS1-16394.

† Department of Mathematics, Southern Methodist University, Dallas, Texas 75275.

In order to solve the inverse problem, data in the form of fluid pressure (or flow rate) is collected at the wells and used in a numerical parameter estimation process. There have been a large number of substantial contributors to the development of theoretical concepts and numerical algorithms for the history matching problem. An exhaustive list of related references would be too lengthy to include here; instead we refer the reader to [23] for an excellent survey of the outstanding efforts in this area. One numerical approach commonly taken involves subdividing the reservoir into a grid of smaller blocks; constant-valued parameters (which are allowed to vary independently from block to block) are then estimated. Unfortunately, if accurate solutions are desired, the grid size often must be quite small and thus the number of unknown parameters, as well as the dimension of the state space, can be very large—as many as 50,000 parameters or more [22]. (This is an unfortunate consequence of the fact that the parameters of interest—as well as state variables—are infinite-dimensional yet computations must be performed in a finite-dimensional setting.) Our goal here is to avoid some of the difficulties associated with the approach described above. Specifically, our ideas involve separating the order of state approximation from that of parameter estimation, so that the need for an approximate state space of high dimension does not impose the same requirements on the dimension of an approximate parameter space; this is accomplished by searching for parameters in classes of functions with quite general spatially varying representations. In order to focus attention on the problems associated with estimating spatially varying discontinuous coefficients in this context, we consider an archetypical model of (parabolic) distributed type that admittedly is a simplified version of the fluid pressure equations associated with reservoir simulation (see [22], [23], and the references therein); nevertheless the model selected here is a prototype that contains the essential parameter-dependent terms for which we may begin our investigations. In the sections that follow we define the model equations of interest and construct an approximation framework in which we wish to consider the parameter estimation problem. Convergence results are presented for problems associated with either spatially distributed or “discrete” sample data. Finally, we discuss numerical implementation in general, and in the context of particular examples. It is our intent in this report to examine convergence properties and implementation problems associated with these methods; we do not address such important questions as identifiability, observability, or general underlying properties of the governing partial differential equation system.

The notation used throughout is standard: For $I \subseteq \mathbb{R}$ (the real line), we shall denote by $C(I; X)$ the space of continuous functions $f: I \rightarrow X$ with uniform norm $|\cdot|_\infty$; by $L_2(I; X)$ we mean the usual space of square-integrable “functions” $f: I \rightarrow X$ with L_2 norm $|\cdot|_{L_2(I; X)}$ and inner product $\langle \cdot, \cdot \rangle_{L_2(I; X)}$. The Sobolev spaces $H^p(I; X)$ and $H_0^p(I; X)$ are defined as usual (see, for example, [1]). Whenever $X = \mathbb{R}$, we shall simplify notation by writing $C(I)$ and $L_2(I)$, respectively, and, where no confusion results, by writing $|\cdot|$ (and $\langle \cdot, \cdot \rangle$) for the norm (and inner product) on $L_2(0, 1)$. In addition, no notational distinction will be made between a function $f: I \rightarrow \mathbb{R}$ and its restriction to $I_1 \subseteq I$.

2. The parameter estimation problem. As our fundamental state system we consider the scalar parabolic distributed system

$$\begin{aligned}
 (2.1) \quad & \frac{\partial u}{\partial t}(t, x) = \frac{1}{\rho(x)} \frac{\partial}{\partial x} \left(q(x) \frac{\partial u}{\partial x}(t, x) \right) + f(t, x; r(x)), \quad (t, x) \in (0, T) \times (0, 1), \\
 & u(t, 0) = u(t, 1) = 0, \\
 & u(0, x) = u_0(x).
 \end{aligned}$$

Here q and ρ are discontinuous (positive) functions representing the permeability and porosity properties, respectively, of the fluid and surrounding rock; the points of discontinuity in these functions correspond to abrupt spatial changes in the physical flow region (such as might be associated with layered media). Both q and ρ are typically unknown so we shall consider the problem of estimating these parameters, as well as the function r , $r(x) \in R^p$, and the initial condition u_0 , from observations of the state variable u .

To simplify notation, we assume that q is discontinuous at one point only, $x = \xi$, and that q is represented by

$$q = \phi_1 + H_\xi \phi_2$$

where ϕ_1 and ϕ_2 are continuous on $[0, 1]$; here H_ξ is the usual Heaviside function on $[0, 1]$ given by $H_\xi = 1$ on $[\xi, 1]$, $H_\xi = 0$ otherwise. There is a straightforward extension of our ideas to the case where

$$q = \phi_1 + \sum_{i=2}^{\mu} H_{\xi_{i-1}} \phi_i$$

$0 = \xi_0 < \xi_1 < \xi_2 < \dots < \xi_\mu = 1$, except that notational difficulties become excessive. (We later demonstrate our approximation and estimation techniques for multiple discontinuity problems in the section on numerical findings.) In addition, we simplify our presentation by assuming $\rho \equiv 1$ and note that there is no difficulty in extending our ideas to the case $\rho = \kappa_1 + \sum_{i=2}^{\mu} H_{\xi_{i-1}} \kappa_i$, where $\kappa_1, \dots, \kappa_\mu$ are unknown [30]. Given the parameterization chosen for q we define the parameter vector $\gamma = (\xi, \phi_1, \phi_2, r, u_0) = (s, u_0)$ as an element of the parameter set $\Gamma \subseteq \tilde{\Gamma} = \mathcal{S} \times L_2(0, 1)$, where, for m, \bar{m} fixed,

$$\mathcal{S}(m, \bar{m}) = \{s = (\xi, \phi_1, \phi_2, r) \in R \times C[0, 1] \times C[0, 1] \times L_2((0, 1); R^p) \mid \xi \in (0, 1), \\ \phi_i \in C^1[0, 1], \text{ and } 0 < m \leq \phi_i(x) \leq \bar{m} \text{ for } i = 1, 2, \text{ and } x \in [0, 1]\}.$$

Concerning Γ and the applied force f , we make the following (standing) hypotheses:

(H1) The parameter set Γ is compact;

(H2) For every $r \in L_2((0, 1); R^p)$, the map $t \rightarrow f(t, \cdot; r(\cdot)): [0, T] \rightarrow L_2(0, 1)$ is Hölder continuous with exponent α , $0 < \alpha < 1$.

(H3) The map $r \rightarrow f(\cdot, \cdot; r(\cdot))$ is continuous from $L_2((0, 1); R^p)$ to $L_2((0, T) \times (0, 1))$.

The parameter estimation problem associated with (2.1) consists of finding a parameter $\gamma^* \in \Gamma$ that is “optimal” in the sense of providing the best match between observed data and model solutions to (2.1). Although a number of criteria may be used to measure “fit to data,” we consider first a least squares criterion J that is defined in conjunction with distributed data: That is, given distributed observations $\hat{u}_i \in L_2(0, 1)$ at discrete times $t_i \in (0, T)$, $i = 1, \dots, n$, we seek $\gamma^* \in \Gamma$ that minimizes

$$(2.2) \quad J(\gamma) = \sum_{i=1}^n \int_0^1 |\mathcal{C}(t_i, x; \gamma)u(t_i, x; \gamma) - \hat{u}_i(x)|^2 dx$$

over all $\gamma \in \Gamma$. For each (t_i, x) , the output map $\mathcal{C}(t_i, x; \gamma): R \rightarrow R$ is assumed to be continuous in γ and such that the mapping $x \rightarrow \mathcal{C}(t_i, x; \gamma)\psi(x)$ is in $L_2(0, 1)$ whenever $\psi \in L_2(0, 1)$. We note that data generally is not available in the distributed form given here; often this difficulty can be handled by fitting a curve (using linear interpolation, for example) to discrete data.

We also treat the problem of truly discrete data, i.e., $\hat{u}_{ij} \in R$ is observed sample data at (t_i, x_j) , $j = 1, \dots, \tilde{n}$. In this case the parameter estimation problem consists of determining $\tilde{\gamma}^* \in \Gamma$ that minimizes a “pointwise” fit-to-data criterion,

$$(2.3) \quad \tilde{J}(\gamma) = \sum_{i=1}^n \sum_{j=1}^{\tilde{n}} |\mathcal{C}(t_i, x_j; \gamma)u(t_i, x_j; \gamma) - \hat{u}_{ij}|^2$$

over $\gamma \in \Gamma$. The use of discrete sample data leads to increased technical detail and additional smoothness hypotheses on u_0 and f . We consider this particular estimation problem in § 3.1.

The parameter estimation problems described above are examples of a large class of problems (“inverse” problems) that are widely known to be ill-posed (see, for example, [26] and [31]) from both a theoretical and computational standpoint. The difficulties arise from several sources: a lack of continuous dependence of parameter estimates γ on observed data \hat{u}_i , \hat{u}_{ij} ; the fact that one cannot, in general, assume the existence of a unique parameter γ^* that provides a perfect match of model to data (i.e., $J(\gamma^*) = 0$) due to noise typically present in observed data and the inability of the model to exactly describe the physical problem [19], [20]; and the numerical instabilities (for example, the problem of “highly oscillatory” parameter estimates [26]) that often appear in parameter estimation schemes when the infinite-dimensional parameter space is replaced by a space of (fairly large) finite dimension. A number of approaches have been studied in an effort to alleviate some of these difficulties and restore a type of problem stability. Among these are regularization techniques (see, for example, [26], [27]), imbedding techniques (e.g., [2]), and the parameter set compactness criteria taken here and in a number of related papers (e.g., [4]–[18], [21]). Compactness alone however is not sufficient to avoid the before-mentioned highly oscillatory behavior sometimes seen in parameter approximations; we do not observe such behavior in examples presented here (and in related efforts) due to the fact that we use higher order (cubic, quintic) spline approximations for parameters and thus feel justified in practice in keeping the dimension of the parameter approximation space relatively low. Indeed, we have successfully used this approach in a number of applications and in problems using “real” collected data [12], [13], [14]. It is easy to envision a number of applications, in particular the large oil reservoir problems of interest here, where it is possible that one will need to increase the dimension of the approximate parameter space and thus potentially have to confront such numerical instability problems. In this situation it seems appropriate to add a regularization term to the least squares functionals J (or \tilde{J}).

We turn now to consideration of the parameter estimation problem of interest here (where $\gamma \in \Gamma$ is unknown and to be determined) and first consider the existence of solutions u of (2.1) for a given parameter $\gamma = (\xi, \phi_1, \phi_2, r, u_0) \in \Gamma$. Defining $u(t) \equiv u(t, \cdot) \in L_2(0, 1)$, we may rewrite (2.1) as an initial value problem in u ,

$$(2.4) \quad \begin{aligned} u_t &= \mathcal{A}(q)u(t) + F(t; r), & t \in (0, T), \\ u(0) &= u_0. \end{aligned}$$

Here $q = \phi_1 + H_\xi \phi_2$, $F(t; r) = f(t, \cdot; r(\cdot))$ and the operator $\mathcal{A}(q)$ is defined by $\mathcal{A}(q)\psi = \mathcal{D}(q\mathcal{D}\psi)$ for $\psi \in \text{dom } \mathcal{A}(q) = V_q$, where $V_q = \{\psi \in H_0^1(0, 1) | q\mathcal{D}\psi \in H^1(0, 1)\}$ (throughout we shall use \mathcal{D} to denote the spatial differentiation operator $\partial/\partial x$). We note that solutions u satisfy the continuity equation

$$(2.5) \quad (q\mathcal{D}u)(\xi^-) = (q\mathcal{D}u)(\xi^+),$$

which represents continuity of stress across a transition point, ξ , between distinct spatial regions (layers of porous media, for example).

Our first result is a statement of existence, uniqueness, and regularity properties of solutions of (2.1). In [30] we provide details for a proof of this theorem that uses properties of $\mathcal{A}(q)$ (a self-adjoint, densely defined operator with $\sigma(\mathcal{A}(q)) \subseteq (-\infty, 0)$) to argue that $\mathcal{A}(q)$ generates an analytic semigroup [25], [35] on $L_2(0, 1)$.

THEOREM 2.1. *Let $\gamma = (\xi, \phi_1, \phi_2, r, u_0)$ be given in Γ and let $q = \phi_1 + H_\xi \phi_2$. There exists a unique (classical) solution u to (2.1) with the property that $u(t) \in V_q$ for any $t > 0$. In addition, if $u_0 \in V_q$, then the map $t \mapsto \mathcal{A}(q)u(t)$ is in $C([0, T]; L_2(0, 1))$.*

We note that if u is a solution of (2.1) then u also satisfies (2.1) in a weak sense; i.e., u satisfies

$$(2.6) \quad \begin{aligned} \langle u_t(t), v \rangle &= -\langle q \mathcal{D}u(t), \mathcal{D}v \rangle + \langle F(t; r), v \rangle, \quad t \in (0, T), \\ u(0) &= u_0 \end{aligned}$$

for every $v \in H_0^1(0, 1)$. Applying the Gronwall inequality directly to this formulation, we may argue the continuous dependence of solutions on (possibly unknown) initial data (see [31]) to obtain the following.

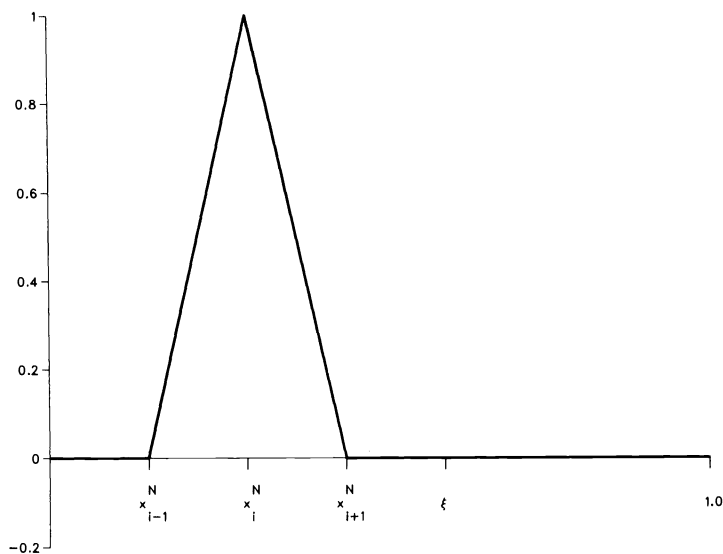
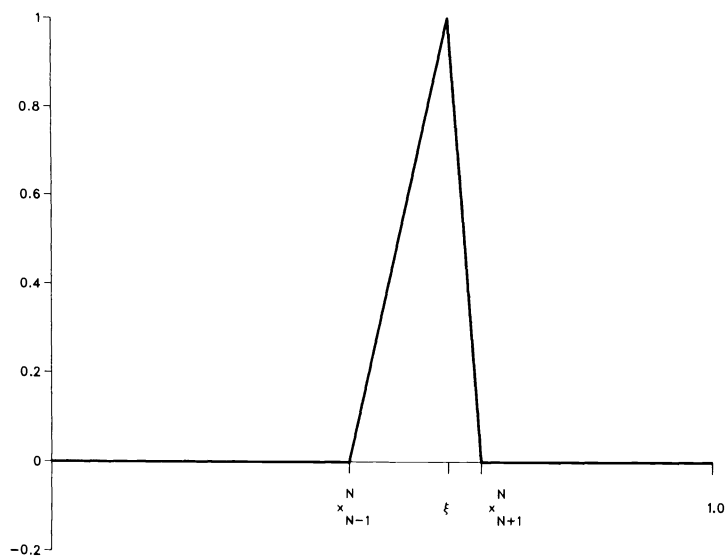
COROLLARY 2.1. *The mapping $u_0 \mapsto u(t; \xi, \phi_1, \phi_2, r, u_0): L_2(0, 1) \rightarrow L_2(0, 1)$ is continuous, uniform in $(\xi, \phi_1, \phi_2, r) \in \mathcal{S}$ and $t \in (0, T)$.*

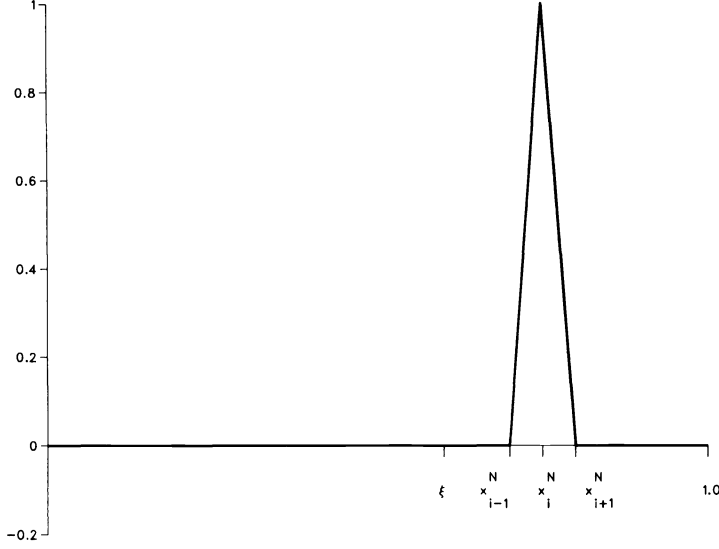
3. A spline-based approximation scheme. Standard numerical optimization schemes applied to the problem of minimizing J (or \tilde{J}) over Γ typically generate a minimizing sequence of parameter iterates, starting from an initial guess, γ^0 . However, schemes of this type generally require that $u(\gamma)$ (the solution of (2.1)) be evaluated as the parameter γ is updated; it is therefore desirable to combine estimation of an optimal parameter γ^* with approximation techniques for solving (2.1). With this goal in mind, we describe a spline-based state/parameter approximation scheme in the same spirit of the ideas found in [7], [11]–[15], [17], [29] to name a few of the related references in this area for (continuous coefficient) parabolic problems.

The convergence arguments developed below are similar to standard variational-type estimates often used in association with finite element approximations (see, for example, [35, p. 129]; [3], [24], [36]) although the estimates given here are complicated somewhat by the presence of unknown parameters. This variational approach was taken in [12], [13], [14] for the problem of estimating continuous coefficients in parabolic systems; we require a somewhat different treatment here primarily due to the fact that we allow discontinuous coefficients, where the points of discontinuity are unknown (necessitating parameter-dependent approximation spaces $X^N(q)$). Thus, an interesting aspect of our approach (and often a source of difficulties) involves the fact that our approximation spaces change with every choice of parameter iterate. We note that although the theoretical problems are quite different, our construction of approximating spaces $X^N(q)$ is somewhat similar to the ideas found in [4], [16], [21]; there the problem was to estimate unknown delays appearing in functional differential equations (there is a correspondence between our treatment of an unknown point of discontinuity and the approach taken in those references to handle an unknown delay, at least from the standpoint of numerical approximation schemes). We turn now to a precise statement of the approximation scheme under consideration.

For any $\gamma = (\xi, \phi_1, \phi_2, r, u_0) \in \Gamma$, we construct parameter-dependent spaces and operators as follows: For $q = \phi_1 + H_\xi \phi_2$ and $N = 1, 2, \dots$, we define $X^N(q) = \text{span} \{B_i^N(q), i = 1, \dots, 2N - 1\}$, where $B_i^N(q)$ denotes the i th continuous piecewise-linear B -spline basis element (satisfying homogeneous boundary conditions) with knots

at $\{x_k^N(q), k=0, \dots, 2N\}$. Here $x_k^N(q) = k\xi/N$, $k=0, \dots, N$, and $x_k^N(q) = \xi + (k-N)(1-\xi)/N$, for $k=N+1, \dots, 2N$. The piecewise linear elements are characterized by $B_i^N(q)(x_k^N) = \delta_{ik}$ for $i, k=1, \dots, 2N-1$ ($B_i^N(q)(0) = B_i^N(q)(1) = 0$); see Figs. 1-3. We remark that in general, for $\gamma, \tilde{\gamma} \in \Gamma$ and $q = \phi_1 + H_\xi \phi_2$, $\tilde{q} = \tilde{\phi}_1 + H_{\tilde{\xi}} \tilde{\phi}_2$, we do not have $X^N(q) \subseteq X^N(\tilde{q})$, nor do we have $X^N(q) \subseteq V_q$ (note that although an element $\psi^N \in X^N(q)$ does have a discontinuity in its first derivative at ξ , ψ^N does not satisfy the continuity equation (2.5) associated with q). As will be illustrated in § 4, this particular ξ -dependent structure for $X^N(q)$ has been chosen in order to minimize computational difficulties that arise when ξ is unknown (and ξ is thus changing throughout an iterative estimation scheme). We take a general Galerkin approach to

FIG. 1. B_i^N , $i=1, \dots, N-1$.FIG. 2. B_N^N .

FIG. 3. B_i^N , $i = N+1, \dots, 2N-1$.

define approximating state systems and then obtain convergence findings by working *directly* with the weak form of these equations. As an alternative approach to that taken here one could define approximating operators $\mathcal{A}^N(q)$ for $\mathcal{A}(q)$ and investigate the sense in which $\mathcal{A}^N(q)$ “converges” to $\mathcal{A}(q)$ (see, for example, Example 2.2 of [28], or [29], for approximating operators that might be used in this context).

For $\gamma \in \Gamma$ fixed and $N=1, 2, \dots$, we seek an approximation to $u(t; \gamma)$ of the form $u^N(t; \gamma) = \sum_{i=1}^{2N-1} w_i^N(t; \gamma) B_i^N(q)$, where the coefficients w_i^N are determined by the system of ordinary differential equations (ODE),

$$(3.1) \quad \begin{aligned} \langle u_i^N(t; \gamma), B_i^N(q) \rangle &= -\langle q \mathcal{D} u^N(t; \gamma), \mathcal{D} B_i^N(q) \rangle + \langle F(t; r), B_i^N(q) \rangle, \quad t \in (0, T), \\ \langle u^N(0; \gamma), B_i^N(q) \rangle &= \langle u_0, B_i^N(q) \rangle, \end{aligned}$$

for $i=1, \dots, 2N-1$. Alternatively, u^N satisfies

$$(3.2) \quad \begin{aligned} \langle u_i^N(t; \gamma), v \rangle &= -\langle q \mathcal{D} u^N(t; \gamma), \mathcal{D} v \rangle + \langle F(t; r), v \rangle, \quad t \in (0, T), \\ u^N(0; \gamma) &= P^N(q) u_0 \end{aligned}$$

for all $v \in X^N(q)$; here $P^N(q): L_2(0, 1) \rightarrow X^N(q)$ denotes the orthogonal projection (with respect to the usual L_2 topology) along $X^N(q)^\perp$. Associated with (3.1) is an approximate estimation problem, namely that of finding $\tilde{\gamma}^N \in \gamma$ that minimizes

$$(3.3) \quad J^N(\gamma) = \sum_{i=1}^n \int_0^1 |\mathcal{E}(t_i, x; \gamma) u^N(t_i; \gamma)(x) - \hat{u}_i(x)|^2$$

over Γ , where $u^N(\gamma)$ is the solution of (3.1) corresponding to $\gamma \in \Gamma$.

Our initial findings concerning the N th approximate problem (3.1), (3.3) are immediate consequences of the fact that (3.1) is an ODE on $X^N(q)$ and that the basis elements $B_i^N(q)$ (and their spatial derivatives) are continuous in ξ (see § 4 for a more detailed examination of (3.1)).

THEOREM 3.1. *For each N and any $\gamma \in \Gamma$, there exists a unique solution $u^N(\gamma)$ of (3.1), $u^N(t; \gamma) \in X^N(q)$ with the property that the mapping $\gamma \rightarrow u^N(t; \gamma): \Gamma \rightarrow L_2(0, 1)$ is continuous for each $t \in (0, T)$. In addition, the mapping $u_0 \rightarrow u^N(t; (s, u_0)): L_2(0, 1) \rightarrow L_2(0, 1)$ is continuous, uniform in N , $s = (\xi, \phi_1, \phi_2, r) \in \mathcal{S}$, and $t \in (0, 1)$.*

COROLLARY 3.1. *For each N , there exists a solution $\bar{\gamma}^N \in \Gamma$ for the problem of minimizing J^N over Γ .*

An essential step in the process of correlating state variable approximation with the problem of estimating an optimal parameter $\gamma^* \in \Gamma$ (for the original parameter identification problem) is the establishment of the convergence of $u^N(t; \gamma^N)$ to $u(t; \bar{\gamma})$ for any sequence $\{\gamma^N\}$ in Γ that converges to $\bar{\gamma} \in \Gamma$. In what follows we assume that $\{\gamma^N\}$ is given in Γ , $\gamma^N = (\xi^N, \phi_1^N, \phi_2^N, r^N, u_0^N)$, with $\gamma^N \rightarrow \bar{\gamma} = (\bar{\xi}, \bar{\phi}_1, \bar{\phi}_2, \bar{r}, \bar{u}_0) \in \Gamma$ (in the usual product topology on Γ); in addition, we assume that $0 < \bar{\xi} < 1$ (and, in the case of multiple discontinuities, $|\bar{\xi}_k - \bar{\xi}_{k-1}| > 0$, $k = 1, \dots, \mu$). Given $q^N = \phi_1^N + H_{\bar{\xi}^N} \phi_2^N$, we shall henceforth simplify notation and abbreviate $P^N \equiv P^N(q^N)$, $X^N \equiv X^N(q^N)$, and $x_k^N \equiv x_k^N(q^N)$, $k = 0, \dots, 2N$.

LEMMA 3.1. *Let ψ be given in $V_{\bar{q}}$, where $\bar{q} = \bar{\phi}_1 + H_{\bar{\xi}} \bar{\phi}_2$. There exist constants c_1 and c_2 , independent of N , such that*

$$(3.4) \quad |\psi - P^N \psi| \leq c_1 N^{-2} |\mathcal{A}(\bar{q}) \psi|$$

and, for N sufficiently large,

$$(3.5) \quad |\mathcal{D}(\psi - P^N \psi)| \leq c_2 N^{-1} |\mathcal{A}(\bar{q}) \psi|.$$

The proof of these basic linear spline estimates is detailed in [30, pp. 17, 18] where standard arguments (see, for example, [35, pp. 16, 17, 78]) are modified to take into account that $\mathcal{D}\psi$, $\psi \in V_{\bar{q}}$, is discontinuous at $\bar{\xi}$. We note that since $V_{\bar{q}}$ is dense in both $L_2(0, 1)$ and $H_0^1(0, 1)$ (in the respective topologies), we may actually weaken the assumptions on ψ and obtain

$$(3.6) \quad |\psi - P^N \psi| \rightarrow 0 \quad \text{as } N \rightarrow \infty \quad \text{for } \psi \in L_2(0, 1),$$

$$(3.7) \quad |\mathcal{D}(\psi - P^N \psi)| \rightarrow 0 \quad \text{as } N \rightarrow \infty \quad \text{for } \psi \in H_0^1(0, 1)$$

where the rates of convergence depend on ψ . Spline estimates (3.4)–(3.7) are used to establish the convergence of state variable approximations in the result that follows.

THEOREM 3.2. *Assume $\{\gamma^N\}$ is given in Γ such that $\gamma^N \rightarrow \bar{\gamma}$ in Γ . Then for each $t \in (0, T)$,*

$$u^N(t; \gamma^N) \rightarrow u(t; \bar{\gamma}) \quad \text{in } L_2(0, 1)$$

as $N \rightarrow \infty$, where u^N is the solution of (3.1) associated with γ^N and u is the solution of (2.1) associated with $\bar{\gamma}$.

Proof. Due to the dense inclusions of $V_{\bar{q}}$ in $L_2(0, 1)$, and the continuous dependence of u , u^N on initial data (Corollary 2.1 and Theorem 3.1), it suffices to argue convergence for the case that $\bar{\gamma} = (\bar{\xi}, \bar{\phi}_1, \bar{\phi}_2, \bar{r}, \bar{u}_0)$ satisfies $\bar{u}_0 \in V_{\bar{q}}$ so that $u(t; \bar{\gamma}) \in V_{\bar{q}}$, $t \in (0, 1)$.

Denoting $u(t) \equiv u(t; \bar{\gamma})$, $u^N(t) \equiv u^N(t; \gamma^N)$, we note that

$$|u^N(t) - u(t)| \leq |u^N(t) - P^N u(t)| + |P^N u(t) - u(t)|$$

where the second term is $\mathcal{O}(N^{-2})$ from (3.4). To consider the first term, we observe that solutions u , u^N of (2.1), (3.1), respectively, satisfy (2.6) and (3.2) for $v \in X^N$ so that these latter equations may be used to establish that

$$\begin{aligned} \left\langle \frac{d}{dt}(u^N(t) - P^N u(t)), v \right\rangle &= -\langle q^N \mathcal{D}(u^N(t) - P^N u(t)), \mathcal{D}v \rangle + \left\langle \frac{d}{dt}(u(t) - P^N u(t)), v \right\rangle \\ &\quad + \langle \bar{q} \mathcal{D}u(t) - q^N \mathcal{D}P^N u(t), \mathcal{D}v \rangle + \langle F^N(t) - F(t), v \rangle, \end{aligned}$$

for $v \in X^N$, $F(t) \equiv F(t; \bar{r})$, $F^N(t) \equiv F(t, r^N)$, and $q^N = \phi_1^N + H_{\bar{\xi}}^N \phi_2^N$. Letting $v = u^N(t) - P^N u(t) \in X^N$, we argue that

$$\begin{aligned} \frac{1}{2} \frac{d}{dt} |u^N(t) - P^N u(t)|^2 &\leq \frac{1}{2} \left| \frac{d}{dt} (u(t) - P^N u(t)) \right|^2 \\ &\quad + \frac{1}{4m} |\bar{q} \mathcal{D}u(t) - q^N \mathcal{D}(P^N u(t))|^2 + \frac{1}{2} |F^N(t) - F(t)|^2 \\ &\quad + |u^N(t) - P^N u(t)|^2, \end{aligned}$$

where we have repeatedly used the inequality $ab \leq \frac{1}{2}(a^2 + b^2)$. An application of the Gronwall inequality thus yields

$$|u^N(t) - P^N u(t)|^2 \leq e^{2T} \{\tau_1^N + \tau_2^N + \tau_3^N + \tau_4^N\}$$

where

$$\begin{aligned} \tau_1^N &= |u^N(0, \gamma^N) - P^N u(0; \bar{\gamma})|^2, \\ \tau_2^N &= \int_0^T \left| \frac{d}{dt} (u(s; \bar{\gamma}) - P^N u(s; \bar{\gamma})) \right|^2 ds, \\ \tau_3^N &= \frac{1}{2m} \int_0^T |\bar{q} \mathcal{D}u(s, \bar{\gamma}) - q^N \mathcal{D}(P^N u(s; \bar{\gamma}))|^2 ds, \\ \tau_4^N &= \int_0^T |F(s; r^N) - F(s; \bar{r})|^2 ds. \end{aligned}$$

We obtain $\tau_4^N \rightarrow 0$ as $N \rightarrow \infty$ from hypothesis (H3) and note that $\tau_1^N \rightarrow 0$ since $\tau_1^N = |P^N u_0^N - P^N \bar{u}_0| \leq |u_0^N - \bar{u}_0|$. Further τ_2^N may be rewritten [30]

$$\tau_2^N = \int_0^T |u_t(s) - P^N u_t(s)|^2 ds,$$

so that dominated convergence of the integrand is guaranteed from Theorem 2.1 ($s \rightarrow u_t(s)$ is in $L_2((0, T), L_2(0, 1))$ since $u_0 \in V_{\bar{q}}$).

Finally, for N sufficiently large,

$$\begin{aligned} m\tau_3^N &\leq \int_0^T |(\bar{q} - q^N) \mathcal{D}u(s)|^2 + \int_0^T |q^N \mathcal{D}(u(s) - P^N u(s))|^2 \\ &\leq 2 \int_0^T |(\bar{\phi}_1 - \phi_1^N) \mathcal{D}u(s)|^2 + 4 \int_0^T |H_{\bar{\xi}}(\bar{\phi}_2 - \phi_2^N) \mathcal{D}u(s)|^2 \\ &\quad + 4 \int_0^T |(H_{\bar{\xi}} - H_{\xi^N}) \phi_2^N \mathcal{D}u(s)|^2 + \bar{m}^2 \int_0^T |\mathcal{D}(u(s) - P^N u(s))|^2 \\ &\leq 4(|\bar{\phi}_1 - \phi_1^N|_{\infty}^2 + |\bar{\phi}_2 - \phi_2^N|_{\infty}^2) \int_0^T |\mathcal{D}u(s)|^2 \\ &\quad + 2\bar{m}^2 |\bar{\xi} - \xi^N| \int_0^T |\mathcal{D}u(s)|^2 + (\bar{m}c_2 N^{-1})^2 \int_0^T |\mathcal{A}(\bar{q})u(s)|^2 \end{aligned}$$

where we have used (3.5) in the last inequality. Further,

$$\begin{aligned} \int_0^T |\mathcal{D}u(s)|^2 &= \int_0^T \langle \mathcal{D}u(s), \mathcal{D}u(s) \rangle \leq m^{-1} \int_0^T \langle \bar{q} \mathcal{D}u(s), \mathcal{D}u(s) \rangle \\ &\leq m^{-1} \int_0^T |\mathcal{A}(\bar{q})u(s)| |u(s)| < \infty \end{aligned}$$

so that $\tau_3^N \rightarrow 0$ as $N \rightarrow \infty$.

Finally, we turn to parameter convergence (a compactness result) and the parameter estimation problem.

THEOREM 3.3. *For each N let $\bar{\gamma}^N$ denote a solution for the problem of minimizing J^N over Γ . There exists $\gamma^* \in \Gamma$ and a subsequence $\{\bar{\gamma}^{N_k}\}$ of $\{\bar{\gamma}^N\}$ such that*

- (i) $\bar{\gamma}^{N_k} \rightarrow \gamma^*$ in the product topology on Γ ,
- (ii) $u^{N_k}(t; \bar{\gamma}^{N_k}) \rightarrow u(t; \gamma^*)$ for each $t \in (0, T)$,
- (iii) $J^{N_k}(\bar{\gamma}^{N_k}) \rightarrow J(\gamma^*)$,
- (iv) γ^* is a solution to the original parameter estimation problem, namely that of minimizing $J(\gamma)$ over Γ .

Proof. Parts (i)–(iii) are immediate consequences of hypothesis (H1), and Theorem 3.2. To prove part (iv), it suffices to note that

$$J(\gamma^*) = \lim_{N_k \rightarrow \infty} J^{N_k}(\bar{\gamma}^{N_k}) \leq \lim_{N_k \rightarrow \infty} J^{N_k}(\gamma) = J(\gamma) \quad \text{for any } \gamma \in \Gamma$$

($\bar{\gamma}^{N_k}$ is a minimizer for J^{N_k} over Γ), so that γ^* is a solution for the problem of minimizing J over Γ .

Remark 3.1. Our efforts to this point have been focused on state variable approximation only; we must also consider discretization of the infinite-dimensional parameter space Γ in order to implement numerical optimization algorithms. The problem of further approximating parameter sets has been the subject of recent studies (see [12], [17], and [29]); we shall summarize, in particular, the results of [17] as they pertain to the problem at hand. For ease of presentation we shall assume that r, u_0 , are known (there is an easy extension of these ideas to the case where these functional parameters are unknown) so that $\gamma = (\xi, \phi_1, \phi_2)$ is the vector of parameters to be estimated. Since we use cubic B -spline approximations to approximate the functional parameters in our numerical examples (§ 4), we shall restrict our attention to a theory based on cubic splines only; a more general theory may be found in [17]. To this end, we assume the (more regular) parameter set Γ satisfies the following additional hypothesis:

(H4) $\Gamma \subseteq \tilde{\mathcal{F}}(m, \bar{m})$ is compact in the $R \times C^1[0, 1] \times C^1[0, 1]$ topology,

where $\tilde{\mathcal{F}}(m, \bar{m}) \equiv \{s = (\xi, \phi_1, \phi_2) \in [\delta, 1 - \delta] \times C^1[0, 1] \times C^1[0, 1] \mid 0 < m \leq \phi_i(x) \leq \bar{m} \text{ for } x \in [0, 1], \phi_i \in H^2(0, 1), \text{ and } |\mathcal{D}^2 \phi_i| \leq \bar{m}, i = 1, 2\}$, ($\delta \in (0, 1)$ is fixed).

For each M we define the finite-dimensional (approximate) parameter sets Γ^M by $\Gamma^M \equiv i^M(\Gamma)$; here $i^M: R \times C^1[0, 1] \times C^1[0, 1] \rightarrow R \times C^2[0, 1] \times C^2[0, 1]$ is given by $i^M(\xi, \phi_1, \phi_2) \equiv (\xi, \mathcal{J}^{M, \xi} \phi_1, \mathcal{J}^{M, \xi} \phi_2)$ where $\mathcal{J}^{M, \xi} \phi$ is defined to be the (unique) cubic spline function $\tilde{\phi}(x)$ satisfying $\tilde{\phi}(x_k^M) = \phi(x_k^M)$, $k = 0, 1, \dots, 2M$, and $\mathcal{D}\tilde{\phi}(x_0^M) = \mathcal{D}\phi(x_0^M)$, $\mathcal{D}\tilde{\phi}(x_{2M}^M) = \mathcal{D}\phi(x_{2M}^M)$ (see, for example, [35, Chap. 4]). The knots x_k^M are the ξ -dependent knots described earlier in this section. As is true with the approximation of state variables, the resulting numerical scheme is greatly simplified if the mesh depends on ξ , as well as on M . We shall defer to § 4 a more detailed discussion on computational features of the resulting algorithm.

Numerical implementation of our scheme to estimate functional parameters thus amounts to an optimization problem in the setting of a “double approximation” (approximation of the state variable and parameter spaces) framework, namely the problem of minimizing J^N over Γ^M . Theoretical findings relevant to this task are summarized in the theorem below, the proof of which may be found in [30, pp. 26, 27].

THEOREM 3.4. *Let $\Gamma^M \equiv i^M(\Gamma)$, where Γ satisfies (H4), and let $\bar{\gamma}^{N, M}$ denote a solution to the problem of minimizing J^N over Γ^M . Then there is a subsequence $\{\bar{\gamma}^{N_k, M^j}\}$ of $\{\bar{\gamma}^{N, M}\}$ such that $\bar{\gamma}^{N_k, M^j} \rightarrow \gamma^*$, where γ^* is a solution to the problem of minimizing J over Γ . In fact, any convergent subsequence has as its limit a solution to the original estimation problem.*

3.1. Approximate estimation problems associated with “discrete” data. It is possible, under additional smoothness assumptions on solutions, to use variational-type estimates (similar to those found above or in [12], [37]) to argue H^1 convergence of state variables. Results of this type lead naturally to a statement about the approximation of a solution for the problem of minimizing the “pointwise” fit-to-data criterion \tilde{J} (see (2.3)) over Γ .

For this formulation we follow the ideas of [37] and add an assumption (which in general may impose additional conditions on parameters and the applied force f) to the hypotheses (H1)–(H3) already assumed:

(H5) For any $\gamma \in \Gamma$, the mapping $s \rightarrow u_t(s; \gamma)$ is in $L_2((0, T); H_0^1(0, 1))$.

Although we define approximate state spaces $X^N(q)$ as before, we now seek approximations u^N to u that satisfy

$$(3.8) \quad \begin{aligned} \langle u_t^N(t), v \rangle &= -\langle q \mathcal{D} u^N(t), \mathcal{D} v \rangle + \langle F(t; r), v \rangle, \quad t \in (0, T), \\ u^N(0) &= \mathcal{P}^N(q) u_0 \end{aligned}$$

for all $v \in X^N(q)$; here \mathcal{P}^N differs from P^N defined in (3.2) in that $\mathcal{P}^N: H_0^1(0, 1) \rightarrow X^N(q)$ is the orthogonal projection in the $H_0^1(0, 1)$ (rather than $L_2(0, 1)$) topology. In addition, the parameter set Γ is taken to be a subset of $\hat{\mathcal{S}} \times H_0^1(0, 1)$, where $\hat{\mathcal{S}} \equiv \{(\xi, \phi_1, \phi_2, r) \in \mathcal{S} \mid \|\mathcal{D}\phi_i\|_\infty \leq \bar{m}, i = 1, 2\}$.

It is easily shown that there exists a unique solution $u^N(\gamma)$ to (3.8) that depends continuously on parameters; in addition, we obtain convergence findings that are H^1 analogues of earlier approximation results. We state only these results and refer the reader to [30, pp. 29–35] for proofs. Extensions of these findings are also possible in order to include ρ as a parameter or to prove a “double approximation” theorem similar to Theorem 3.4.

THEOREM 3.5. *Let $\{\gamma^N\}$ be given in Γ with $\gamma^N \rightarrow \bar{\gamma} \in \Gamma$ in the $\mathcal{S} \times H_0^1(0, 1)$ topology. Then, for each $t \in [0, T)$,*

$$u^N(t; \gamma^N) \rightarrow u(t; \bar{\gamma}) \quad \text{in } H_0^1(0, 1) \quad \text{as } N \rightarrow \infty.$$

THEOREM 3.6. *For each N , let $\tilde{\gamma}^N$ denote a solution for the problem of minimizing \tilde{J}^N over Γ , where*

$$\tilde{J}^N(\gamma) = \sum_{i=1}^n \sum_{j=1}^{\tilde{n}} |\mathcal{C}(t_i, x_j; \gamma) u^N(t_i, x_j; \gamma) - \hat{u}_{ij}|^2$$

and u^N is the solution of (3.8) associated with $\gamma \in \Gamma$. Then there exists $\tilde{\gamma}^ \in \Gamma$ and a subsequence $\{\tilde{\gamma}^{N_k}\}$ of $\{\tilde{\gamma}^N\}$ such that $\tilde{\gamma}^{N_k} \rightarrow \tilde{\gamma}^*$, $\tilde{J}^{N_k}(\tilde{\gamma}^{N_k}) \rightarrow \tilde{J}(\tilde{\gamma}^*)$, and $\tilde{\gamma}^*$ is a solution for the problem of minimizing \tilde{J} over Γ .*

4. Implementation and numerical findings. A desirable feature of the spline-based scheme developed in preceding sections is the ease of implementation of the approximation ideas, especially when the points of discontinuity ξ_i , $i = 1, \dots, \mu - 1$, for coefficients are unknown and to be estimated. In what follows we describe how the particular *state* approximation framework chosen here serves to facilitate (from a computational standpoint) the *parameter* estimation/approximation process. We conclude the section by presenting our findings for some representative test examples.

We begin by examining the approximating ordinary differential equation (3.1) rewritten here in terms of $w^N(t; \gamma) \equiv (w_1^N(t; \gamma), w_2^N(t; \gamma), \dots, w_{2N-1}^N(t; \gamma))^T$, where

the w_i^N , defined in § 3, are the coefficients in the expansion $u^N(t; \gamma) = \sum_{i=1}^{2N-1} w_i^N(t; \gamma) B_i^N(q)$. Using this notation, the ODE may be written

$$(4.1) \quad \begin{aligned} Q^N w^N(t) &= -K^N w^N(t) + G^N(t), \quad t \in (0, T), \\ w^N(0) &= w_0^N; \end{aligned}$$

here the $(2N-1)$ -square matrices $Q^N = Q^N(\gamma)$ and $K^N = K^N(\gamma)$ have entries

$$Q_{i,j}^N = \langle B_j^N(q), B_i^N(q) \rangle, \quad K_{i,j}^N = \langle q \mathcal{D} B_j^N(q), \mathcal{D} B_i^N(q) \rangle,$$

while the perturbation term and initial condition satisfy

$$G^N(t) = G^N(t; \gamma) \equiv (\langle F(t; r), B_1^N(q) \rangle, \dots, \langle F(t; r), B_{2N-1}^N(q) \rangle)^T$$

and

$$w_0^N = w_0^N(\gamma) \equiv (Q^N)^{-1} (\langle u_0, B_1^N(q) \rangle, \dots, \langle u_0, B_{2N-1}^N(q) \rangle)^T,$$

respectively.

In order to best indicate some of the advantages of the chosen approximation framework, we first consider the special case where only q is unknown, where $q = \phi_1 + H_\xi \phi_2$, and ϕ_1 and ϕ_2 are constants. First, it is easy to see how our choice of a linear spline approximation scheme yields matrices K^N and Q^N that are quite simple in structure: For a given value of q , the inner products appearing in these matrices may be determined from explicit formulas (depending on N and ξ), a few of which are given here. For example, diagonal entries in the (tridiagonal) matrices Q^N and K^N are given by

$$(4.2) \quad \begin{aligned} Q_{i,i}^N &= 2\xi/3N, \quad i = 1, \dots, N-1, \\ Q_{N,N}^N &= 1/3N, \\ Q_{i,i}^N &= 2(1-\xi)/3N, \quad i = N+1, \dots, 2N-1, \end{aligned}$$

$$(4.3) \quad K_{i,i}^N = 2N\phi_1/\xi, \quad i = 1, \dots, N-1,$$

$$(4.4) \quad K_{N,N}^N = N\phi_1/\xi + N(\phi_1 + \phi_2)/(1-\xi),$$

$$(4.4) \quad K_{i,i}^N = 2N(\phi_1 + \phi_2)/(1-\xi), \quad i = N+1, \dots, 2N-1,$$

with similar representations for off-diagonal elements. We note that we are able to avoid time-consuming and error-producing numerical quadratures; in addition, our approach is more desirable (from a computational point of view) than a method based on a uniform mesh size. For example, if for each N we simply subdivide $[0, 1]$ into units of length $1/N$ (so that position of ξ is not taken into account) the matrix Q^N will be fixed throughout the estimation process; this however is at the expense of considerable added difficulties associated with evaluating entries in K^N . Using a uniform mesh, some of the inner products must be “broken up” at the point ξ , e.g.,

$$\langle q \mathcal{D} B_j^N, \mathcal{D} B_i^N \rangle = \phi_1 \int_0^\xi \mathcal{D} B_j^N \mathcal{D} B_i^N + (\phi_1 + \phi_2) \int_\xi^1 \mathcal{D} B_j^N \mathcal{D} B_i^N,$$

requiring (multiple) numerical quadratures every time that q (and thus ξ) is updated. In contrast, with the ξ -dependent structure chosen here we need only recombine simple algebraic expressions (such as those given in (4.2)–(4.4)) to obtain the elements of K^N .

Many of these computational advantages are still present in the case where the ϕ_i are not assumed to be constant. If, for example, M and N are fixed and Γ^M consists

of cubic spline element approximations for ϕ_1, ϕ_2 (defined on a ξ -dependent mesh of points x_k^M) many of the quadratures may still be performed in advance of the iterative process. In particular, if we let $\phi_n^M(x) = \sum_{m=1}^{k(M)} \gamma_{n,m}^M \mathcal{G}_m^M(x)$, for $n = 1, 2$, where \mathcal{G}_m^M are the usual cubic B -spline basis elements defined using the mesh points $\{x_k^M, k = 0, \dots, 2M\}$, we find that $K_{i,j}^N = \langle q^M \mathcal{D}B_j^N, \mathcal{D}B_i^N \rangle$ may now be written as

$$(4.5) \quad K_{i,j}^N = \sum_{m=1}^{k(M)} \gamma_{1,m}^M \langle \mathcal{G}_m^M \mathcal{D}B_j^N, \mathcal{D}B_i^N \rangle_{L_2(0,\xi)} + \sum_{m=1}^{k(M)} (\gamma_{1,m}^M + \gamma_{2,m}^M) \langle \mathcal{G}_m^M \mathcal{D}B_j^N, \mathcal{D}B_i^N \rangle_{L_2(\xi,1)}.$$

Since simple explicit algebraic expressions (in terms of ξ and M) exist for \mathcal{G}_m^M , the quadrature in (4.5) may also be evaluated analytically (yielding expressions involving ξ, M , and N), avoiding numerical integration.

We consider here numerical examples where γ is known and we have generated synthetic data for use in testing our ideas. In all examples presented here, we assume that r and u_0 are known and fixed at their true values so that only $q = \phi_1 + H_\xi \phi_2$ is unknown (i.e., $\gamma = (\xi, \phi_1, \phi_2)$) and to be determined. The special problems associated with estimating this discontinuous coefficient have been the focus of our efforts throughout; the problem of identifying continuous functional parameters and initial conditions has been considered elsewhere [16], [17], [21]. For each example that follows, both γ^* and $u(\gamma^*)$ are selected in advance while the appropriate forcing function f is artificially determined by substituting $\gamma^*, u(\gamma^*)$ into (2.1). For chosen sample times $t_i, i = 1, \dots, n$, and sampling locations $x_j, j = 1, \dots, \tilde{n}$ (discrete data is used for these examples), data is generated by setting $\hat{u}_{ij} = u(t_i, x_j; \gamma^*)$, with random noise added in some cases. We note that the sample data is *not* generated using our spline-based scheme; rather, the data is constructed from an analytic expression for the solution and thus is independent of the methods we illustrate here.

We begin the parameter estimation process by supplying an initial guess of γ^0 to IMSL's minimization routine ZXSSQ (a Levenberg-Marquardt algorithm) which numerically attempts to determine a minimum, for given N , to \tilde{J}^N (using $\mathcal{C} \equiv 1$ in (2.3)) over a fixed constraint set Γ^M . Here $u^N(\gamma)$ is the solution to (3.1) calculated using IMSL's DGEAR, an ODE solver, where the known values of u_0 and f are used in the equations. We note that although we are actually using the cost functional associated with discrete observations \hat{u}_{ij} , the approximating equations (4.1) differ somewhat from those given in (3.8) (where an H^1 projection \mathcal{P}^N is used instead of the usual L_2 projection P^N). Indeed it is not surprising that, in practice, we obtain pointwise convergence of the approximating states under hypotheses more general than those needed in § 3.1 so that we may, in fact, relax some of the restrictions on the approximating system.

Example 4.1. In our first example we take

$$q^*(x) = \begin{cases} 15, & 0 \leq x < .6, \\ 50, & .6 \leq x \leq 1, \end{cases}$$

and define $u(t, \cdot; \gamma^*) \in \text{dom } \mathcal{A}(q^*)$ by

$$u(t, x; \gamma^*) = \begin{cases} x(70 - 100x)(t^2 + 2), & 0 \leq x < .6, \\ (15 - 15x)(t^2 + 2), & .6 \leq x \leq 1. \end{cases}$$

In Examples 4.1(a)–4.1(c) below we seek to estimate $\gamma = (\xi, \phi_1, \phi_2) \in \Gamma \subseteq \mathbb{R}^3$ (with true value $\gamma^* = (.6, 15, 50)$) using an initial guess of $\gamma^0 = (.8, 30, 30)$. In each case we obtain the converged values $\tilde{\gamma}^N$ for $N = 4, 8, 16$, and 24, using γ^0 to start the iterative scheme

for $N = 4$, and previous converged values as start-up for $N = 8, 16, 24$ (e.g., $\bar{\gamma}^4$ is used as initial guess for the $N = 8$ run). We note that convergence was also obtained (in this and other examples) for other choices of start-up values γ^0 . In the present example we were also able to obtain similar findings using $\gamma^0 = (.99, 30, 30)$ (so that ξ^0 is far from the true value, ξ^*).

Example 4.1(a). Data is generated for this example using $\hat{u}_{ij} = u(t_i, x_j; \gamma^*)$ for $t_i = .5i, i = 1, \dots, 4$, and $x_j = .1j, j = 1, \dots, 9$. Our findings are reported in Table 4.1(a).

Example 4.1(b). We repeat the last example except that spatial sampling locations are now given by $x_j = .1j + .05, j = 0, 1, \dots, 9$ (so that there is no spatial observation point at ξ^* , the point of discontinuity). We summarize our results in Table 4.1(b) and note there is little change between this example and Example 4.1(a).

Example 4.1(c). We repeat Example 4.1(a), but add noise to the data. In this case we define $\hat{u}_{ij} = u(t_i, x_j; \gamma^*) + r_{ij}$ where $\{r_{ij}\}$ are Gaussian random numbers which (with 98% certainty) fall in the range $[-.06\bar{u}, .06\bar{u}]$, $\bar{u} = \sum_{i,j} \hat{u}_{ij} / (n\tilde{n})$. Our findings for this example are summarized in Table 4.1(c).

In the examples that follow we shall shorten our discussion by abbreviating the length (and number) of tables and by displaying some results graphically. The rather detailed presentation given for Example 4.1 was provided simply for the purpose of observing if noise in the data or changes in the placement of data affected the outcome.

TABLE 4.1(a)
Example 4.1(a).

N	$\bar{\xi}^N$	$\bar{\phi}_1^N$	$\bar{\phi}_2^N$	\bar{J}^N	CP time (secs)	No. of iterates
4	.623	14.669	51.950	1.5×10^2	28	13
8	.602	14.845	50.672	1.5×10^0	54	7
16	.600	14.961	50.095	8.8×10^{-2}	202	7
24	.600	15.000	50.000	5.6×10^{-9}	141	4

TABLE 4.1(b)
Example 4.1(b).

N	$\bar{\xi}^N$	$\bar{\phi}_1^N$	$\bar{\phi}_2^N$	\bar{J}^N	CP time (secs)	No. of iterates
4	.621	14.956	48.494	9.0×10^1	32	20
8	.607	15.009	50.063	3.8×10^1	35	7
16	.601	14.991	49.728	1.2×10^{-1}	355	13
24	.600	15.000	50.000	5.7×10^{-9}	239	5

TABLE 4.1(c)
Example 4.1(c) (noisy data).

N	$\bar{\xi}_1^N$	$\bar{\phi}_1^N$	$\bar{\phi}_2^N$	\bar{J}_2^N	CP time (secs)	No. of iterates
4	.621	14.730	51.573	1.6×10^2	27	10
8	.599	14.887	50.434	8.1×10^0	68	10
16	.598	14.991	50.296	8.0×10^0	178	5
24	.597	15.006	50.149	8.3×10^0	733	8

In Example 4.2 below, we illustrate the use of our methods in problems with two discontinuities ξ_1, ξ_2 , in q ; the example also serves to illustrate that we are able to accurately estimate ξ_i even when the forcing function f does not contain discontinuities at each of those points.

Example 4.2. We seek here the “true” value of q given by

$$q^* = \begin{cases} 1.0, & 0 \leq x < .2, \\ 6.0, & .2 \leq x < .6, \\ 0.5, & .6 \leq x \leq 1. \end{cases}$$

In this case, $\gamma^* = (\xi_1^*, \xi_2^*, \phi_1^*, \phi_2^*, \phi_3^*) = (.2, .6, 1, 6, .5)$ and the true solution corresponding to γ^* is

$$u(t, x; \gamma^*) = \begin{cases} 30x, & 0 \leq x < .2, \\ 5x + 5, & .2 \leq x < .6, \\ -200x^2 + 300x - 100, & .6 \leq x \leq 1, \end{cases}$$

with data available at $t_i = .5i, i = 1, \dots, 4$, and $x_j = .1j, j = 1, \dots, 9$. For an initial guess of $\gamma^0 = (.3, .7, 5, 5, 5)$ we determined $\bar{\gamma}^8 = (.200, .600, 1.000, 6.000, .5000)$ after 501 CP seconds, with $\bar{J}^8 = 8.3 \times 10^{-6}$.

We consider now two examples where the “true” $q^* = \phi_1^* + H_{\xi^*} \phi_2^*$ involves non-constant values of ϕ_1^* and ϕ_2^* . In each case we search for approximate ϕ_1 and ϕ_2 in the cubic spline space constructed using an $M = 1$ level of approximation (see § 3).

Example 4.3. Here we seek to estimate the “true” parameter

$$q^* = \begin{cases} 2x + 12, & 0 \leq x < .6, \\ 1100x^2/9, & .6 \leq x < 1, \end{cases}$$

starting from the initial guess for q of $q^0 \equiv 3$ on $[0, 1]$ (with start-up value for ξ of $\xi = .5$). The solution

$$u(t, x; \gamma^*) = \begin{cases} (70x - 100x^2)(t^2 + 2), & 0 \leq x < .6, \\ 15(1 - x)(t^2 + 2), & .6 \leq x \leq 1, \end{cases}$$

is used to generate data at $t_i = .5i, i = 1, \dots, 4$, and $x_j = .1j, j = 1, \dots, 9$. In Fig. 4a we compare the estimated $\bar{q}^{N,M} = \bar{\phi}_1^{N,M} + H_{\bar{\xi}^N} \bar{\phi}_2^{N,M}$ ($N = 16, M = 1$) with the “true” coefficient q^* . Figure 4b is the same graph that has been enlarged and restricted to the interval $[.4, .63]$ in order to better distinguish between “true” and approximate curves.

Example 4.4. Again we estimate a functional parameter with true representation given by

$$q^* = \begin{cases} 27.424 - 40x, & 0 \leq x < .3, \\ 90(x - .3)^2 + 18, & .3 \leq x \leq 1. \end{cases}$$

Data is generated as in Example 4.3, using instead the solution

$$u(t, x; \gamma^*) = \begin{cases} 200xt^2(.5 - x), & 0 \leq x < .3, \\ 17.143t^2(1 - x), & .3 \leq x \leq 1. \end{cases}$$

The start-up guess of

$$q^0 = \begin{cases} 18, & 0 \leq x < .2, \\ 48, & .2 \leq x \leq 1, \end{cases}$$

is depicted in Fig. 2, along with the “converged” value of $\bar{q}^{N,M}$ for $N = 24, M = 1$. In this particular example, ϕ_1 and ϕ_2 were estimated easily but we were unable, at the

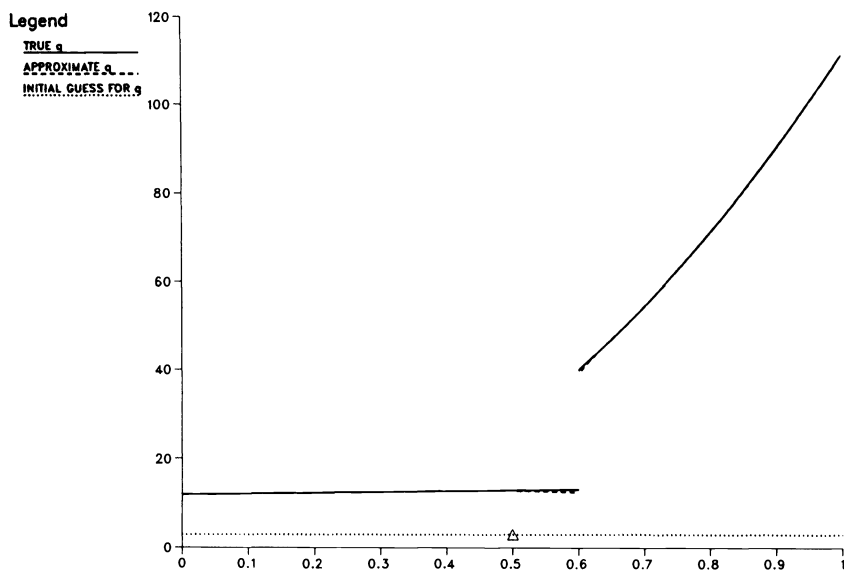


FIG. 4a. Example 4.3.

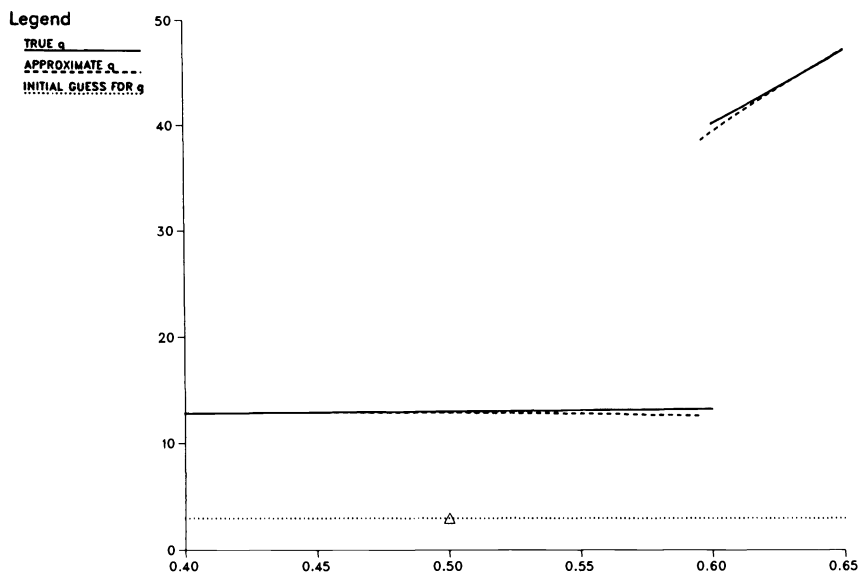


FIG. 4b. Example 4.3.

outset, to move ξ from its initial guess of $\xi^0 = .2$. The results displayed in Fig. 5 are actually the outcome of a multistep process whereby the ϕ_i were held fixed while we iterated on ξ and then the process was repeated with ξ fixed and ϕ_i changing (see [30] for a full discussion of this example). This somewhat adaptive algorithm is a common approach taken, often of necessity, when real data is used in connection with model-building applications, e.g. [13], [14]. The difficulties experienced in this particular example may be due to some well-known limitations of the optimization scheme (Levenberg-Marquardt) we chose to use with our approximation ideas. It has been our experience that difficulties sometimes arise when this scheme is used in conjunction

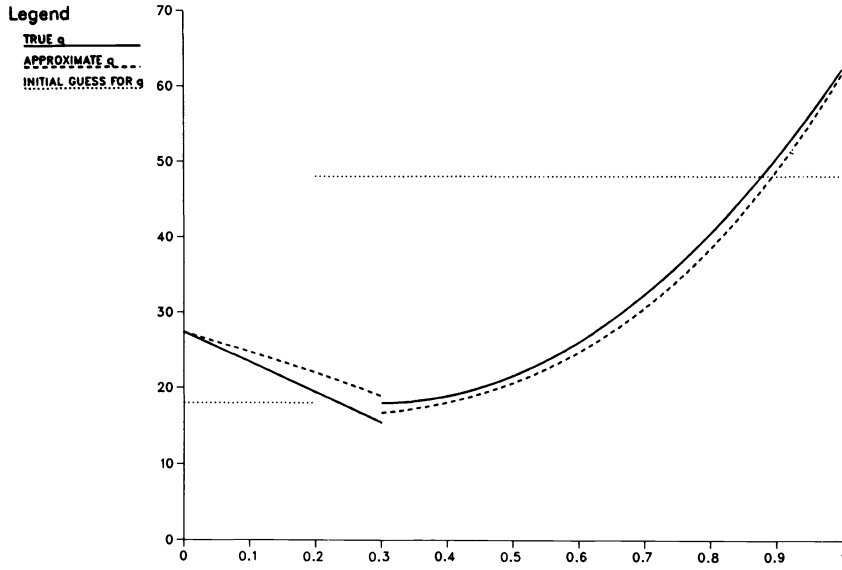


FIG. 5. Example 4.4.

with more than 7 or 8 unknown parameters (there are 9 degrees of freedom in this example). We note that we did *not* experience such difficulties in the last example (also with 9 degrees of freedom) possibly due to the fact that the difference between $q(\xi^+)$ and $q(\xi^-)$ is much greater in that example, making the problem more sensitive to the placement of ξ .

Finally, we remark that a drawback of our approximation framework is that we must specify the *number* of discontinuities in advance of the estimation process. Fortunately, it is possible to overestimate and underestimate this number and still obtain useful information. This will be the focus of our last two examples.

Example 4.5. We repeat Example 4.2 except that we assume throughout that q is discontinuous at only one point (while two discontinuities are actually present in q^*); we also allow spatial variation in ϕ_1 and ϕ_2 and approximate using cubic splines. An initial guess for q and a converged estimate $\bar{q}^{N,M}$ ($N = 24$, $M = 1$) are depicted in Fig. 6 where it is interesting to note that the initial guess of $\xi^0 = .4$ converges to a value close to that of the true (second) discontinuity, $\xi_2^* = .6$. In addition, to the right of this point the estimated shape of q begins to approximate the constant function ϕ_3^* , while to the left of that point the rapidly increasing estimated shape gives an indication that we have underestimated the number of discontinuities present.

Example 4.6. We repeat Example 4.1, except that now we *overestimate* the number of discontinuities in q . We assume throughout that $q = \phi_1 + H_{\xi_1} \phi_2 + H_{\xi_2} \phi_3$ where ϕ_1 , ϕ_2 , and ϕ_3 are constants. For an initial guess of

$$q^0 = \begin{cases} 25, & 0 \leq x < .5, \\ 5, & .5 \leq x < .7, \\ 20, & .7 \leq x \leq 1, \end{cases}$$

we obtained ($N = 8$, 291 CP seconds)

$$\bar{q}^8 = \begin{cases} 14.95, & 0 \leq x \leq .503, \\ 14.99, & .503 \leq x \leq .600, \\ 50.05, & .600 \leq x \leq 1; \end{cases}$$

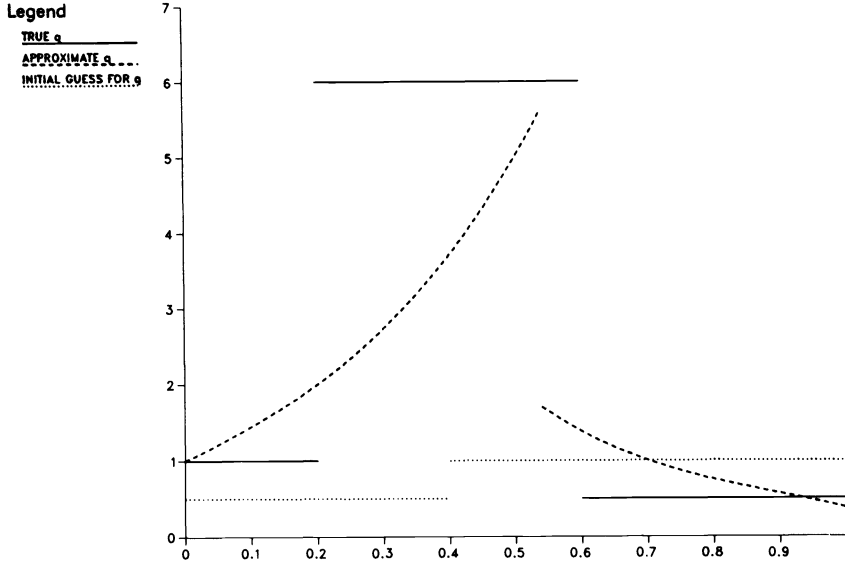


FIG. 6. Example 4.5.

repeating the same example but with a different initial guess,

$$q^0 = \begin{cases} .001, & 0 \leq x < .333, \\ .001, & .333 \leq x < .667, \\ .001, & .667 \leq x \leq 1, \end{cases}$$

we observed the following converged values ($N = 16$)

$$\bar{q}^{16} = \begin{cases} 2.44, & 0 \leq x < .0001, \\ 15.05, & .0001 \leq x < .6001, \\ 49.88, & .6001 \leq x \leq 1. \end{cases}$$

A close inspection of either result reveals that we were, in fact, able to accurately estimate q^* (as defined in Example 4.1), even though a two-discontinuity approximation structure was incorrectly used throughout.

Remark 4.1. We note that our success in numerical examples given here is not due to the fact that “true” parameters happen to lie in the approximating spaces Γ^M . We have also observed equally good results in examples where the parameters cannot be fit exactly by elements of Γ^M for any M (e.g., see Example 4.2 of [17], or see [6], [9], [32] for a number of examples in the context of several applications).

Remark 4.2. We recognize that some of the success seen in our test examples is due to our choice of approximate parameter spaces Γ^M , where M is small. Indeed, an advantage of our approach from a computational standpoint is that the choice of M need not depend on the size (i.e., N) of the approximate state space. Because M is small (and thus the number of unknowns is kept small) we are able to achieve success with an optimization scheme like the Levenberg–Marquardt algorithm. In applications where one might desire to increase the size of the parameter space, it is likely that a direct gradient algorithm will be required to handle the larger number of parameter degrees of freedom.

5. Concluding remarks. In conclusion, we have developed a spline-based approximation framework for the problem of estimating discontinuous functional coefficients

(including locations of discontinuities) in one-dimensional parabolic equations. We feel that some positive aspects of our approach include the separate treatment of approximations for state variables and for parameters. Computational advantages of this framework are seen in the combination of approximate state spaces of *large* dimension with parameter spaces of *small* dimension; when such an approach is justified, the latter often facilitates the numerical parameter search. In addition, our scheme has been developed specifically to minimize any computational difficulties associated with unknown points of discontinuity that are continually being updated throughout an iterative estimation algorithm.

We are currently working to further develop these ideas and to extend the theory to other applications, e.g., hyperbolic (seismic) equations and higher order (elastic beam) systems. We are also working to develop a related theory for two-dimensional domains, although for obvious reasons this is not simply a trivial extension of the ideas presented thus far.

Acknowledgment. The author would like to express appreciation to Professor H. T. Banks for numerous insightful discussions during the course of this work.

REFERENCES

- [1] R. A. ADAMS, *Sobolev Spaces*, Academic Press, New York, 1975.
- [2] H. W. ALT, K.-H. HOFFMAN AND J. SPREKELS, *A numerical procedure to solve certain identification problems*, Internat. Ser. Numer. Math., 68 (1984), pp. 11-43.
- [3] I. BABUSKA AND K. AZIZ, *Survey lecture on the mathematical foundations of the finite element method*, in *The Mathematical Foundation of the Finite Element Method with Applications to Partial Differential Equations*, A. K. Aziz, ed., Academic Press, New York, 1972, pp. 3-363.
- [4] H. T. BANKS, J. A. BURNS AND E. M. CLIFF, *Parameter estimation and identification for systems with delays*, this Journal, 19 (1981), pp. 791-828.
- [5] H. T. BANKS AND J. M. CROWLEY, *Parameter estimation for distributed systems arising in elasticity*, in *Proc. Symposium on Engineering Sciences and Mechanics* (National Cheng Kung University, Tainan, Taiwan, Dec. 28-31, 1981), pp. 158-177; LCDS Tech. Rep. 81-24, Brown Univ., November, 1981.
- [6] ———, *Parameter identification in continuum models*, LCDS Rep. M-83-1, Brown Univ., March, 1983; in *Proc. Amer. Cont. Conf.*, San Francisco, June, 1983, pp. 997-1001.
- [7] H. T. BANKS, J. M. CROWLEY AND K. KUNISCH, *Cubic spline approximation techniques for parameter estimation in distributed systems*, IEEE Trans. Automat. Control, 28 (1983), pp. 773-786.
- [8] H. T. BANKS, P. L. DANIEL (LAMM) AND E. S. ARMSTRONG, *A spline-based parameter and state estimation technique for static models of elastic surfaces*, ICASE Rep. No. 82-25, NASA LRC, Hampton, VA, June, 1983.
- [9] ———, *Spline-based estimation techniques for parameters in elliptic distributed systems*, in *Proc. Fourth VPI & SU/AIAA Symposium on Dynamics and Control of Large Structures* (Blacksburg, June, 1983), to appear; LCDS Rep. No. 83-22, Brown Univ., June 1983.
- [10] H. T. BANKS, K. ITO AND K. A. MURPHY, *Computational methods for estimation of parameters in hyperbolic systems*, in *Proc. Conf. on Inverse Scattering: Theory and Application* (Tulsa, May 16-18, 1983), Society for Industrial and Applied Mathematics, Philadelphia, PA, 1983.
- [11] H. T. BANKS AND P. KAREIVA, *Parameter estimation techniques for transport equations with application to population dispersal and tissue bulk flow models*, J. Math. Biol., 17 (1983), pp. 253-273.
- [12] H. T. BANKS, P. M. KAREIVA AND P. K. DANIEL LAMM, *Estimation techniques for transport equations*, in *Proc. Int'l. Conf. on Mathematics in Biology and Medicine* (Bari, July 18-22, 1983), to appear; LCDS Tech. Rep. No. 83-23, Brown Univ., July, 1983.
- [13] ———, *Estimation of temporally and spatially varying coefficients in models for insect dispersal*, LCDS Tech. Rep. No. 83-14, Brown Univ., June, 1983.
- [14] ———, *Modeling insect dispersal and estimating parameters when mark-release techniques may cause initial disturbances*, J. Math. Biol., to appear.
- [15] H. T. BANKS AND K. KUNISCH, *An approximation theory for nonlinear partial differential equations with applications to identification and control*, this Journal, 20 (1982), pp. 815-849.

- [16] H. T. BANKS AND P. K. DANIEL LAMM, *Estimation of delays and other parameters in nonlinear functional differential equations*, this Journal, 21 (1983), pp. 895-915.
- [17] H. T. BANKS AND PATRICIA DANIEL LAMM, *Estimation of variable coefficients in parabolic distributed systems*, IEEE Trans. Automat. Control, 30 (1985), pp. 386-398.
- [18] H. T. BANKS AND K. A. MURPHY, *Estimation of coefficients and boundary parameters in hyperbolic systems*, LCDS Rep. 84-5, February, 1984, Brown University; this Journal, submitted.
- [19] G. CHAVENT, *About the stability of the optimal control solution of inverse problems*, in Inverse and Improperly Posed Problems in Differential Equations, G. Anger, ed., Akademie-Verlag, Berlin, 1979, pp. 45-58.
- [20] ———, *Local stability of the output least square parameter estimation technique*, INRIA Rapport No. 136, Domaine de Voluceau, Rocquencourt, B.P. 105, 78150 Le Chesnay, France, 1982; Math. Appl. Comput., 2 (1983), pp. 3-22.
- [21] P. L. DANIEL (LAMM), *Spline-based approximation methods for the identification and control of nonlinear functional differential equations*, Ph.D. dissertation, Brown Univ., Providence, RI, June, 1981.
- [22] R. E. EWING, *Determination of coefficients in reservoir simulation*, in Numerical Treatment of Inverse Problems in Differential and Integral Equations, P. Deufhard and E. Hairer, eds., Birkhäuser, Boston, 1983, pp. 206-226.
- [23] ———, *The mathematics of reservoir simulation*, SIAM Frontiers in Appl. Math., 1 (1984), to appear.
- [24] G. FAIRWEATHER, *Finite Element Galerkin Methods for Differential Equations*, Marcel Dekker, New York, 1978.
- [25] J. A. GOLDSTEIN, *Semigroups of Operators and Abstract Cauchy Problems*, Lecture Notes, Tulane Univ., 1970.
- [26] C. KRAVARIS AND J. SEINFELD, *Identification of parameters in distributed parameter systems*, this Journal, to appear.
- [27] ———, *Identification of spatially-varying parameters in distributed parameter systems by discrete regularization*, this Journal, submitted.
- [28] K. KUNISCH, *Identification and estimates of parameters in abstract Cauchy problems*, Preprint, No. 11 (1981), Technical Univ. Graz.
- [29] K. KUNISCH AND L. W. WHITE, *The parameter estimation problem for parabolic equations in multi-dimensional domains in the presence of point evaluations*, Preprint No. 17 (1983), Technical Univ. Graz.
- [30] P. DANIEL LAMM, *Estimation of discontinuous coefficients in parabolic systems: Applications to reservoir simulation*, ICASE Rep. 84-7, NASA Langley Research Center, Hampton, VA, Feb., 1984.
- [31] J. L. LIONS, *Some aspects of modeling problems in distributed parameter systems*, in Proc. IFIP Working Conf. in Rome, 1976, A. Ruberti, ed., Lecture Notes in Control and Information Sciences 1, Springer-Verlag, Berlin, 1978, pp. 11-41.
- [32] K. A. MURPHY, *A spline-based approximation method for inverse problems for a hyperbolic system including unknown boundary parameters*, Proc. 22nd IEEE Conf. on Decision and Control, San Antonio, Dec., 1983, to appear.
- [33] A. PAZY, *Semigroups of Linear Operators and Applications to Partial Differential Equations*, Applied Mathematical Sciences 44, Springer-Verlag, New York, 1983.
- [34] M. P. POLIS, *The distributed system parameter identification problem: a survey of recent results*, in Proc. 3rd Symp. on Control of Distributed Parameter Systems, Pergamon Press, New York, 1982, pp. 45-58.
- [35] M. H. SCHULTZ, *Spline Analysis*, Prentice-Hall, Englewood Cliffs, NJ, 1973.
- [36] G. STRANG AND G. J. FIX, *An Analysis of the Finite Element Method*, Prentice-Hall, Englewood Cliffs, NJ, 1973.
- [37] M. F. WHEELER, *L_∞ estimates of optimal orders for Galerkin methods for one-dimensional second order parabolic and hyperbolic equations*, SIAM J. Numer. Anal., 10 (1973), pp. 908-913.

AN APPROACH TO DISCRETE-TIME STOCHASTIC CONTROL PROBLEMS UNDER PARTIAL OBSERVATION*

GIOVANNI B. DI MASI† AND WOLFGANG J. RUNGGLALDIER‡

Abstract. We consider a general class of discrete-time nonlinear stochastic control problems with partial observation, for which in general only ε -optimal controls exist, and provide a method for explicitly computing them. Transforming, as usual, these problems into equivalent ones with complete observation leads to various difficulties, in particular to a nonlinear filtering problem. We first define a subclass of the given problems such that the associated nonlinear filtering problem can be explicitly solved and, for each $\delta > 0$, a δ -optimal control computed. We then show that, under suitable assumptions, for each original problem and each given $\varepsilon > 0$, a problem in the particular class and a $\delta > 0$ can be found, such that a δ -optimal control for the latter is ε -optimal for the former.

Key words. stochastic control, nonlinear filtering, ε -optimal controls, approximate dynamic programming

AMS(MOS) subject classifications. 93E26, 93E11, 93E25

1. Introduction. Consider the following discrete-time stochastic control problem: a partially observable process $\{x_t, y_t\}$, $x_t, y_t \in \mathbb{R}$, with x_t the unobservable and y_t the observable components, is given for $t = 0, 1, \dots, T$ on some probability space (Ω, \mathcal{F}, P) by

$$(1.1a) \quad x_{t+1} = a(x_t, u_t) + \sigma(x_t)v_{t+1}, \quad x_0 = v_0,$$

$$(1.1b) \quad y_t = c(x_t) + w_t, \quad y_0 = w_0,$$

where $\{v_t\}$ and $\{w_t\}$ are independent standard white Gaussian noises and $\{u_t\}$ is a sequence of admissible controls, namely such that u_t takes values in a given set $U \subset \mathbb{R}$ and depends only on past and present observations $y^t := \{y_0, \dots, y_t\}$ and past controls $u^{t-1} := \{u_0, \dots, u_{t-1}\}$. Defining the value function

$$(1.2) \quad v(u) := E \left\{ \sum_{t=0}^{T-1} r(x_t, u_t) + b(x_T) \right\}$$

where $r(x, u)$ and $b(x)$ are given cost functions, it is desired to find, for any given $\varepsilon > 0$, an ε -optimal control $\{u_t^\varepsilon\}$, i.e. an admissible control such that

$$v(u^\varepsilon) \leq \inf_u v(u) + \varepsilon$$

where the inf is over all admissible controls u .

The usual first step in approaching a stochastic control problem with partial observation as (1.1)–(1.2) is (see e.g. [2], [4]) to transform it into an “equivalent” problem with complete observation by taking as new state at time t the conditional density of x_t given y^t . The equivalence is in the sense that to each optimal (ε -optimal) control in one problem there corresponds an optimal (ε -optimal) control in the other. The major difficulty that arises with this approach is that the new state takes values

* Received by the editors October 15, 1983, and in revised form February 15, 1985.

† LADSEB-CNR and Istituto di Elettrotecnica, Università di Padova, I-35100 Padova, Italy.

‡ Seminario Matematico, Università di Padova, I-35100 Padova, Italy.

in an infinite-dimensional space, namely the Borel space [4, App. 5.2] of all probability densities over the real line.

Our approach here consists of two steps. First, in analogy to an approach used by the authors [3] to obtain approximate solutions to discrete-time nonlinear filtering problems in additive white Gaussian noise, we define a particular class of problems of the type (1.1)–(1.2) such that, given any $\delta > 0$, a δ -optimal control can be explicitly computed. Second, we show under suitable assumptions that for each problem (1.1)–(1.2) and each $\varepsilon > 0$, it is possible to construct a problem in the particular class so that a δ -optimal control for the latter is ε -optimal for the former. This second step is the main subject of § 2, where we also describe the particular class of problems of type (1.1)–(1.2). A method for actually obtaining a δ -optimal control for the problems in the particular class is then presented in § 3.

2. ε -optimal control for the original problem. In this section we shall consider a particular class of problems of the type (1.1)–(1.2) and show that, given $\varepsilon > 0$, it is possible to obtain an ε -optimal control for the original problem (1.1)–(1.2) provided that for a suitable δ , a δ -optimal control for a problem in the particular class has been obtained.

The procedure followed here consists of constructing for each problem (1.1)–(1.2) a corresponding problem in the particular class and in showing that a δ -optimal control for the latter is ε -optimal for the former. The problem in the particular class will be called the approximating problem and a δ -optimal control for it will be called a solution to the approximating problem.

2.1. The particular class. Consider the following particular case of problem (1.1)–(1.2):

$$(2.1a) \quad x_{t+1} = \sum_{i,k=1}^n a_i(k) I_{D_i}(x_t) I_{U_k}(u_t) + \sum_{i=1}^n \sigma_i I_{D_i}(x_t) v_{t+1},$$

$$(2.1b) \quad y_t = \sum_{i=1}^n c_i I_{D_i}(x_t) + w_t,$$

$$(2.2) \quad v(u^{T-1}) = E \left\{ \sum_{t=0}^{T-1} \left[\sum_{i,k=1}^n r_i(k) I_{D_i}(x_t) I_{U_k}(u_t) \right] + \sum_{i=1}^n b_i I_{D_i}(x_T) \right\}$$

where $a_i(k)$, $r_i(k)$, σ_i , b_i , c_i ($i, k = 1, \dots, n$) are given real numbers, $\{D_i\}$ is a finite partition of the real line into intervals, $\{U_k\}$ is a class of disjoint intervals on \mathbb{R} and the admissible control set is given by $U = \bigcup_k U_k$. In other words, this class consists of problems (1.1)–(1.2) with a , r , σ , b , c step functions, and U a finite union of intervals.

It is clear from the particular structure of (2.1)–(2.2) that the conditional probabilities $\pi_t^i := P\{x_t \in D_i | y^t, u^{t-1}\}$, $i = 1, \dots, n$ contain all the information on the past history (y^t, u^{t-1}) which is relevant for control purposes so that the vector $\pi_t = [\pi_t^1, \dots, \pi_t^n]$ can be taken as state variable of the equivalent complete-observation control problem.

It is also clear from (2.1)–(2.2) that the choice of a particular value for the control u_t reduces to the choice of a $k = 1, \dots, n$, so that in this subsection we shall consider $U = \{1, \dots, n\}$.

Furthermore, exploiting the particular structure of (2.1), it is possible to determine the transition law for π_t ; in fact, using the recursive Bayes formula, it is easily seen that

$$(2.3) \quad \pi_{t+1}^j = G^j(\pi_t, y_{t+1}, u_t) := \frac{\sum_{i=1}^n \pi_t^i p_{ij}(u_t) f_j(y_{t+1})}{\sum_{i=1}^n \sum_{h=1}^n \pi_t^i p_{ih}(u_t) f_h(y_{t+1})}$$

where

$$(2.4) \quad p_{ij}(u_i) = \frac{1}{\sqrt{2\pi}\sigma_i} \int_{D_i} \exp \left[-\frac{(x - a_i(u_i))^2}{2\sigma_i^2} \right] dx,$$

$$(2.5) \quad f_j(y_{t+1}) = \frac{1}{\sqrt{2\pi}} \exp \left[-\frac{1}{2}(y_{t+1} - c_j)^2 \right]$$

and the initial condition is given by

$$(2.6) \quad \pi_0^j = P\{x_0 \in D_j\}, \quad j = 1, \dots, n.$$

Notice that the denominator in (2.3) is the conditional density $g(y_{t+1}|y^t, u^t)$.

The equivalent complete-observation problem is characterized by the state space $\Pi = \{\pi | \pi^i \in [0, 1], i = 1, \dots, n; \sum_i \pi^i = 1\}$ by the state-transition law

$$(2.7) \quad \pi_{t+1} = G(\pi_t, y_{t+1}, u_t)$$

as given in (2.3), by admissible control sequences $\{u_t\}$ such that $u_t \in U$ depends only on π_t , and cost functions given, with abuse of notation, by

$$(2.8a) \quad r(\pi_t, u_t) = E\{r(x_t, u_t) | y^t, u^{t-1}\} = \sum_{i=1}^n \pi_t^i r_i(u_t),$$

$$(2.8b) \quad b(\pi_T) = E\{b(x_T) | y^T, u^{T-1}\} = \sum_{i=1}^n \pi_T^i b_i.$$

2.2. Construction of the approximating problem. The main purpose of this subsection is to construct for each problem (1.1)–(1.2) a corresponding problem in the particular class (2.1)–(2.2) such that a δ -optimal control for the latter is ε -optimal for the former. We shall use the following assumptions (where it is implicit that U can be partitioned into intervals):

A.1. There exist sequences of step functions $a^{(n)}(x, u)$, $\sigma^{(n)}(x)$, $c^{(n)}(x)$, $r^{(n)}(x, u)$, $b^{(n)}(x)$ such that $\|a^{(n)}(x, u) - a(x, u)\|$, $\|\sigma^{(n)}(x) - \sigma(x)\|$, $\|c^{(n)}(x) - c(x)\|$, $\|r^{(n)}(x, u) - r(x, u)\|$, $\|b^{(n)}(x) - b(x)\| \rightarrow 0$ as $n \rightarrow \infty$ where $\|\cdot\|$ denotes the sup norm.

A.2. The functions $\sigma(x)$, $c(x)$ and $b(x)$ are Lipschitz continuous with Lipschitz constants L_σ , L_c and L_b respectively. Furthermore, $a(x, u)$ and $r(x, u)$ are Lipschitz continuous in x , uniformly in u , with Lipschitz constants L_a and L_b respectively.

Furthermore, let $B, C, B^{(n)}, C^{(n)}$ denote constants such that $\|b(x)\| \leq B$, $\|b^{(n)}(x)\| \leq B$, $\|r(x, u)\| \leq B$, $\|r^{(n)}(x, u)\| \leq B$, $\|c(x)\| \leq C$, $\|c^{(n)}(x)\| \leq C$, $\|b^{(n)}(x) - b(x)\| \leq B^{(n)}$, $\|r^{(n)}(x, u) - r(x, u)\| \leq B^{(n)}$, $\|c^{(n)}(x) - c(x)\| \leq C^{(n)}$.

The approximating problem is constructed by first approximating the process $\{x_t\}$, defined for a given admissible control u in (1.1a), by a process $\{x_t^{(n)}\}$ satisfying (2.1a) for the same control u . For this purpose, given $n \in \mathbb{N}$, let $a^{(n)}(x, u)$ and $\sigma^{(n)}(x)$ be step functions according to assumption A.1. The approximating process is then defined by

$$(2.9) \quad x_{t+1}^{(n)} = a^{(n)}(x_t^{(n)}, u_t) + \sigma^{(n)}(x_t^{(n)})v_{t+1}, \quad x_0^{(n)} = x_0 = v_0.$$

Furthermore, since the original observation process $\{y_t\}$ defined in (1.1b) provides the only available data, we want the approximating problem to generate this same observation process. To this end we use an absolutely continuous transformation of probability measures. Given $n \in \mathbb{N}$, let $c^{(n)}(x)$ be a step function according to assumption A.1 and consider the processes

$$(2.10) \quad \lambda_t = \prod_{s=1}^t \exp [c(x_s)y_s - \frac{1}{2}c^2(x_s)], \quad \lambda_0 = 1,$$

$$(2.11) \quad \lambda_t^{(n)} = \prod_{s=1}^t \exp [c^{(n)}(x_s^{(n)})y_s - \frac{1}{2}c^{(n)2}(x_s^{(n)})], \quad \lambda_0^{(n)} = 1.$$

Letting $\mathcal{F}_t = \sigma\{w^t, v^t\}$, define the probability measures P_0 and $P^{(n)}$ by

$$(2.12) \quad \left. \frac{dP_0}{dP} \right|_{\mathcal{F}_t} = \lambda_t^{-1}, \quad \left. \frac{dP^{(n)}}{dP_0} \right|_{\mathcal{F}_t} = \lambda_t^{(n)}$$

and denote by E_0 and $E^{(n)}$ the corresponding expectations. We have the following.

PROPOSITION 2.1. *The process y_t defined in (1.1b) is under P_0 standard white Gaussian noise independent of $\{v_t\}$, while under $P^{(n)}$ it satisfies*

$$(2.13) \quad y_t = c^{(n)}(x_t^{(n)}) + w_t^{(n)}$$

where $w_t^{(n)}$ is a $P^{(n)}$ -standard white Gaussian noise independent of $\{v_t\}$. Furthermore the process v_t has the same distribution under P , P_0 and $P^{(n)}$.

Proof. It is easily seen that for an \mathcal{F}_t -adapted process z_t

$$E_0\{z_t | \mathcal{F}_{t-1}\} = E \left\{ \frac{z_t \lambda_t^{-1}}{\lambda_{t-1}^{-1}} \middle| \mathcal{F}_{t-1} \right\}.$$

Then, using (1.1b) and the fact that x_t is \mathcal{F}_{t-1} -measurable, we have for $\alpha \in \mathbb{R}$

$$\begin{aligned} E_0\{\exp[i\alpha y_t] | \mathcal{F}_{t-1}\} &= E\{\exp[i\alpha y_t] \exp[c(x_t)y_t - \frac{1}{2}c^2(x_t)]^{-1} | \mathcal{F}_{t-1}\} \\ &= \exp[i\alpha c(x_t) - \frac{1}{2}c^2(x_t)] E\{\exp[w_t(i\alpha - c(x_t))] | \mathcal{F}_{t-1}\} \\ &= \exp[-\frac{1}{2}\alpha^2], \end{aligned}$$

which shows that y_t is a (P_0, \mathcal{F}_t) -standard white Gaussian noise and is therefore independent of $\{v_t\}$. Furthermore, since $\lambda_0 = 1$, v_t preserves its distribution under P_0 .

In an analogous way it is possible to show that under $P^{(n)}$ (2.13) holds and v_t again preserves its distribution. \square

For each $n \in \mathbb{N}$ we now have an approximating problem defined on $(\Omega, \mathcal{F}, P^{(n)})$ by

$$(2.14a) \quad x_{t+1}^{(n)} = a^{(n)}(x_t^{(n)}, u_t) + \sigma^{(n)}(x_t^{(n)})v_{t+1}, \quad x_0^{(n)} = x_0 = v_0,$$

$$(2.14b) \quad y_t = c^{(n)}(x_t^{(n)}) + w_t^{(n)}, \quad y_0 = w_0^{(n)},$$

$$(2.15) \quad v^{(n)}(u) = E^{(n)} \left\{ \sum_{t=0}^{T-1} r^{(n)}(x_t^{(n)}, u_t) + b^{(n)}(x_T^{(n)}) \right\}$$

where $\{v_t\}$ and $\{w_t^{(n)}\}$ are independent $P^{(n)}$ -standard Gaussian white noises and $r^{(n)}$ and $b^{(n)}$ are step functions according to assumption A.1.

2.3. ε -optimality of the solution of the approximating problem. In this subsection we assume that for every $\delta > 0$ a δ -optimal control for problem (2.14)–(2.15) has already been obtained. Such control, hereafter denoted by $u^{n,\delta}$, can then be applied to the original problem (1.1)–(1.2) and we shall show that for $\varepsilon > 0$ given, an $n \in \mathbb{N}$ and a $\delta > 0$ can be found such that

$$(2.16) \quad v(u^{n,\delta}) \leq \inf_u v(u) + \varepsilon.$$

In Proposition 2.2 below we shall first give a sufficient condition for (2.16) to hold.

PROPOSITION 2.2. *If for all admissible controls u*

$$(2.17) \quad |v^{(n)}(u) - v(u)| \leq \frac{\varepsilon}{2} - \delta \quad \left(0 < \delta < \frac{\varepsilon}{2} \right),$$

then

$$(2.18) \quad v(u^{n,\delta}) \leq \inf_u v(u) + \varepsilon.$$

Proof. Denoting by u^δ a δ -optimal control for the original problem (1.1)–(1.2) and using (2.17) we have

$$(2.19) \quad \begin{aligned} \inf_u v(u) &\geq v(u^\delta) - \delta \geq v^{(n)}(u^\delta) - \frac{\varepsilon}{2} \geq \inf_u v^{(n)}(u) - \frac{\varepsilon}{2} \\ &\geq v^{(n)}(u^{n,\delta}) - \frac{\varepsilon}{2} - \delta \geq v(u^{n,\delta}) - \varepsilon. \end{aligned} \quad \square$$

Using Proposition 2.2, our problem now reduces to that of finding for $\varepsilon > 0$ given, an $n \in \mathbb{N}$ and a δ ($0 < \delta < \varepsilon/2$) such that for any admissible control (2.17) holds. This will follow from Theorem 2.1 below, which shows that for each n we can explicitly construct a bound $K^{(n)}$ such that $|v^{(n)}(u) - v(u)| \leq K^{(n)}$ for all u and $\lim_{n \rightarrow \infty} K^{(n)} = 0$. In what follows we shall assume that an admissible control sequence u has been fixed; the results then hold for any such u . In order to prove the main Theorem 2.1 we need some preliminary results.

LEMMA 2.1. *Under A.1 and A.2 we have for $t \geq 1$*

$$(2.20) \quad E|x_t^{(n)} - x_t| \leq A_t^{(n)}$$

where

$$(2.21) \quad A_t^{(n)} = (\|a^{(n)} - a\| + \sqrt{2/\pi} \|\sigma^{(n)} - \sigma\|) \left[\sum_{s=0}^{t-1} (L_a + \sqrt{2/\pi} L_\sigma)^s \right]$$

is independent of the control u and $\rightarrow 0$ as $n \rightarrow \infty$.

Proof. From (1.1a) and (2.14a)

$$\begin{aligned} |x_{t+1}^{(n)} - x_{t+1}| &\leq |a^{(n)}(x_t^{(n)}, u_t) - a(x_t^{(n)}, u_t)| + |a(x_t^{(n)}, u_t) - a(x_t, u_t)| \\ &\quad + |\sigma^{(n)}(x_t^{(n)}) - \sigma(x_t^{(n)})| |v_{t+1}| + |\sigma(x_t^{(n)}) - \sigma(x_t)| |v_{t+1}| \\ &\leq \|a^{(n)} - a\| + L_a |x_t^{(n)} - x_t| + \|\sigma^{(n)} - \sigma\| |v_{t+1}| + L_\sigma |x_t^{(n)} - x_t| |v_{t+1}| \end{aligned}$$

and, taking into account that $E|v_{t+1}| = \sqrt{2/\pi}$, we have

$$E|x_{t+1}^{(n)} - x_{t+1}| \leq (L_a + \sqrt{2/\pi} L_\sigma) E|x_t^{(n)} - x_t| + \|a^{(n)} - a\| + \sqrt{2/\pi} \|\sigma^{(n)} - \sigma\|$$

from which the result follows. \square

Remark 2.1. From the proof of Lemma 2.1 it follows immediately using Proposition 2.1 that (2.20) holds also when the expectation there is computed with respect to P_0 or $P^{(n)}$.

LEMMA 2.2. *Under A.1 and A.2 we have*

$$(2.22) \quad E_0|\lambda_T - \lambda_T^{(n)}| \leq \Lambda^{(n)}$$

where

$$(2.23) \quad \Lambda^{(n)} = 2T(L_c A_T^{(n)} + C^{(n)})(2C + \sqrt{2/\pi})$$

with $A_T^{(n)}$ given by (2.21), so that $\Lambda^{(n)} \rightarrow 0$ as $n \rightarrow \infty$.

Proof. Using the assumptions and the inequality $|e^x - e^y| \leq (e^x + e^y)|x - y|$ we have

$$\begin{aligned} |\lambda_T - \lambda_T^{(n)}| &\leq (\lambda_T + \lambda_T^{(n)}) \cdot \sum_{s=1}^T \left(|c(x_s) - c^{(n)}(x_s^{(n)})| |y_s| + \frac{1}{2} |c^2(x_s) - c^{(n)2}(x_s^{(n)})| \right) \\ &\leq (\lambda_T + \lambda_T^{(n)}) \sum_{s=1}^T (L_c |x_s - x_s^{(n)}| + C^{(n)}) (|y_s| + C). \end{aligned}$$

Then, using (1.1b), (2.14b), the independence of $(x_s - x_s^{(n)})$ and w_s , Lemma 2.1 and Remark 2.1,

$$\begin{aligned} E_0 |\lambda_T - \lambda_T^{(n)}| &\leq E \left\{ \sum_{s=1}^T (L_c |x_s - x_s^{(n)}| + C^{(n)})(2C + |w_s|) \right\} \\ &\quad + E^{(n)} \left\{ \sum_{s=1}^T (L_c |x_s - x_s^{(n)}| + C^{(n)})(2C + |w_s^{(n)}|) \right\} \\ &\leq 2T(L_c A_t^{(n)} + C^{(n)})(2C + \sqrt{2/\pi}). \end{aligned} \quad \square$$

Finally, the following Theorem 2.1 provides the sufficient condition in Proposition 2.2.

THEOREM 2.1. *Under A.1 and A.2 we have for all admissible controls u*

$$(2.24) \quad |v(u) - v^{(n)}(u)| \leq K^{(n)}$$

where

$$(2.25) \quad K^{(n)} = (T+1)(B\Lambda^{(n)} + L_b A_T^{(n)} + B^{(n)})$$

with $A_T^{(n)}$ and $\Lambda^{(n)}$ given by (2.21) and (2.23), so that $K^{(n)} \rightarrow 0$ as $n \rightarrow \infty$.

Proof. Using Lemmas 2.1 and 2.2 we have

$$\begin{aligned} |v(u) - v^{(n)}(u)| &= \left| E_0 \left\{ \lambda_T \left[\sum_{s=0}^{T-1} r(x_s, u_s) + b(x_T) \right] - \lambda_T^{(n)} \left[\sum_{s=0}^{T-1} r(x_s^{(n)}, u_s) + b(x_T^{(n)}) \right] \right\} \right| \\ &\leq E_0 \left\{ |\lambda_T - \lambda_T^{(n)}| \left[\sum_{s=0}^{T-1} |r(x_s, u_s)| + |b(x_T)| \right] \right\} \\ &\quad + E_0 \left\{ \lambda_T^{(n)} \left[\sum_{s=0}^{T-1} |r(x_s, u_s) - r(x_s^{(n)}, u_s)| + |b(x_T) - b(x_T^{(n)})| \right] \right\} \\ &\quad + E_0 \left\{ \lambda_T^{(n)} \left[\sum_{s=0}^{T-1} |r(x_s^{(n)}, u_s) - r^{(n)}(x_s^{(n)}, u_s)| + |b(x_T^{(n)}) - b^{(n)}(x_T^{(n)})| \right] \right\} \\ &\leq (T+1)(\Lambda^{(n)} B + L_b A_T^{(n)} + B^{(n)}). \end{aligned} \quad \square$$

3. δ -optimal control for the approximating problem. The purpose of this section is to determine a δ -optimal control for the approximating problem (2.14)–(2.15) or, equivalently, for the corresponding complete-observation problem described in 2.1, which in the sequel will be referred to as problem (P) and is given by

$$(3.1) \quad \begin{cases} \pi_{t+1} = G(\pi_t, y_{t+1}, u_t), & \pi_0^j = P\{x_0 \in D_j\}, \quad j = 1, \dots, n, \end{cases}$$

$$(3.2) \quad \begin{cases} v^{(n)}(u) = E^{(n)} \left\{ \sum_{t=0}^{T-1} r(\pi_t, u_t) + b(\pi_T) \right\}. \end{cases}$$

Problem (P) has the states taking values in a finite dimensional space, but its possible values are still infinite. The approach used here to obtain a δ -optimal control for problem (P) consists in approximating it by a problem (\bar{P}) whose state space is finite and for which, as the control set is already finite, an optimal control can be actually computed. This latter control, when suitably extended, will then be shown to be δ -optimal for (P). In this sense the approach to be presented here can be considered an extension of some recent approaches concerning “approximate dynamic programs” which have been proposed for complete-observation problems, in particular Markovian

decision problems (see e.g. [1], [5], [6]). The main idea in [1], [5], [6] is to partition the state space into a finite number of subsets, each characterized by a representative element, and to approximate the original problem by one whose state space is given by the representative elements. A direct application of these methods to our problem (P) leads to various difficulties, in particular that of determining the transition law of the approximating finite state problem. It therefore appears more convenient to exploit the partially observable nature of the original problem (2.14)–(2.15) and to first discretize the observation process y_t .

For $Q, q \in \mathbb{N}$ given, consider the following partition $\{Y_i : i = 1, \dots, \bar{q}; \bar{q} = 2Qq + 2\}$ of \mathbb{R} :

$$(3.3) \quad \begin{cases} Y_1 = (-\infty, -Q), \\ Y_i = [-Q + (i-2)/q, -Q + (i-1)/q), & i = 2, \dots, \bar{q} - 1, \\ Y_{\bar{q}} = [Q, +\infty), \end{cases}$$

and let $\{\eta_i : \eta_i \in Y_i; i = 1, \dots, \bar{q}\}$ be a set of representative elements of the partition. Then, defining the projection

$$(3.4) \quad \bar{y}(y) = \sum_{i=1}^{\bar{q}} \eta_i I_{Y_i}(y),$$

consider the finite valued process $\{\bar{y}_t\}$ (discretized observation process) given by

$$(3.5) \quad \bar{y}_t = \bar{y}(y_t)$$

and set, with abuse of notation,

$$(3.6) \quad \bar{y}(y^t) := \bar{y}^t.$$

Furthermore, let

$$(3.7) \quad \bar{f}_j(\eta_i) := P^{(n)}\{\bar{y}_t = \eta_i \mid x_t \in D_j\} = \frac{1}{\sqrt{2\pi}} \int_{Y_i} \exp\left[-\frac{1}{2}(y - c_j)^2\right] dy.$$

It is clear that the vector $\bar{\pi}_t = [\bar{\pi}_t^1, \dots, \bar{\pi}_t^n]$ of conditional probabilities $\bar{\pi}_t^i = P^{(n)}\{x_t^{(n)} \in D_i \mid \bar{y}^t, u^{t-1}\}$ is a sufficient statistic for the problem with the discretized observation process $\{\bar{y}_t\}$ and it satisfies, analogously to (2.3), the recursive relation

$$(3.8) \quad \bar{\pi}_{t+1}^j = \bar{G}^j(\bar{\pi}_t, \bar{y}_{t+1}, u_t) := \sum_{i=1}^n \bar{\pi}_t^i p_{ij}(u_t) \bar{f}_j(\bar{y}_{t+1}) \Big/ \sum_{i=1}^n \sum_{h=1}^n \bar{\pi}_t^i p_{ih}(u_t) \bar{f}_h(\bar{y}_{t+1})$$

where $p_{ij}(u_t)$ and $\bar{f}_j(\bar{y}_{t+1})$ are given by (2.4) and (3.7), and the initial condition is $\bar{\pi}_0^j = P^{(n)}\{x_0 \in D_j\} = \pi_0^j$.

The finite state problem (\bar{P}) approximating (P) is now given by

$$(3.9) \quad \begin{cases} \bar{\pi}_{t+1} = \bar{G}(\bar{\pi}_t, \bar{y}_{t+1}, u_t), & \bar{\pi}_0 = \pi_0, \\ \bar{v}^{(n)}(u) = E^{(n)}\left\{\sum_{s=0}^{T-1} r(\bar{\pi}_s, u_s) + b(\bar{\pi}_T)\right\}. \end{cases} \quad (\bar{P})$$

Let

$$(3.11) \quad \bar{u}_t(\bar{\pi}_t) = \bar{u}_t(\bar{y}^t, \bar{u}^{t-1})$$

be an optimal control for problem (\bar{P}) and extend it to all the histories of problem (P) by setting

$$(3.12) \quad \bar{u}_t(y^t, \bar{u}^{t-1}) := \bar{u}_t(\bar{y}(y^t), \bar{u}^{t-1}).$$

It remains to show that by choosing Q and q sufficiently large, the control $\{\bar{u}_t\}$ defined in (3.12) is δ -optimal for problem (P), i.e. for the original problem (2.14)–(2.15).

This is the purpose of the rest of this section, where the main result is given in the final Theorem 3.1.

LEMMA 3.1. *Letting*

$$(3.13) \quad p := \min_{i,h,u} p_{ih}(u) > 0, \quad L := n/p^2,$$

we have for all π_1, π_2, y, u

$$(3.14) \quad \max_j |G^j(\pi_1, y, u) - G^j(\pi_2, y, u)| \leq L \max_i |\pi_1^i - \pi_2^i|.$$

Proof. Using the relations $p_{ih}(u) \leq 1$ ($i, h = 1, \dots, n$), $\sum_i \pi_i^l = 1$ ($l = 1, 2$) and $\sum_{i,h} \pi_1^i p_{ih}(u) f_h(y) \geq \sum_i \pi_1^i p_{ij}(u) f_j(y)$, we have

$$\begin{aligned} & |G^j(\pi_1, y, u) - G^j(\pi_2, y, u)| \\ &= \frac{|\sum_{i,h} \pi_1^i p_{ij}(u) f_j(y) \pi_2^i p_{ih}(u) f_h(y) - \sum_{i,h} \pi_2^i p_{ij}(u) f_j(y) \pi_1^i p_{ih}(u) f_h(y)|}{(\sum_{i,h} \pi_1^i p_{ih}(u) f_h(y))(\sum_{i,h} \pi_2^i p_{ih}(u) f_h(y))} \\ &\leq \frac{f_j(y) \sum_{i,h} \pi_1^i f_h(y) |\sum_l (\pi_1^l \pi_2^i - \pi_2^l \pi_1^i)|}{(\sum_i \pi_1^i p_{ij}(u) f_j(y))(\sum_{i,h} \pi_2^i p_{ih}(u) f_h(y))} \\ &\leq \frac{\sum_{i,h} f_h(y) |\pi_1^i - \pi_2^i|}{p^2 \sum_{i,h} \pi_2^i f_h(y)} \end{aligned}$$

from which (3.14) follows. \square

In what follows we shall assume, without loss of generality, $c_1 \leq c_2 \leq \dots \leq c_n$.

LEMMA 3.2. *If $y \notin [-Q, Q]$, we have for every $j = 1, \dots, n$ and all π, u*

$$(3.15) \quad |G^j(\pi, y, u) - \bar{G}^j(\pi, \bar{y}(y), u)| \leq G(Q)$$

where

$$(3.16) \quad G(Q) = \left[1 + p / \max \left\{ \sum_{h \neq n} \exp \left[(c_h - c_n)Q - \frac{1}{2}(c_h^2 - c_n^2) \right], \sum_{h \neq 1} \exp \left[(c_1 - c_h)Q - \frac{1}{2}(c_h^2 - c_1^2) \right] \right\} \right]^{-1}$$

so that $G(Q) \rightarrow 0$ as $Q \rightarrow +\infty$.

Proof. Assuming first $y \geq Q$, we have for $j < n$

$$\begin{aligned} (3.17) \quad 0 &\leq G^j(\pi, y, u) \leq \sum_i \sum_{h \neq n} \pi^i p_{ih}(u) f_h(y) / \sum_i \sum_h \pi^i p_{ih}(u) f_h(y) \\ &= \left[1 + \sum_i \pi^i p_{in} \exp \left[-\frac{1}{2}(y - c_n)^2 \right] / \sum_i \sum_{h \neq n} \pi^i p_{ih} \exp \left[-\frac{1}{2}(y - c_h)^2 \right] \right]^{-1} \\ &\leq \left[1 + p / \sum_{h \neq n} \exp \left[(c_h - c_n)y - \frac{1}{2}(c_h^2 - c_n^2) \right] \right]^{-1} \\ &\leq \left[1 + p / \sum_{h \neq n} \exp \left[(c_h - c_n)Q - \frac{1}{2}(c_h^2 - c_n^2) \right] \right]^{-1} \end{aligned}$$

and for $h = j$

$$\begin{aligned} (3.18) \quad 0 &\leq 1 - G^n(\pi, y, u) = \sum_i \sum_{h \neq n} \pi^i p_{ih}(u) f_h(y) / \sum_i \sum_h \pi^i p_{ih}(u) f_h(y) \\ &\leq \left[1 + p / \sum_{h \neq n} \exp \left[(c_h - c_n)Q - \frac{1}{2}(c_h^2 - c_n^2) \right] \right]^{-1}. \end{aligned}$$

Noticing now that

$$\begin{aligned} & \int_Q^{+\infty} \exp \left[-\frac{1}{2}(y - c_h)^2 \right] dy \Big/ \int_Q^{+\infty} \exp \left[-\frac{1}{2}(y - c_n)^2 \right] dy \\ & \leq \exp \left[-\frac{1}{2}(Q - c_h)^2 \right] \Big/ \exp \left[-\frac{1}{2}(Q - c_n)^2 \right] \end{aligned}$$

we obtain the same bounds (3.17) and (3.18) for $\bar{G}^j(\pi, \bar{y}(y), u)$, so that (3.15) is proved for $y \geq Q$. The case $y < -Q$ can be proved analogously, simply replacing n by 1 and Q by $-Q$. \square

Notice now that with $f_j(y)$ as defined in (2.5) we have that if $|y - \bar{y}| < 1/q$, then for all j

$$(3.19) \quad |f_j(y) - f_j(\bar{y})| < 1/q.$$

Furthermore, notice that if $y \in [-Q, Q]$ and if $Y(y)$ denotes the set of the partition (3.3) containing y , then by the mean value theorem for integrals we have for all π and u

$$\begin{aligned} (3.20) \quad \bar{G}^j(\pi, \bar{y}(y), u) &= \sum_i \pi^i p_{ij}(u) \int_{Y(y)} \exp \left[-\frac{1}{2}(y - c_j)^2 \right] dy \\ &\quad \Big/ \sum_{i,h} \pi^i p_{ih}(u) \int_{Y(y)} \exp \left[-\frac{1}{2}(y - c_h)^2 \right] dy \\ &= \sum_i \pi^i p_{ij}(u) f_j(y_j(y)) \Big/ \sum_{i,h} \pi^i p_{ih}(u) f_h(y_h(y)) \end{aligned}$$

with $y_j(y)$ and $y_h(y)$ suitable elements in $Y(y)$. Letting

$$(3.21) \quad f(Q) := [\min_h \min \{f_h(-Q), f_h(Q)\}]^{-1}$$

so that for all $y \in [-Q, Q]$ and all h

$$(3.22) \quad f_h^{-1}(y) \leq f(Q),$$

we have

LEMMA 3.3. *If $y \in [-Q, Q]$, we have for every $j = 1, \dots, n$ and all π, u*

$$(3.23) \quad |G^j(\pi, y, u) - \bar{G}^j(\pi, \bar{y}(y), u)| \leq 2f(Q)/q$$

where $f(Q)$ is given by (3.21).

Proof. Using (3.19) through (3.22) we have

$$\begin{aligned} & |G^j(\pi, y, u) - \bar{G}^j(\pi, \bar{y}(y), u)| \\ & \leq \sum_{i,i,h} \pi^i \pi^i p_{ij}(u) p_{ih}(u) f_j(y) |f_h(y) - f_h(y_h(y))| \\ & \quad \Big/ \sum_i \pi^i p_{ij}(u) f_j(y) \sum_{i,h} \pi^i p_{ih}(u) f_h(y_h(y)) \\ & \quad + \sum_{i,i,h} \pi^i \pi^i p_{ij}(u) p_{ih}(u) f_h(y) |f_j(y) - f_j(y_j(y))| \\ & \quad \Big/ \sum_{i,h} \pi^i p_{ih}(u) f_h(y) \sum_i \pi^i p_{ij}(u) f_j(y_j(y)) \\ & \leq \sum_{i,h} \pi^i p_{ih}(u) |f_h(y) - f_h(y_h(y))| \Big/ \sum_{i,h} \pi^i p_{ih}(u) f_h(y_h(y)) \\ & \quad + \sum_i \pi^i p_{ij}(u) |f_j(y) - f_j(y_j(y))| \Big/ \sum_i \pi^i p_{ij}(u) f_j(y_j(y)) \\ & \leq 2f(Q)/q. \end{aligned}$$

\square

Now let

$$(3.24) \quad G(Q, q) = \max \{G(Q), 2f(Q)/q\}$$

so that

$$(3.25) \quad \lim_{\substack{Q, q \rightarrow \infty \\ q \geq f(Q)^{1+\gamma}, \gamma > 0}} G(Q, q) = 0$$

and, making explicit the dependence of π_t and $\bar{\pi}_t$ on (y^t, u^{t-1}) and (\bar{y}^t, u^{t-1}) respectively, define

$$(3.26) \quad \Delta_t := \sup_{y^t} \max_{u^{t-1}} \max_j |\pi_t^j(y^t, u^{t-1}) - \bar{\pi}_t^j(\bar{y}(y^t), u^{t-1})|.$$

We then have

PROPOSITION 3.1. *With L as defined in (3.13) we have for all t*

$$\Delta_t \leq G(Q, q)(L^T - 1)/(L - 1).$$

Proof. $\Delta_0 = 0$; furthermore, by definition,

$$\begin{aligned} \pi_t^j(y^t, u^{t-1}) &= G^j(\pi_{t-1}(y^{t-1}, u^{t-2}), y_t, u_{t-1}), \\ \bar{\pi}_t^j(\bar{y}(y^t), u^{t-1}) &= \bar{G}^j(\bar{\pi}_{t-1}(\bar{y}(y^{t-1}), u^{t-2}), \bar{y}_t, u_{t-1}). \end{aligned}$$

Therefore, using Lemmas 3.1, 3.2 and 3.3,

$$\begin{aligned} &|\pi_t^j(y^t, u^{t-1}) - \bar{\pi}_t^j(\bar{y}(y^t), u^{t-1})| \\ &\leq |G^j(\pi_{t-1}, y_t, u_{t-1}) - G^j(\bar{\pi}_{t-1}, y_t, u_{t-1})| \\ &\quad + |G^j(\bar{\pi}_{t-1}, y_t, u_{t-1}) - \bar{G}^j(\bar{\pi}_{t-1}, \bar{y}_t, u_{t-1})| \\ &\leq L \max_i |\pi_{t-1}^i - \bar{\pi}_{t-1}^i| + G(Q, q) \end{aligned}$$

so that $\Delta_{t+1} \leq L\Delta_t + G(Q, q)$, from which the result follows. \square

THEOREM 3.1. *Let $\{\bar{u}_t\}$ be the control sequence given by (3.12). Then*

$$\begin{aligned} &\left| \inf_{\{u_t\}} E^{(n)} \left\{ \sum_{s=0}^{T-1} r(\pi_s, u_s) + b(\pi_T) \right\} - E^{(n)} \left\{ \sum_{s=0}^{T-1} r(\pi_s, \bar{u}_s) + b(\pi_T) \right\} \right| \\ &\leq nBTG(Q, q)(L^T - 1)/(L - 1), \end{aligned}$$

where the inf is over all control sequences $\{u_t\}$ such that u_t depends on π_t or equivalently, on (y^t, u^{t-1}) and where $G(Q, q)$ is defined in (3.24) and goes to zero according to (3.25).

Proof. From (2.8) and the definition of π_t and $\bar{\pi}_t$ we immediately have

$$(3.27) \quad E^{(n)} \left\{ \sum_{s=0}^{T-1} r(\pi_s, \bar{u}_s) + b(\pi_T) \right\} = E^{(n)} \left\{ \sum_{s=0}^{T-1} r(\bar{\pi}_s, \bar{u}_s) + b(\bar{\pi}_T) \right\}.$$

Furthermore

$$\begin{aligned} &\inf_{\{u_t\}} E^{(n)} \left\{ \sum_{s=0}^{T-1} r(\pi_s, u_s) + b(\pi_T) \right\} - E^{(n)} \left\{ \sum_{s=0}^{T-1} r(\bar{\pi}_s, \bar{u}_s) + b(\bar{\pi}_T) \right\} \\ (3.28) \quad &\leq E^{(n)} \left\{ \sum_{s=0}^{T-1} r(\pi_s, \bar{u}_s) + b(\pi_T) \right\} - E^{(n)} \left\{ \sum_{s=0}^{T-1} r(\bar{\pi}_s, \bar{u}_s) + b(\bar{\pi}_T) \right\} = 0. \end{aligned}$$

On the other hand, letting u^γ be a γ -optimal control for problem (P) ($\gamma > 0$), we have by the definition of \bar{u}_t , by (2.8) and by Proposition 3.1

$$\begin{aligned}
 (3.29) \quad & E^{(n)} \left\{ \sum_{s=0}^{T-1} r(\bar{\pi}_s, \bar{u}_s) + b(\bar{\pi}_T) \right\} - \inf_{\{u_t\}} E^{(n)} \left\{ \sum_{s=0}^{T-1} r(\pi_s, u_s) + b(\pi_T) \right\} \\
 & \leq \left| E^{(n)} \left\{ \sum_{s=0}^{T-1} r(\bar{\pi}_s, u_s^\gamma) + b(\bar{\pi}_T) \right\} - E^{(n)} \left\{ \sum_{s=0}^{T-1} r(\pi_s, u_s^\gamma) + b(\pi_T) \right\} \right| + \gamma \\
 & \leq E^{(n)} \left\{ \sum_{s=0}^{T-1} \sum_{i=1}^n |\bar{\pi}_s^i - \pi_s^i| |r_i(u_s^\gamma)| + \sum_{i=1}^n |\bar{\pi}_T^i - \pi_T^i| |b_i| \right\} + \gamma \\
 & \leq nBTG(Q, q)(L^T - 1)/(L - 1) + \gamma.
 \end{aligned}$$

Combining (3.27), (3.28) and (3.29), we have the result.

REFERENCES

- [1] D. P. BERTSEKAS, *Convergence of discretization procedures in dynamic programming*, IEEE Trans. Automat. Control, AC-20 (1975), pp. 415-419.
- [2] D. P. BERTSEKAS AND S. E. SHREVE, *Stochastic Optimal Control: The Discrete-Time Case*, Academic Press, New York, 1978.
- [3] G. B. DI MASI AND W. J. Runggaldier, *Approximations and bounds for discrete-time nonlinear filtering*, in Analysis and Optimization of Systems, A. Bensoussan and J. L. Lions, eds., Lecture Notes in Control and Information Sciences 44, Springer-Verlag, Berlin, 1982, pp. 191-202.
- [4] E. B. DYNKIN AND A. A. YUSHKEVICH, *Controlled Markov Processes*, Springer-Verlag, New York, 1979.
- [5] K. HINDERER, *On approximate solutions of finite-stage dynamic programs*, in Dynamic Programming and Its Applications, M. Puterman, ed., Academic Press, New York, 1979, pp. 289-317.
- [6] W. WHITT, *Approximations of dynamic programs I*, Math. Oper. Res., 3 (1978), pp. 231-243.

DISCOUNTED MDP'S: DISTRIBUTION FUNCTIONS AND EXPONENTIAL UTILITY MAXIMIZATION*

KUN-JEN CHUNG†‡ AND MATTHEW J. SOBEL†§

Abstract. The present value of the rewards associated with a discrete-time Markov process has a probability distribution which depends on the initial state. The first part of the paper applies fixed point theory to a system of equations for the distribution functions of the present value. The second part of the paper expands the model to a Markov decision process (MDP) and considers the maximization of the expected utility of the present value when the utility function is exponential.

Key words. Markov decision process, distribution of present value, exponential utility, fixed point

AMS(MOS) subject classifications. 90C39, 90C47, 90A06

1. Introduction. Let X_1, X_2, \dots be the sequence of single period rewards of a Markov decision process (hereafter called MDP). The present value of this sequence is

$$(1) \quad B = \sum_{n=1}^{\infty} \beta^{n-1} X_n$$

where $0 < \beta < 1$ is discount factor. We emphasize that B is a random variable.¹

The criterion *maximize* $E(B)$ is the focus of the literature concerned with discounted MDP's. However practical applications of stochastic models often require more information about a criterion than merely its expected value. This paper elicits other characteristics of B .

Von Neumann and Morgenstern's expected utility theorem concerns a "rational" decision maker who compares alternative random payoffs, i.e. random variables. The theorem states necessary and sufficient conditions for a partial ordering of the random variables to be equivalent to the existence of a "utility" function $u(\cdot)$ such that X is preferred to Y if, and only if, $E[u(X)] > E[u(Y)]$. See Fishburn [6] for details.

If the utility function $u(\cdot)$ is linear then the partial ordering of random variables depends only on their expected values. Hence, most of the MDP literature assumes implicitly that a decision maker has a linear utility function, i.e. is risk-neutral. As in [9] and [15], we use the label *risk-sensitive* for the complement of risk-neutral. Thus, a risk-sensitive model assumes implicitly that a decision maker has a *nonlinear* utility function. Most of the consequent optimization problems are more difficult than in the linear case.

The next section of the paper assumes that a particular stationary policy has already been chosen and proceeds to partially characterize the distribution function of B . Thus, the model can be regarded as a Markov chain in which a random reward

* Received by the editors March 6, 1984, and in revised form October 1, 1985. This material is based on work supported by the National Science Foundation under grant ECS-8305963.

† Georgia Institute of Technology, Atlanta, Georgia 30332.

‡ Present address, Department of Industrial Management, National Taiwan Institute of Technology, Taipei, Taiwan, Republic of China.

§ Part of this work was performed at the University of Arizona, Tucson, Arizona 85287.

¹ B is a random variable under the usual measurability assumptions.

X_n is associated with occupying a state during period n . The probability distribution of X_n depends on the identity of the state occupied during period n .

The last section of the paper examines the optimization of the expected value of a particular nonlinear utility function, namely

$$(2) \quad \text{maximize } E(-e^{-\lambda B}) \quad (\lambda > 0).$$

Exponential utility functions have attractive properties which are mentioned in § 3 of the paper. There we show that problem (2) satisfies an optimality equation. However, the equation is intrinsically an approximation problem and § 3 discusses some monotone approximations. The discounting of rewards distinguishes § 3 from past work on the optimization of exponential utility functions.

Suppose that S is the set of states and that a stationary policy is used. Let $F_s(\cdot)$ be the distribution function of B when s is the initial state. Section 2 of the paper concerns a system of equations satisfied by $\{F_s(\cdot) : s \in S\}$. The major result is negative. The solution of the equations is a fixed point of a mapping which is nonexpansive but generally not a contraction.

Section 2.1 presents notation which is used throughout the paper. Section 2.2 contains preliminary results and § 2.3 has the fixed-point results. The negative results concerning contraction mappings are presented in § 2.4. For background on MDP's see Bertsekas [1], Denardo [3], or Heyman and Sobel [8].

2. Distribution function of the present value.

2.1. Notation. Consider a Markov decision process with discount factor β ($0 < \beta < 1$). Let S be the state space. Let A_s be the set of actions available in state s , and $C = \{(s, a) : s \in A_s, s \in S\}$. Throughout the paper, we assume that C contains only finitely many elements. Such a model is called a *finite* MDP. Let s_n , a_n and X_n indicate the state, action and reward in the n th period. We assume that s_{n+1} and X_n are random variables which depend only on s_n and a_n . Suppose that there is a countable² sample space K such that $P\{X_n \in K\} = 1$ for all n . Let

$$p_{sjk}^a = P\{s_{n+1} = j, X_n = k | s_n = s, a_n = a\}.$$

We assume that K lies in a compact set; this corresponds to the assumption that there is $b < \infty$ such that

$$P\{0 \leq X_1 \leq b | s_1 = s, a_1 = a\} = 1 \quad \text{for all } (s, a) \in C.$$

For simplicity of exposition, we refrain from adding "with probability one" to statements that depend on the rewards being nonnegative and bounded with probability one rather than everywhere.

The notation p_{sjk}^a permits a nondegenerate conditional distribution of X_n given s_n , a_n and s_{n+1} . The notation p_{sj}^a is common in the MDP literature and corresponds to replacement of X_n by its conditional expected value. The Appendix contains a simple example which illustrates that the notation p_{sjk}^a is essential in the first half of this paper.

The present value of the single period rewards is $B = \sum_{n=1}^{\infty} \beta^{n-1} X_n$ which takes values only in $[0, u]$ where $u = b/(1 - \beta)$. Suppose that the stationary policy δ is used to choose actions, i.e. $a_n = \delta(s_n)$ for all n . Let B_s denote the random variable B if $s_1 = s$ and $a_n = \delta(s_n)$ for all n . Let $F_s(x) = P(B_s \leq x)$ be the distribution function of B_s and let $p_{sjk} = p_{sjk}^{\delta(s)}$.

² We assume that K is countable only for expository convenience.

The main purpose of § 2 of the paper is to use fixed point theory to explore the behavior of the vector of distribution functions $F(\cdot) = (F_s(\cdot), s \in S)$. Therefore, δ is suppressed in the notation for the remainder of § 2.

2.2. Preliminary results. Let $\mathcal{X} = [0, u]$ and let \mathcal{B} be the smallest σ -algebra of subsets of \mathcal{X} which contains all the open subsets of \mathcal{X} . It is well known that there is a one-to-one correspondence between measures on \mathcal{B} and distribution functions supported by $[0, u]$. Therefore, we identify measures on \mathcal{B} with the corresponding distribution functions; let Y denote the latter set. Let V be the space of all real-valued continuous functions on \mathcal{X} . For any $f \in V$ we write $\|f\| = \sup \{|f(x)|: x \in \mathcal{X}\}$. V is a Banach space under the norm $\|\cdot\|$. A linear functional Λ on V is a mapping $f \rightarrow \Lambda(f)$ of V into the real line such that for any two constants α and γ and any two elements f and g of V the equation $\Lambda(\alpha f + \gamma g) = \alpha \Lambda(f) + \gamma \Lambda(g)$ is valid. Let V^* denote the set of all continuous linear functionals on V . Therefore, if $\Lambda \in V^*$, we have $\|\Lambda\| = \sup \{|\Lambda(f)|: f \in V, \|f\| \leq 1\}$.

Let $E = \times_{i \in S} V^*$. Then E is a Banach space if its norm is

$$\|H\| = \max \{\|H_i\|: i \in S\} \quad \text{where } H = [H_i(\cdot); i \in S] \in E.$$

The following result, whose proof is in [2], connects linear functionals on V with measures on \mathcal{B} .

LEMMA 1. *Let $\mathcal{X} = [0, u]$ and Λ be a nonnegative linear functional on V (i.e. $\Lambda(f) \geq 0$ if $f \geq 0$) with $\Lambda(1) = 1$. Then there exists a unique measure w on \mathcal{B} such that $\Lambda(f) = \int_0^u f dw$, $f \in V$.*

This lemma shows that there is a map from a subset of V^* into Y . Trivially, this map is one-to-one, i.e., each measure w in Y corresponds to only one linear functional on V , defined by $\alpha(f) = \int f dw$ for all $f \in V$. Thus, one can identify Y with the corresponding subset of V^* . In particular, we have the following metric on Y . If v and w are probability measures on \mathcal{B} , then

$$(1) \quad \|w - v\| = \sup \left\{ \left| \int f dw - \int f dv \right|; \|f\| \leq 1, f \in V \right\}.$$

Let $\mathcal{X} = \times_{i \in S} Y$. If $H = [H_i(\cdot); i \in S] \in \mathcal{X}$ and $G = [G_i(\cdot); i \in S] \in \mathcal{X}$, we define

$$(2) \quad \begin{aligned} \|H - G\| &= \sup \{\|H_i - G_i\|: i \in S\} \\ &= \sup \left\{ \left| \int_0^u f dH_i - \int_0^u f dG_i \right|; f \in V, \|f\| \leq 1, i \in S \right\}. \end{aligned}$$

It is convenient to define $T = S \times K$ and $F_j^k(x) = F_j[(x - k)/\beta]$. The formula

$$(3) \quad F_s(x) = \sum_{(j,k) \in T} p_{sjk} F_j^k(x) \quad (s \in S)$$

was obtained in [22] (see [8, § 4.5] for this form). Therefore, a mapping $M: \mathcal{X} \rightarrow E$ can be defined via (3), as follows

$$M(G)_s(x) = \sum_{(j,k) \in T} p_{sjk} G_j^k(x)$$

for $s \in S$ if $G = [G_i(\cdot); i \in S] \in \mathcal{X}$ and $G_j^k(x)$ denotes $G_j[(x - k)/\beta]$. It is clear that $M[G]_s$ is a distribution function for $s \in S$. Also, if $x \geq u$, $k \in K$, and $j \in S$, then

$$(4) \quad 1 \geq G_j^k(x) \geq G_j^k[b/(1 - \beta)] \geq G_j^b(u) \geq G_j(u) = 1.$$

If $x < 0$, $k \in K$ and $j \in S$, then

$$(5) \quad G_j\left(\frac{x - k}{\beta}\right) = 0.$$

Therefore, $M(G) \in \mathcal{X}$ so M maps \mathcal{X} into itself. From the definition of M , $G = [G_i(\cdot); i \in S] \in \mathcal{X}$ is a fixed point of M if, and only if, G satisfies (3) for all $s \in S$. The following section establishes several properties of M .

2.3. Fixed-point theorems.

LEMMA 2. *The set Y is a weakly compact convex subset of V^* . Therefore, \mathcal{X} is a weakly compact convex subset of E .*

Proof. It is clear that $\{w_\alpha\} \subset Y$ converges in the weak topology to a distribution $w \in Y$ if, and only if, $\int f dw_\alpha \rightarrow \int f dw$ for every $f \in V$. From [2, Prop. 8.10, 8.15, pp. 162–165], Y is weakly compact. The convexity of Y is trivial so Tyhonov's theorem [23] implies that \mathcal{X} is a weakly compact convex subset of E .

The mapping M has the following properties.

THEOREM 1.

- (a) M is a nonexpansive mapping (i.e. $\|MH - MG\| \leq \|H - G\|$),
- (b) M is weakly continuous,
- (c) M is an affine mapping (i.e. $M[\alpha H + (1 - \alpha)G] = \alpha M(H) + (1 - \alpha)M(G)$ for all $H, G \in \mathcal{X}$; $0 \leq \alpha \leq 1$),
- (d) $M: \mathcal{X} \rightarrow \mathcal{X}$ has a fixed point in \mathcal{X} .

Proof. (a) Let $G = [G_i(\cdot); i \in S] \in \mathcal{X}$ and $H = [H_i(\cdot); i \in S] \in \mathcal{X}$. We employ the notation $H_j^k(x) = H_j[(x - k)/\beta]$ and $G_j^k(x) = G_j[(x - k)/\beta]$.

$$\begin{aligned}
 & \|M[H]_s - M[G]_s\| \\
 &= \left\| \sum_{(j,k) \in T} p_{sjk} H_j^k(x) - \sum_{(j,k) \in T} p_{sjk} G_j^k(x) \right\| \\
 &\leq \sum_{(j,k) \in T} p_{sjk} \|H_j^k(x) - G_j^k(x)\| \\
 &= \sum_{(j,k) \in T} p_{sjk} \sup \left\{ \left| \int_0^u f(x) dH_j^k(x) - \int_0^u f(x) dG_j^k(x) \right| : \|f\| \leq 1, f \in V \right\} \\
 &\leq \sum_{(j,k) \in T} p_{sjk} \|H_j - G_j\| \\
 &\leq \sum_{(j,k) \in T} p_{sjk} \|H - G\| = \|H - G\|.
 \end{aligned}$$

Since $s \in S$ is arbitrary, $\|MH - MG\| \leq \|H - G\|$.

(b) Suppose $G_\alpha = [G_{\alpha i}(\cdot); i \in S]$ converges weakly to $G = [G_i(\cdot); i \in S]$. Then $G_{\alpha i}(x)$ weakly converges to $G_i(x)$ for all $i \in S$, and $\int f(x) dG_{\alpha i}(x) \xrightarrow{\alpha} \int f(x) dG_i(x)$ for all $i \in S$, and $f \in V$. Let $G_{\alpha j}^k(x)$ denote $G_{\alpha j}[(x - k)/\beta]$. Since $M[G_\alpha]_s(x) = \sum_{(j,k) \in T} p_{sjk} G_{\alpha j}^k(x)$,

$$\begin{aligned}
 \int f(x) dM[G_\alpha]_s &= \int f(x) dM[G_\alpha]_s(x) = \sum_{(j,k) \in T} p_{sjk} \int f(x) dG_{\alpha j} \left(\frac{x - k}{\beta} \right) \\
 &= \sum_{(j,k) \in T} p_{sjk} \int f(k + \beta y) dG_{\alpha j}(y) \\
 &\xrightarrow{\alpha} \sum_{(j,k) \in T} p_{sjk} \int f(k + \beta y) dG_j(y) \\
 &= \sum_{(j,k) \in T} p_{sjk} \int f(x) dG_j^k(x).
 \end{aligned}$$

Hence $M[G_\alpha]_s(x) \xrightarrow{\alpha} M[G]_s(x)$ for all $s \in S$ so M is weakly continuous.

(c) Let $H, G \in \mathcal{Z}$ and $0 \leq \alpha \leq 1$.

$$\begin{aligned} M[\alpha H + (1 - \alpha)G]_s(x) &= \sum_{(j,k) \in T} p_{sjk} [\alpha H_j^k(x) + (1 - \alpha)G_j^k(x)] \\ &= \alpha \sum_{(j,k) \in T} p_{sjk} H_j^k(x) + (1 - \alpha) \sum_{(j,k) \in T} p_{sjk} G_j^k(x) \\ &= \alpha M[H]_s(x) + (1 - \alpha)M[G]_s(x). \end{aligned}$$

So M is affine.

(d)³ Now $M: \mathcal{Z} \rightarrow \mathcal{Z}$ is weakly continuous, and \mathcal{Z} is a weakly compact subset of E . Therefore, the Schauder and Tyhonov fixed point theorems [20], [23] imply that M has a fixed point in \mathcal{Z} .

2.4. Contraction mappings. Although M is a nonexpansive mapping, it is not generally a contractive mapping. Moreover, in general it is not true that $\|MH - MG\| < \|H - G\|$. For a simple counterexample, let

$$\begin{aligned} S &= \{1, 2\}, p_{112} = p_{121} = p_{211} = p_{222} = 0.5, \beta = 0.9, b = 2, u = 20, \\ K &= \{1, 2\}, G = [G_1(\cdot), G_2(\cdot)] \quad \text{and} \quad H = [H_1(\cdot), H_2(\cdot)], \\ G_s(x) &= \begin{cases} 0 & \text{if } x < 0, \\ 1 & \text{if } x \geq 0, \end{cases} \quad H_s(x) = \begin{cases} 0 & \text{if } x < 20, \\ 1 & \text{if } x \geq 20, \end{cases} \end{aligned}$$

for $s = 1, 2$. So, $\|MG - MH\| = 2 = \|G - H\|$.

The preceding example is not isolated. We now characterize a family of counterexamples. Suppose $D_i = \{k \in K: \text{there exists some } j \in S \text{ such that } p_{ijk} > 0\}$, $W_i = \{|k_1 - k_2| > 0: k_1, k_2 \in D_i\}$, and $W = \bigcap_{i \in S} W_i$. For the remainder of § 2.4 it is convenient to denote the mapping M as $M(\beta, b)$.

THEOREM 2. *If $\beta u \notin W_s$ for some s , then $M(\beta, b)$ is not a contractive mapping.*

Proof. Since $\sum_{(j,k) \in T} p_{sjk} = 1$, for every $\varepsilon > 0$ there exist finitely many points k_1, k_2, \dots, k_n in D_s such that $\sum_{i=1}^n \sum_{j \in S} p_{sjk_i} \geq 1 - \varepsilon/4$. By assumption $(\beta u + D_s) \cap D_s = \emptyset$. Therefore, if $L_1 = \{k_1, k_2, \dots, k_n\}$ and $L_2 = \{\beta u + k_1, \beta u + k_2, \dots, \beta u + k_n\}$, then $L_1 \cap L_2 = \emptyset$. Define a continuous function g on $L_1 \cup L_2$ as follows:

$$g(x) = \begin{cases} 1 & \text{if } x \in L_1, \\ -1 & \text{if } x \in L_2. \end{cases}$$

A well-known theorem of Tietze [18] implies that g has a continuous extension $\bar{g}: [0, u] \rightarrow [-1, 1]$. Let $G = [G_s(\cdot); s \in S]$ and $H = [H_s(\cdot); s \in S]$ where $G_s(\cdot)$ and $H_s(\cdot)$ are the unit jump functions at 0 and u , respectively, for all $s \in S$. Then

$$\begin{aligned} &\|(MG)s - (MH)s\| \\ &= \sup \left\{ \left| \sum_{(j,k) \in T} p_{sjk} \int f(x) d \left[G_j \left(\frac{x-k}{\beta} \right) - H_j \left(\frac{x-k}{\beta} \right) \right] \right| : \|f\| \leq 1, f \in V \right\} \\ &\geq \left| \sum_{(j,k) \in T} p_{sjk} \int \bar{g}(x) d[G_j^k(x) - H_j^k(x)] \right| \\ &= \left| \sum_{\substack{(j,k) \in T \\ k \in L_1}} p_{sjk} \int \bar{g}(x) d[G_j^k(x) - H_j^k(x)] \right. \\ &\quad \left. + \sum_{\substack{(j,k) \in T \\ k \notin L_1}} p_{sjk} \int \bar{g}(x) d[G_j^k(x) - H_j^k(x)] \right| \end{aligned}$$

³ This proof of (d) depends only on the properties of M and \mathcal{Z} . Alternatively, one could argue from (3) that the vector of distribution functions $F(\cdot) = (F_s(\cdot); s \in S)$ is a fixed point.

$$\begin{aligned}
& \geq \left| \sum_{\substack{(j,k) \in T \\ k \in L_1}} p_{sjk} \int \bar{g}(x) d[G_j^k(x) - H_j^k(x)] \right| - \sum_{\substack{(j,k) \in T \\ k \notin L_1}} p_{sjk} \int \bar{g}(x) d[G_j^k(x) - H_j^k(x)] \\
& = \left| \sum_{\substack{(j,k) \in T \\ k \in L_1}} p_{sjk} [\bar{g}(k) - \bar{g}(\beta u + k)] \right| - \left| \sum_{\substack{(j,k) \in T \\ k \notin L_1}} p_{sjk} [\bar{g}(k) - \bar{g}(\beta u + k)] \right| \\
& = 2 \sum_{\substack{(j,k) \in T \\ k \in L_1}} p_{sjk} - \sum_{\substack{(j,k) \in T \\ k \notin L_1}} p_{sjk} [|\bar{g}(k)| + |\bar{g}(\beta u + k)|] \\
& \geq 2(1 - \varepsilon/4) - 2(\varepsilon/4) = 2 - \varepsilon.
\end{aligned}$$

But ε is an arbitrary positive number; so $\|(MG)_s - (MH)_s\| \geq 2$, i.e., $\|MG - MH\| \geq 2$.

The nonexpansiveness of $M(\beta, b)$ implies $2 \leq \|MG - MH\| \leq \|G - H\| = 2$; so $\|MG - MH\| = \|G - H\| = 2$ and $M(\beta, b)$ is not a contractive mapping. \square

Since K is countable, it follows from Theorem 2 that, except for countably many values of β , $M(\beta, b)$ is not a contractive mapping. Furthermore, if W is a finite set, then, except for finitely many values of β , $M(\beta, b)$ is not a contractive mapping.

COROLLARY 1. *If W is empty, then for all $0 < \beta < 1$, $M(\beta, b)$ is not a contractive mapping.*

COROLLARY 2. *Let $w_0 = \sup W$. If $\beta > w_0/(b + w_0)$, then $M(\beta, b)$ is not a contractive mapping.*

Proof. If $\beta u \in W$, there is a number $w \in W$ such that $\beta u = w$, i.e., $\beta b/(1 - \beta) = w$. So $\beta = w/(b + w) \leq w_0/(b + w_0)$. Hence, if $\beta > w_0/(b + w_0)$, then $\beta u \notin W$. Theorem 2 implies that $M(\beta, b)$ is not a contractive mapping.

The inequality $w_0 \leq b$ implies $w_0/(b + w_0) \leq \frac{1}{2}$. Hence, if $\beta > \frac{1}{2}$, then $M(\beta, b)$ is not a contractive mapping.

COROLLARY 3. *For each β , $0 < \beta < 1$, there exists $c > 0$ such that $P\{0 \leq X_1 \leq c | s_1 = s, a_1 = s\} = 1$ for all $(s, a) \in C$ and $M(\beta, c)$ is not a contractive mapping.*

Proof. For each $0 < \beta < 1$, there is a number $c > 0$ such that $w_0/(c + w_0) < \beta$ and $P\{0 \leq X_1 \leq c | s_1 = s, a_1 = a\} = 1$ for all $(s, a) \in C$ determines when $M(\beta, c)$ is not a contractive mapping. \square

It follows from Corollary 3 that the value of c which satisfies $P\{0 \leq X_1 \leq c | s_1 = s, a_1 = a\} = 1$ for all $(s, a) \in C$ determines when $M(\beta, c)$ is not a contractive mapping. That is, let $J = \{c: P\{0 \leq X_1 \leq c | s_1 = s, a_1 = a\} = 1 \text{ for all } (s, a) \in C\}$, and $J_c = \{\beta \in (0, 1) | M(\beta, c) \text{ is not a contractive mapping and } c \in J\}$; then $J_{c_1} \subseteq J_{c_2}$ if $c_2 \geq c_1$ and $c_1, c_2 \in J$.

For all $1 \leq i, j \in S$, let $H_{ij} = \sup \{k \in K: p_{ijk} > 0\}$ and $h_{ij} = \inf \{k \in K: p_{ijk} > 0\}$, and let $\alpha = \max \{(H_{ij} - h_{ij})/(2H_{ij} - h_{ij}): i \in S, j \in S\} = (H_{s_p} - h_{s_p})/[2(H_{s_p} - h_{s_p}) + h_{s_p}]$ for some $s \in S$ and $p \in S$. Then $0 < \alpha \leq \frac{1}{2}$.

COROLLARY 4. *If $\beta > \alpha$, then $M(\beta, b)$ is not a contractive mapping.*

Proof. If $\beta > \alpha$, then $\min \{\beta u + h_{ij}: i \in S, j \in S\} > \max \{H_{ij}: i, j \in S\}$ which implies $(\beta u + D) \cap D = \emptyset$ where $D = \bigcup_{i \in S} D_i$. Therefore $\beta u \notin W$. Theorem 2 implies that $M(\beta, b)$ is not a contractive mapping.

Comments. 1°. In fact, $0 \leq \alpha \leq \frac{1}{2}$, so, from Corollary 4, $\beta > \frac{1}{2}$ also implies that $M(\beta, b)$ is not a contractive mapping.

2°. Let $0 \leq \varepsilon < \frac{1}{2}$ and $\alpha = \varepsilon$; so $H_{s_p} = (1 - \varepsilon)h_{s_p}/(1 - 2\varepsilon) \geq h_{s_p}$. From the above argument, for any $0 < \beta < 1$, $M(\beta, b)$ may fail to be a contractive mapping.

2.5. Convergence.

DEFINITION. Let $x_1 \in \mathcal{X}$ and let $\{x_n\}_{n=1}^\infty$ be the sequence defined by $x_{n+1} = (1 - t_n)x_n + t_n Mx_n$, where $\{t_n\}_{n=1}^\infty$ is a real sequence. If a sequence $\{t_n\}_{n=1}^\infty$ satisfies the following two conditions, $\{x_n\}_{n=1}^\infty$ will be said to satisfy condition:

$$(a) \quad \sum_{n=1}^{\infty} t_n = \infty,$$

(b) there exists $0 < h < 1$ such that $0 < t_n < h < 1$ for all n .

Note that if $t_n \in [g, h]$ for all n and $0 < g < h < 1$, then it is obvious that the sequence $\{t_n\}_{n=1}^\infty$ satisfies (a) and (b).

Ishikawa [10] and others have investigated iterative methods to compute a fixed point of a nonexpansive mapping. The following result is [10, Lemma 2].

LEMMA 3. *If there exist $x_1 \in \mathcal{X}$ and $\{t_n\}_{n=1}^\infty$ that satisfy Condition A, then $x_n - Mx_n$ converges to zero in norm as $n \rightarrow \infty$.*

THEOREM 3. *Suppose $x_1 \in \mathcal{X}$ and $\{t_n\}_{n=1}^\infty$ satisfy Condition A.*

(a) *Then every subsequence $\{x_{n_i}\}$ of $\{x_n\}$ contains a subsequence that converges weakly to a fixed point of M .*

(b) *If M has a unique fixed point ν in \mathcal{X} , then $\{x_n\}_{n=1}^\infty$ converges weakly to ν .*

(c) *If there are a strongly compact subset of D of \mathcal{X} and a subsequence $\{x_{n_i}\}$ of $\{x_n\}$ such that $x_{n_i} \in D$ for all n_i then $\{x_n\}_{n=1}^\infty$ converges in norm to a fixed point of M .*

Proof. (a) Since \mathcal{X} is a weakly compact convex subset of E , $\{x_{n_i}\}$ contains a subsequence $\{x_{n_i}\}_{i=1}^\infty$ which converges weakly to a point ν . A well-known theorem of Mazur [14] implies that there is a sequence of convex combinations $\{\nu_n\}$ such that $\nu_n = \sum_{i=n}^N \lambda_i x_{n_i}$ where $\sum_{i=n}^N \lambda_i = 1$, and $\lambda_i = \lambda_i(n) \geq 0$, $n \leq i \leq N = N(n)$ which converges to ν in norm. Since M is nonexpansive and affine,

$$\begin{aligned} \|M\nu - \nu\| &= \|M\nu - M\nu_n + M\nu_n - \nu_n + \nu_n - \nu\| \\ &\leq 2\|\nu - \nu_n\| + \|M\nu_n - \nu_n\| \leq 2\|\nu - \nu_n\| + \sum_{i=n}^N \lambda_i \|Mx_{n_i} - x_{n_i}\| \end{aligned}$$

which implies that ν is a fixed point of M since $\lim_{n \rightarrow \infty} \|\nu_n - \nu\| = 0$ and $\lim_{i \rightarrow \infty} \|Mx_{n_i} - x_{n_i}\| = 0$.

(b) If M has a unique fixed point ν_0 in \mathcal{X} , (a) implies that $\{x_n\}$ converges weakly to ν_0 .

(c) Strong convergence of $\{x_n\}$ is a direct consequence of [10, Thm. 1].

3. Exponential utility.

3.1. Background. This section concerns maximization of $E[u_\lambda(B)]$ for the utility function $u_\lambda(x) = -e^{-\lambda x}$. This utility function has several attractions. First, the local risk aversion coefficient, i.e. $-u''(x)/u'(x)$, is constant with respect to x if, and only if, either $u(x) = c + dx$ or $u(x) = c + d e^{-\lambda x}$. If $u(\cdot)$ exhibits risk aversion and is increasing i.e. $u''(\cdot) < 0$ and $u'(\cdot) > 0$, then $u(x) = x + d e^{-\lambda x}$ with $d < 0$ and $\lambda > 0$. Second, $u(x) = c + d e^{-\lambda x}$ with $d < 0$ and $\lambda > 0$ is the only risk-averse increasing utility function whose risk premium is invariant with respect to wealth. That is, let $\alpha_X = E(X)$ and let π_X be the "risk-premium" for X : $u(\alpha_X - \pi_X) = E[u(X)]$. Then $\pi_X = \pi_{X+k}$ for all real numbers k and random variables X such that $|\alpha_X| < \infty$ if, and only if, either $u(\cdot)$ is linear or $u(x) = c + d e^{-\lambda x}$. If $d > 0$ then $\lambda > (<) 0$ implies risk aversion (risk preference).

Several writers have analyzed dynamic models with exponential utility functions. Let $B_n = \sum_{i=1}^n X_i$. Howard and Matheson [9] studied the maximization of $E[u_\lambda(B_n)]$ both for fixed n and as $n \rightarrow \infty$. Let $B_\infty = \lim_{n \rightarrow \infty} B_n$ if the limit exists. Denardo and Rothblum [4] studied the maximization of $E[u_\lambda(B_\infty)]$ in a stopping problem. That is,

the model in [4] is an MDP in which each set A_s includes an action which “stops” the decision process. The models in [9] and [4] exhibit risk-sensitivity but lack time-preference; the absence of discounting is the principal difference between their models and ours.

Jaquette [11], [12] studies the same problem as ours, namely maximization of $E[u_\lambda(B)]$. The analysis in [11] exploits the fact that $E[u_\lambda(B)]$ is the negative of the Laplace transform of B . As a result, there is a $\lambda_0 > 0$ and a stationary policy which is optimal for all $0 < \lambda < \lambda_0$. It is shown in [11] that such a stationary policy is also “moment optimal,” that is, it lexicographically maximizes the sequence of signed moments of B . The sign is positive (negative) if the moment is odd (even). In [12], λ is fixed and, perhaps, $\lambda \geq \lambda_0$. Let an optimal policy consist of a sequence $\delta_1, \delta_2, \dots$ of single-period decision rules. Jaquette shows that there is an optimal policy such that there exists $N < \infty$ with $\delta_n = \delta_N$ for all $n \geq N$. He also presents an example due to J. M. Harrison which implies that generally $N > 1$; that is, it is possible for all stationary policies to be suboptimal.

Porteus [15], [16] and Eagle [5] present dynamic choice theories which are compatible with exponential *intra*-period utility functions. Porteus describes a process in which preferences are based on the evolution of a decision-maker’s wealth. Eagle’s model begins with preferences based on consumption which is constrained by wealth. Thus, he infers a utility function for wealth.

References [15], [16] and [5] assume that in each period the decision-maker would accept a certainty equivalent as a terminal payment in lieu of continuing the dynamic choice process. They observe that their processes can be interpreted as having different intertemporal resolutions of uncertainty than in [11], [12] (hence, different also from ours). The models in [15], [16] and [5] share the property that there is a stationary optimal policy. Following the proof of Theorem 4 (below) we comment on this property of their models.

The models in [4], [5], [9], [11], [12], [15], [16] all have the following property. The conditional distribution of the reward X_n , given s_n , a_n , and s_{n+1} is degenerate. In our notation, the transition probabilities in those papers are written p_{sj}^a rather than p_{sjk}^a as in this paper. A simple example in the Appendix illustrates the manner in which nondegenerate conditional distributions occur in practice. However, the mathematics of exponential utility optimization does not depend on whether or not the conditional distributions (of the X_n ’s) are degenerate. A transformation (below between (7) and (7’)) permits suppression of the triple-subscript notation p_{sjk}^a for the remainder of the paper’s main text. Although the within-period randomness is not relevant to the analysis of MDP’s with risk-neutral criteria, ignoring its presence in risk-sensitive MDP’s would alter numerical results (such as which policy is optimal).

The results in [4], [5], [9], [11], [12], [15], [16] (such as optimality of a stationary policy in [5], [15], [16]) are preserved if the models in those papers are enlarged to nondegenerate conditional distributions of X_n . Thus, we shall ignore this difference between other models and ours when we compare results. Moreover, every MDP can be made to possess degenerate conditional distributions by suitably enlarging the state space.

3.2. Optimality equations. Let $B(n) = \sum_{i=1}^n \beta^{i-1} X_i$ with B continuing to denote $B(\infty)$. For each $s \in S$ and $\lambda \geq 0$ let $f_0(s, \lambda) = -1$,

$$(6) \quad \begin{aligned} f_n(s, \lambda) &= \sup \{E(-e^{-\lambda B(n)} | s_1 = s)\} & (n = 1, 2, \dots), \\ f(s, \lambda) &= \sup \{E(-e^{-\lambda B} | s_1 = s)\} \end{aligned}$$

where the suprema are over all *policies*, i.e. nonanticipative decision rules for choosing the actions a_1, a_2, \dots .

It can be shown that

$$(7) \quad \begin{aligned} f_n(s, \lambda) &= \max \{E[e^{-\lambda X_1} f_{n-1}(s_2, \beta\lambda) | s_1 = s, a_1 = a]; a \in A_s\} \\ &= \max \left\{ \sum_{j \in S} q_{sj}^a(\lambda) f_{n-1}(j, \beta\lambda) : a \in A_s \right\} \end{aligned}$$

where $q_{sj}^a(\lambda) = \sum_{k \in K} p_{sjk}^a e^{-\lambda k}$. Note that (7) can be written

$$(7') \quad f_n^\lambda = \max_{\delta} \{Q_\delta(\lambda) f_{n-1}^\lambda : \delta \in \Delta\}$$

where f_n^λ is the vector with coordinates $f_n(s, \lambda)$, $Q_\delta(\lambda)$ is the matrix whose (s, j) th element is $q_{sj}^{\delta(s)}(\lambda)$, and $\Delta = \times_{s \in S} A_s$. Let π be a policy and

$$v(\pi, s, \lambda) = E_\pi(-e^{-\lambda B} | s_1 = s)$$

where E_π denotes the expectation with respect to the probability distribution of B induced by π . Then π^* is said to be λ -optimal if

$$v(\pi^*, s, \lambda) \geq v(\pi, s, \lambda) \quad \text{for all } s \in S \text{ and } \pi.$$

Let R denote the set of real numbers.

THEOREM 4.⁴ (a) For each $s \in S$ and $\theta > 0$,

$$\lim_{n \rightarrow \infty} f_n(s, \theta) = f(s, \theta)$$

with $f_n(s, \theta) \leq f_{n+1}(s, \theta)$ for all n .

(b) For each $s \in S$ and $\theta > 0$,

$$(8) \quad f(s, \theta) = \max \left\{ \sum_{j \in S} q_{sj}^a f(j, \beta\theta) : a \in A_s \right\}.$$

That is,

$$(8') \quad f^\theta = \max_{\delta} Q_\delta(\theta) f^{\beta\theta}$$

where f^θ is the vector with components $f(s, \theta)$.

(c) Let $a = \delta_{n\lambda}(s) \in A_s$ attain the maximum in (8) when $\theta = \beta^{n-1}\lambda$ and let $\pi(\lambda) = (\delta_{1\lambda}, \delta_{2\lambda}, \dots)$ be the policy which uses the single period rule $\delta_{n\lambda}$ in period n . Then $\pi(\lambda)$ is λ -optimal.

Proof. Fix $\lambda > 0$. Since $B(n) \geq 0$ for all n , $-1 \leq -\exp[-\lambda B(n)] \leq 0$ so $-1 \leq f_n(s, \lambda) \leq 0$ for all n, s , and λ . Therefore, for all $\theta \geq 0$, $-e \leq f_0^\theta \leq f_1^\theta$ where e is the S -vector in which all components are one. Induction leads to $f_n^\theta \leq f_{n+1}^\theta$ for all $\theta \geq 0$. It follows from

$$(9) \quad P\{0 \leq X_i \leq u \text{ for all } i\} = 1$$

that $P\{0 \leq B - B(n) \leq \beta^n u / (1 - \beta)\} = 1$. Therefore, $f_n^\theta \leq f^\theta \leq f_n^\theta \exp[-\theta \beta^n u / (1 - \beta)] \leq f^\theta \exp[-\theta \beta^n u / (1 - \beta)]$; so $f^\theta = \lim_{n \rightarrow \infty} f_n^\theta$.

In order to prove that f satisfies (8), fix s and θ and let

$$J_n(\theta, \delta) = Q_\delta(\theta) f_{n-1}^{\beta\theta}.$$

⁴ We are grateful to a referee for suggesting much shorter proofs of parts (a) and (c) than our original ones.

Pointwise monotone convergence of f_n implies the same property for J_n : $J_n(\theta, \delta) \leq J_{n+1}(\theta, \delta)$ and there exists $J(\theta, \delta) = \lim_{n \rightarrow \infty} J_n(\theta, \delta)$. Let $n \rightarrow \infty$ in (7') to obtain

$$(10) \quad f_n^\theta \leq \max \{Q_\delta(\theta) f^{\beta\theta} : \delta \in \Delta\}$$

so

$$(11) \quad f^\theta \leq \max \{Q_\delta(\theta) f^{\beta\theta} : \delta \in \Delta\}.$$

In order to obtain the reverse inequality let $n \rightarrow \infty$ in (7'):

$$f^\theta \geq \lim_{n \rightarrow \infty} \max \{Q_\delta(\theta) f_{n-1}^{\beta\theta} : \delta \in \Delta\}.$$

For each s , A_s is a finite set; so $J_n(\theta, \cdot)$ is trivially upper semicontinuous and converges pointwise uniformly to $J(\theta, \cdot)$. Therefore

$$\lim_{n \rightarrow \infty} \max \{Q_\delta(\theta) f_{n-1}^{\beta\theta} : \delta \in \Delta\} = \max \{Q_\delta(\theta) f^{\beta\theta} : \delta \in \Delta\}$$

so reversing the inequality in (11) yields a valid statement and (8) is true.

In order to establish (c), define $\pi(\lambda)$ as in the statement of (c). An induction which employs (8) and starts at $n = 1$ establishes

$$f(s, \lambda) = E_{\pi(\lambda)}[-e^{-\lambda B(n)} | f(s_{n+1}, \beta^n \lambda) | | s_1 = s]$$

for all $n = 1, 2, \dots$. However, $|f(s_{n+1}, \beta^n \lambda)| \rightarrow 1$ as $n \rightarrow \infty$ because (9) implies $\exp[-\beta^n \lambda u / (1 - \beta)] \leq f(s_{n+1}, \beta^n \lambda) \leq 1$ for all n (all with probability one). Therefore, $f(s, \lambda) = E_{\pi(\lambda)}(-e^{-\lambda B} | s_1 = s)$.

Comments. 1°. A result analogous to Theorem 4 is valid for minimization problems. That is, if “inf” replaces “sup” in (6), then (7) and (8) are valid with “min” replacing “max” and parts (a) and (c) of Theorem 4 remain true.

2°. Theorem 4 is valid under more general conditions concerning the sets of actions. For each $s \in S$, suppose that A_s is a compact set and $J_n(\theta, \cdot)$ is continuous on Δ . Parts (a) and (b) of the theorem are valid with “sup” instead of “max” in (8) but the supremum is always attained. Part (c) is valid if “supremum” replaces “maximum.” Dini’s theorem [18] is invoked in order to establish (8).

3°. The inductive proof of $f_n^\theta \leq f_{n+1}^\theta$ depends on rewards being nonnegative and salvage values not depending on the terminal state. Nonnegativity entails no loss of generality beyond the compactness assumptions already made. Also, (7) and (7') remain valid if $f_0(s, \theta) = -\exp[-\lambda L(s)]$ replaces $f_0(s, \theta) = -1$ when $L(s)$ is the salvage value of terminal state s .

4°. It is insightful to compare (8) with the functional equations corresponding to the infinite horizon models in Porteus [15] and [16] and Eagle [5]. Our equation is

$$(8) \quad f(s, \theta) = \max \left\{ \sum_{j \in S} q_{sj}^a(\theta) f(j, \beta\theta) : a \in A_s \right\},$$

the equation corresponding to [15] and [16] is

$$(12a) \quad f(s, \theta) = \max \left\{ -\frac{\beta}{\theta} \log \left[\sum_{j \in S} q_{sj}^a(\theta/\beta) e^{-\theta f(j, \theta)} \right] : a \in A_s \right\},$$

and the equation corresponding to [5] is

$$(12b) \quad -e^{-\theta f(s, \theta)} = \max \left\{ -\sum_{j \in S} q_{sj}^a(\theta) e^{-\theta \beta f(j, \theta)} : a \in A_s \right\}.$$

Notice in (12a) and (12b) (which are implicit but do not appear in [5], [15], [16]) that wherever $f(\cdot, \cdot)$ appears on the right side, the second argument (θ) is the same quantity which appears in the second argument of $f(\cdot, \cdot)$ on the left side. With an infinite horizon, in (12a) and (12b) the decision-maker expects to have the same preference ordering next period as this period. In (8), however, the second argument is θ on the left side and $\beta\theta$ on the right side. Since $\theta > \beta\theta$, the decision-maker expects to be less risk-averse next period than at present. It should not be surprising that it may be optimal to use a different decision rule next period than at present, i.e., a nonstationary policy.

5°. The results in [21] can be used to construct proofs of the existence of a stationary optimal policy in the models in [5], [15] and [16]. Our comment in 4° above corresponds to the "stationarity of preference" axiom in [21]. If C is at most denumerable (i.e. if S and A_s for each $s \in S$ are at most denumerable), then a stationary policy is optimal if any policy is optimal. Moreover, if C is finite then policy improvement algorithms for [5], [15], [16] can be inferred from [21]. Such an algorithm is described in [5] but the result for [15] and [16] is new.

6°. Theorem 4 is closely related to results obtained by Furukawa and Iwamoto [7], Kreps [13], and Schäl [19]. First, $u_\lambda(x) \leq 0$ for all x so it follows from [19] that all policies are " u_λ -equalizing" and that a policy is λ -optimal if, and only if, it is " u_λ -thrifty." Second, $u_\lambda(\cdot)$ would be encompassed by the results in [7] (cf. Example 3) if it were replaced by $+e^{-\lambda x}$ (or if one minimized $E[u_\lambda(B)]$ instead of maximizing it.) Third, parts (a) and (b) of Theorem 4 show that part (e) of the corollary and Proposition 1 in [13] are valid for $u_\lambda(\cdot)$ with a restriction to Markov policies. Last, an alternative proof of part (c) of Theorem 4 could be based on part (c) of the corollary in [13].

3.3. Approximations. We continue to assume that the MDP is finite, i.e. $\# C < \infty$ where $C = \{(s, a); a \in A_s, s \in S\}$. Suppose δ attains the maximum on the right side of (8') when $\theta = \lambda$, \mathcal{D} denotes the set of bounded real-valued functions on $S \times (0, \infty)$, and

$$g_\lambda(s) = g(s, \lambda) \quad \text{for } g \in \mathcal{D}, \quad g_\lambda = [g_\lambda(s); s \in S], \quad (g_\lambda)_m = \min \{g_\lambda(s); s \in S\}, \\ (g_\lambda)^M = \max \{g_\lambda(s); s \in S\} \quad \text{and} \quad Lg_\lambda = \max \{Q_\delta(\lambda)g_{\lambda\beta}; \delta \in \Delta\}.$$

The following result is similar to [17, Prop. 1].

LEMMA 4. *For $g \in \mathcal{D}$ and $h \in \mathcal{D}$, suppose α attains Lg_λ , i.e. $Lg_\lambda = Q_\alpha(\lambda)g_{\lambda\beta}$. Then $Lh_\lambda - Lg_\lambda \geq Q_\alpha(h_{\beta\lambda} - g_{\beta\lambda})$.*

Proof.

$$Lh_\lambda - Lg_\lambda = \max \{Q_\delta h_{\beta\lambda}; \delta \in \Delta\} - Q_\alpha g_{\beta\lambda} \geq Q_\alpha h_{\beta\lambda} - Q_\alpha g_{\beta\lambda} = Q_\alpha(h_{\beta\lambda} - g_{\beta\lambda}). \quad \square$$

It is convenient to employ the notation $\xi = Lv_\lambda - v_{\beta\lambda} - f_\lambda + f_{\beta\lambda}$. Let

$$\gamma = \max \{E(e^{-\lambda X_1} | s_1 = s, a_1 = a); (s, a) \in C\}.$$

We assume $\gamma < 1$ which is true if for every $(s, a) \in C$ there is $(j, k) \in T$ such that $k > 0$ and $p_{sjk}^a > 0$. The following bounds are not useful but are the basis (in Corollary 5, below) of practical bounds.

THEOREM 5. *If $\gamma < 1$, then for all $v \in \mathcal{D}$,*

$$(13) \quad Lv_\lambda - \xi + e(\xi)_m / (1 - \gamma) \leq Lv_\lambda + \gamma e(\xi)_m / (1 - \gamma) \leq f_\lambda \\ \leq Lv_\lambda + \gamma e(\xi)^M / (1 - \gamma) \leq Lv_\lambda - \xi + e(\xi)^M / (1 - \gamma).$$

Proof. Let Q be induced by a single stage rule which attains Lf_λ . Then

$$Lv_\lambda - Lf_\lambda \geq Q(v_{\beta\lambda} - f_{\beta\lambda}), \quad \xi \geq (Q - I)(v_{\beta\lambda} - f_{\beta\lambda})$$

because

$$Lf_\lambda = f_\lambda, \quad (I - Q)^{-1}\xi \geq -(v_{\beta\lambda} - f_{\beta\lambda}),$$

$f_{\beta\lambda} \leq v_{\beta\lambda} + (I - Q)^{-1}\xi = v_{\beta\lambda} + (I - Q)^{-1}Q\xi + (I - Q)^{-1}(\xi - Q\xi) \leq \xi + v_{\beta\lambda} + (I - Q)^{-1}Q\xi$,
that is,

$$f_\lambda \leq Lv_\lambda + (I - Q)^{-1}Q\xi \leq Lv_\lambda + \mathbf{e}\gamma(\xi)^M/(1 - \gamma)$$

which is the third inequality in (13). The fourth inequality is implied by $\xi \leq (1 - \gamma)\mathbf{e}(\xi)^M/(1 - \gamma)$.

Suppose Q_α is induced by a single-stage rule which attains Lv_λ . Then

$$\begin{aligned} Lf_\lambda - Lv_\lambda &\geq Q_\alpha(f_{\beta\lambda} - v_{\beta\lambda}), & -\xi &\geq (Q_\alpha - I)(f_{\beta\lambda} - v_{\beta\lambda}), \\ & & -(I - Q_\alpha)^{-1}(-\xi) &\leq f_{\beta\lambda} - v_{\beta\lambda}, \\ & & v_{\beta\lambda} + (I - Q_\alpha)^{-1}\xi &\leq f_{\beta\lambda}. \end{aligned}$$

That is,

$$f_{\beta\lambda} \geq v_{\beta\lambda} + \sum_{l=0}^{\infty} Q_\alpha^l \xi \geq v_{\beta\lambda} + \xi + \mathbf{e}\gamma(\xi)_m/(1 - \gamma) = \xi + v_{\beta\lambda} + \mathbf{e}\gamma(\xi)_m/(1 - \gamma).$$

So $f_\lambda \geq Lv_\lambda + \mathbf{e}\gamma(\xi)_m/(1 - \gamma)$ which is the second inequality in (13). The first inequality is implied by $\xi \geq \mathbf{e}(\xi)_m$. \square

The bounds in Theorem 5 are impractical because they depend on the function whose bounds are sought. Corollary 5 (below) removes this defect. Let

$$c_\theta(s) = -\min \left\{ \sum_{j \in S} q_{sj}^a(\theta): a \in A_s \right\} = \max \{ E(-e^{-\theta X_1} | s_1 = s, a_1 = a): a \in A_s \}$$

and let $c_\theta = [c_\theta(s); s \in S]$. Then $X_1 \leq B \leq X_1 + \beta b + \beta^2 b + \cdots$ implies

$$(14) \quad c_\theta \leq f^\theta \leq c_\theta e^{-\beta\theta u}.$$

COROLLARY 5. *If $\gamma < 1$ and $v \in \mathcal{D}$ then*

$$Lv_\lambda + \gamma \mathbf{e}(Lv_\lambda - v_{\beta\lambda} - e^{-\beta u} c_\lambda + c_{\beta\lambda})_m/(1 - \gamma) \leq f_\lambda \leq Lv_\lambda + \gamma \mathbf{e}(Lv_\lambda - v_{\beta\lambda})^M/(1 - \gamma).$$

Proof. The upper bound follows from the upper bound in Theorem 5 and the fact that each component of f^θ is nondecreasing in θ ; so $f^{\lambda\beta} \leq f^\lambda$. The lower bound follows from the lower bound in Theorem 5 and (14) which implies

$$f^\lambda - f^{\lambda\beta} \leq c_\lambda e^{-\beta u} - c_{\beta\lambda}.$$

3.4. A pointless procedure. Let \mathcal{F} be the set of all bounded real-valued functions on S . Fix $\lambda (\lambda > 0)$. For each $u \in \mathcal{F}$ and $v \in \mathcal{F}$, let $d(u, v) = \sup \{|u(s, \lambda) - v(s, \lambda)|: s \in S\}$. Then $(\mathcal{F}, d(\cdot, \cdot))$ is a complete metric space.

Without loss of generality, we assume that $P\{X_1 > 1 | s_1 = s, s_2 = j, a_1 = a\} = 1$ so $\sum_{j \in S} q_{sj}^a(\theta) < 1$, $(s, a) \in C$, $\theta > 0$. Define a mapping $\Gamma: \mathcal{F} \rightarrow \mathcal{F}$ where $\Gamma u(s) = \max \{\sum_{j \in S} q_{sj}^a(\theta) u(j, \theta): a \in A_s\}$ for $s \in S$. Then $d(\Gamma u, \Gamma v) \leq e^{-\theta} d(u, v)$ for all $u, v \in \mathcal{F}$. Hence, Γ is a contraction mapping, and the fixed-point theorem for contraction mappings guarantees that Γ has a unique fixed point. Since $\Gamma 0 = 0$ it follows that 0 is the fixed point of Γ . Therefore, the equation

$$g(s, \theta) = \max \left\{ \sum_{j \in S} q_{sj}^a(\theta) g(j, \theta): a \in A_s \right\}$$

has the unique solution $g(s, \theta) = 0$ for all $s \in S$ and $\theta > 0$. Therefore, iterating Γ from any initial function does not necessarily yield improving approximations of f in (8).

Appendix. This example illustrates that the notation p_{sjk}^a is essential in some risk-sensitive models. The following MDP models a single product's inventory process in which excess demand is lost, there is storage space for at most two items, demand each period is equally likely to be 0, 1, 2, or 3, and demands in successive periods are independent random variables. We assume that the replenishment decision is made each period before demand is known, but ordered goods are delivered immediately. Suppose that the unit purchase cost is \$1, the unit holding cost for end-of-period inventory is \$1, and there is a \$1 penalty for each unit of demand which exceeds supply.

Let a denote the inventory level after replenishment; nonnegative replenishment quantities are equivalent to $a_n \geq s_n$. Thus, $S = \{0, 1, 2\}$ and $C = \{(0, 0), (0, 1), (0, 2), (1, 1), (1, 2), (2, 2)\}$. Since excess demand is lost, $s_{n+1} = (a_n - D_n)^+$ where D_n denotes the quantity demanded in period n . The net profit in period n is

$$\begin{aligned} X_n &= a_n - s_n + (a_n - D_n)^+ + (D_n - a_n)^+ \\ &= a_n - s_n + s_{n+1} + D_n - a_n + s_{n+1} = -s_n + 2s_{n+1} + D_n \end{aligned}$$

which uses the identity $\min\{x, y\} = x - (x - y)^+$.

Suppose for some n that $s_n = 1$, $a_n = 2$, and $s_{n+1} = 0$. Then $X_n = -1 + D_n$. We know $D_n \geq 2$ because $s_{n+1} = 0 = (a_n - D_n)^+ = (2 - D_n)^+$; so $D_n \in \{2, 3\}$. That is, $X_n = 1$ and $X_n = 2$ are equally likely and $p_{101}^2 = p_{102}^2 = 0.5$. The main point is that there are k and l , $k \neq l$, such that $p_{10k}^2 > 0$ and $p_{10l}^2 > 0$.

In this example, X_n cannot be given as a deterministic function of s_n , a_n , and s_{n+1} .

REFERENCES

- [1] D. P. BERTSEKAS, *Dynamic Programming and Stochastic Control*, Academic Press, New York, 1976.
- [2] L. BREIMAN, *Probability*, Addison-Wesley, Reading, MA, 1968.
- [3] E. V. DENARDO, *Dynamic Programming*, Prentice-Hall, Englewood Cliffs, NJ, 1982.
- [4] E. V. DENARDO AND U. G. ROTHBLUM, *Optimal stopping, exponential utility, and linear programming*, Math. Programming, 16 (1979), pp. 228-244.
- [5] J. N. EAGLE II, *A utility criterion for the Markov decision process*, Ph.D. thesis, Stanford Univ., Stanford, CA, 1975.
- [6] P. C. FISHBURN, *Utility Theory for Decision Making*, John Wiley, New York, 1970.
- [7] N. FURUKAWA AND S. IWAMOTO, *Markovian decision processes with recursive reward functions*, Bull. Math. Statist., 15 (1972), pp. 79-91.
- [8] D. P. HEYMAN AND M. J. SOBEL, *Stochastic Models in Operation Research*, Volume II, McGraw-Hill, New York, 1984.
- [9] R. S. HOWARD AND J. E. MATHESON, *Risk-sensitive Markov decision processes*, Management Sci., 8 (1972), pp. 356-369.
- [10] S. ISHIKAWA, *Fixed points and iteration of a nonexpansive mapping in a Banach space*, Proc. Amer. Math. Soc., 59 (1976) pp. 65-71.
- [11] S. C. JAQUETTE, *Markov decision processes with a new optimality criterion: Discrete time*, Ann. Stat., 1 (1973), pp. 496-505.
- [12] ———, *A utility criterion for Markov decision processes*, Management Sci., 23 (1976) pp. 43-49.
- [13] D. M. KREPS, *Decision problems with expected utility criteria, I: upper and lower convergent utility*, Math. Oper. Res., 2 (1977) pp. 45-53.
- [14] S. MAZUR, *Über knovexe Menger in linearen normierten rauman*, Studia Math., 4 (1933) pp. 70-84.
- [15] E. L. PORTEUS, *On the optimality of structured policies in countable stage decision processes*, Management Sci., 22 (1975) pp. 148-157.
- [16] ———, *On the optimality of structured policies in countable stage decision processes*, Research Paper No. 141 Rev., Graduate School of Business, Stanford Univ., Stanford, CA, 1975.

- [17] M. L. PUTERMAN AND S. L. BRUMELLE, *Policy iteration in stationary dynamic programming*, Math. Oper. Res., 4 (1979) pp. 60–69.
- [18] H. L. ROYDEN, *Real Analysis*, Macmillan, New York, 1963.
- [19] M. SCHÄL, *Utility functions and optimal policies in sequential decision problems*, in Game Theory and Mathematical Economics, O. Moeschlin and D. Pallaschke, eds., North-Holland, Amsterdam, 1981, pp. 357–365.
- [20] J. SCHAUDER, *Der fixpunktsatz in funktional raumen*, Studia Math., 2 (1930), pp. 171–180.
- [21] M. J. SOBEL, *Ordinal dynamic programming*, Management Sci., 21 (1975) pp. 967–975.
- [22] ———, *The variance of discounted MDP's*, J. Appl. Probab., 19 (1982), pp. 794–802.
- [23] A. TYHONOV, *Ein fixpunktsatz*, Math. Ann., III (1935) pp. 767–776.

REALIZATION THEORY FOR HAMILTONIAN SYSTEMS*

J. BASTO GONÇALVES†

Abstract. We prove that control systems of the Hamiltonian type have a quasi-minimal realization on a state space of minimal dimension, of one of two types: a Hamiltonian system or the suspension of a time-dependent Hamiltonian system.

These realizations are unique up to a symplectomorphism, in the first case, or a canonical transformation, in the second one.

When the Lie algebra of the initial system Σ , assumed to be analytic, is finite dimensional, we obtain a characterization of the state space and the dynamics of the quasi-minimal realization in terms of the coadjoint actions of Lie groups associated with Σ ; an example is given applying these results.

Using control theoretic methods, we prove the Kostant-Kirillov-Souriau theorem; the technique used for systems with finite dimensional Lie algebra is closely related to the theory developed by those authors.

Key words. Hamiltonian systems, realization, quasi-minimality, nonlinear systems

AMS(MOS) subject classifications. 93B20, 93C10, 93C15, 3H05, 58F05, 70G99

Introduction. The work on Hamiltonian control systems began with Brockett [5] as an attempt to present a control theoretic point of view of problems in analytical mechanics involving external forces. The precise basic formalism and first results on their realization are due to van der Schaft [13]–[16].

Our definition of Hamiltonian control system corresponds to a special case of [13], and was independently introduced in [3], along with a result about the equivalence of Hamiltonian systems (Theorem 4.1 here). Using this definition we obtain a more complete description of the construction of realizations of minimal dimension (quasi-minimal realizations) containing the results of [13], [14], [15], [16], where strong accessibility is assumed.

We extend the results to the nonstrongly accessible case, where the quasi-minimal realization is a suspension of a time-dependent Hamiltonian system, thereby proving that any complete smooth or analytic Hamiltonian control system has a quasi-minimal realization, either a Hamiltonian system or the suspension of a time dependent Hamiltonian system. The constructions are based on the theory introduced in [11] and developed in [4].

We prove that the symplectic (contact) structures obtained are characteristic of the given system, in the sense that any equivalence between the two quasi-minimal realizations preserves them, being a symplectomorphism (canonical transformation).

If the initial system has a finite dimensional Lie algebra, the constructions use the moment map [1], [2], [9], [17], [22] (general references for symplectic geometry and Hamiltonian vector fields), under quite natural assumptions, and the resulting state spaces are either an orbit of the coadjoint action of G (the associated group of diffeomorphisms) on the dual of its Lie algebra, or $\mathbb{R} \times S$, where S is an orbit of the coadjoint action of G_0 (a subgroup of G) on the dual of its Lie algebra; the dynamics are given by that action, in the first case, or its suspension, in the second case. For related work, developing some of the techniques presented here, see [7]. As an

* Received by the editors December 12, 1984, and in revised form August 13, 1985.

† Grupo de Matemática Aplicada, Faculdade de Ciências da Universidade do Porto, 4000 Porto, Portugal.

interesting by-product we give an easy proof of the Kostant–Kirillov–Souriau theorem [17], [9].

1. Basic definitions. Let M be a smooth or analytic connected manifold; a symplectic form w on M is a closed nondegenerate 2-form, and the pair (M, w) is called a symplectic manifold.

Let $s: V(M) \rightarrow \Omega^1(M)$ be defined by $s(X) = w(X, \cdot)$, where $V(M)$ and $\Omega^1(M)$ are the set of smooth (or analytic) vector fields, respectively 1-forms, on M . Then s is a linear isomorphism, and we can define the map $j: C(M) \rightarrow V(M)$ by $j(h) = s^{-1}(dh)$, $C(M)$ being the set of smooth, or analytic, functions on M ; clearly if $X_h = j(h)$ then $w(X_h, \cdot) = dh$.

X_h is said to be a Hamiltonian vector field, and h the corresponding Hamiltonian (function). We denote by $H(M)$ the set of Hamiltonian vector fields, the image of j .

A map $f: M' \rightarrow M''$ between two symplectic manifolds is a symplectomorphism if $f^*w'' = w'$. In particular, this implies f is a local diffeomorphism [1, p. 177].

PROPOSITION 1.1 [1, p. 188]. *If X is a Hamiltonian vector field then:*

(i) $L_X w = 0$.

(ii) *If $\{X_i\}$ is the associated pseudo-group of diffeomorphisms, then $X_i^* w = w$.*

The Poisson bracket of two functions g and h in $C(M)$ is defined by $\{g, h\} = w(X_h, X_g)$.

PROPOSITION 1.2 [1, p. 194]. (i) $C(M)$ with the Poisson bracket $\{\cdot, \cdot\}$ is a Lie algebra.

(ii) $X_{\{g, h\}} = [X_g, X_h]$.

(iii) $H(M)$ is a Lie subalgebra of $V(M)$.

PROPOSITION 1.3. [10, p. 261]. *The sequence $0 \rightarrow \mathbb{R} \xrightarrow{i} C(M) \xrightarrow{j} H(M) \rightarrow 0$, where i is the natural inclusion as a subspace of constants, is an exact sequence of Lie algebras.*

Let $P: M \times \mathbb{R}^m \rightarrow \mathbb{R}$ be a C^2 function such that P_u defined by $P_u(x) = P(x, u)$ is a smooth (analytic) function on the symplectic manifold M . We define $f: M \times \mathbb{R}^m \rightarrow TM$ by $f(\cdot, u)$ being the Hamiltonian vector field corresponding to the function P_u , and take U as the family of piecewise constant maps from $[0, +\infty[$ into \mathbb{R}^m ; then $\Sigma = (M, \Omega = \mathbb{R}^m, f, U)$ is a control system, as considered in [20], the Hamiltonian control system defined by P .

We denote by D the set of associated vector fields $f(\cdot, u)$ with u in Ω —which we assume to be complete—and by $G(S)$ the group (semi-group) of diffeomorphisms generated by finite products of diffeomorphisms associated to vector fields in D (corresponding to nonnegative time).

Σ is said to be reachable or orbit-minimal if the orbit of any point in M under the action of G is M .

Some difficulties arise when we try to define the corresponding system with outputs within the framework of [20], since the output maps should be $y: M \times \Omega \rightarrow \Omega$, with $y(x, u) = \partial/\partial u P(x, u)$. We define the output space to be Ω ; instead of one output map we consider a family of maps $(y_u, u \in \Omega)$, $y_u(x) = y(x, u)$, and the system with outputs will be $\Sigma = (M, \Omega, f, U, \Omega, \{y_u\})$. It has been proved in [3] that the relevant properties concerning observability and quasi-minimal realizations, as stated in [4], are not altered if we define indistinguishability by: x' and x'' are (weakly) indistinguishable if for every $g \in S(G)$ and $u \in \Omega$ we have $y_u(gx') = y_u(gx'')$.

The observability codistribution Q is the smallest G -invariant codistribution containing the differentials of the components of the output maps; if $\dim Q = n$ then Σ is locally weakly observable [4]. The observability distribution Δ is defined by $\Delta = \{X \in V(M), Q \cdot X = 0\}$.

Let $P(D)$ be the smallest G -invariant distribution containing D [19]. A realization is said to be quasi-minimal [4] if $\dim P(D) = n$ (reachable realization) and $\dim Q = n$ (weakly observable realization); its state space has minimal dimension for all possible realizations.

2. Quasi-minimality. Given a system Σ , we denote by P_0 the distribution generated by vector fields of the form $g_*(X - X')$, with $g \in G$ and $X, X' \in D$; P_0 is the smallest G -invariant distribution containing the differences of vector fields in D .

PROPOSITION 2.1. *If Σ is a reachable Hamiltonian system and $\dim P_0 = n$ then Σ is quasi-minimal.*

Proof. If we prove that $Q = w(P_0, \cdot)$ then the condition $\dim P_0 = n$ implies that $\dim Q = n$, and as $\dim P(D) = n$, Σ will be quasi-minimal.

Let g be in G and $X^i \lrcorner w = dP(\cdot, u_i)$, $i = 1, 2$. Then $g_*(X^1 - X^2) \lrcorner w = g^{-1*}(dP(\cdot, u_1) - dP(\cdot, u_2))$ as can be verified by an easy calculation bearing in mind that g is a symplectomorphism, and so $P_0 \lrcorner w$ is generated by these elements.

On the other hand, Q is generated by elements of the form

$$g^* d \frac{\partial}{\partial u_i} P(\cdot, u) = g^* \frac{\partial}{\partial u_i} dP(\cdot, u).$$

Essentially the situation is as follows: given a smooth (or analytic) map $q: \mathbb{R}^m \rightarrow \mathbb{R}^n$ we compare the subspace V generated by the vectors $\partial/\partial u_i q(u)$ with the subspace V' generated by the vectors $q(u) - q(u')$.

From $\partial/\partial u_i q(u) = \lim (q(u+h) - q(u))/h$ as $h \rightarrow 0$, and since V' is closed, we see that V is contained in V' .

Now let u, u' be two distinct points; we can write the difference $q(u') - q(u)$ as the sum over $i = 1, 2, \dots, m$ of $q(u+ih) - q(u+(i-1)h)$, with $h = (u' - u)/m$; also $q(u+ih) - q(u+(i-1)h) = \partial/\partial u q(u+(i-1)h) \cdot h + o(h)$, and the first term in the right-hand side is in V . Therefore $q(u') - q(u)$ is the sum of a term in V plus a term $m \times o(h)$, which tends to zero as m tends to ∞ .

Again closeness of subspaces in finite dimensions allows us to conclude that V contains V' , and thus $V' = V$. \square

Given a Hamiltonian system Σ and x in M , an equivalent reachable realization Σ' can be obtained by restricting Σ to the orbit M' of G through x ; in general Σ' is not Hamiltonian.

We can define a new system Π from Σ by just altering the output maps: instead of $\partial/\partial u P(\cdot, u)$ we consider $P(\cdot, u)$. The two systems have the same dynamics, and the restriction of Π to M' is a reachable realization.

Let Δ^* be the observability distribution of Π' , and TM'^w the w -orthogonal of TM' : $w(x)(T_x M', v) = 0$ if $v \in TM'^w(x)$.

To obtain a quasi-minimal realization from Π' we have to quotient out the observability distribution [4]. Since here Δ^* is shown to be the intersection of TM' and its w -orthogonal, the resulting state space M_1 is a symplectic manifold on which the projection of Π' is quasi-minimal.

In fact, as $P(\cdot, u)$ is constant on the integral submanifolds of Δ^* so is $\partial/\partial u P(\cdot, u)$, and therefore we can also project Σ' on the same state space, obtaining a reachable system Σ_1 .

Σ_1 is not locally weakly observable (and a fortiori is not quasi-minimal) in general, but we shall prove that it is Hamiltonian and give a necessary and sufficient condition for being a quasi-minimal realization.

LEMMA 2.2. $M_1 = M'/\Delta^*$ is a symplectic manifold.

Proof. From [4] we know that M_1 is a smooth or analytic Hausdorff manifold and Δ^* a sub-bundle of TM' . Therefore if Δ^* is the intersection of TM' and its w -orthogonal, M_1 is a symplectic manifold [22].

By definition, Δ^* is contained in TM' ; it is easy to see that for the system Π' the observability codistribution can be obtained as $Q^* = w(P(D), \cdot)$: Q^* is generated by elements of the form $g^* dh$, where g is in G and h is a component of an output map, which can also be written as $w(g_*^{-1}X, \cdot)$. Then $\Delta^* = \{X \in V(M'), Q^* \cdot X = w(P(D), X) = 0\}$, and thus Δ^* is also contained in the w -orthogonal of $P(D)$; but $P(D) = TM'$ because Π' is reachable system, and the result follows from [22]. \square

THEOREM 2.3. *Any initialized Hamiltonian system (Σ, x) has an equivalent reachable Hamiltonian realization (Σ_1, x_1) . Σ_1 is a quasi-minimal realization if the condition of Proposition 2.1 is verified for Σ_1 .*

Proof. We take Σ_1 as defined before, and x_1 as the projection of x on M_1 . Since we have already seen that Σ_1 is reachable we need only to prove it is Hamiltonian and then the last assertion follows from Proposition 2.1.

It is enough to show that X is the Hamiltonian vector field corresponding to h_1 in the diagram below:

$$\begin{array}{ccc} TM & \xleftarrow{X_h} & M \xrightarrow{h} \mathbb{R} \\ \downarrow \pi_* & & \downarrow \pi \nearrow h_1 \\ TM_1 & \xleftarrow{X} & M_1 \end{array}$$

From the definition of w_1 on the quotient manifold M_1 and if $\pi(x) = x_1$, $\pi_*v = v_1$, we have:

$$\pi^*w_1(X(x_1), v_1) = w(X_h(x), v) = dh(x) \cdot v = \pi^*dh_1(x) \cdot v = \pi^*(dh_1(x_1) \cdot v_1).$$

As v_1 is arbitrary, $w_1(X, \cdot) = dh_1$. \square

This is a generalization of a similar result of van der Schaft [15], [16], obtained for affine systems, using essentially the same technique.

A time-dependent Hamiltonian control system on a symplectic manifold M is a system of the form:

$$\dot{x} = X_{P(t, \cdot, u)}(x), \quad y_{u,t}(x) = \frac{\partial}{\partial u} P(t, \cdot, u)$$

where P is a smooth or analytic function defined on $\mathbb{R} \times M \times \mathbb{R}^m$. To study this system as an autonomous system we can construct its suspension to the contact manifold $(\mathbb{R} \times M, p^*w)$ where $p: \mathbb{R} \times M \rightarrow M$ is the canonical projection. We have then

$$(t, x) = \left(\frac{\partial}{\partial t}, X_{P(t, \cdot, u)}(x) \right), \quad y_u(t, x) = \frac{\partial}{\partial u} P(t, x, u).$$

Let us assume now that we have a reachable Hamiltonian system Σ for which $\text{codim } P_0 = 1$; the approach used in [21] for the nonstrongly accessible case allows us to view Σ as the suspension of a time-dependent system defined on a maximal integral submanifold W of P_0 . In fact, there exists a covering projection $\pi: \mathbb{R} \times W \rightarrow M$ such that if we pull-back the function P , the associated vector fields and the symplectic form on M , we get an equivalent reachable Hamiltonian system, defined on $\mathbb{R} \times W$;

for this system the associated vector fields have the form $\partial/\partial t + q((t, x), u)$, with $q((t, x), u') = 0$ if u' in Ω corresponds to the flow used in the definition of π [21].

We can assume then that Σ is defined on a symplectic manifold $M = \mathbb{R} \times W$, and $f((t, x), u) = \partial/\partial t + q((t, x), u)$ with $\partial/\partial t(t, x) \perp w = dP(t, x, 0)$ and

$$q(t, \cdot, u) \perp dP'(t, \cdot, u), \quad P'(\cdot, u) = P(\cdot, u) - P(\cdot, 0);$$

note that $\partial/\partial u P = \partial/\partial u P'$.

The vector fields $q(\cdot, u)$ belong to TW (to be rigorous to $T\{t\} \times W$) and we can identify P_0 with TW by the same abuse of language.

LEMMA 2.4. *If Δ is the observability distribution of Σ , then $M_2 = M/\Delta$ has the form $M_2 = \mathbb{R} \times W_2$ where W_2 is a symplectic manifold.*

Proof. We have $P_0 = TW$, as explained before, and since $\dim \Delta + \dim P_0 = 2n$ [22] as they are w -orthogonal, $\dim \Delta = \text{codim } P_0 = 1$. Also, as $P_0/P_0 \cap \Delta$ has even dimension [22], $\dim P_0 \cap \Delta = 1$ and thus Δ is contained in $P_0 = TW$ and coincides with its intersection with TW^w , i.e. Δ is the intersection of TW with its w -orthogonal.

We can then write $M_2 = M/\Delta = (\mathbb{R} \times W)/\Delta = \mathbb{R} \times (W/TW \cap TW^w) = \mathbb{R} \times W_2$, and W_2 is a symplectic manifold [22]. \square

THEOREM 2.5. *A reachable Hamiltonian system Σ , with $\text{codim } P_0 = 1$, has a quasi-minimal realization Σ_2 which is the suspension of a time-dependent Hamiltonian system.*

Proof. Quasi-minimality results from the construction of Σ_2 as the projection of Σ on M_2 according to the procedure put forward in [11], [4] and used in the construction of Σ_1 . We have only to prove that $w(q_2(t, \cdot, u), \cdot) = dP'_2(t, \cdot, u)$ as in Theorem 3, writing $f_2(\cdot, u) = \partial/\partial t + q_2(\cdot, u)$ in the diagram below:

$$\begin{array}{ccccc} T\mathbb{R} \times TW & \xleftarrow{f(\cdot, u)} & \mathbb{R} \times W & \xrightarrow{P(\cdot, u)} & \mathbb{R} \\ \downarrow \text{id} \times \pi_* & & \downarrow \text{id} \times \pi & \nearrow P_2(\cdot, u) & \\ T\mathbb{R} \times TW_2 & \xleftarrow{f_2(\cdot, u)} & \mathbb{R} \times W_2 & & \end{array}$$

Since $\pi^*(q_2 \perp w_2) = q \perp w = dP'(t, \cdot, u) = \pi^*(dP'_2(t, \cdot, u))$ we can conclude that $q_2(t, \cdot, u) \perp w_2 = dP'_2(t, \cdot, u)$ as needed. \square

3. Hamiltonian systems with finite dimensional Lie algebra. In this section we consider only complete analytic Hamiltonian systems, such that T , the smallest Lie algebra in $V(M)$ containing D , is finite dimensional. G denotes the group of diffeomorphisms generated by the associated vector fields, as before.

THEOREM 3.1 [12]. (i) *G is a Lie transformation group on M , with action $\mu(g, x) = gx$.*

(ii) *The Lie algebra \mathfrak{g} of G is isomorphic to T through the isomorphism $\varphi: \mathfrak{g} \rightarrow T$, $\varphi(v)(x) = \mu_{x*}(v)$ where $\mu_x(g) = \mu(g, x)$.*

(iii) *Given $v \in \mathfrak{g}$ the vector field $\varphi(v)$ in T corresponding to v is complete.*

LEMMA 3.2. *$\text{Im } \varphi$ belongs to $H(M)$.*

Proof. If $\varphi(v)$ is in D it is a Hamiltonian vector field, and since $\text{im } \varphi = T$ every vector field in $\text{im } \varphi$ is a linear combination of Lie brackets of Hamiltonian vector fields, therefore a Hamiltonian vector field. \square

The action of G on M is symplectic, from Proposition 1.1(ii); it is a Poisson action if there exists a Lie algebra homomorphism β making the following diagram

commutative:

$$\begin{array}{ccc}
 & & \mathfrak{g} \\
 & \swarrow b & \downarrow \varphi \\
 0 \rightarrow \mathbb{R} \rightarrow C(M) & \xrightarrow{j} & H(M) \rightarrow 0
 \end{array}$$

From now on, we assume the action of G on M to be Poisson. Then we can define the moment map J of μ by $J: M \rightarrow \mathfrak{g}^*$, $J(x)v = \varphi(v)(x)$.

THEOREM 3.3 [1], [2], [10]. *If Ad_g^* is the coadjoint action of G on \mathfrak{g}^* , the following diagram is commutative:*

$$\begin{array}{ccc}
 M & \xrightarrow{\mu_g} & M \\
 \downarrow J & & \downarrow J \\
 \mathfrak{g}^* & \xrightarrow{\text{Ad}_g^*} & \mathfrak{g}^*
 \end{array}$$

THEOREM 3.4 [1], [2], [10]. *Let W be an orbit of the coadjoint action of G . Then:*

(i) *If $\Phi(v)$ is the vector field in \mathfrak{g}^* corresponding to v we have $J_*\varphi(v) = \Phi(v)$ and $\Phi(v)(a) \cdot v' = a \cdot [v, v']$, with $a \in \mathfrak{g}^*$ and $v, v' \in \mathfrak{g}$.*

(ii) *W has a canonical symplectic form defined by $w'(a)(\Phi(v), \Phi(v')) = a \cdot [v, v']$.*

(iii) *The vector field $\Phi(v)$ in W has Hamiltonian $h_v: W \rightarrow \mathbb{R}$, $h_v(a) = a \cdot v$.*

Example. Let $M = \mathbb{R}^6$, with w the usual symplectic form, and P given by $P(x, u) = g_0(x) + u_1 g_1(x) + u_2 g_2(x)$, where $g_0(x) = x_3 x_5 - x_2 x_6$, $g_1(x) = x_1 x_5 - x_2 x_4$, $g_2(x) = x_1 x_6 - x_3 x_4$.

An easy computation shows that, if X^i is defined by $X^i \lrcorner w = dg_i$, we have $[X^1, X^2] = X^0$, $[X^2, X^0] = X^1$, $[X^0, X^1] = X^2$. Therefore $\mathbf{T} = \mathbf{T}_0$ and is finite dimensional; from the above relations we see that \mathbf{T} is isomorphic to $\mathfrak{so}(3)$ through:

$$X^i \rightarrow A_i = \begin{pmatrix} 0 & e_{1i} & e_{2i} \\ \vdots & \vdots & \vdots \\ -e_{1i} & 0 & -e_{0i} \\ \vdots & \vdots & \vdots \\ -e_{2i} & e_{0i} & 0 \end{pmatrix}$$

with $i = 0, 1, 2$ and $e_{ij} = 0$ if $i \neq j$ and 1 if $i = j$.

Moreover, we can identify $\mathfrak{so}(3)$ with \mathbb{R}^3 by means of the map 1 defined by $1(e_1) = A_1$, $1(e_2) = A_2$, $1(e_3) = A_0$, and then the bracket in $\mathfrak{so}(3)$ corresponds to the usual vector product in \mathbb{R}^3 .

Defining an action of $\text{SO}(3)$ in \mathbb{R}^6 by means of

$$(A, x) \rightarrow \begin{pmatrix} A & 0 \\ 0 & A \end{pmatrix} x = A \cdot x$$

we see that $X^i(x) = d/dt (\exp tA_i \cdot x)|_{t=0}$ for $i = 0, 1, 2$. Thus we can assume we have been given an action $\text{SO}(3) \times \mathbb{R}^6 \rightarrow \mathbb{R}^6$, the maps $x \rightarrow A \cdot x$ being symplectomorphisms for every A in $\text{SO}(3)$.

If we identify \mathbb{R}^6 with $T^*\mathbb{R}^3$ the above action is induced by the usual action of $\text{SO}(3)$ on \mathbb{R}^3 , and thus is a Poisson action [2, p. 377]. Having identified $\mathfrak{so}(3)$ with \mathbb{R}^3 , we can identify $\mathfrak{so}(3)^*$ with \mathbb{R}^3 as well, using the Euclidean structure.

It is easy to see that the adjoint action of $SO(3)$ on \mathbb{R}^3 is equivalent to the usual action (but not the same: for instance the usual action of $\exp tA_0$ is the adjoint action of $\exp tA_1$) and therefore they have the same orbits: the orbit of x is the set $\{y \in \mathbb{R}^3, \|y\| = \|x\|\}$. From the definition of coadjoint action, its orbits are exactly the same.

As we know that the moment map does exist, we can compute it as $J(x)e_1 = g_1(x)$, $J(x)e_2 = g_2(x)$, $J(x)e_3 = -g_0(x)$. Denoting by p_i the projection of \mathbb{R}^3 on the i th factor, we have:

$$\begin{array}{ccc} \mathbb{R}^6 \times \mathbb{R}^2 & \xrightarrow{P} & \mathbb{R} \\ J \times \text{id} \downarrow & \nearrow (-p_3 \pm u_1 p_1 + u_2 p_2) & \\ \mathbb{R}^3 \times \mathbb{R}^2 & & \end{array}$$

Therefore we can project the original system in \mathbb{R}^3 , through J , and clearly the dynamics of the new system will be given by the coadjoint action and the output maps will be p_1 and p_2 . To obtain a reachable system we have to consider its restriction to an orbit of the coadjoint action; this is a symplectic manifold, and later we shall prove that we get a Hamiltonian system Σ by the procedure just outlined.

In this case it will be a quasi-minimal realization since we have already seen that $[X^1, X^2] = X^0$ and we can then apply Proposition 2.1.

In conclusion, the quasi-minimal realization Σ is defined on a sphere in \mathbb{R}^3 with the canonical symplectic form; the associated vector fields corresponding to X^i are rotations around the axis, their Hamiltonians being the restriction to that sphere of the canonical projections of \mathbb{R}^3 . \square

In the remaining part of this section, we shall generalize and justify the above construction, including the nonstrongly accessible case.

From the definition of J and as we have already done in the previous example, we see that, for every v in \mathfrak{g} , $\beta(v)$ factors as follows:

$$\begin{array}{ccc} M & \xrightarrow{\beta(V)} & \mathbb{R} \\ J \downarrow & \nearrow h_v & \\ \mathfrak{g}^* & & \end{array}$$

This means that we can define P' by

$$\begin{array}{ccc} M \times \mathbb{R}^m & \xrightarrow{P} & \mathbb{R} \\ J \times \text{id} \downarrow & \nearrow P' & \\ \mathfrak{g}^* \times \mathbb{R}^m & & \end{array}$$

and let P_1 be the restriction of P' to $M_1 \times \mathbb{R}^m$.

THEOREM 3.5. *Let x in M be such that $J(x)$ belongs to M_1 . The initialized Hamiltonian system (Σ, x) corresponding to the function P projects down on M_1 through J , defining a reachable Hamiltonian system $(\Sigma_1, J(x))$ corresponding to the function P_1 , with dynamics given by the coadjoint action of G . Σ_1 is quasi-minimal if it is strongly accessible.*

Proof. We first remark Σ_1 is well defined: if $P(\cdot, u)$ is constant on the fibres of J so is $\partial/\partial u P(\cdot, u)$. Defining J' as the restriction of J to M' , the orbit of D through x , we have $P = (J', \text{id})^* P_1$ and $X_{P(\cdot, u)} = J'_* X_{P(\cdot, u)}$ in view of Theorem 3.4.

If we define y_1 from $\partial/\partial u P$ as we have defined P_1 from P , it is trivial to verify that $y_1 = \partial/\partial u P_1$, therefore Σ_1 is a Hamiltonian system, the other properties following directly from the construction. Note that for analytic systems $\dim P_0 = n$ is equivalent to strong accessibility [21]. \square

We consider now a reachable but not strongly accessible Hamiltonian system Σ defined on an orbit M of the coadjoint action of G on \mathfrak{g}^* . The action of G on M is symplectic, and has a moment given by the inclusion $i: M \rightarrow \mathfrak{g}^*$; the submanifold W considered in the previous section is the orbit of G_0 , the Lie subgroup of G corresponding to the Lie subalgebra T_0 of T isomorphic to a certain subalgebra \mathfrak{g}_0 of \mathfrak{g} with codimension one.

Considering $\mathfrak{g}^* = \mathbb{R}^* \times \mathfrak{g}_0^*$ and $p_0: \mathfrak{g}^* \rightarrow \mathfrak{g}_0^*$ the canonical projection, let i_0 be the map $i_0: M \rightarrow \mathfrak{g}_0^*$ defined by $i_0 = p_0 \circ i$.

G_0 acts by symplectomorphisms on M , and the action has a moment map given by i_0 ; therefore if $x \in W$, $i_0(W)$ is the orbit W_2 of $i_0(x)$ under the coadjoint action of B_0 on \mathfrak{g}_0^* .

If we consider the map $\pi_2: W \rightarrow W_2$, restriction of i_0 to W , we see π_2 is an analytic submersion, since W_2 can be interpreted as a maximal integral submanifold of an analytic distribution on \mathfrak{g}_0^* [6].

LEMMA 3.6. *The kernel of π_{2*} is the w -orthogonal of TW .*

Proof. Identifying the tangent space at a point of \mathfrak{g}^* with \mathfrak{g}^* , we can represent a vector in $T_x W$ by $\{x, v\}$, with v in \mathfrak{g} , and $\{\cdot, \cdot\}$ defined by the relation $\{x, v\} \cdot v' = x \cdot [v, v']$ for any $v \in \mathfrak{g}$ [2].

Now, if $\{x, v\}$ is in the w -orthogonal of TW at x and $\{x, v'\}$ belongs to TW , we have $w(\{x, v\}, \{x, v'\}) = x \cdot [v, v'] = 0$; as $\{x, v'\}$ belonging to TW is the same as v' belonging to \mathfrak{g}_0^* , that can be written as $\{x, v\} \cdot v' = 0$ for any v' in \mathfrak{g}_0^* and, from the definition of π_2 , $\{x, v\}$ belongs to the kernel of π_{2*} . \square

As in the previous section, we can take the system as being defined on $\mathbb{R} \times W$ and pull-back everything by π .

THEOREM 3.7. *If Σ is reachable but not strongly accessible, it has a quasi-minimal realization which is the suspension of a time-dependent Hamiltonian system on an orbit of the coadjoint action of G_0 on \mathfrak{g}_0^* , with dynamics given by the suspension of that action.*

Proof. Taking into account the previous lemma we obtain a quasi-minimal realization Σ_2 from Σ by projecting on $\mathbb{R} \times W_2$ through $(\text{id}, \pi^* \pi_2)$. The resulting system is the suspension of a time-dependent Hamiltonian system since this was the construction already used in Theorem 2.5.

Now if g is an element of G we can write it as a product $g_1 \circ \dots \circ g_k$ with g_i the flow of an associated vector field X^i during the time t_i ; let $t' = t_1 + \dots + t_k$.

If (t, x) belongs to $\mathbb{R} \times W$ then $g(t, x) = (t + t', g'x)$ where g' is an element of G_0 given by $g' = (\partial/\partial t)_{-t'} \circ g$; in the projection on $\mathbb{R} \times W_2$ the action of g becomes $(t, x_2) \rightarrow (t + t', \text{Ad}_g^* x_2)$ where Ad^* refers to the coadjoint action of G_0 ; but this is exactly what we mean by the suspension of that action. \square

Now we turn to the Kostant-Kirillov-Souriau theorem.

THEOREM 3.8 [17], [9]. *If a Lie group G acts transitively on a symplectic manifold M , and the action is Poisson, then M is a covering space of the corresponding orbit W of the coadjoint action of G .*

Proof. The action induces a Lie algebra T of Hamiltonian vector fields in M isomorphic to \mathfrak{g} , and we consider a complete Hamiltonian control system Σ for which the set D of associated vector fields generates T , with the corresponding Hamiltonians as output maps: if v_1, \dots, v_k is a basis of \mathfrak{g} and X^i their respective Hamiltonian vector fields in M , $X^i \lrcorner w = df_i$, we define Σ from the map $P(x, u) = u_1 f_1(x) + \dots + u_k f_k(x)$.

The system Σ is reachable and strongly accessible, therefore quasi-minimal; if Σ' is the minimal realization of Σ defined on the state space M' , from [20] we know the projection $p: M \rightarrow M'$ is a covering projection, since Σ satisfies the observability rank condition.

If Σ_1 is the realization on W equivalent to Σ , constructed by using the moment map, the projection $p_1: W \rightarrow M'$ is also a covering projection by exactly the same argument as before, and we have the following commutative diagram:

$$\begin{array}{ccc} M & \xrightarrow{\quad} & M' \\ & \searrow J & \nearrow p_1 \\ & W & \end{array}$$

From [18, p. 64] J is a covering projection (being obviously surjective), therefore M is a covering space of W . \square

4. Equivalence of Hamiltonian systems. If Σ and Σ' are minimal (quasi-minimal) realizations and F is a homomorphism between them, F is a diffeomorphism (local diffeomorphism), smooth or analytic according to the class of the given systems [20], [4].

For Hamiltonian systems we can also prove:

THEOREM 4.1. *Let Σ and Σ' be two Hamiltonian systems with outputs, Σ reachable, and $F: M \rightarrow M'$ a homomorphism; then F is a (smooth, analytic) symplectomorphism.*

Proof. We need to prove that $F^*w_2 = w_1$ or equivalently $w' = F^*w_2 - w_1 = 0$; we know [20] that F is smooth or analytic, according to the class we are working in, and it has also been proved in [3] that if X is in D_1 , the family of associated vector fields of Σ_1 , then $w'(X, \cdot) = dS$ for some function S independent of X .

Clearly $L_X S = 0$, since $dS \cdot X = w'(X, X) = 0$, and therefore S is constant along trajectories of vector fields in D_1 ; as Σ_1 is reachable, S is constant on M_1 and $dS = 0$.

Let g_1 belong to G_1 . By Lemma 5(ii) in [15] there exists g_2 in G_2 such that $F \circ g_1 = g_2 \circ F$; then

$$g_1^* F^* w_2 = (F \circ g_1)^* w_2 = (g_2 \circ F)^* w_2 = F^* g_2^* w_2 = F^* w_2,$$

and also

$$g_1^* w' = g_1^* (F^* w_2 - w_1) = F^* w_2 - w_1 = w'.$$

Now if X belongs to D_1 , we have

$$g_1^* X \lrcorner w' = g_1^{-1*} (X \lrcorner g_1^* w') = g_1^{-1*} (X \lrcorner w') = 0$$

and since Σ_1 is reachable the vectors of the form $g_1^* X$ span the tangent space of M_1 at every point, thus $w' = 0$. \square

Now let Σ be the suspension of a time-dependent Hamiltonian system on M ; then $\mathbb{R} \times M$ has a contact structure (a closed two-form of maximal rank) w defined by $w = p^* w$, where w is the symplectic form on M and p is the projection $\mathbb{R} \times M \rightarrow M$. The analogue of a symplectomorphism for contact structures is a canonical transformation [1, p. 384], a map $\mathbb{R} \times M_1 \rightarrow \mathbb{R} \times M_2$ satisfying:

- (i) F is a diffeomorphism.

(ii) The diagram below is commutative:

$$\begin{array}{ccc} \mathbb{R} \times M_1 & \xrightarrow{F} & \mathbb{R} \times M_2 \\ & \searrow \pi_1 & \swarrow \pi_2 \\ & \mathbb{R} & \end{array}$$

(iii) There exists a (smooth, analytic) map $K: \mathbb{R} \times M_1 \rightarrow \mathbb{R}$ such that $F^* \underline{w}_2 = \underline{w}_1 + dK \wedge dt$.

We can now state an analogue of the previous theorem:

THEOREM 4.2. *Let Σ_1, Σ_2 be quasi-minimal realizations as in Theorem 3.8; if the homomorphism F between them is the suspension of a diffeomorphism $\Phi: M_1 \rightarrow M_2$, then F is a canonical transformation and Φ is a symplectomorphism.*

Proof. It is clear that F satisfies (i) and (ii), and we only have to prove (iii).

As in the previous proof we can obtain, for X in D_1 , $\underline{w}' = F^* \underline{w}_2 - \underline{w}_1$ that $X \lrcorner \underline{w}'(t, x) = d_x K(t, x)$ for some function on $\mathbb{R} \times M_1$ independent of X , and thus $L_X K = \partial/\partial t K$ as $X \lrcorner d_x K = \underline{w}'(X, X) = 0$; in particular this means K is independent of x in M_1 , and $X \lrcorner \underline{w}' = 0$.

If g_1 and g_2 are corresponding elements in G_1 and G_2 respectively we have $g_1^* X \lrcorner \underline{w}' = g_1^{-1*} (X \lrcorner g_1^* \underline{w}')$, and from [1] also:

$$\begin{aligned} g_1^* \underline{w}_1 &= g_1^* p_1^* \underline{w}_1 = (p_1 \circ g_1)^* \underline{w}_1 = \underline{w}_1 - dP_1 \wedge dt_1, \\ g_1^* F^* \underline{w}_2 &= F^* g_2^* p_2^* \underline{w}_2 = F^* (\underline{w}_2 - dP_2 \wedge dt_2) \end{aligned}$$

where dP_1 and dP_2 are taken at the value of u , supposed constant, corresponding to g_1 and g_2 .

Then, as $F^* dP_2 = dP_1 + \partial/\partial t K dt_1$ and $F^* dt_2 = dt_1$, we have $g_1^* \underline{w}' = \underline{w}'$ and therefore $g_1^* X \lrcorner \underline{w}' = 0$. It is easy to see, by induction, that this remains true if g_1 corresponds to a piecewise-constant control, because then g_1 is a finite product of elements of G_1 of the type just considered. We can conclude that $\underline{w}' = 0$ as in the previous proof and this means F is a canonical transformation. \square

Acknowledgments. I would like to thank P. Crouch for his many suggestions and the encouragement he gave me, and Professor Krener for his advice; P. Barros contributed also with a very valuable suggestion.

Part of this work was done at the Control Theory Centre of the University of Warwick, and the continued financial support of the Calouste Gulbenkian Foundation has been invaluable.

REFERENCES

- [1] R. ABRAHAM AND J. MARSDEN, *Foundations of Mechanics*, 2nd ed., W. A. Benjamin, Reading, MA, 1978.
- [2] V. ARNOLD, *Méthodes mathématiques de la mécanique classique*, Editions MIR, Moscou, 1974.
- [3] J. BASTO GONÇALVES, *Equivalence of gradient systems*, Control Theory Centre Report no. 84, University of Warwick, 1980.
- [4] ———, *Nonlinear observability and duality*, Systems Control Lett., 4 (1984), pp. 97–101.
- [5] R. BROCKETT, *Control theory and analytic dynamics*, in Geometric Control Theory, Math. Sci. Press, 1977.
- [6] C. CHEVALLEY, *Lie Groups*, Princeton Univ. Press, Princeton, NJ, 1946.
- [7] P. CROUCH AND M. IRVING, *Dynamical realizations of homogeneous Hamiltonian systems*, Control Theory Centre Report no. 122, University of Warwick, 1984.
- [8] D. ELLIOTT AND N. KALOUPSIDIS, *Accessibility properties of smooth nonlinear control systems*, in Geometric Control Theory, Math. Sci. Press, 1977.

- [9] V. GUILLEMIN AND S. STERNBERG, *Geometric Asymptotics*, American Mathematical Society, Providence, RI, 1977.
- [10] A. KIRILLOV, *Elements de la théorie des représentations*, Editions MIR, Moscou, 1974.
- [11] R. HERMANN AND A. KRENER, *Nonlinear controllability and observability*, IEEE Trans. Automat. Control, AC-22 (1977), pp. 728-740.
- [12] R. PALAIS, *A Global Formulation of the Lie Theory of Transformation Groups*, American Mathematical Society, Providence, RI, 1957.
- [13] A. VAN DER SCHAFT, *Hamiltonian dynamics with external forces and observations*, Math. Systems Theory, 15 (1982), pp. 145-168.
- [14] ———, *Observability and controllability for smooth nonlinear systems*, this Journal, 20 (1982), pp. 338-354.
- [15] ———, *Controllability and observability for affine nonlinear Hamiltonian systems*, IEEE Trans. Automat. Control, AC-27 (1982), pp. 490-492.
- [16] ———, *System theoretic descriptions of physical systems*, CWI Tract no. 3, CWI, Amsterdam, 1984.
- [17] J. SOURIAU, *Structure des systèmes dynamiques*, Dunod, Paris, 1970.
- [18] E. SPANIER, *Algebraic Topology*, McGraw-Hill, New York, 1966.
- [19] H. SUSSMANN, *Orbits of families of vector fields and integrability of distributions*, Trans. Amer. Math. Soc., 180 (1973), pp. 171-188.
- [20] ———, *Existence and uniqueness of minimal realizations of nonlinear systems*, Math. Systems Theory, 10 (1977), pp. 263-284.
- [21] H. SUSSMANN AND V. JURDJEVIC, *Controllability of nonlinear systems*, J. Differential Equations, 12 (1972), pp. 95-116.
- [22] A. WEINSTEIN, *Lectures on Symplectic Manifolds*, CBMS Regional Conference Series in Applied Mathematics, American Mathematical Society, Providence, RI, 1977.

DISTRIBUTED ASYNCHRONOUS RELAXATION METHODS FOR CONVEX NETWORK FLOW PROBLEMS*

DIMITRI P. BERTSEKAS† AND DIDIER EL BAZ‡

Abstract. We consider the solution of the single commodity strictly convex network flow problem in a distributed asynchronous computation environment. The dual of this problem is unconstrained, differentiable, and well suited for solution via Gauss-Seidel relaxation. We show that the structure of the dual allows the successful application of a distributed asynchronous method whereby relaxation iterations are carried out in parallel by several processors in arbitrary order and with arbitrarily large interprocessor communication delays.

Key words. parallel computation, distributed algorithms, network flows, asynchronous relaxation, coordinate descent

1. Introduction. Consider a directed graph with set of nodes N and set of arcs A . Each arc (i, j) has associated with it a cost function $g_{ij}: R \rightarrow (-\infty, +\infty]$. We denote by f_{ij} the flow of the arc (i, j) and consider the problem of minimizing total cost subject to a conservation of flow constraint at each node

$$(1) \quad \begin{aligned} &\text{minimize} \quad \sum_{(i,j) \in A} g_{ij}(f_{ij}) \\ &\text{subject to} \quad \sum_{(m,i) \in A} f_{mi} - \sum_{(i,j) \in A} f_{ij} = 0 \quad \forall i \in N. \end{aligned}$$

We assume that problem (1) has at least one feasible solution. We also make the following standing assumptions on g_{ij} :

- (a) g_{ij} is strictly convex, and lower semicontinuous;
- (b) the conjugate convex function of g_{ij} , defined by

$$(2) \quad g_{ij}^*(t_{ij}) = \sup_{f_{ij}} \{t_{ij}f_{ij} - g_{ij}(f_{ij})\},$$

is real valued, i.e. $-\infty < g_{ij}^*(t_{ij}) < \infty$ for all real t_{ij} . (Because of the strict convexity assumed in (a) above, g_{ij}^* is also continuously differentiable and its gradient denoted $\nabla g_{ij}^*(t_{ij})$ is the unique f_{ij} attaining the supremum in (2) (see [7, pp. 218, 253]).)

It is easily seen from (2) that assumption (b) implies that $\lim_{|f_{ij}| \rightarrow \infty} g_{ij}(f_{ij}) = \infty$. Therefore the objective function of the primal problem (1) has bounded level sets [7, § 8]. It follows that there exists an optimal solution for problem (1) which must be unique in view of the strict convexity assumed in (a).

The problem above is of great practical interest and has been studied for a long time. Except for strict convexity our assumptions are not overly restrictive. For example they are satisfied in the following two cases:

- 1) The *constrained case* where g_{ij} is of the form

$$(3) \quad g_{ij}(f_{ij}) = \begin{cases} \infty & \text{if } f_{ij} \notin [l_{ij}, c_{ij}], \\ \hat{g}_{ij}(f_{ij}) & \text{otherwise,} \end{cases}$$

* Received by the editors November 5, 1984; accepted for publication (in revised form) October 25, 1985. This work was supported by the National Science Foundation under contract NSF ECS-8217668 and the Defense Advanced Research Projects Agency under contract ONR-N00014-84-K-0357.

† Laboratory for Information and Decision Systems, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139.

‡ This author is on leave from LAAS, Toulouse, France. The work of this author was supported by an Institut National de Recherche en Informatique et en Automatique grant.

where l_{ij} and c_{ij} are given lower and upper bounds on the arc flow, and \hat{g}_{ij} is a strictly convex, real valued function defined on the real line R .

2) The *unconstrained case* where g_{ij} is strictly convex, real valued and its right and left derivatives g_{ij}^+ and g_{ij}^- satisfy

$$(4) \quad \lim_{f_{ij} \rightarrow \infty} g_{ij}^+(f_{ij}) = \infty, \quad \lim_{f_{ij} \rightarrow -\infty} g_{ij}^-(f_{ij}) = -\infty.$$

A dual problem for (1) is given by

$$(5) \quad \begin{aligned} &\text{minimize } q(p) \\ &\text{subject to no constraints on the vector } p = \{p_i | i \in N\}, \end{aligned}$$

where q is the dual functional given by

$$(6) \quad q(p) = \sum_{(i,j) \in A} g_{ij}^*(p_i - p_j).$$

We refer to p as a *price vector* and its components p_i as *prices*. The i th price is really a Lagrange multiplier associated with the i th conservation of flow constraint. The duality between problems (1) and (5) is well known and is explored in great detail in the recent book by Rockafellar [1]. The earlier book by Rockafellar [7] gives the necessary and sufficient condition for optimality of a pair (f, p) . A feasible flow vector $f = \{f_{ij} | (i, j) \in A\}$ is optimal for (1) and a price vector $p = \{p_i | i \in N\}$ is optimal for (5) if and only if for all arcs (i, j) [7, pp. 337-338]

$$p_i - p_j \text{ is a subgradient of } g_{ij} \text{ at } f_{ij}.$$

An equivalent condition is

$$(7) \quad f_{ij} = \nabla g_{ij}^*(p_i - p_j) \quad \forall (i, j) \in A.$$

Any one of these equivalent relations is referred to as the *complementary slackness condition*, and is shown in Fig. 1.

Since the dual problem is unconstrained and differentiable it is natural to consider algorithmic solution by a descent iterative method. The Gauss-Seidel relaxation method is particularly interesting in this respect since it admits a simple implementation. Given a price vector p , a node i is selected and its price p_i is changed (relaxed) to a value \hat{p}_i such that

$$(8) \quad \sum_{(m,i) \in A} \nabla g_{mi}^*(p_m - \hat{p}_i) = \sum_{(i,j) \in A} \nabla g_{ij}^*(\hat{p}_i - p_j).$$

It is easily seen (compare with the definition (6) of the dual cost q) that this equation is equivalent to $\partial q / \partial p_i = 0$, so the dual cost is minimized at \hat{p}_i with respect to the i th price, all other prices being kept constant. The algorithm proceeds by relaxing the prices of all nodes in cyclic order and repeating the process. The convergence of this algorithm does not follow immediately from standard results on relaxation methods [2], [3], [4] since these results require some assumption that is akin to strict convexity of the dual objective function which does not hold here (for a counterexample, see Powell [5]). However Cottle and Pang [6] have shown convergence of a network algorithm based on relaxation. It applies to transportation problems with quadratic cost function, and involves certain restrictions in the way relaxation is carried out. Their result is substantially extended in Bertsekas, Hosein and Tseng [19].

Our main objective in this paper is to explore the convergence properties of distributed versions of the relaxation method just described. Here we assume that each price p_i is under the control of a separate processor who changes p_i to \hat{p}_i on the basis

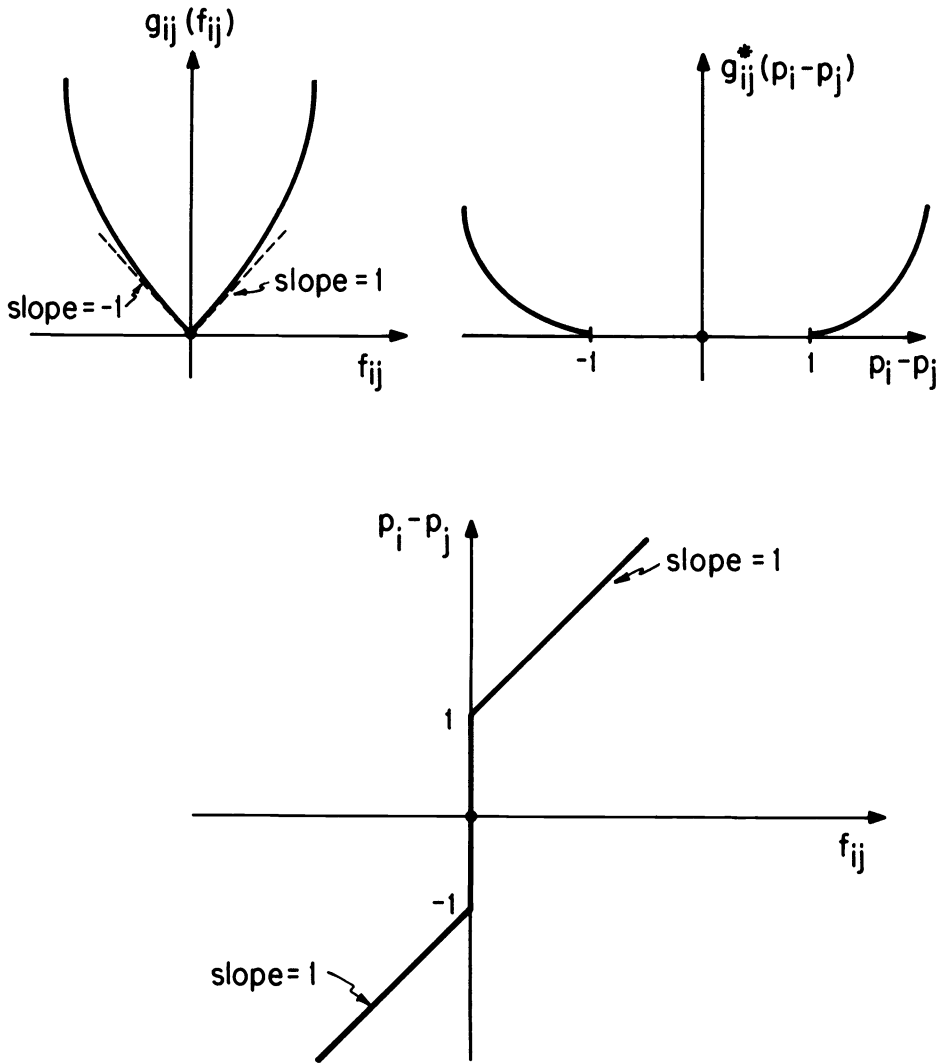


FIG. 1. Complementary slackness condition diagram for cost function $g_{ij}(f_{ij}) = |f_{ij}| + \frac{1}{2}(f_{ij})^2$.

of (8) and communicates the new value to the other processors. One can consider a parallel computation procedure carried out in an orderly manner whereby all processors exchange their current prices before carrying out their relaxation iteration. Mathematically this would be equivalent to a Jacobi type of relaxation procedure. We would like to consider, however, a much more general procedure whereby the communication between processors is not regular, and the information available at some processors regarding prices of other processors may be arbitrarily out-of-date. In addition we allow some processors to iterate more frequently than others. Models of such asynchronous algorithms have been formulated some time ago and by now there is considerable understanding of their convergence properties (see [8]–[16]; [17] is a survey). It turns out that the dual problem (5) has structure that allows us to show that the asynchronous relaxation method has satisfactory convergence properties. This is particularly true when the dual problem (5) has an essentially unique optimal solution. Otherwise satisfactory convergence depends on the starting point. These results are all

new and are shown in § 3. The next section analyzes the structure of the dual solution set and provides some preliminary analysis.

The results of this paper carry over verbatim to the case where the conservation of flow constraint has the form

$$\sum_{(m,i) \in A} f_{mi} - \sum_{(i,j) \in A} f_{ij} = b_i \quad \forall i \in N$$

where b_i are given scalars with $\sum_{i \in N} b_i = 0$. The dual cost of (6) must then include the term $\sum_{i \in N} b_i p_i$, and the relaxation equation (8) must include an additional term b_i in its right side. This extension is important from the practical point of view, but we have restricted attention to the case where $b_i = 0, \forall i \in N$ in order to simplify notation.

The results of this paper can also be extended in a simple manner to network problems with positive gains and strictly convex arc costs. This extension was mentioned to us by P. Tseng who also showed [20] two additional interesting facts. First that Proposition 2 holds even if the strict convexity assumption of (a) is removed thereby including the important class of linear minimum cost flow problems. Second that, within the class of monotropic programming problems, the largest class for which the monotonicity property of Proposition 1 holds is the class of network flow problems with positive gains.

Our notational conventions are that a subscript denotes a node or processor index, and a superscript denotes a time or iteration index. All vector inequalities should be interpreted in a coordinatewise sense. In order to simplify notation we have implicitly assumed that there is at most one arc associated with any ordered pair of nodes i and j , so that the arc notation (i, j) has a unambiguous meaning. However this assumption is not essential to any of our results.

2. Structure of the optimal dual solution set. Our standing assumptions, (a) and (b), guarantee that the primal problem (1) has a unique optimal solution. Existence of an optimal solution of the dual problem can be guaranteed under an additional (mild) regular feasibility assumption in which case the existence theorem of [1, p. 360] applies. On the other hand the optimal solution of the dual problem is never unique since adding the same constant to all coordinates of a price vector p leaves the dual cost unaffected. We can remove this degree of freedom by constraining the price of one node, say node N , to be zero. (With slight abuse of notation we number nodes as $1, 2, \dots, N$.) Thus we consider the *reduced dual optimal solution set* P^* defined by

$$(9) \quad P^* = \{p^* \mid q(p^*) = \min_p q(p), p_N^* = 0\}$$

where q is the dual objective function

$$(10) \quad q(p) = \sum_{(i,j) \in A} g_{ij}^*(p_i - p_j).$$

For the most part of the paper, we will operate under the following assumption.

Assumption 1. The reduced dual optimal solution set P^* is nonempty and compact.

Assumption 1 is not overly restrictive. For example let $\{f_{ij}^* \mid (i, j) \in A\}$ be the unique primal optimal solution, and consider the set of arcs

$$(11) \quad \hat{A} = \{(i, j) \mid f_{ij}^* \text{ lies in the interior of the set } \{f_{ij} \mid g_{ij}(f_{ij}) < \infty\}\}.$$

Then Assumption 1 is satisfied if the subgraph (N, \hat{A}) is connected. To see this note that for all arcs $(i, j) \in \hat{A}$ we have a bounded set of subgradients of g_{ij} at f_{ij}^* thereby implying a bounded set of price differences $p_i - p_j$ corresponding to dual optimal solutions [cf. (6)]. Note that in the unconstrained case mentioned in the previous

section every arc belongs to \hat{A} ; so, if the original graph is connected, Assumption 1 is satisfied. The constrained case of the previous section can be converted to the unconstrained case by replacing constraints by nondifferentiable penalty functions (see [18, § 5.5]). For example, assuming a dual optimal solution exists, a constraint $f_{ij} \geq 0$ can be eliminated by adding to the cost g_{ij} a penalty $c \max\{0, -f_{ij}\} + [\max\{0, -f_{ij}\}]^2$ with c positive and sufficiently large.

Consider now the set

$$(12) \quad P = \{p \mid p_N = 0\}$$

and for $i = 1, \dots, N-1$, the point-to-set mapping R_i which assigns to a price vector $p \in P$ the set of all prices \hat{p}_i that minimize the dual cost along the i th price starting from p , i.e. (cf. (8))

$$(13) \quad R_i(p) = \left\{ \hat{p}_i \mid \sum_m \nabla g_{mi}^*(p_m - \hat{p}_i) = \sum_j \nabla g_{ij}^*(\hat{p}_i - p_j) \right\}.$$

It is well known that a real valued convex function having one compact level set, has all its level sets compact [7, p. 70]. Therefore under Assumption 1 the sets $R_i(p)$, $p \in P$ are all nonempty, compact intervals. It follows that under Assumption 1 the (point-to-point) mappings

$$(14) \quad \bar{R}_i(p) = \max_{\hat{p}_i \in R_i(p)} \hat{p}_i,$$

$$(15) \quad \underline{R}_i(p) = \min_{\hat{p}_i \in R_i(p)} \hat{p}_i,$$

are well defined on the set P . We call \bar{R}_i (\underline{R}_i) the i th *maximal (minimal) relaxation mapping*. It gives the maximal (minimal) minimizing point of the dual cost along the i th coordinate starting from its argument. The point-to-set mapping R_i is called the *i*th *relaxation mapping*.

Some key facts are given in the following proposition.

PROPOSITION 1. *Let Assumption 1 hold. The mappings \bar{R}_i and \underline{R}_i are continuous on P . They are also monotone on P in the sense that for any $p, p' \in P$, $i = 1, \dots, N-1$ we have*

$$(16) \quad \bar{R}_i(p) \leq \bar{R}_i(p') \quad \text{if } p \leq p',$$

$$(17) \quad \underline{R}_i(p) \leq \underline{R}_i(p') \quad \text{if } p \leq p'.$$

Proof. To show continuity of \bar{R}_i we argue by contradiction. Suppose there exists a convergent price vector sequence $p^k \rightarrow p$ such that the corresponding sequence $\{\bar{R}_i(p^k)\}$ does not converge to $\bar{R}_i(p)$. By passing to a subsequence if necessary suppose that for some $\delta > 0$ we have

$$(18) \quad \bar{R}_i(p) \geq \bar{R}_i(p^k) + \delta \quad \forall k$$

(the proof is very similar if $\delta < 0$ and the inequality is reversed). By the definition of \bar{R}_i we have

$$(19) \quad \sum_m \nabla g_{mi}^*(p_m - \bar{R}_i(p)) = \sum_j \nabla g_{ij}^*(\bar{R}_i(p) - p_j),$$

$$(20) \quad \sum_m \nabla g_{mi}^*(p_m^k - \bar{R}_i(p^k)) = \sum_j \nabla g_{ij}^*(\bar{R}_i(p^k) - p_j^k) \quad \forall k.$$

Since $p^k \rightarrow p$ it follows using (18) that for sufficiently large k we have

$$\begin{aligned} p_m^k - \bar{R}_i(p^k) &> p_m - \bar{R}_i(p) \quad \forall (m, i) \in A, \\ \bar{R}_i(p^k) - p_j^k &< \bar{R}_i(p) - p_j \quad \forall (i, j) \in A. \end{aligned}$$

Therefore for sufficiently large k we have using the convexity of g_{mi}^* , g_{ij}^*

$$\begin{aligned}\nabla g_{mi}^*(p_m^k - \bar{R}_i(p^k)) &\geq \nabla g_{mi}^*(p_m - \bar{R}_i(p)) \quad \forall (m, i) \in A, \\ \nabla g_{ij}^*(\bar{R}_i(p^k) - p_j^k) &\leq \nabla g_{ij}^*(\bar{R}_i(p) - p_j) \quad \forall (i, j) \in A.\end{aligned}$$

Using these relations together with (19), (20) we obtain for all sufficiently large k

$$(21) \quad f_{mi} \triangleq \nabla g_{mi}^*(p_m - \bar{R}_i(p)) = \nabla g_{mi}^*(p_m^k - \bar{R}_i(p^k)) \quad \forall (m, i) \in A,$$

$$(22) \quad f_{ij} \triangleq \nabla g_{ij}^*(\bar{R}_i(p) - p_j) = \nabla g_{ij}^*(\bar{R}_i(p^k) - p_j^k) \quad \forall (i, j) \in A.$$

Consider the intervals I_{mi} and I_{ij} given by

$$\begin{aligned}I_{mi} &= \{t \mid \nabla g_{mi}^*(t) = f_{mi}\} \quad \forall (m, i) \in A, \\ I_{ij} &= \{t \mid \nabla g_{ij}^*(t) = f_{ij}\} \quad \forall (i, j) \in A.\end{aligned}$$

For k sufficiently large so that (21), (22) hold we have

$$\begin{aligned}\bar{R}_i(p) &= \max \{\hat{p}_i \mid \hat{p}_i \in p_m - I_{mi}, (m, i) \in A, \hat{p}_i \in I_{ij} - p_j, (i, j) \in A\}, \\ \bar{R}_i(p^k) &= \max \{\hat{p}_i \mid \hat{p}_i \in p_m^k - I_{mi}, (m, i) \in A, \hat{p}_i \in I_{ij} - p_j^k, (i, j) \in A\}.\end{aligned}$$

Since $p^k \rightarrow p$, it is evident from these relations that $\bar{R}_i(p^k) \rightarrow \bar{R}_i(p)$ thereby contradicting (18).

To show monotonicity of \bar{R}_i we again argue by contradiction. Suppose there exist p and p' such that $p'_j \geq p_j$, $\forall j = 1, \dots, N-1$ but $\bar{R}_i(p) > \bar{R}_i(p')$. It follows then that

$$\begin{aligned}p'_m - \bar{R}_i(p') &> p_m - \bar{R}_i(p) \quad \forall (m, i) \in A, \\ \bar{R}_i(p') - p'_j &< \bar{R}_i(p) - p_j \quad \forall (i, j) \in A.\end{aligned}$$

Therefore

$$(23) \quad \nabla g_{mi}^*(p'_m - \bar{R}_i(p')) \geq \nabla g_{mi}^*(p_m - \bar{R}_i(p)) \quad \forall (m, i) \in A,$$

$$(24) \quad \nabla g_{ij}^*(\bar{R}_i(p') - p'_j) \leq \nabla g_{ij}^*(\bar{R}_i(p) - p_j) \quad \forall (i, j) \in A.$$

Since by definition we have

$$(25) \quad \sum_m \nabla g_{mi}^*(p'_m - \bar{R}_i(p')) = \sum_i \nabla g_{ij}^*(\bar{R}_i(p') - p'_j),$$

$$(26) \quad \sum_m \nabla g_{mi}^*(p_m - \bar{R}_i(p)) = \sum_j \nabla g_{ij}^*(\bar{R}_i(p) - p_j),$$

it follows that equality holds in (23), (24), i.e.

$$\begin{aligned}f_{mi} &\triangleq \nabla g_{mi}^*(p'_m - \bar{R}_i(p')) = \nabla g_{mi}^*(p_m - \bar{R}_i(p)), \\ f_{ij} &\triangleq \nabla g_{ij}^*(\bar{R}_i(p') - p'_j) = \nabla g_{ij}^*(\bar{R}_i(p) - p_j).\end{aligned}$$

Consider the intervals

$$I_{mi} = \{t \mid \nabla g_{mi}^*(t) = f_{mi}\}, \quad I_{ij} = \{t \mid \nabla g_{ij}^*(t) = f_{ij}\},$$

and let

$$\delta = \bar{R}_i(p) - \bar{R}_i(p').$$

We have for all $(m, i) \in A$

$$p'_m - \bar{R}_i(p') \in I_{mi}, \quad p_m - \bar{R}_i(p) \in I_{mi}$$

and since $p_m \leq p'_m$ we obtain

$$p_m - \bar{R}_i(p) \leq p'_m - \bar{R}_i(p') - \delta \leq p'_m - \bar{R}_i(p').$$

Therefore

$$p'_m - \bar{R}_i(p') - \delta \in I_{mi} \quad \forall (m, i) \in A$$

and similarly

$$\bar{R}_i(p') + \delta - p_j \in I_{ij} \quad \forall (i, j) \in A.$$

It follows that

$$\bar{R}_i(p') + \delta \in R_i(p')$$

thereby contradicting the maximal nature of \bar{R}_i [cf. (14)].

The proof of continuity and monotonicity of \bar{R}_i is analogous with the one just given for \bar{R}_i and is omitted. Q.E.D.

The monotonicity and continuity of the mappings \bar{R}_i and \bar{R}_i imply a thus far unreported and somewhat surprising property of the optimal dual solution set.

PROPOSITION 2. *Let Assumption 1 hold. There exist a maximal and a minimal optimal solution of the dual problem, i.e. there exist $\bar{p} \in P^*$ and $\underline{p} \in P^*$ such that*

$$(27) \quad \underline{p} \leq p \leq \bar{p} \quad \forall p \in P^*.$$

Proof. Since P^* is nonempty and compact it contains a noninferior element \bar{p} for which there is no vector $p \in P^*$ such that $p \neq \bar{p}$ and $p_i \geq \bar{p}_i$ for all i . From the definition of \bar{R}_i and the optimality of \bar{p} we have $\bar{p}_i \leq \bar{R}_i(\bar{p})$ for all i . Furthermore for all i the vector $(\bar{p}_1, \dots, \bar{p}_{i-1}, R_i(\bar{p}), \bar{p}_{i+1}, \dots, \bar{p}_N)$ belongs to P^* so from noninferiority of \bar{p} it follows that $\bar{R}_i(\bar{p}) \leq \bar{p}_i$. Therefore we have $\bar{p}_i = \bar{R}_i(\bar{p})$ for all i . Let now \tilde{p} be a price vector obtained from \bar{p} according to

$$\tilde{p}_i = \begin{cases} \bar{p}_i + \delta, & i = 1, \dots, N-1, \\ 0, & i = N, \end{cases}$$

where $\delta > 0$ is sufficiently large so that

$$(28) \quad \tilde{p} \geq p \quad \forall p \in P^*.$$

It is easily seen that we have $\bar{R}_i(\tilde{p}) \leq \tilde{p}_i$, for all i so, using the monotonicity of \bar{R}_i shown in Proposition 1, we obtain

$$(29) \quad \bar{p} \leq \bar{R}^{k+1}(\tilde{p}) \leq \bar{R}^k(\tilde{p}) \quad \forall k$$

where $\bar{R}: R^{N-1} \rightarrow R^{N-1}$ is the mapping

$$(30) \quad \bar{R}(p) = [\bar{R}_1(p), \dots, \bar{R}_{N-1}(p)]$$

and \bar{R}^k is the composition of \bar{R} with itself k times. From (29) we see that the sequence $\bar{R}^k(\tilde{p})$ converges to some \hat{p} and by continuity of \bar{R} we must have $\hat{p} = \bar{R}(\hat{p})$ as well as $\hat{p} \geq \bar{p}$. Since $\hat{p} = \bar{R}(\hat{p})$ implies that $\hat{p} \in P^*$ it follows from the choice of \bar{p} that $\hat{p} = \bar{p}$. Also from (28), (29) and the fact $p \leq \bar{R}(p)$ for all $p \in P^*$ we obtain $\hat{p} = \bar{p} \geq p$ for all $p \in P^*$ which shows that \bar{p} is a maximal element of P^* . The proof for existence of a minimal element \underline{p} is entirely similar. Q.E.D.

3. Convergence analysis of asynchronous relaxation. The model of distributed asynchronous computation we adopt is described in [11], [12]. With each node $i = 1, \dots, N-1$ we associate a processor that computes from time to time some element

of $R_i(p)$ (here p is the latest price vector available to processor i), and sets the price p_i to this element. This price is then communicated at some later time to all other processors. Computation and communication at the various processors need not be synchronized. The precise model is as follows.

At each time instant, node i can be in one of three possible states *compute*, *transmit*, or *idle*. In the compute state node i computes a new price p_i . In the transmit state node i communicates the price p_i obtained from its own latest computation to one or more nodes m ($m \neq i$). In the idle state node i does nothing related to the solution of the problem.

We assume that computation and transmission for each node takes place in time intervals $[t_1, t_2]$ with $t_1 < t_2$, but do not exclude the possibility that a node may be simultaneously transmitting to more than one node nor do we assume that the transmission intervals to these nodes have the same origin and/or termination. We also make no assumptions on the length, timing and sequencing of computation and transmission intervals other than the following.

Assumption 2. For every node i and time $t \geq 0$ there exists a time $t' > t$ such that $[t, t']$ contains at least one computation interval for i and at least one transmission interval from i to each node m such that $(m, i) \in A$ or $(i, m) \in A$.

Assumption 2 is very natural. It states in essence that no node “drops out of the algorithm” permanently—perhaps due to a hardware failure. Without this assumption there is hardly anything we can hope to prove.

Each node i has a buffer B_{im} for each $m \neq i$ where it stores the latest transmission from m , as well as a buffer B_{ii} where it stores its own price estimate p_i . The contents for each buffer B_{im} at time t are denoted $p_m^t(i)$. Thus $p_m^t(i)$ is, for every t, i and m an estimate of the price p_m available at node i at time t . It is important to realize in what follows that *the buffer contents $p_m^t(i)$, and $p_m^t(i')$ at two different nodes i and i' need not coincide at all times. If $i \neq m$ and $i' \neq m$ the buffer contents $p_m^t(i)$, and $p_m^t(i')$ need not coincide at any time t .* The vector of all buffer contents of node i is denoted $p'(i)$, i.e.,

$$p'(i) = \{p_m^t(i) \mid m = 1, \dots, N-1\}.$$

The rules according to which the buffer contents $p_m^t(i)$ are updated are as follows:

(1) If $[t_1, t_2]$ is a transmission interval from node m to node i , the contents of the buffer B_{im} at time t_1 are transmitted and entered in the buffer B_{im} at time t_2 , i.e.

$$(31) \quad p_m^{t_2}(i) = p_m^{t_1}(m).$$

(2) If $[t_1, t_2]$ is a computation interval for node i , the content of the buffer B_{ii} is replaced at time t_2 with an element of $R_i(p^t(i))$, i.e.

$$(32) \quad p_i^{t_2}(i) \in R_i(p^t(i)).$$

(3) The contents of a buffer B_{ii} can change only at the end of a computation interval for node i . The contents of a buffer B_{im} , $i \neq m$ can change only at the end of a transmission interval from m to i .

The algorithm based on (32) will be called *Asynchronous Relaxation Method (ARM)*.

Our objective is to derive conditions under which limit points of the sequences $\{p^t(i)\}$ are optimal solutions of the dual problem (5). The following proposition is our main result. The proof is based on a general convergence theorem given in [12] (see also [17]) and applicable to asynchronous iterative algorithms such as the one just described. The key property that makes asynchronous convergence possible is the monotonicity of the mappings \bar{R}_i and \underline{R}_i shown in Proposition 1. This property is also

present in dynamic programming models and has been similarly exploited to show the validity of asynchronous versions of the successive approximation method [11].

PROPOSITION 3. *Let Assumptions 1 and 2 hold. For any initial buffer contents $p^0(i) \in P$, $i = 1, \dots, N-1$, each limit point of the sequences $\{p'(i)\}$ generated by the ARM belongs to the set*

$$(33) \quad \bar{P} = \{p \mid p \leq p \leq \bar{p}\}$$

where \bar{p} and p are the maximal and minimal dual optimal solutions. In particular, if the reduced dual optimal solution set P^* consists of a unique vector p^* we have

$$(34) \quad \lim_{i \rightarrow \infty} p'(i) = p^*, \quad i = 1, \dots, N-1.$$

Proof. Let $p, \tilde{p} \in P$ be price vectors such that

$$p \leq p^0(i) \leq \tilde{p} \quad \forall i = 1, \dots, N-1$$

and such that

$$\begin{aligned} p &\leq \underline{R}(p) \leq p \leq \bar{p} \leq \bar{R}(\tilde{p}) \leq \tilde{p}, \\ \lim_{k \rightarrow \infty} \underline{R}^k(p) &= \underline{p}, \quad \lim_{k \rightarrow \infty} \bar{R}^k(\tilde{p}) = \bar{p}. \end{aligned}$$

(The existence of such vectors was established in the proof of Proposition 2.) Consider the sets

$$(35) \quad \bar{P}^k = \{p \mid \underline{R}^k(p) \leq p \leq \bar{R}^k(\tilde{p})\}, \quad k = 1, 2, \dots.$$

Note that the sequence $\{\bar{P}^k\}$ is nested and that the common intersection of the sequence is the set \bar{P} of (33).

We will apply now a convergence theorem given in [12, § 3] (or [17, Prop. 3.1]). According to this theorem the desired result will be proved if the following three conditions are satisfied. (Rather than consulting the references just cited, the reader may wish to think through the proof of this since it is rather simple.)

(a) If $p \in \bar{P}^k$ then for every i the vector p' with coordinates

$$p'_j = \begin{cases} p_j & \text{if } j \neq i, \\ R_i(p) & \text{if } j = i \end{cases}$$

(cf. equation (32) associated with computation at node i) also belongs to \bar{P}^k .

(b) If $p \in \bar{P}^k$ and $\hat{p} \in \bar{P}^k$ then, for every i and m , the vector p' with coordinates

$$p'_j = \begin{cases} p_j & \text{if } j \neq m, \\ \hat{p}_m & \text{if } j = m \end{cases}$$

(cf. equation (31) associated with transmission from node m to node i) also belongs to \bar{P}^k .

(c) If $p(1), \dots, p(N-1)$ belong to \bar{P}^k then the vector p' with coordinates

$$\begin{aligned} p'_j &= R_j(p(j)), \quad j = 1, \dots, N-1, \\ p'_N &= 0 \end{aligned}$$

(cf. a computation (32) at each node followed by a transmission to every other node) belongs to \bar{P}^{k+1} .

It is easily seen that all the conditions stated above are satisfied in our case so the desired conclusion follows. Q.E.D.

Proposition 3 shows that the ARM has satisfactory convergence when P^* has a unique element. One way to guarantee this is to consider the optimal solution f^* of the primal problem (1) and the set of arcs

$$\tilde{A} = \{(i, j) \in A \mid g_{ij} \text{ is differentiable at } f_{ij}^*\}.$$

Then, if the graph (N, \tilde{A}) is connected, P^* consists of a unique point in view of the complementary slackness condition (7). In order to improve the convergence properties when P^* has more than one point it is necessary to modify the ARM so that a computation at node i replaces p_i with $\bar{R}_i(p)$ (not just any element of $R_i(p)$). We call this the *maximal* ARM. If in place of $\bar{R}_i(p)$ we use $\underline{R}_i(p)$ the resulting method is called the *minimal* ARM.

PROPOSITION 4. *Let Assumptions 1 and 2 hold. Assume that the starting buffer contents satisfy*

$$(36) \quad p^0(i) \geq \bar{p} \quad \forall i = 1, \dots, N-1.$$

Then if $\{p'(i)\}$ is generated by the maximal ARM we have

$$(37) \quad \lim_{t \rightarrow \infty} p^t(i) = \bar{p}, \quad i = 1, \dots, N-1.$$

Proof. The proof is identical to the one of Proposition 3 except that the set \bar{P}^k of (35) should be replaced by

$$\bar{P}^k = \{p \mid \bar{p} \leq p \leq \bar{R}^k(\bar{p})\}. \quad \text{Q.E.D.}$$

There is a similar result for the minimal ARM whereby \bar{p} is replaced by \underline{p} and condition (36) is replaced by $p^0(i) \leq \underline{p}$ for all i . The following example demonstrates that the results of Proposition 3 and 4 cannot be improved.

Example. Consider the 3-node network shown in Fig. 2. The arc costs are

$$g_{12}(f_{12}) = (f_{12})^2, \quad g_{23}(f_{23}) = |f_{23}| + (f_{23})^2, \quad g_{31}(f_{31}) = |f_{31}| + (f_{31})^2$$

and the optimal primal solution is

$$f_{12}^* = f_{23}^* = f_{31}^* = 0.$$

The reduced dual optimal solution set is derived from condition (7) and is given by

$$P^* = \{p \mid p_3 = 0, p_1 = p_2, -1 \leq p_1 \leq 1, -1 \leq p_2 \leq 1\}.$$

The results of Proposition 3 and 4 are illustrated in Fig. 3. To see that the ARM as well as the maximal and minimal ARM may not converge to a dual optimal solution, let the buffer contents of processors 1 and 2 be both equal to $(-1, 1)$ and let both processors update the respective price coordinates and then exchange the results of the computation. Then the buffer contents will be $(1, -1)$, and by repeating this process one more time the buffer contents will become again $(-1, 1)$ thereby completing a cycle. Therefore in general we cannot expect convergence of the ARM to the optimal solution set if the latter contains more than one element. Similarly the maximal and

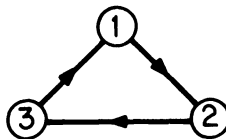


FIG. 2

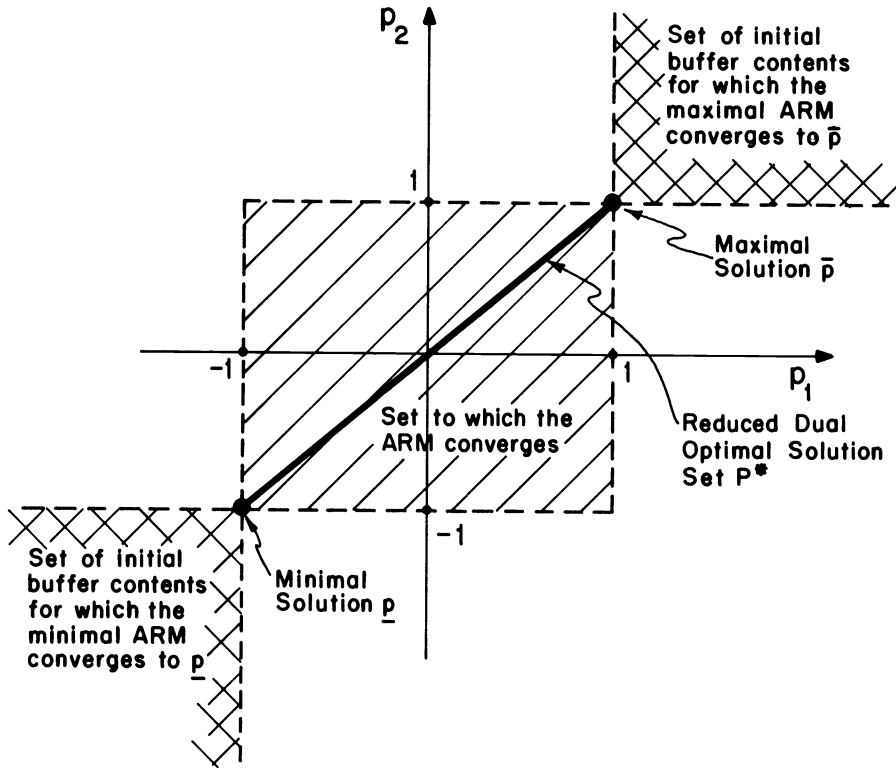


FIG. 3. Structure of the optimal solution set, and convergence regions of the ARM, the maximal ARM, and the minimal ARM.

minimal ARM need not converge to \bar{p} and \tilde{p} respectively if the initial buffer contents do not belong to the appropriate regions [cf. (36)]. Note that this counterexample applies also to a synchronous Jacobi method.

REFERENCES

- [1] R. T. ROCKAFELLAR, *Network Flows and Monotropic Optimization*, John Wiley, New York, 1984.
- [2] W. J. ZANGWILL, *Nonlinear Programming*, Prentice-Hall, Englewood Cliffs, NJ, 1969.
- [3] R. W. H. SARGENT AND D. J. SEBASTIAN, *On the convergence of sequential minimization algorithms*, J. Optim. Theory and Appl., 12 (1973), pp. 567-575.
- [4] E. POLAK, *Computational Methods in Optimization: A Unified Approach*, Academic Press, New York, 1971.
- [5] M. J. D. POWELL, *On search directions for minimization algorithms*, Math. Programming, 4 (1973), pp. 193-201.
- [6] R. W. COTTLE AND J. S. PANG, *On the convergence of a block successive overrelaxation method for a class of linear complementarity problems*, Math. Programming Study, 17 (1982), pp. 126-138.
- [7] R. T. ROCKAFELLAR, *Convex Analysis*, Princeton University Press, Princeton, NJ, 1970.
- [8] D. CHAZAN AND W. MIRANKER, *Chaotic relaxation*, Linear Algebra Appl., 2 (1969), pp. 199-222.
- [9] J. C. MIELLOU, *Iterations Chaotiques a Retards, Etude de la Convergence dans le Cas d'Espaces Partiellement Ordonnes*, Comptes Rendus de l'Academie des Sciences, Paris, Serie A, 280 (1975), pp. 233-236.
- [10] G. M. BAUDET, *Asynchronous iterative methods for multiprocessors*, J. Assoc. Comput. Mach., 2 (1978), pp. 226-244.
- [11] D. P. BERTSEKAS, *Distributed dynamic programming*, IEEE Trans. Automat. Control, AC-27 (1982), pp. 610-616.

- [12] D. P. BERTSEKAS, *Distributed asynchronous computation of fixed points*, Math. Programming, 27 (1983), pp. 107-120.
- [13] J. N. TSITSIKLIS, D. P. BERTSEKAS AND M. ATHANS, *Distributed asynchronous deterministic and stochastic gradient optimization algorithms*, IEEE Trans. Automat. Control (1986), to appear.
- [14] J. N. TSITSIKLIS, *Problems in decentralized decision making and computation*, Ph.D. thesis, Dept. of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge, MA, 1984.
- [15] D. EL BAZ, *Etude d'algorithmes iteratifs de calcul parallele application a la resolution distribuee du probleme du routage optimal dans un resau maille a commutation de paquets*, These de Docteur Ingenieur, Toulouse, 1984.
- [16] G. AUTHIE, J. BERNUSSOU AND D. EL BAZ, *Distributed asynchronous iterative control algorithms, optimal routing application*, IFAC Symposium: Components and Instruments for Distributed Control Systems, December 9-11, 1982, Paris.
- [17] D. P. BERTSEKAS, J. N. TSITSIKLIS AND M. ATHANS, *Convergence theories of distributed iterative processes: a survey*, Laboratory for Information and Decision Systems Report LIDS-P-1342, Mass. Inst. Tech., September 1984.
- [18] D. P. BERTSEKAS, *Constrained Optimization and Lagrange Multiplier Methods*, Academic Press, New York, 1982.
- [19] D. P. BERTSEKAS, P. HOSEIN AND P. TSENG, *Relaxation methods for network flow problems with convex arc costs*, LIDS Report P-1523, Mass. Inst. Tech., Dec. 1985.
- [20] P. TSENG, *Relaxation methods for monotropic programming problems*, Ph.D. thesis, Operation Research Center, Massachusetts Institute of Technology, Cambridge, MA, 1986.

MODULE THEORETIC ZERO STRUCTURES FOR SYSTEM MATRICES*

BOSTWICK F. WYMAN† AND MICHAEL K. SAIN‡

Abstract. The foundation for a coordinate-free theory of the poles of a linear dynamical system was laid in 1965 by the module-theoretic work of Kalman. However, although the theoretical and application importance of transfer function zeros for feedback system design was widely known for single input, single output systems by 1955, it remained for Rosenbrock in 1970 to propose the ideas of transmission zeros, input-decoupling zeros, and output-decoupling zeros for multi-input, multi-output systems, by means of matrix theoretic methods. A coordinate-free, module-theoretic treatment of transmission zeros for a multi-input, multi-output transfer function was given in 1981 by Wyman and Sain. This paper extends these coordinate-free, module-theoretic studies to include systems which need not be controllable or observable. Interpretation of the Rosenbrock system matrix is given on three levels: rational, finitely generated free-modular, and torsion divisible. On the second level, an Ω -Zero Module Z_Ω is defined and imbedded in a short exact sequence showing that the input-decoupling zero module is contained as a factor module in Z_Ω . On the third level, a Γ -Zero Module Z_Γ is defined and imbedded in a short exact sequence showing that the output-decoupling zero module is contained as a submodule in Z_Γ . Both structures are studied further in regard to transmission zero module information with emphasis on lumped zeros, and the cases of right and left invertible transfer functions are given in detail. Not surprisingly, these investigations can support considerable fine detail, which is in accord with the widely held belief that questions on the nature of multivariable zeros must be broadly based.

Key words. zeros, zero modules, system zeros

AMS(MOS) subject classifications. 93B25, 13C10

1. Introduction. This paper is a contribution to the algebraic theory of zeros of a linear multivariable system. The importance of zeros of transfer functions was well understood in classical design, reaching a high level of refinement in the well-known text of Truxal [27]. It can perhaps be said that the “modern era” of the theory of zeros of multivariable systems began with Rosenbrock [23]. In this work Rosenbrock defined the *transmission zeros*, *input-decoupling zeros* and *output-decoupling zeros* of a linear system. The transmission zeros describe properties of the input/output map associated with a system, and the decoupling zeros measure obstructions to the controllability and observability of the system itself.

Since Rosenbrock’s pioneering work, a number of authors have considered multivariable zeros from various points of view. Francis and Wonham [8] contains a summary and comparison of several different approaches, and another survey can be found in MacFarlane and Karcnias [19]. The book edited by Fallside [7] contains several papers dealing with both poles and zeros. An important dynamical interpretation appears in Desoer and Schulman [6]. There is a close connection between multivariable zeros and geometric control theory; see, for example, Wonham [30, Chap. 4]. Numerical computation of multivariable zeros has been treated in Davison and Wang [5], Laub and Moore [18] and Van Dooren [28]. Zeros of square systems are important in the work of Bart, et al. [2].

* Received by the editors July 3, 1984, and in revised form July 2, 1985. A brief preliminary report of some of the ideas, without proof, was presented in the Proceedings of the 20th Allerton Conference on Communication, Control, and Computing, which was supported by the National Aeronautics and Space Administration under grant NAG 2-34/35.

† Department of Mathematics, The Ohio State University, Columbus, Ohio 43210.

‡ Department of Electrical and Computer Engineering, University of Notre Dame, Notre Dame, Indiana 46556. The work of this author was supported by the National Science Foundation under grants ECS 81-02891 and ECS 84-05714.

Following Kalman's module-theoretic work on the poles of a linear system, the authors introduced the notion of the multivariable *zero module* to capture the structure of the transmission zeros of a system in an economical, coordinate-free way; see Wyman and Sain [32], [34]. The zero module has also been considered by Fuhrmann and Hautus [11] and Conte and Perdon [4]. All of this work has been restricted to the module-theoretic study of *transmission* zeros. Horan [14] contains module-theoretic studies of certain types of decoupling zeros.

The note by Fuhrmann and Hautus [11] is particularly important as an inspiration for the present paper. Namely these authors establish that if $G(s) = C(sI - A)^{-1}B$, and if (A, B, C) is controllable and observable, then the zero module of the Rosenbrock system matrix is isomorphic to the zero module of $G(s)$ itself. This result, together with the earlier work of Rosenbrock and others, suggests the task of examining the zero module structures associated with the system matrix in general.

The goal of the present paper is to study the overall zeros of a system which need not be controllable or observable. Particular attention needs to be paid to the possible interactions of transmission and decoupling zeros. In order to study the internal structure of a system, it is first necessary to choose a method of describing the system structure. For the purposes of this paper we deal with the classical representation $(A, B, C, D(s))$ leading to the (not necessarily proper) transfer function matrix $G(s) = C(sI - A)^{-1}B + D(s)$. See § 2 for more details.

The *system matrix*

$$\Sigma = \begin{bmatrix} sI - A & B \\ -C & D(s) \end{bmatrix}$$

introduced in Rosenbrock [23] combines the crucial data of a linear multivariable system into an extraordinarily convenient and useful package. The goal of this paper is to study the coordinate-free meaning of the system matrix and, in so doing, to obtain a unified description of the different sorts of zeros a system can have and of how they fit together.

From a commutative algebra point of view, the main idea is that the system matrix really defines three very different abstract functions. First, it defines a linear transformation between two vector spaces over the field of rational functions. In this form it contains zero information of the most coarse kind, namely information about the kernel and cokernel of the transfer function $G(s)$. The second appearance of the system matrix is as a map between two free modules, and in this case the *cokernel* is an appropriate zero module construction. Third, the system matrix defines a module homomorphism between two (infinitely generated) torsion divisible modules. In this case the *kernel* of the mapping contains the appropriate zero information.

Section 2 of this paper contains preliminaries and notation; and it gives a brief review of abstract realization theory. Section 3 contains some required material about modules. Section 4 introduces the system matrix, presents some basic definitions, and discusses the $k(s)$ -vector space theory. Section 5 discusses the divisible module case. Highlights are a discussion of a natural module theory setting for Rosenbrock's output decoupling zeros and a discussion of the Desoer-Schulman blocked transmission philosophy extending Wyman and Sain [34] to nonminimal systems. A general discussion of the zero-module associated to the free-module system matrix interpretation is given in § 6. This section also contains a detailed discussion of Rosenbrock's input decoupling zeros. Section 7 studies right-invertible systems from the free-module point of view, and § 8 discusses left-invertible systems from the torsion-divisible module point of view.

Prerequisites and general references. To understand the technical results of this paper, the reader should be familiar with the theory of principal ideal domains and the general tools of commutative diagrams and exact sequences. Kalman [16], [17] introduced modules into system theory, emphasizing the role of poles and realization theory. More recent expositions can be found in Sain [25, Chap. 7] and Fuhrmann [9], [12]. The study of multivariable zeros by module-theoretic tools was begun by Wyman and Sain [32], [34]. The present work can best be viewed as a continuation of these papers.

A thorough introduction to finitely generated modules (used for states and inputs) appears in Hartley and Hawkes [13], and divisible modules (used for outputs) are discussed in Sharpe and Vamos [26]. More advanced material on modules, with emphasis on diagrams and exact sequences, can be found in Atiyah and MacDonald [1].

The system matrix was introduced in Rosenbrock [23] which contained precise definitions of zeros from a polynomial “invariant-factor” point of view. The system matrix has figured extensively in the subsequent literature, including Morf [22], Molinari [21], Fuhrmann [10], Verghese, et al. [29], Rosenbrock [24], and Kailath [15, Chap. 8].

2. Systems, modules, and the realization diagram. Suppose that k is a field, and that X , U , and Y are finite dimensional vector spaces over k . Let $A: X \rightarrow X$, $B: U \rightarrow X$, $C: X \rightarrow Y$, and $D_i: U \rightarrow Y$, $i = 0, \dots, l$ be k -linear transformations. The algebraic results of this paper are motivated by the study of the continuous-time system

$$\begin{aligned}\dot{x} &= Ax + Bu, \\ y &= Cx + D_0u + D_1\dot{u} + \dots + D_lu^{(l)}\end{aligned}$$

(if k is the real or complex field), and the discrete-time system

$$\begin{aligned}x(t+1) &= Ax(t) + Bu(t), \\ y(t) &= Cx(t) + D_0u(t) + D_1u(t+1) + \dots + D_lu(t+l)\end{aligned}$$

over an arbitrary field k .

Let s be an indeterminate, and denote by $k[s]$ the ring of polynomials and by $k(s)$ the field of rational functions in s over k . We will describe the dynamical behavior of systems in terms of modules over $k[s]$ and vector spaces over $k(s)$. If V is a finite dimensional vector space over k , then $V(s)$ is the $k(s)$ -vector space defined abstractly as a tensor product $V(s) = V \otimes_k k(s)$, or concretely as column vectors $k(s)^n$ if V is k^n .

Each system gives a *transfer function* $G(s): U(s) \rightarrow Y(s)$ which is a $k(s)$ -linear transformation given by

$$G(s) = C(sI - A)^{-1}B + D(s),$$

where $D(s) = D_0 + D_1(s) + \dots + D_ls^l$ considered as a $k(s)$ -linear transformation which can be defined by matrices with polynomial entries.

A detailed algebraic study of systems begins with the introduction of three quite different types of $k[s]$ -modules, which we may call modules of “state-type,” “input-type,” and “output-type.”

State modules are finite-dimensional vector spaces X over k which are equipped with a k -linear map $A: X \rightarrow X$, so that a $k[s]$ -module action can be defined by $p(s)x = p(A)x$ for every polynomial $p(s)$ in $k[s]$ and every x in X . Conversely, if V is a $k[s]$ -module on a finite-dimensional vector space over k , then multiplication by s induces a k -linear transformation A on V . Modules of this kind are *finitely-generated* and *torsion*.

Input modules are finitely-generated free modules. If U is an m -dimensional space over k , we denote by ΩU the free $k[s]$ -module given by $\Omega U = U \otimes_k k[s] \cong k[s]^m$. This module can be used to study inputs, since any k -linear map $B: U \rightarrow X$ extends in a natural way to a $k[s]$ -module map $\tilde{B}: \Omega U \rightarrow X$ defined by

$$\tilde{B}(u_0 + u_1 s + \cdots + u_r s^r) = Bu_0 + ABu_1 + \cdots + A^r Bu_r.$$

The image $\tilde{B}(\Omega U)$ of this map is called the “controllability subspace” of X , and the pair (A, B) is called “controllable” if $\tilde{B}(\Omega U) = X$. In any case, the image is an A -invariant subspace (or a submodule) and the cokernel $X/\tilde{B}(\Omega U)$ is called the “module of input-decoupling zeros” of (A, B) .

Output modules are less familiar. If Y is a p -dimensional space over k , we denote by ΓY the factor module $Y(s)/\Omega Y$, which makes sense since ΩY is a $k[s]$ -submodule of the set $Y(s)$ of rational vectors. An element of ΓY can be thought of as an equivalence class containing a unique strictly proper representative

$$y = y_1 s^{-1} + y_2 s^{-2} + \dots + y_n s^{-n} + \dots, y_i \text{ in } Y.$$

If $[y]$ denotes the equivalence class in ΓY containing y , then the action of s is given by

$$s[\mathbf{v}] = [v_2s^{-1} + v_3s^{-2} + \dots + v_ns^{-n+1} + \dots].$$

To use the module ΓY to study outputs, first define $\pi: \Gamma Y \rightarrow Y$ by $\pi[y] = y_1$ (the coefficient of s^{-1}). Then any k -linear map $C: X \rightarrow Y$ extends naturally to a $k[s]$ -module map $\tilde{C}: X \rightarrow \Gamma Y$ such that $\pi\tilde{C} = C$. Namely, $\tilde{C}x$ is the equivalence class in ΓY of $C(sI - A)^{-1}x = \sum_{i=0}^{\infty} CA^i x s^{-i-1}$. The kernel of \tilde{C} is called the submodule of unobservable states, or the module of "output-decoupling-zeros." Note that ΓY is a torsion divisible module. However, ΓY is not finitely generated, and there is no single polynomial which annihilates every member of ΓY .

All of the modules and spaces discussed so far fit together in the fundamental commutative *realization diagram* (Fig. 1). Here we denote by $i: \Omega U \rightarrow U(s)$ the inclusion map, and by $p: Y(s) \rightarrow \Gamma Y$ the natural projection, and we define $G^*: \Omega U \rightarrow \Gamma Y$ by $G^* = pGi$.

3. Module structure and the snake lemma. The theory of finitely generated (f.g.) modules has been used in algebraic system theory for some time. Here we only recall

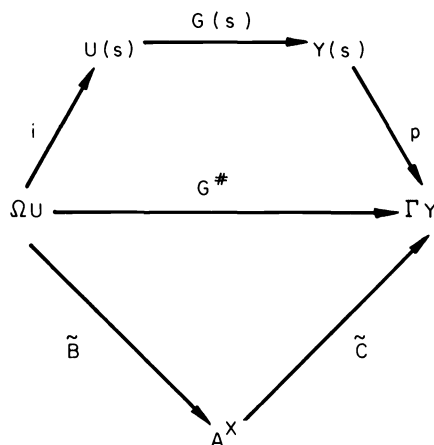


FIG. 1

that every f.g. $k[s]$ -module is isomorphic to a direct sum $M \cong F \oplus T$, where F is a free module and T is a finite dimensional torsion module (that is, of state type). A thorough treatment of the corresponding structure theory for output modules (which are not finitely generated) can be found in Sharpe and Vamos [26], but the present discussion together with Conte and Perdon [3] should be adequate for system-theoretic applications. We will be considering modules of the form ΓW (W a finite dimensional space over k) together with their submodules. A module of the form ΓW is torsion and divisible, but not f.g. A submodule M of ΓW is necessarily torsion but not necessarily divisible, and M may be f.g. or not. A single element m of M is called divisible (in M) if for any polynomial $p(s)$ the equation $p(s)m' = m$ can be solved for m' (in M ! We know it can be solved in ΓW). The set of all divisible elements of M form a submodule M_{div} , the maximal divisible submodule of M . A general result assures that M is isomorphic to a direct sum: $M \cong M_{\text{div}} \oplus M/M_{\text{div}}$ (since divisible modules over a polynomial ring are *injective*). The factor module M/M_{div} is not f.g. in general, but it will be in the system-theoretic applications.

We proceed to an abstract discussion of "polynomial matrices." Suppose given two vector spaces V and W and a $k[s]$ -module map $T_\Omega: \Omega V \rightarrow \Omega W$. Then T_Ω induces a $k(s)$ -linear map $T(s): V(s) \rightarrow W(s)$. If bases are chosen, T_Ω can be represented by a polynomial matrix, and the same matrix gives the $k(s)$ -linear map $T(s)$. Finally, the map $T(s)$ maps ΩV to ΩW (by construction), so it gives a map $T_\Gamma: \Gamma V \rightarrow \Gamma W$ on equivalence classes. The moral is that a polynomial matrix defines three mappings with three different pairs of domains and ranges. These three mappings, T_Ω , $T(s)$, and T_Γ have very different algebraic properties. They appear in Fig. 2 whose horizontal rows are exact sequences. The two center columns are short exact sequences.

The kernels and cokernels of these maps are related by the "Snake Lemma" which provides a six term exact sequence:

$$\begin{aligned} 0 \rightarrow \ker T_\Omega &\xrightarrow{\bar{i}_v} \ker T(s) \xrightarrow{\bar{\pi}_v} \ker T_\Gamma \xrightarrow{\Delta} \text{coker } T_\Omega \\ &\xrightarrow{\bar{i}_w} \text{coker } T(s) \xrightarrow{\bar{\pi}_w} \text{coker } T_\Gamma \rightarrow 0. \end{aligned}$$

$$\begin{array}{ccccccc} & & 0 & & 0 & & \\ & & \downarrow & & \downarrow & & \\ 0 & \longrightarrow & \ker T_\Omega & \longrightarrow & \Omega V & \xrightarrow{T_\Omega} & \Omega W \longrightarrow \text{coker } T_\Omega \longrightarrow 0 \\ & & \downarrow \bar{i}_v & & \downarrow i_v & & \downarrow i_w & & \downarrow \bar{i}_w \\ 0 & \longrightarrow & \ker T(s) & \longrightarrow & V(s) & \xrightarrow{T(s)} & W(s) \longrightarrow \text{coker } T(s) \longrightarrow 0 \\ & & \downarrow \bar{\pi}_v & & \downarrow \pi_v & & \downarrow \pi_w & & \downarrow \bar{\pi}_w \\ 0 & \longrightarrow & \ker T_\Gamma & \longrightarrow & \Gamma V & \xrightarrow{T_\Gamma} & \Gamma W \longrightarrow \text{coker } T_\Gamma \longrightarrow 0 \\ & & & & \downarrow & & \downarrow & & \\ & & & & 0 & & 0 & & \end{array}$$

FIG. 2

The proof, which can be found, for example, in Atiyah and MacDonald [1, p. 23], supplies explicit constructions for all the mappings.

4. The system matrix and the internal zeros of a system. Suppose given, as in § 2, a field k , vector spaces U , X , and Y , k -linear maps $A: X \rightarrow X$, $B: U \rightarrow X$, $C: X \rightarrow Y$, and a $k[s]$ -module map $D_\Omega: \Omega U \rightarrow \Omega Y$. The map D_Ω gives also a $k(s)$ -linear map $D(s): U(s) \rightarrow Y(s)$. We refer to these data as a system $\Sigma = (A, B, C, D(s))$ with transfer function $G(s) = C(sI - A)^{-1}B + D(s)$, $G(s): U(s) \rightarrow Y(s)$.

The (Rosenbrock) *System Matrix* of this system, written symbolically

$$\Sigma = \begin{bmatrix} sI - A & B \\ -C & D(s) \end{bmatrix}$$

is a $k(s)$ -linear map

$$\Sigma: X(s) \oplus U(s) \rightarrow X(s) \oplus Y(s).$$

Recalling that $D(s)$ comes from a polynomial matrix map D_Ω , we see that the system matrix also defines two $k[s]$ -module maps

$$\Sigma_\Omega: \Omega X \oplus \Omega U \rightarrow \Omega X \oplus \Omega Y,$$

$$\Sigma_\Gamma: \Gamma X \oplus \Gamma U \rightarrow \Gamma X \oplus \Gamma Y.$$

These maps are shown together in the exact commutative diagram of Fig. 3. These two mappings can be used to define two kinds of "internal zero module" for the given system.

MAIN DEFINITION. Let $\Sigma = (A, B, C, D(s))$ be a system with System Matrix Σ .

(a) The Ω -Zero Module of Σ is $Z_\Omega = \text{coker } \Sigma_\Omega$.

(b) The Γ -Zero Module of Σ is $Z_\Gamma = \ker \Sigma_\Gamma$.

According to the Snake Lemma, we have a six term exact sequence which relates these two modules:

$$0 \rightarrow \ker \Sigma_\Omega \rightarrow \ker \Sigma \rightarrow \ker \Sigma_\Gamma \xrightarrow{\Delta} \text{coker } \Sigma_\Omega \rightarrow \text{coker } \Sigma \rightarrow \text{coker } \Sigma_\Gamma \rightarrow 0.$$

As a corollary of the Snake Lemma, we can write

$$Z_\Gamma = (\text{torsion divisible module}) \oplus M_1.$$

$$Z_\Omega = (\text{f.g. free module}) \oplus M_2.$$

Here M_1 and M_2 are isomorphic f.g. torsion modules (that is, of state-type) which should be thought of as the "lumped" zeros of the system. The other components

$$\begin{array}{ccccccccc} 0 & \longrightarrow & \ker \Sigma_\Omega & \longrightarrow & \Omega X \oplus \Omega U & \xrightarrow{\Sigma_\Omega} & \Omega X \oplus \Omega Y & \longrightarrow & \text{coker } \Sigma_\Omega \longrightarrow 0 \\ & & & & \downarrow i_{XU} & & \downarrow i_{XY} & & \\ 0 & \longrightarrow & \ker \Sigma & \longrightarrow & X(s) \oplus U(s) & \xrightarrow{\Sigma} & X(s) \oplus Y(s) & \longrightarrow & \text{coker } \Sigma \longrightarrow 0 \\ & & & & \downarrow \pi_{XU} & & \downarrow \pi_{XY} & & \\ 0 & \longrightarrow & \ker \Sigma_\Gamma & \longrightarrow & \Gamma X \oplus \Gamma U & \xrightarrow{\Sigma_\Gamma} & \Gamma X \oplus \Gamma Y & \longrightarrow & \text{coker } \Sigma_\Gamma \longrightarrow 0 \end{array}$$

FIG. 3

represent different sorts of “generic zeros” and deserve a great deal of further study. In this section we characterize the rational vector spaces $\ker \Sigma$ and $\operatorname{coker} \Sigma$ as a first step in our analysis. In later sections we study the zero modules Z_Ω and Z_Γ .

THEOREM 1. *Suppose given a system $\Sigma = (A, B, C, D(s))$ with system matrix Σ and transfer function $G(s)$. Then there are $k(s)$ -vector space isomorphisms*

$$\ker \Sigma \cong \ker G(s),$$

$$\operatorname{coker} \Sigma \cong \operatorname{coker} G(s).$$

In particular, Σ is monic if and only if $G(s)$ is monic, and Σ is epic if and only if $G(s)$ is epic.

Proof. We define a mapping $\alpha: U(s) \rightarrow X(s) \oplus U(s)$ by $\alpha(u(s)) = (-(sI - A)^{-1}Bu(s), u(s))$. Then α is $k(s)$ -linear and monic, and a straightforward calculation shows that α on $\ker G(s)$ is epic on to $\ker \Sigma$.

To establish the isomorphism of cokernels, we first define a map

$$\beta: X(s) \oplus Y(s) \rightarrow Y(s)/G(s)U(s)$$

by $\beta(x(s), y(s)) \equiv y(s) + C(sI - A)^{-1}x(s) \pmod{G(s)U(s)}$. Since β is obviously epic, we must show that $\beta(x(s), y(s)) = 0$ if and only if $(x(s), y(s)) = \Sigma(x_0(s), u(s))$ for some $(x_0(s), u(s))$ in $X(s) \oplus U(s)$. We omit this calculation, which shows that the following sequence is exact:

$$X(s) \oplus U(s) \xrightarrow{\Sigma} X(s) \oplus Y(s) \xrightarrow{\beta} Y(s)/G(s)U(s) \rightarrow 0.$$

In other words, $\operatorname{coker} \Sigma \cong Y(s)/G(s)U(s) = \operatorname{coker} G(s)$.

5. Blocked transmissions and the Γ -zero module. In this section we relate the Γ -Zero Module to a set of strictly proper input signals which can be blocked by the output of a suitably chosen initial state. As a dividend, the output-decoupling zeros of the system are identified as a submodule of the Γ -zeros. In the case of minimal systems, the blocking result follows from Wyman and Sain [34].

The system $\Sigma = (A, B, C, D(s))$ gives rise to a $k(s)$ -linear transfer function $G(s): U(s) \rightarrow Y(s)$, and also (from Fig. 1) a $k[s]$ -module map $G^*: \Omega U \rightarrow \Gamma Y$. In this section we must consider a third version of the transfer function adapted to strictly proper input signals.

Consider the exact $k[s]$ -module sequence

$$0 \rightarrow \Omega U \rightarrow U(s) \xrightarrow{\pi_u} \Gamma U \rightarrow 0.$$

Let γ be an equivalence class in ΓU , with strictly proper representative u_γ . The map $\sigma: \Gamma U \rightarrow U(s)$ defined by $\sigma(\gamma) = u_\gamma$ is a k -linear map satisfying $\pi_u \circ \sigma = \operatorname{id}_{\Gamma U}$, but it is not $k[s]$ or $k(s)$ -linear. Nevertheless, we define the map $G_\# = \pi_y \circ G(s) \circ \sigma$, yielding the commutative diagram (Fig. 4):

$$\begin{array}{ccc} & U(s) & \xrightarrow{G(s)} & Y(s) \\ \sigma \nearrow & & & \searrow \pi_y \\ \Gamma U & \xrightarrow{G_\#} & & \Gamma Y \end{array}$$

FIG. 4

We say that a strictly proper input $[\mathbf{u}]$ can be *blocked* by the output of a state x_0 in X (compare Wyman and Sain [34] and Desoer and Schulman [6]) if $G_*[\mathbf{u}] + \tilde{C}x_0 = 0$, where $\tilde{C}: X \rightarrow \Gamma Y$ is the output map from Fig. 1. We denote the set of all blockable-signal equivalence classes by $Z_{\text{signal}} \subset \Gamma U$. That is,

$$Z_{\text{signal}} = \{\gamma: G_*\gamma + \tilde{C}x_0 = 0 \text{ for some } x_0 \text{ in } X\}.$$

Since G_* is only k -linear, a little work is required to show that Z_{signal} is a module. Calculations very similar to results in [34] establish that Z_{signal} is a $k[s]$ -submodule of ΓU .

Consider the projection map $p_2: \Gamma X \oplus \Gamma U \rightarrow \Gamma U$ defined by $p_2([\mathbf{x}], [\mathbf{u}]) = [\mathbf{u}]$. Recall that $Z_\Gamma = \ker \Sigma_\Gamma$ is a submodule of $\Gamma X \oplus \Gamma U$. We claim that p_2 maps Z_Γ onto $Z_{\text{signal}} \subset \Gamma U$. Suppose that $([\mathbf{x}], [\mathbf{u}])$ lies in $\ker \Sigma_\Gamma$, or that

$$\begin{bmatrix} sI - A & B \\ -C & D(s) \end{bmatrix} \begin{bmatrix} \mathbf{x} \\ \mathbf{u} \end{bmatrix} \equiv 0 \pmod{\Omega X \oplus \Omega Y}.$$

In particular there is a polynomial vector x_{poly} in ΩX such that $(sI - A)\mathbf{x} + B\mathbf{u} = x_{\text{poly}}$. In fact, a degree argument shows that $x_{\text{poly}} = x_0$ in X since \mathbf{x} and \mathbf{u} are strictly proper. Then $\mathbf{x} = -(sI - A)^{-1}B\mathbf{u} + (sI - A)^{-1}x_0$. Also,

$$\begin{aligned} -C\mathbf{x} + D(s)\mathbf{u} &= y_{\text{poly}}, \\ G(s)\mathbf{u} - C(sI - A)^{-1}x_0 &= y_{\text{poly}}, \\ G_*[\mathbf{u}] - \tilde{C}x_0 &\equiv 0 \pmod{\Omega Y} \end{aligned}$$

which shows that $[\mathbf{u}]$ lies in Z_{signal} .

Conversely, if γ in ΓU satisfies $G_*\gamma - \tilde{C}x_0 = 0$, write $\mathbf{u} = \sigma(\gamma)$ and let $\mathbf{x} = -(sI - A)^{-1}B\mathbf{u} + (sI - A)^{-1}x_0$. Then it follows that $([\mathbf{x}], \gamma)$ lies in $\ker \Sigma_\Gamma$ and $p_2([\mathbf{x}], \gamma) = \gamma$. That is, $p_2: \ker \Sigma_\Gamma \rightarrow Z_{\text{signal}}$ is a $k[s]$ -module epimorphism.

To identify the kernel of p_2 , suppose $p_2([\mathbf{x}], [\mathbf{u}]) = 0$ in ΓU . Then $[\mathbf{u}] = 0$ in ΓU , and

$$\begin{bmatrix} sI - A & B \\ -C & D(s) \end{bmatrix} \begin{bmatrix} \mathbf{x} \\ 0 \end{bmatrix} \equiv \begin{bmatrix} 0 \\ 0 \end{bmatrix} \pmod{\Omega X \oplus \Omega Y}.$$

From $(sI - A)[\mathbf{x}] = 0$ in ΓX , it follows as above that $(sI - A)\mathbf{x} = x_0$ for some state x_0 in X , or that $\mathbf{x} = (sI - A)^{-1}x_0$. Also $\tilde{C}x_0 = C(sI - A)^{-1}x_0 = 0$ in ΓY , so that x_0 is an *unobservable* state. That is

$$\ker p_2 = \{(\mathbf{x}, 0): \mathbf{x} = (sI - A)^{-1}x_0 \text{ and } \tilde{C}x_0 = 0\}.$$

Following Rosenbrock [23, p. 65] it is reasonable to call the set of unobservable states the *output-decoupling zero module*

$$Z_{\text{o.d.}} = \{x \text{ in } X: \tilde{C}x = 0\}.$$

To establish the connection between this module and Rosenbrock's definition, consider the polynomial matrix

$$\mathbf{C} = \begin{bmatrix} sI - A \\ C \end{bmatrix},$$

which defines two $k[s]$ -module maps $C_\Omega: \Omega X \rightarrow \Omega X \oplus \Omega Y$, and $C_\Gamma: \Gamma X \rightarrow \Gamma X \oplus \Gamma Y$. It is not difficult to prove that $\ker C_\Gamma \cong \ker \tilde{C}$ as $k[s]$ -modules.

Rosenbrock defined output decoupling zeros in terms of the invariant factors of \mathbf{C} . The usual view is that the invariant factors of \mathbf{C} describe the structure of the module $\text{coker } C_\Omega$. However, an application of the Snake Lemma shows immediately that $\ker C_\Gamma$

is isomorphic to the torsion submodule of $\text{coker } C_\Omega$. This justifies the name $Z_{\text{o.d.}}$ for $\ker \tilde{C}$, and allows us to summarize this section with the following theorem.

THEOREM 2. *Suppose given a system $\Sigma = (A, B, C, D(s))$ with transfer function $G(s) = C(sI - A)^{-1}B + D(s)$. Let $\tilde{C}: X \rightarrow \Gamma Y$ be the observability map, let $Z_{\text{o.d.}} = \ker \tilde{C}$ be the output-decoupling zero module. Let $Z_\Gamma = \ker \Sigma_\Gamma$ be the Γ -zero module defined by the system matrix Σ , and let Z_{signal} be the module of strictly proper input signals which can be blocked by states of X . Then there is an exact sequence*

$$0 \rightarrow Z_{\text{o.d.}} \rightarrow Z_\Gamma \rightarrow Z_{\text{signal}} \rightarrow 0.$$

In particular, if Σ is observable, then Z_Γ and Z_{signal} are isomorphic $k[s]$ -modules.

6. The Ω -zero module. We begin with a brief discussion of input-decoupling zeros (Rosenbrock [23, p. 64]). Suppose given $A: X \rightarrow X$ and $B: U \rightarrow X$ as usual. Consider the $k[s]$ -module map $[sI - A \ B]: \Omega X \oplus \Omega U \rightarrow \Omega X$ defined by

$$[sI - A \ B] \begin{bmatrix} x(s) \\ u(s) \end{bmatrix} = (sI - A)x(s) + Bu(s).$$

An exercise in diagram chasing establishes that the cokernel of $[sI - A \ B]$ is isomorphic to the cokernel of the controllability map $\tilde{B}: \Omega U \rightarrow X$, or that

$$\Omega X / [sI - A \ B](\Omega X \oplus \Omega U) \cong X / \tilde{B}(\Omega U).$$

Since Rosenbrock refers to the invariant factors of $[sI - A \ B]$ as “input-decoupling zeros” it is reasonable to call $\text{coker } [sI - A \ B]$ or, equivalently, $X / \tilde{B}(\Omega U)$, the *input-decoupling zero module* $Z_{\text{i.d.}}$ of the system.

The following theorem, whose proof parallels § 5 and is therefore omitted, summarizes the Ω -results.

THEOREM 3. *Suppose given a system $(A, B, C, D(s))$ with transfer function $G(s) = C(sI - A)^{-1}B + D(s)$. Let $\tilde{B}: \Omega U \rightarrow X$ be the reachability map, and consider $\ker \tilde{B} \subset \Omega U$ and $\text{coker } \tilde{B} = Z_{\text{i.d.}}$, the input-decoupling zero module. Let $Z_\Omega = \text{coker } \Sigma_\Omega$ be the Ω -Zero Module, defined using the system matrix Σ . Then there is an exact sequence of $k[s]$ -modules*

$$0 \rightarrow \Omega Y / G(s)(\ker \tilde{B}) \xrightarrow{i} Z_\Omega \xrightarrow{j} Z_{\text{i.d.}} \rightarrow 0$$

where i and j are induced from the maps

$$i(y) = \begin{bmatrix} 0 \\ y \end{bmatrix} \bmod \Sigma_\Omega(\Omega X \oplus \Omega U),$$

$$j \begin{bmatrix} x \\ y \end{bmatrix} = x \bmod \tilde{B}(\Omega U).$$

The first term of the exact sequence can be put in a more familiar form using coprime factorization theory. Suppose $G(s) = P(s)Q^{-1}(s)$ is a coprime factorization with $Q(s): \Omega U \rightarrow \Omega U$ and $P(s): \Omega U \rightarrow \Omega Y$, with $\ker \tilde{B} = Q(s)\Omega U$ and $G(s)(\ker \tilde{B}) = P(s)\Omega U \subset \Omega Y$. Then $\Omega Y / G(s)(\ker \tilde{B}) \cong \Omega Y / P(s)\Omega U$. For minimal systems, this calculation agrees with an earlier description of the (transmission) zero module given in Wyman and Sain [32] for transfer functions (and therefore minimal systems).

7. The Ω -zero module for right-invertible systems. Suppose the system $\Sigma = (A, B, C, D(s))$ has a right-invertible transfer function $G(s) = C(sI - A)^{-1}B + D(s)$. That is, $G(s)$ is epic as a $k(s)$ -linear transformation, so according to Theorem 1, $\text{coker } \Sigma = 0$, where Σ is the $k(s)$ -form of the system matrix, Z_Ω is a torsion module.

According to the exact sequence of Theorem 3, with $Z_1 = \Omega Y / G(s)(\ker \tilde{B})$,

$$0 \rightarrow Z_1 \xrightarrow{i} Z_\Omega \xrightarrow{j} Z_{i.d.} \rightarrow 0.$$

To study Z_1 , consider the transmission zeros of a transfer function $G(s)$ (Wyman and Sain [32])

$$Z(G) = \frac{G^{-1}(\Omega Y) + \Omega U}{\ker G + \Omega U}.$$

If $G(s)$ is right-invertible, an alternative description holds. There is an isomorphism of $k[s]$ -modules

$$Z(G) \cong \Omega Y / G(s)(\ker G^*)$$

where $G^* : \Omega U \rightarrow \Gamma Y$ is the restricted input/output map of realization theory as described in Fig. 1. To verify this isomorphism, consider the basic result

$$Z(G) \cong G^{-1}(\Omega Y) / G^{-1}(\Omega Y) \cap (\ker G + \Omega U).$$

Roughly speaking, the map $G : G^{-1}(\Omega Y) \rightarrow \Omega Y$ induces the required isomorphism: $Z(G) \rightarrow \Omega Y / G(s)(\ker G^*)$. The technical details are omitted.

Since $G^* = \tilde{C} \circ \tilde{B}$, from Fig. 1, it follows that $\ker \tilde{B} \subset \ker G^*$, giving a natural projection

$$\pi : \Omega Y / G(s)(\ker \tilde{B}) \rightarrow \Omega Y / G(s)(\ker G^*),$$

or, using earlier notation:

$$\pi : Z_1 \rightarrow Z(G).$$

If the system is observable, then \tilde{C} is monic and $\ker \tilde{B} = \ker G^*$, showing that π is the identity map in this case. We can summarize the work so far:

THEOREM 4. *Suppose $\Sigma = (A, B, C, D(s))$ is an observable system with right-invertible transfer function $G(s)$. Then there is an exact sequence*

$$0 \rightarrow Z(G) \rightarrow Z_\Omega \rightarrow Z_{i.d.} \rightarrow 0,$$

in which Z_Ω is the (finitely generated torsion) Ω -Zero Module, $Z(G)$ is the transmission zero module, and $Z_{i.d.}$ is the input decoupling zero module.

If the system is not observable, it will be necessary to examine the kernel of $\pi : Z_1 \rightarrow Z(G)$. Call the kernel Z_2 and consider the exact sequence

$$0 \rightarrow Z_2 \rightarrow Z_1 \xrightarrow{\pi} Z(G) \rightarrow 0.$$

It is easy to establish that

$$Z_2 \cong \frac{G(s)(\ker G^*)}{G(s)(\ker \tilde{B})}.$$

The study of Z_2 is technical, but if $\ker \tilde{C} \cap X_{rch}$ is the space of reachable, unobservable states, there is an epimorphism

$$\beta : \ker \tilde{C} \cap X_{rch} \rightarrow Z_2$$

defined by $\beta(x) \equiv G(s)u \pmod{G(s)(\ker \tilde{B})}$, whenever $x = \tilde{B}u$.

We do not proceed to compute $\ker \beta$ in general, although several examples are included at the end of this section to show that β is not necessarily monic. The results can be summarized in the following refinement of Theorem 4.

THEOREM 5. *Suppose a system is given with right invertible transfer function $G(s)$. Then Z_Ω is a torsion module, and there are two exact sequences:*

$$\begin{aligned} 0 \rightarrow Z_1 \rightarrow Z_\Omega \rightarrow Z_{i.d.} \rightarrow 0, \\ 0 \rightarrow Z_2 \rightarrow Z_1 \rightarrow Z(G) \rightarrow 0 \end{aligned}$$

and an epimorphism

$$\beta : Z_{o.d.} \cap X_{rch} \rightarrow Z_2 \rightarrow 0.$$

The map β is an isomorphism if $G(s)$ is a square invertible transfer function, but otherwise it may have a kernel.

Although Theorems 4 and 5 are quite technical, several intuitive conclusions can be drawn. If $G(s)$ is right invertible, then Z_Ω and the modules related to it have no free part, so they are intuitive system-theoretic objects. The module Z_Ω contains the input-decoupling zeros as a factor module, and it also contains the transmission zeros as a subfactor (that is, as a factor module of a submodule). This means in particular that all the numerical input decoupling and transmission zeros really correspond to rank drops of the system matrix. On the other hand, the output decoupling zeros affect Z_Ω in a more complicated way. Modes of the system which are both uncontrollable and unobservable occur in the Ω -Zero Module only once (by virtue of being input zeros). Finally, since the map β discussed above need not be monic, under some conditions output decoupling zeros do not appear as rank drops in the system matrix. This phenomenon is related to the kernel of $G(s)$ as a map of vector spaces.

The rest of this section consists of examples illustrating some of the phenomena.

Example 1. Consider the system (A, b, c^T) with

$$A = \begin{bmatrix} a_1 & 0 \\ 0 & a_2 \end{bmatrix}, \quad b = \begin{bmatrix} 0 \\ 1 \end{bmatrix}, \quad c^T = [0 \quad 1],$$

where $a_1 \neq a_2$. The corresponding transfer function is $g(s) = 1/(s - a_2)$, with transmission zero module $Z(g) = (0)$. Smith-form calculations give (up to isomorphism) $Z_\Omega \cong Z_{i.d.} \cong Z_{o.d.} \cong k[s]/(s - a_1)$. That is, the mode corresponding to a_1 is both unreachable and unobservable, but it is only counted once in Z_Ω .

Since $Z_{o.d.} \cap X_{rch} = (0)$, Theorem 5 implies that $Z_2 = (0)$, so that $Z_1 \cong Z(g) = (0)$, which implies that $j: Z_\Omega \rightarrow Z_{i.d.}$ is an isomorphism. This calculation is an explicit alternative to Smith form calculations.

Example 2. Consider the system (A, b, c^T) given by

$$A = \begin{bmatrix} -1 & 0 & 0 \\ 0 & -2 & 0 \\ 0 & 0 & -3 \end{bmatrix}, \quad b = \begin{bmatrix} 1 \\ 0 \\ 1 \end{bmatrix}, \quad c^T = [1 \quad 1 \quad 1].$$

The transfer function is

$$g(s) = \frac{2(s+2)}{(s+1)(s+3)}$$

and the transmission-zero module is

$$Z(g) \cong k[s]/(s+2).$$

The system is observable, but the mode with eigenvalue -2 is unreachable, so:

$$Z_{i.d.} \cong k[s]/(s+2); \quad Z_{o.d.} = (0).$$

Theorem 4 applies, so Z_Ω fits into the exact sequence

$$0 \rightarrow Z(g) \rightarrow Z_\Omega \rightarrow Z_{\text{i.d.}} \rightarrow 0.$$

From this information alone, there are just two possibilities for Z_Ω , namely

$$(a) \quad \text{“split”} \quad Z_\Omega \cong k[s]/(s+2) \oplus k[s]/(s+2),$$

or

$$(b) \quad \text{“not split”} \quad Z_\Omega \cong k[s]/(s+2)^2.$$

In this case, Z_Ω is not split. On the other hand, if the output map is changed to $C^T = (1, 0, 1)$, then the analysis is the same up to the exact sequence, but Z_Ω splits!

Example 3. This example shows that the map β of Theorem 5 is not necessarily monic. Consider (A, B, c^T) ,

$$A = \begin{bmatrix} a_1 & 0 \\ 0 & a_2 \end{bmatrix}, \quad B = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \quad c^T = [1 \quad 0]$$

with

$$G(s) = \begin{bmatrix} \frac{1}{s-a_1} & 0 \end{bmatrix},$$

$$Z(G) = (0),$$

$$Z_{\text{i.d.}} = (0),$$

$$Z_{\text{o.d.}} = k[s]/(s-a_2).$$

Direct calculation shows that $Z_\Omega = 0$, so (in the notation of Theorem 5) Z_1 and Z_2 are both zero. On the other hand, $X_{\text{rch}} = X$, so $Z_{\text{o.d.}} \cap X_{\text{rch}} = Z_{\text{o.d.}} \neq (0)$, and β is not monic.

8. The Γ -zero module for left-invertible systems. In this section we state results analogous to those in § 7.

Assume given a system $\Sigma = (A, B, C, D(s))$ which has a left-invertible transfer function $G(s)$. That is, $G(s)$ is monic, and according to § 4 so is the $k(s)$ -linear transformation corresponding to the system matrix Σ . It follows that the Γ -Zero Module Z_Γ is finitely generated and isomorphic to the torsion submodule of Z_Ω . From Theorem 2, § 5, we recall the exact sequence

$$0 \rightarrow Z_{\text{o.d.}} \rightarrow Z_\Gamma \rightarrow Z_{\text{signal}} \rightarrow 0.$$

The inclusion $Z_{\text{o.d.}} \rightarrow Z_\Gamma$ shows that Z_Γ contains information about the output-decoupling zeros. Define $\tilde{B}_1 = p \circ \tilde{B}$, for p the natural projection onto $X/\ker \tilde{C}$.

The analogue of Theorem 5 consists of two exact sequences.

THEOREM 6. *Suppose given a system with left-invertible transfer function $G(s)$. Then Z_Γ is a torsion module, and there are two exact sequences:*

$$0 \rightarrow Z_{\text{o.d.}} \rightarrow Z_\Gamma \rightarrow Z_{\text{signal}} \rightarrow 0,$$

$$0 \rightarrow Z(G) \rightarrow Z_{\text{signal}} \xrightarrow{\bar{\alpha}} X_{\text{obs}}/\tilde{B}_1(\Omega U).$$

If $G(s)$ is square and invertible, then $\bar{\alpha}$ is epic, but in general, $\bar{\alpha}$ may not be epic.

Example. Consider the system (A, b, C)

$$A = \begin{bmatrix} a_1 & 0 \\ 0 & a_2 \end{bmatrix}, \quad b = \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \quad C = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix},$$

with transfer function

$$G(s) = \begin{bmatrix} 1/(s - a_1) \\ 0 \end{bmatrix}$$

which is monic but not epic. It is easy to see that

$$Z(G) = Z_{o.d.} = (0),$$

$$Z_{i.d.} \cong k[s]/(s - a_2).$$

A Smith-form calculation gives

$$Z_\Omega \cong k[s],$$

a free module of rank 1. Since Z_Γ is isomorphic to the torsion submodule of Z_Ω , we get

$$Z_\Gamma = (0).$$

Also (see Theorem 6 above) $Z_{\text{signal}} = 0$, but $X_{\text{obs}}/\tilde{B}_1(\Omega U) \cong Z_{i.d.} \neq (0)$, so $\bar{\alpha}$ is not epic in this case.

9. Conclusions. The modern, coordinate-free approach to studying poles of linear dynamical systems was set into motion by Kalman [16]. Wyman and Sain [32], [34] proposed a module-theoretic treatment of transmission zeros for transfer functions in an analogous manner.

This paper extends these module theoretic studies to systems which need not be controllable or observable. The Rosenbrock system matrix is studied on three levels: rational, finitely generated free-modular, and torsion divisible. Two zero modules, Z_Ω and Z_Γ , are defined on the system level; and exact sequences are developed to relate them to modules of input-decoupling zeros and output-decoupling zeros. In the cases for which the system transfer function has a right or a left inverse, certain of the fine zero structures have been specially examined.

The modules Z_Ω and Z_Γ may be expected to capture the key concepts which may be advanced in a coordinate-free study of zero module structures for the Rosenbrock system matrix. Not surprisingly, these concepts admit considerable depth of analysis, which indicates that the notion of zero is a rich system theoretic area for research.

Finally, if $G(s) = P(s)D^{-1}(s)Q(s) + R(s)$ is a polynomial matrix description, then the corresponding system matrix is

$$\begin{bmatrix} D(s) & Q(s) \\ -P(s) & R(s) \end{bmatrix}.$$

Most of the present paper can be extended immediately to this case.

REFERENCES

- [1] M. F. ATIYAH AND I. G. MACDONALD, *Introduction to Commutative Algebra*, Addison-Wesley, Reading, MA, 1969.
- [2] H. BART, I. GOHBERG AND M. A. KAASHOEK, *Minimal Factorization of Matrix and Operator Functions*, Birkhauser, Basel, Switzerland, 1979.
- [3] G. CONTE AND A. M. PERDON, *On polynomial matrices and finitely generated torsion $K[z]$ -modules*, in *Algebraic and Geometric Methods in Linear System Theory*, C. I. Byrnes and C. Martin, eds., American Mathematical Society, Lectures in Applied Mathematics, Vol. 18, Providence, RI, 1980.
- [4] ———, *Infinite pole module and infinite zero module*, in *Proceedings VII International Conference on Analysis and Optimization of Systems*, Nizza, 1984.
- [5] E. J. DAVISON AND S. H. WANG, *Properties and calculation of transmission zeros of linear multivariable systems*, *Automatica*, 10 (1974), pp. 643–658.

- [6] C. A. DESOER AND J. D. SCHULMAN, *Zeros and poles of matrix transfer functions and their dynamical interpretation*, IEEE Trans. Circuits and Systems, CAS-21 (1974), pp. 3-8.
- [7] F. FALLSIDE, ed., *Control System Design by Pole-Zero Assignment*, Academic Press, New York, 1977.
- [8] B. A. FRANCIS AND W. M. WONHAM, *The role of transmission zeros in linear multivariable regulators*, Internat. J. Control, 22, No. 5 (1975), pp. 657-681.
- [9] P. A. FUHRMANN, *Algebraic system theory: an analyst's point of view*, J. Franklin Inst., 301 (1976), pp. 521-540.
- [10] ———, *On strict system equivalence and similarity*, Internat. J. Control, 25 (1977), pp. 5-10.
- [11] P. A. FUHRMANN AND M. L. J. HAUTUS, *On the zero module of rational matrix functions*, Proceedings 19th IEEE Conference on Decision and Control, IEEE, New York, 1980, pp. 256-257.
- [12] P. A. FUHRMANN, *Linear Systems and Operators in Hilbert Space*, McGraw-Hill, New York, 1981, Chapter 1.
- [13] B. HARTLEY AND T. O. HAWKES, *Rings, Modules, and Linear Algebra*, Chapman and Hall, London, 1970.
- [14] R. E. HORAN, *On the decoupling zeros and poles of a system*, IEEE Trans. Automat. Control, AC-25, No. 3 (June 1980), pp. 517-521.
- [15] T. KAILATH, *Linear Systems*, Prentice Hall, Englewood Cliffs, NJ, 1980.
- [16] R. E. KALMAN, *Algebraic structure of linear dynamical systems. I. The module of Σ* , in Proceedings National Academy of Science, Vol. 54, 1965, pp. 1503-1508.
- [17] R. E. KALMAN, P. FALB AND M. ARBIB, *Topics in Mathematical System Theory*, McGraw-Hill, New York, 1969, Chapter 10.
- [18] A. J. LAUB AND B. MOORE, *Calculation of transmission zeros using QZ techniques*, Automatica, 14 (1978), pp. 557-566.
- [19] A. G. J. MACFARLANE AND N. KARCANIAS, *Poles and zeros of linear multivariable systems: A survey of the algebraic, geometric and complex variable theory*, Internat. J. Control, 24 (1976), pp. 33-74.
- [20] S. MACLANE, *Homology*, Springer, Berlin, 1963, Chapter V.
- [21] B. P. MOLINARI, *Zeros of the system matrix*, IEEE Trans. Automat. Control, AC-21 (October 1976), pp. 795-797.
- [22] M. MORF, *Extended system and transfer function matrices and system equivalence*, Proceedings 14th IEEE Conference on Decision and Control, IEEE, New York, 1975, pp. 199-206.
- [23] H. H. ROSENBROCK, *State-Space and Multivariable Theory*, John Wiley, New York, 1970.
- [24] ———, *Systems and polynomial matrices*, in Geometrical Methods in the Theory of Linear Systems, C. Byrnes and C. Martin, eds., Reidel, Boston, 1980, pp. 233-256.
- [25] M. K. SAIN, *Introduction to Algebraic System Theory*, Academic Press, New York, 1981.
- [26] D. W. SHARPE AND P. VAMOS, *Injective Modules*, University Press, Cambridge, 1972.
- [27] J. G. TRUXAL, *Automatic Feedback Control System Synthesis*, McGraw-Hill, New York, 1955.
- [28] P. VAN DOOREN, *The generalized eigenstructure problem in linear system theory*, IEEE Trans. Automat. Control, AC-26 (February 1981), pp. 111-129.
- [29] G. VERGHESE, P. VAN DOOREN AND T. KAILATH, *Properties of the system matrix of a generalized state-space system*, Internat. J. Control, 1979.
- [30] W. M. WONHAM, *Linear Multivariable Control: A Geometric Approach*, 2nd ed. Springer, New York, 1979.
- [31] B. F. WYMAN, *Pole placement over integral domains*, Communications in Algebra, 6 (1980), pp. 969-983.
- [32] B. F. WYMAN AND M. K. SAIN, *The zero module and essential inverse systems*, IEEE Transactions on Circuits and Systems, CAS-28 (1981), pp. 112-126.
- [33] ———, *Internal zeros and the system matrix*, Proceedings 20th Allerton Conference on Communication, Control, and Computing, University of Illinois, Urbana, IL, 1982, pp. 153-158.
- [34] ———, *The zero module of a minimal realization*, Linear Algebra Appl., 50 (1983), pp. 621-637.

PARAMETER ESTIMATION, REGULARITY AND THE PENALTY METHOD FOR A CLASS OF TWO POINT BOUNDARY VALUE PROBLEMS*

K. KUNISCH[†] AND L. W. WHITE[‡]

Abstract. We study a penalization technique for a class of parameter estimation problems associated with elliptic equations. Moreover, a regularization phenomenon of the minimizers is discussed for various parameters in the equation and for different observation operators. Finally it is shown that under weak assumptions the norm constraints are active.

Key words. inverse problems, least squares, regularity of solution

AMS(MOS) subject classifications. 35R25, 35R30

1. Introduction. In this paper we study parameter estimation problems and their approximations for a simple class of two point boundary value problems. We use the output-least-squares approach and formulate the parameter estimation problem as a constrained optimization problem. Two distinct kinds of constraints arise: certain pointwise constraints on the coefficients guarantee well-posedness of the equation, whereas norm bounds are used to argue existence of the solution of the optimization problem. When the original infinite dimensional problem is approximated by finite dimensional problems, then these constraints need to be translated to conditions on the elements in the subspaces that approximate the sets of admissible coefficients. This can lead to quite involved technicalities [8]. When solving the approximate optimization problems with a computer it is natural at first to try to ignore these constraints and to use one of the many available software packages for unconstrained optimization. This approach is successful for some problems ([2], [3] et al.) but can lead to serious difficulties with others. Some examples illustrating this point are given in § 6, where it is shown that the numerical solution or the optimal (i.e. "identified") parameter may exhibit extraneous oscillations or may not converge at all, when only unconstrained optimization is carried out. One of the possibilities to overcome such difficulties is to use a regularization approach, see [4], [7] for example. Other possibilities include the use of constrained optimization routines or a penalty function approach. In this paper we investigate the latter. More specifically, we transform the norm bounds to penalization terms and keep the pointwise bounds (which are more readily implemented on a computer) as explicit constraints. Such a penalty function technique is shown to be effective both theoretically and numerically; we show for our model equation that a satisfactory convergence theory can be obtained and that in problems that did not converge or contained the above mentioned oscillations, the implementation of the norm constraint as a penalty term led to successful numerical results.

* Received by the editors October 14, 1983, and in revised form May 22, 1985.

[†] Institut für Mathematik Technische Universität Graz, A-8010 Graz, Austria, and Department of Mathematics, University of Oklahoma, Norman, Oklahoma 73019. This author gratefully acknowledges support from the Max Kade Foundation. This research was supported in part by the Fonds zur Förderung der Wissenschaftlichen Forschung, Austria, no. P4534, and by the Steiermärkische Wissenschafts und Forschungsförderungsfonds.

[‡] Department of Mathematics, University of Oklahoma, Norman, Oklahoma 73019. The work of this author was supported in part by National Science Foundation grant MCS-7902037 and by the Cooperative Institute for Mesoscale Meteorological Studies.

The motivation for considering a boundary value problem is that it serves as a technically simple example to demonstrate some of the special behavior that optimization problems arising from parameter estimation problems have, in addition to being of practical relevance, see [2], for instance, for a discussion of the use of elliptic equations in modelling large space antennas. We show that under weak assumptions (guaranteeing the nonvanishing of a Lagrange multiplier), the minimizers of the optimization problems enjoy a certain regularity phenomenon; they are necessarily smoother than the functions in the class over which the minimization is carried out. To our knowledge, this is the first time that a regularization phenomenon is documented for parameter estimation problems. It can be used advantageously in the penalization method. Without regularity, when the coefficients in the equation as well as the state equation are approximated, these two limit processes, together with the limit process appearing due to the penalization term, must go to infinity in a certain order to achieve convergence. In the presence of additional regularity of the solution of the parameter estimation problem, these limits can be taken essentially independently of each other. We expect that regularity results of this kind will also be useful for the study of the rate of convergence in parameter estimation problems; see [5], for instance, where smoothness requirements are made without further discussion.

Finally we obtain the result that under the same conditions that guarantee the regularity effect of the minimizers, the norm constraints must be active. This implies that the solution will change as the norm bounds change and that it will not exist, in general, without these norm bounds. This result is remarkable from the modeling point of view and has stimulated that, in a further investigation on the sensitivity of inverse problem, sensitivity is not only considered with respect to the observations but also with respect to the constraints defining the admissible parameter set [4].

The paper is organized as follows: In § 2 we summarize some results on two point boundary value problems that are needed in the remainder of the paper. A penalization method is discussed in § 3 and in § 4 we establish the above mentioned regularity results and the fact that the norm constraints may be active. Section 5 is devoted to employing these results to improve upon the approximation theorems of § 3. The numerical examples are discussed in § 6.

2. Generalities. Let us consider the equation

$$(2.1) \quad \begin{aligned} & -(a(x)u_x)_x + b(x)u_x + c(x)u = f \quad \text{for } 0 < x < 1, \\ & u(0) = u(1) = 0, \end{aligned}$$

where $f \in L^2(0, 1)$ and

$$q = (a, b, c) \in Q \subset \tilde{Q} = W^{1,2}(0, 1) \times W^{1,2}(0, 1) \times L^2(0, 1),$$

with \tilde{Q} endowed with the Hilbert-space product topology. By employing the mean value theorem for integrals one can show that

$$(2.2) \quad |\varphi|_{L^\infty} \leq d |\varphi|_{W^{1,2}} \quad \text{for } \varphi \in W^{1,2}(0, 1) \quad \text{where } d = \sqrt{2},$$

and $W^{1,2}(0, 1)$ is endowed with the $(\int_0^1 |\varphi(s)|^2 ds + \int_0^1 |\dot{\varphi}(s)|^2 ds)^{1/2}$ -norm. Let $\tilde{\alpha} > 0$ and define

$$Q = \left\{ q \in \tilde{Q}: \tilde{\alpha} \leq a(x), |a|_{W^{1,2}} \leq \mu, |b|_{W^{1,2}} \leq \mu, |c|_{L^2} \leq \mu, c(x) \geq \tilde{c} > \frac{d^2 \mu^2}{4 \tilde{\alpha}} \right\}.$$

In $W_0^{1,2}(0, 1) \times W_0^{1,2}(0, 1)$ we define for each $q \in \tilde{Q}$ the form

$$(2.3) \quad l(u, v; q) = (au_x, v_x) + (bu_x, v) + (cu, v),$$

where (\cdot, \cdot) denotes the inner product in $L^2(0, 1)$. It is simple to verify that

$$(2.4) \quad |l(u, v; q)| \leq C_1 |u|_{W_0^{1,2}} |v|_{W_0^{1,2}},$$

and

$$(2.5) \quad l(u, u; q) \geq C_2 |u|_{W_0^{1,2}}^2,$$

for all $q \in Q$ and $u, v \in W_0^{1,2}$, with $C_1 > 0$ and $C_2 > 0$. Here $W_0^{1,2}(0, 1)$ is endowed with the $(\int_0^1 |\dot{\varphi}(x)|^2 dx)^{1/2}$ -norm. We give the details for (2.5) and take $u \in W_0^{1,2}$. Further τ_1 is chosen so that $\tilde{c} - \tau_1(d^2\mu^2/4\tilde{\alpha}) > 0$. Then

$$\begin{aligned} l(u, u; q) &= (au_x, a_x)_{L^2} + (bu_x, u) + (cu, u) \\ &\geq \tilde{\alpha} |u_x|_{L^2}^2 + \tilde{c} |u|_{L^2}^2 - \sup_x |b(x)| |u_x|_{L^2} |u|_{L^2} \\ &\geq \tilde{\alpha} |u_x|_{L^2}^2 + \tilde{c} |u|_{L^2}^2 - d\mu |u_x|_{L^2} |u|_{L^2} \\ &\geq \tilde{\alpha} \left(1 - \frac{1}{\tau_1}\right) |u_x|_{L^2}^2 + \left(\tilde{c} - \frac{d^2\mu^2\tau_1}{4\tilde{\alpha}}\right) |u|_{L^2}^2 \\ &\geq \tilde{\alpha} \left(1 - \frac{1}{\tau_1}\right) |u_x|_{L^2}^2, \end{aligned}$$

which implies (2.5) with $C_2 = \tilde{\alpha}(1 - 1/\tau_1)$.

We shall frequently use the following

Remark 2.1. There exist positive constants c_1, c_2 and δ such that for $q \in Q_\delta$ with

$$Q_\delta = \left\{ q \in \tilde{Q}: |a|_{W^{1,2}} \leq \mu + \delta, |b|_{W^{1,2}} \leq \mu + \delta, \right. \\ \left. |c|_{L^2} \leq \mu + \delta, a(x) \geq \tilde{\alpha}, c(x) \geq \tilde{c} \geq \frac{d^2\mu^2}{4\alpha} \right\}$$

we have $|l(u, v; q)| \leq c_1 |u|_{W^{1,2}} |v|_{W^{1,2}}$, and $l(u, u; q) \geq c_2 |u|_{W^{1,2}}^2$.

The Lax-Milgram theorem implies for each $q \in Q_\delta$ the existence of a unique solution $u(q) \in W_0^{1,2}(0, 1)$ of (2.1) satisfying

$$(2.6) \quad l(u(q), v; q) = (f, v) \quad \text{for all } v \in W_0^{1,2}(0, 1).$$

By Remark 2.1 we also have

$$|u(q)|_{W^{1,2}}^2 \leq c_2^{-1} |(u(q), f)| \leq c_2^{-1} |u(q)|_{L^2} |f|_{L^2}$$

and therefore

$$(2.7) \quad |u(q)|_{W^{1,2}} \leq c_2^{-1} |f|_{L^2},$$

for every $q \in Q_\delta$ and $f \in L^2(0, 1)$. Since $a \in W^{1,2}(0, 1)$ we further have $au_{xx} = -a_x u_x + bu_x + cu - f$ and consequently $u \in W^{2,2}(0, 1)$. By (2.7) there exists a constant $c_3 = c_3(\tilde{c}, d, \mu, \alpha, \delta)$ such that

$$(2.8) \quad |u(q)|_{W^{2,2}} \leq c_3 |f|_{L^2}.$$

Remark 2.2. Let

$$Q_P = \{q \in \tilde{Q}: \tilde{\alpha} \leq a(x), |b(x)| \leq d\mu, \tilde{c} \leq c(x) \text{ for almost all } x \in [0, 1]\}.$$

The calculation after (2.5) reveals that $l(u, u; q) \geq C_2 |u|_{W_0^{1,2}}^2$ for all $q \in Q_P$ and $u \in W_0^{1,2}(0, 1)$. Moreover, for each $q \in Q_P$ there exists a constant $C_3 = C_3(q)$ such that $|l(u, v; q)| \leq C_3 |u|_{W^{1,2}} |v|_{W^{1,2}}$ for all u and v in $W_0^{1,2}(0, 1)$. Therefore there exists a solution $u(q) \in W^{2,2}(0, 1)$ of (2.1) for every $q \in Q_P$. Note also, that $Q \subset Q_P$, since $|b|_{L^\infty} \leq d|b|_{W^{1,2}} \leq d\mu$. This remark will frequently be used in § 3.

If the dependence of u on q and f is relevant, we denote the solution of (2.1) by $u(q, f)$.

LEMMA 2.1. (a) *The set Q is weakly closed and weakly compact in \tilde{Q} . If Q_{ad} is the intersection of Q with a closed convex set in \tilde{Q} , then Q_{ad} is weakly closed and weakly compact as well.*

(b) *If $f^k \rightharpoonup f$ in $L^2(0, 1)$ and $q^k \rightharpoonup q^0$ in \tilde{Q} with $q^0 \in Q$ and $q^k \in Q_P$, then $u(q^k, f^k) \rightharpoonup u(q^0, f)$ in $W^{2,2}(0, 1)$.*

Remark 2.3. If $Q_{\text{ad}} = Q \cap \{b \in W_0^{1,2}(0, 1)\}$, then $d = \frac{1}{2}$.

Proof of Lemma 2.1. Since (a) is obvious, we immediately turn to (b). Since $q^k \rightharpoonup q^0$ there exists a constant k_0 , such that $q^k \in Q_\delta$ for $k \geq k_0$ and by Remark 2.1 $u(q^k, f^k)$ is uniformly bounded in $W^{2,2}(0, 1)$ for $k \geq k_0$. Consequently there exists a subsequence $\{k_j\}$ and $z \in W^{2,2}(0, 1)$ so that $u(q^{k_j}, f^{k_j}) \rightharpoonup z$ in $W^{2,2}(0, 1)$. We have for every $v \in W_0^{1,2}(0, 1)$

$$(a^{k_j} u_x(q^{k_j}, f^{k_j}), v_x) + (b^{k_j} u_x(q^{k_j}, f^{k_j}), v) + (c^{k_j} u(q^{k_j}, f^{k_j}), v) = (f^{k_j}, v).$$

Taking the limit as $k_j \rightarrow \infty$, we find with $q^0 = (a^0, b^0, c^0)$

$$(a^0 z_x, v_x) + (b^0 z_x, v) + (c^0 z, v) = (f, v),$$

for all $v \in W_0^{1,2}(0, 1)$. This implies $z = u(q^0, f)$. The usual subsequence argument implies $u(q^k, f^k) \rightharpoonup u(q^0, f)$ [11, p. 116].

In the following sections we shall use finite dimensional (Galerkin) approximations of (2.1) that converge uniformly in $q \in Q$. So let $H^N \subset W_0^{1,2}$ be a sequence of finite dimensional linear subspaces and consider

$$(2.9) \quad l(u, v^N; q) = (f, v^N) \quad \text{for all } v^N \in H^N.$$

Again the Lax-Milgram theorem implies the existence of unique solutions $u^N(q)$ of (2.9) for each $q \in Q_\delta$. By (2.6) and (2.9) we have

$$l(u(q) - u^N(q), v^N; q) = 0 \quad \text{for all } v^N \in H^N.$$

Therefore for all $v^N \in H^N$

$$\begin{aligned} c_2 \|u(q) - u^N(q)\|_{W^{1,2}}^2 &\leq l(u(q) - u^N(q), u(q) - v^N; q) \\ &\leq c_1 \|u(q) - u^N(q)\|_{W^{1,2}} \|u(q) - v^N\|_{W^{1,2}}, \end{aligned}$$

so that

$$(2.10) \quad \|u(q) - u^N(q)\|_{W^{1,2}} \leq c_1 c_2^{-1} \inf_{v^N \in H^N} \|u(q) - v^N\|_{W^{1,2}},$$

for all $q \in Q_\delta$.

We make the following standard assumption.

(H1) For each $N = 1, 2, \dots$ the subspaces satisfy $H^N \subset W_0^{1,2}(0, 1)$ and $\|\varphi - P_1^N \varphi\|_{W^{1,2}} \leq \rho(N) \|\varphi\|_{W^{2,2}}$,

where $P_1^N: W^{1,2}(0, 1) \rightarrow H^N$ is the orthogonal projection in the $W^{1,2}(0, 1)$ norm, and $\lim_{N \rightarrow \infty} \rho(N) = 0$.

From (2.8) and (2.10) we derive the following

LEMMA 2.2. *Let (H1) hold and take q in Q_δ . Then*

$$\|u(q) - u^N(q)\|_{W^{1,2}} \leq \tilde{\rho}(N) \|f\|_{L^2},$$

with $\lim_{N \rightarrow \infty} \tilde{\rho}(N) = 0$ independent of $q \in Q_\delta$.

COROLLARY 2.1. *Let (H1) hold and let $q^N \rightharpoonup q^0$, with $q^0 \in Q$ and $q^N \in Q_P$. Then*

$$u^N(q^N) \rightarrow u(q^0) \quad \text{in } W^{1,2}(0, 1).$$

This corollary follows directly from Lemmas 2.1 and 2.2. In the final lemma of this section we discuss the dependence of u^N on q and f .

LEMMA 2.3. *Let $H^N \subset W_0^{1,2}(0, 1)$ for each $N = 1, 2, \dots$. Then*

(a) $|u^N(q)|_{W^{1,2}} \leq c_2^{-1} |f|_{L^2}$ for each $q \in Q_\delta$.

(b) *If moreover $q^k \rightharpoonup q^0$ in \tilde{Q} with $q^0 \in Q_\delta$ and $q^k \in Q_P$, and $f^k \rightharpoonup f$ in $L^2(0, 1)$, then*

$$u^N(q^k, f^k) \rightarrow u^N(q^0, f) \quad \text{in } W^{1,2}(0, 1).$$

The proof of this lemma is simple and will therefore not be included.

3. Penalization. Let $z \in Z$, with Z the observation space which is assumed to be a Banach space, and let $C: W^{1,2}(0, 1) \rightarrow Z$ be the associated observation operator. For $q \in Q_P$ we put

$$J(q) = |Cu(q) - z|_Z^2.$$

Note that $u(q)$ is well defined by Remark 2.2.

We then formulate the parameter estimation problem as the optimization problem:

(P) Minimize $J(q)$ over $q \in Q$ subject to $u(q)$ satisfying (2.6).

Throughout we make the following assumption:

(H2) C is a bounded linear operator from $W^{1,2}(0, 1)$ to Z .

In particular we could take $Z = Z_1 = L^2(0, 1)$ and $C_1\varphi = \varphi$, or $Z = Z_2 = \mathbb{R}$ and $C_2\varphi = \varphi(x_0)$, $x_0 \in [0, 1]$, or $Z = Z_3 = L^2(0, 1)$ and $C_3\varphi = \dot{\varphi}$.

Note that (P) is a restricted optimization problem; the *pointwise constraints* guarantee well-posedness of the equation (2.1) whereas the *norm constraints* in the characterization of Q are essential to establish existence of solutions to (P). Several investigations concerning the approximation of (P) by finite dimensional problems have been carried out, both theoretically and numerically for many classes of partial differential equations [2], [3], [8]. For practical purposes the approximating optimization problems have commonly been implemented by employing unconstrained optimization packages, so that the constraints used in defining Q have essentially been ignored. A penalization technique appears to parallel what is done in practice and we consequently study penalization techniques for the simple model problem (P).

First we define the penalization functionals:

$$\psi_A(a) = \begin{cases} 0 & \text{if } |a|_{W^{1,2}} \leq \mu, \\ (|a|_{W^{1,2}} - \mu)^2 & \text{if } |a|_{W^{1,2}} > \mu; \end{cases}$$

and functionals ψ_B and ψ_C are defined in an analogous manner. We also put

$$\Psi(q) = \psi_A(a) + \psi_B(b) + \psi_C(c),$$

and note that

$$(3.1) \quad |q| \rightarrow \infty \quad \text{implies } \Psi(q) \rightarrow \infty.$$

Remark 3.1. The pointwise bounds appearing in Q_P can be more readily implemented in computer algorithms than the norm constraints that are defining Q .

Associated with Ψ and the approximating state equations we consider

$$(P_M) \quad \min_{q \in Q_P} (J(q) + M\Psi(q)),$$

$$(P^N) \quad \min_{q \in Q} J^N(q) = \min_{q \in Q} |Cu^N(q) - z|^2 \quad \text{where } u^N(q) \text{ satisfies (2.9) and}$$

$$(P_M^N) \quad \min_{q \in Q_P} (J^N(q) + M\Psi(q)).$$

LEMMA 3.1. Assume that $H^N \subset W_0^{1,2}$ and that (H2) holds. Then there exist solutions $\bar{q} \in Q$, $\bar{q}^N \in Q$, $\bar{q}_M \in Q_P$ and $\bar{q}_M^N \in Q_P$ of (P), (P^N) , (P_M) and (P_M^N) respectively.

Proof. Since Ψ is radially unbounded and since $J(q) \geq 0$, $J^N(q) \geq 0$, the existence of minimizers \bar{q}_M and \bar{q}_M^N in Q_P of (P_M) and (P_M^N) easily follows (see [9, p. 8]). Next let $q^k \in Q$ be a minimizing sequence for (P). Then there exists a subsequence $\{q^{k_n}\}$ with $q^{k_n} \rightharpoonup \bar{q}$ in \tilde{Q} with $\bar{q} \in Q$. By Lemma 2.1, $u(q^{k_n}) \rightarrow u(\bar{q})$ in $W^{1,2}(0, 1)$ and further $J(q^{k_n}) \rightarrow J(\bar{q})$. Since $\inf_{q \in Q} J(q) = \lim_{k_n \rightarrow \infty} J(q^{k_n})$, it follows that \bar{q} solves (P). The existence of a minimizer for (P^N) is guaranteed by Lemma 2.3(b).

THEOREM 3.1. Let the approximation assumption (H1) and (H2) hold and let q^* be a weak limit point of $\{\bar{q}_M^N\}$, such that $\bar{q}_M^N \rightharpoonup q^*$ in \tilde{Q} with $N_k \rightarrow \infty$, $M_k \rightarrow \infty$ as $k \rightarrow \infty$. Then $q^* \in Q$, q^* is a solution of (P), $u^{N_k}(\bar{q}_M^N) \rightarrow u(q^*)$ in $W^{1,2}(0, 1)$ and $J^{N_k}(\bar{q}_M^N) + M_k \Psi(\bar{q}_M^N) \rightarrow J(q^*)$ as $k \rightarrow \infty$.

Proof. The verification of this result can be given by well-known arguments. Let $q \in Q$ be arbitrary. Then

$$J^{N_k}(\bar{q}_M^N) + M_k \Psi(\bar{q}_M^N) \leq J^{N_k}(q).$$

But $\{J^{N_k}(q): q \in Q, N = 1, 2, \dots\}$ is bounded as a consequence of (H2) and Lemma 2.3(a). Therefore, there exists a constant K such that

$$(3.2) \quad J^{N_k}(\bar{q}_M^N) + M_k \Psi(\bar{q}_M^N) \leq K.$$

Since $J^N(\bar{q}_M^N) \geq 0$ we have by (3.2) that $\lim_{k \rightarrow \infty} \Psi(\bar{q}_M^N) \rightarrow 0$, which implies that $q^* \in Q$. For each $q \in Q$ the following inequalities hold:

$$(3.3) \quad J^{N_k}(\bar{q}_M^N) \leq J^{N_k}(\bar{q}_M^N) + M_k \Psi(\bar{q}_M^N) \leq J^{N_k}(q) + M_k \Psi(q) = J^{N_k}(q).$$

From Corollary 2.1 we conclude that

$$(3.4) \quad \begin{aligned} \lim_{k \rightarrow \infty} u^{N_k}(\bar{q}_M^N) &= u(q^*) \quad \text{in } W^{1,2}(0, 1) \quad \text{and} \\ \lim_{N \rightarrow \infty} u^N(q) &= u(q) \quad \text{in } W^{1,2}(0, 1). \end{aligned}$$

Taking the limit as $k \rightarrow \infty$ in (3.3) and recalling (H2), we arrive at

$$J(q^*) \leq J(q) \quad \text{for all } q \in Q,$$

so that q^* is in fact a solution of (P). The remaining claims follow from (3.4) and (3.3) with $q = q^*$.

Remark 3.2. The conclusion of Theorem 3.1 is not changed if M in $J(q) + M\Psi(q)$ is replaced by $h(M)$ with $\lim_{M \rightarrow \infty} h(M) = \infty$. The proof of Theorem 3.1 reveals that $\Psi(\bar{q}_M^N) = O(h(M_k)^{-1})$ as $k \rightarrow \infty$.

Remark 3.3. Problems (P^N) and (P_M^N) are not finite dimensional, since Q is an infinite dimensional coefficient space. If \tilde{Q} is replaced by \tilde{Q}_F , where \tilde{Q}_F is a finite dimensional subspace of \tilde{Q} , and Q and Q_P are replaced by $Q \cap \tilde{Q}_F$ and $Q_P \cap \tilde{Q}_F$, then Theorem 3.1 remains correct and (P^N) and (P_M^N) are finite dimensional problems (compare also Lemma 2.1). Otherwise a further approximation of the coefficients is required, which we describe next.

Let $Q^m \subset \tilde{Q}$ be a sequence of finite dimensional subspaces of \tilde{Q} satisfying:

(H3) For each $q \in \tilde{Q} \cap Q_P$ and $m = 1, 2, \dots$ there exists $\tilde{q}^m \in Q^m \cap Q_P$ such that $|q - \tilde{q}^m| = \rho(m)$, and $\lim_{m \rightarrow \infty} \rho(m) = 0$.

Consider the problems:

$$(P_M^{N,m}) \quad \min_{q \in Q^m \cap Q_P} J^N(q) + M\Psi(q).$$

Clearly solutions $\bar{q}_M^{N,m}$ of $(P_M^{N,m})$ exist. In the next theorem the notation $w\text{-lim}$ is used to denote the weak limit.

THEOREM 3.2. *Let (H1)–(H3) hold and let q^* be a weak iterated limit point of $\bar{q}_M^{N,m}$, such that $q^* = w\text{-}\lim_{i \rightarrow \infty} w\text{-}\lim_{k \rightarrow \infty} \bar{q}_{M_i}^{N_i, m_k}$, with $N_i \rightarrow \infty$, $M_i \rightarrow \infty$ as $i \rightarrow \infty$ and $m_k \rightarrow \infty$ as $k \rightarrow \infty$. Then q^* is a solution of (P), $\lim_{i \rightarrow \infty} \lim_{k \rightarrow \infty} u^{N_i}(\bar{q}_{M_i}^{N_i, m_k}) = u(q^*)$ in $W^{1,2}$, and*

$$\lim_{i \rightarrow \infty} \lim_{k \rightarrow \infty} J^{N_i}(\bar{q}_{M_i}^{N_i, m_k}) = J(q^*).$$

Proof. Let $q \in Q$ be arbitrary and choose $\tilde{q}^m \in Q^m \cap Q_P$ according to (H3). Then we have

$$(3.5) \quad J^{N_i}(\bar{q}_{M_i}^{N_i, m_k}) + M_i \Psi(\bar{q}_{M_i}^{N_i, m_k}) \leq J^{N_i}(\tilde{q}^{m_k}) + M_i \Psi(\tilde{q}^{m_k}).$$

For each i we have $\lim_{k \rightarrow \infty} (J^{N_i}(\tilde{q}^{m_k}) + M_i \Psi(\tilde{q}^{m_k})) = J^{N_i}(q)$ and consequently

$$(3.6) \quad \lim_{i \rightarrow \infty} \lim_{k \rightarrow \infty} (J^{N_i}(\tilde{q}^{m_k}) + M_i \Psi(\tilde{q}^{m_k})) = J(q)$$

by Lemma 2.3(b) and Lemma 2.2. In particular, the left-hand side of (3.5) is nonnegative and bounded. This implies that $\lim_{i \rightarrow \infty} \lim_{k \rightarrow \infty} \psi(\bar{q}_{M_i}^{N_i, m_k}) = 0$ from which we conclude that $q^* \in Q$. Next we put $\bar{q}_{M_i}^{N_i, 0} = w\text{-}\lim_{k \rightarrow \infty} \bar{q}_{M_i}^{N_i, m_k}$ and note that $\bar{q}_{M_i}^{N_i, 0} \in Q_P$. Taking the limit with respect to k in

$$J^{N_i}(\bar{q}_{M_i}^{N_i, m_k}) \leq J^{N_i}(\tilde{q}^{m_k}) + M_i \Psi(\tilde{q}^{m_k}),$$

we find for sufficiently large i

$$(3.7) \quad J^{N_i}(\bar{q}_{M_i}^{N_i, 0}) \leq J^{N_i}(q),$$

by Lemma 2.3(b) and since $q^* = w\text{-}\lim_{i \rightarrow \infty} \bar{q}_{M_i}^{N_i, 0} \in Q$. Taking the limit as $i \rightarrow \infty$ in (3.7) leads to

$$(3.8) \quad J(q^*) \leq J(q)$$

by Corollary 2.1. Since q was an arbitrary element of Q , this implies that q^* is a solution of (P). The remaining assertions are easily verified and the proof is complete.

Remark 3.4. The proof of the previous result depends strongly on the fact that the limits are taken iteratively; in general (3.6) and (3.8) will not be true for arbitrary order of the limits. This shortcoming cannot be overcome by replacing M by a function $h(M)$ with $\lim_{M \rightarrow \infty} h(M) = \infty$.

Remark 3.5. We give an example for subspaces Q^m , so that (H3) is satisfied. Let $S_1^m(0, 1) = \{\text{set of piecewise linear spline functions with knots at } i/m, i = 0, \dots, m\}$ and $Q^m = \bigotimes_{i=1}^3 S_1^m(0, 1)$. By $P^m: \tilde{Q} \rightarrow Q^m \cap Q_P$ we denote the projection operator which associates to each element $q \in \tilde{Q}$ the unique element $P^m q \in Q^m \cap Q_P$ with shortest distance to q , [6, p. 249]. To verify (H3) for this case, let $q = (a, b, c) \in (W^{2,2}(0, 1) \times W^{2,2}(0, 1) \times W^{1,2}(0, 1)) \cap Q_P$ and note that

$$(3.9) \quad |q - P^m q|_{\tilde{Q}} \leq K_1 m^{-1} |q|_{W^{2,2} \times W^{2,2} \times W^{1,2}},$$

where K_1 is independent of q and N . This estimate is easily obtained by employing the interpolation operation and well-known spline estimates (see e.g. [12, Chap. 2]). For $q \in \tilde{Q} \cap Q_P$ note that $P^m q$ coincides with the Hilbert space projection of q onto Q^m . Moreover, for every $q \in \tilde{Q} \cap Q_P$ there exists a sequence $q^l \in (W^{2,2}(0, 1) \times W^{2,2}(0, 1) \times W^{1,2}(0, 1)) \cap Q_P$ with $\lim_{l \rightarrow \infty} q^l = q$ in \tilde{Q} . These last two facts, together with (3.9) and the triangle inequality imply that (H3) holds for linear spline approximations of the coefficients.

Remark 3.6. The calculations in the previous example depend strongly on the interpolation properties of linear spline functions; this is reflected, for example, in (3.9) where the interpolation operator also realizes the pointwise constraints associated with Q_P . In general norm or pointwise bounds can only be achieved in the limit (as for example, when cubic splines are used to approximate the coefficients). In this case it may be necessary that an additional index is used allowing that the coefficients of the finite dimensional problem satisfy the constraints characterizing Q only approximately (compare [8, § 4]). We shall not follow up on these ideas here but rather continue to develop a technique that allows one to replace the iterated limit in Theorem 3.2 by a result that guarantees that the various limits may be taken independently.

4. Regularity. In this section we consider again the equation

$$(4.1) \quad \begin{aligned} & -(au_x)_x + bu_x + cu = f \quad \text{on } (0, 1), \\ & u(0) = u(1) = 0. \end{aligned}$$

We first assume that a and c are fixed and that there exists at least one \tilde{b} with $(a, \tilde{b}, c) \in Q$. This leaves b to be estimated and we assume either of the two observations C_i , $i = 1$ or 2 , described at the beginning of § 3 to be available. For $i = 1$ we let $z_1 \in L^2(0, 1)$, whereas $z_2 \in \mathbb{R}$. We further put

$$J_i(b) = |C_i u(b) - z_i|^2,$$

where we now denote the parameter dependence of u by $u(b)$. The parameter estimation problems for $i = 1$ or 2 become:

(P_i) Minimize $J_i(b)$ over B subject to $u(b)$ satisfying (4.1).

Here $B = \{b \in W^{1,2}(0, 1) : |b|_{W^{1,2}} \leq \mu\}$. Analogous to the definition of Q_δ in § 2 we put $B_\delta = \{b \in W^{1,2}(0, 1) : |b|_{W^{1,2}} \leq \mu + \delta\}$. The existence of a solution of (P_i) is guaranteed by Lemma 3.1. We next calculate the Fréchet derivative of the function $b \rightarrow u(b)$.

LEMMA 4.1. *The mapping $b \rightarrow u(b)$ from $B \subset W^{1,2}(0, 1)$ into $W^{2,2}(0, 1)$ is Fréchet differentiable with the Fréchet differential with increment h denoted by $\delta_b u(b)h = \eta(h)$ with $\eta(h)$ the unique solution of*

$$(4.2) \quad \begin{aligned} & -(a\eta(h)_x)_x + b\eta_x(h) + c\eta(h) = -hu_x(b), \quad \text{in } (0, 1), \\ & \eta(h)(0) = \eta(h)(1) = 0. \end{aligned}$$

Proof. The verification is quite standard but we include it for the purpose of completeness. Let $h \in W^{1,2}(0, 1)$ and $b \in B$ and note that there exists $\varepsilon(h) > 0$ such that for any $\varepsilon \in (0, \varepsilon(h))$ the element $b + \varepsilon h \in B_\delta$ and $u(b + \varepsilon h)$ exists. Set $u^\varepsilon = \varepsilon^{-1}(u(b + \varepsilon h) - u(b))$ and observe that u^ε must satisfy

$$(4.3) \quad \begin{aligned} & -(au_x^\varepsilon)_x + bu_x^\varepsilon + cu^\varepsilon = -hu_x(b + \varepsilon h), \\ & u^\varepsilon(0) = u^\varepsilon(1) = 0. \end{aligned}$$

From Lemma 2.1 it follows that $u(b + \varepsilon h) \rightarrow u(b)$ in $W^{1,2}(0, 1)$ as $\varepsilon \rightarrow 0$. Applying Lemma 2.1 once again to (4.3), we find that u^ε converges weakly in $W^{2,2}(0, 1)$ and thus strongly in $W^{1,2}(0, 1)$. We denote this limit by $\eta(h)$. As a consequence of u^ε satisfying (4.3) and the limit behavior of u^ε , we observe that $u^\varepsilon \rightarrow \eta(b)$ as $\varepsilon \rightarrow 0$ strongly in $W^{2,2}(0, 1)$ as well. Thus, the limit $\eta(h)$ is the Gateaux derivative of the mapping $b \rightarrow u(b)$ with increment h , i.e. $\eta(h) = \delta u(b)h$, and it satisfies

$$(4.4) \quad \begin{aligned} & -(a\eta_x(h))_x + b\eta_x(h) + c\eta(h) = -hu_x(b) \quad \text{in } (0, 1), \\ & \eta(h)(0) = \eta(h)(1) = 0. \end{aligned}$$

Note that $h \rightarrow \eta(h)$ is a bounded linear operator from $W^{1,2}(0, 1)$ to $W^{2,2}(0, 1)$. That $h \rightarrow \eta(h)$ is the Fréchet differential of u at b with increment h can be seen as follows: let $h \in W^{1,2}(0, 1)$ with $|h|_{W^{1,2}} \leq \delta$, and set

$$\Delta(h) = |h|_{W^{1,2}}^{-1}(u(b+h) - u(b) - \eta(h)).$$

We note that $\Delta(h)$ satisfies

$$(4.5) \quad \begin{aligned} -(a\Delta_x(h))_x + b\Delta_x(h) + c\Delta(h) &= \frac{h}{|h|_{W^{1,2}}}(u_x(b+h) - u_x(b)), \\ \Delta(h)(0) &= \Delta(h)(1) = 0. \end{aligned}$$

From (2.8) we have

$$|\Delta(h)|_{W^{2,2}} \leq c_3 |u_x(b+h) - u_x(b)|_{L^2}.$$

Lemma 2.1 now implies that $|\Delta(h)|_{W^{2,2}} \rightarrow 0$ whenever $|h|_{W^{1,2}} \rightarrow 0$. Thus, we have established the lemma.

Since the functionals $b \rightarrow J_i(b)$ are the composition of Fréchet differentiable functions, they are themselves Fréchet differentiable; we find

$$(4.6) \quad \delta_b J_1(b)h = 2 \int_0^1 (u(b) - z_1) \eta(h) dx$$

and

$$(4.7) \quad \delta_b J_2(b)h = 2(u(b)(x_0) - z_2) \eta(h)(x_0).$$

To get a different representation of the Fréchet differential of J_i , we introduce the adjoint problems

$$(4.8) \quad \begin{aligned} -(ap_x^1)_x - (bp^1)_x + cp^1 &= u(b) - z_1, \\ p^1(0) &= p^1(1) = 0, \end{aligned}$$

and

$$(4.9) \quad \begin{aligned} -(ap_x^2)_x - (bp^2)_x + cp^2 &= (u(b)(x_0) - z_2) \delta(x_0), \\ p^2(0) &= p^2(1) = 0, \end{aligned}$$

where $\delta(x_0)$ denotes the Dirac delta measure with mass at $x_0 \in (0, 1)$. Existence and regularity of unique solutions to equations (4.8) and (4.9) are shown in an analogous manner as for equation (2.1). We find

LEMMA 4.2. *Let $b \in B$. There exist unique solution p^i of (4.8) and (4.9) with $p^1 \in W^{2,2}(0, 1) \cap W_0^{1,2}(0, 1)$ and $p^2 \in W_0^{1,2}(0, 1)$.*

We now obtain the Fréchet differential of J_i in terms of p_i by multiplying (4.8) and (4.9) by $\eta(h)$ and integrating by parts. Thus, we see that

$$(4.10) \quad \delta_b J_i(b)h = -2 \int_0^1 u_x(b) p^i h dx$$

for $i = 1$ or 2 and $h \in W^{1,2}(0, 1)$.

We next investigate *regularity* of the solutions \bar{b}_i of (P_i) . Our approach is to incorporate the explicit optimization constraints into a Lagrangian functional to be minimized over the larger set B_δ . The set B_δ is unavoidable since it is essential to have $b \rightarrow u(b)$ and its Fréchet derivative defined on B . We are able to obtain an Euler equation since the admissible parameters are contained in the interior of B_δ . Another

point to be remembered is that $b \rightarrow u(b)$ is not linear. Hence, the functionals J_i are not convex. Accordingly we use the generalized Kuhn-Tucker theorem for local optimization theory. We define functionals G on $W^{1,2}(0, 1)$ by

$$G(b) = |b|_{W^{1,2}}^2 - \mu^2$$

and observe that the Fréchet differential of G at b with increment h is given by

$$(4.11) \quad \delta_b G(b)h = 2 \int_0^1 (b_x h_x + bh) dx = 2 \int_0^1 (-b_{xx} + b)h dx + [b_x(1)h(1) - b_x(0)h(0)].$$

The right-hand side of (4.11) defines a continuous linear functional on $W^{1,2}(0, 1)$. In terms of G the estimation problems are given by:

(P_i^{*}) Minimize $J_i(b)$ subject to $G(b) \leq 0$ and (4.1).

The Kuhn-Tucker theorem [10, p. 249] asserts that if \bar{b}_i solves (P_i^{*}) and if \bar{b}_i is a *regular point* of the inequality $G(b) \leq 0$, and there exists $\lambda \geq 0$ such that the Lagrangian

$$\Lambda_i(b) = J_i(b) + \lambda G(b)$$

is stationary at \bar{b}_i . We note that $\Lambda_i(b)$ is defined on B_δ which contains B in its interior. To verify that \bar{b}_i is indeed a regular point of $G(b) \leq 0$, we need to verify that if $G(\bar{b}_i) \leq 0$ then there exists an $\tilde{h} \in W^{1,2}(0, 1)$ with $G(\bar{b}_i) + \delta G(\bar{b}_i)h < 0$. Set $h = -\bar{b}_i$. Then

$$\begin{aligned} G(\bar{b}_i) - \delta_b G(\bar{b}_i)\bar{b}_i &= (|\bar{b}_i|_{W^{1,2}}^2 - \mu^2) - 2|\bar{b}_i|_2 W^{1,2} \\ &= -(|\bar{b}_i|_{W^{1,2}}^2 + \mu^2) < 0. \end{aligned}$$

Thus, if \bar{b}_i is a solution of (P_i^{*}) then

$$\delta_b \Lambda_i(\bar{b}_i)h = 0 \quad \text{and} \quad \lambda G(\bar{b}_i) = 0,$$

since $\bar{b}_i \in B \subset \text{int } B_\delta$. From (4.10) and (4.11) it follows that

$$\int_0^1 [-u_x(\bar{b}_i)p^i + \lambda(\bar{b}_i - (\bar{b}_i)_{xx})]h dx + \lambda[\bar{b}_x(1)h(1) - \bar{b}_x(0)h(0)] = 0,$$

for all $h \in W^{1,2}(0, 1)$. Accordingly we have the Euler equation

$$(4.12) \quad \begin{aligned} \lambda(-(\bar{b}_i)_{xx} + \bar{b}_i) &= u_x(\bar{b}_i)p^i \quad \text{in } (0, 1), \\ \lambda(\bar{b}_i)_x(1) &= \lambda(\bar{b}_i)_x(0) = 0. \end{aligned}$$

We observe that since $u_x(\bar{b}_i)$ and p^i both belong to $W_0^{1,2}(0, 1)$, their product belongs to $W_0^{1,2}(0, 1)$ as well. This follows by the product rule or, more generally, by the Banach algebraic properties of $W^{m,p}(\Omega)$ (see [1, p. 115]).

First we consider the case of distributed observations, $i = 1$. Suppose that $\lambda = 0$. Then (4.12) implies

$$(4.13) \quad u_x(\bar{b}_1)p^1 = 0 \quad \text{on } [0, 1].$$

In this case $[0, 1] = \bigcup_{j=1}^\infty I_j$ with $u_x(\bar{b}_1)$ or p^1 identically zero on I_j . If $u_x(\bar{b}_1) = 0$ on I_j , then $u(\bar{b}_1)(x) = f(x) c^{-1}(x) = \text{constant}$ almost everywhere on I_j by (4.1). On the other hand, if $p = 0$ and I_j then $u(\bar{b}_1) = z_1$ on I_j by (4.8). We thus have:

LEMMA 4.3. *If fc^{-1} is nonconstant (a.e.) on every subinterval of $[0, 1]$, then either $u(\bar{b}_1) = z_1$ on $[0, 1]$ and $u(\bar{b}_1)$ is a solution of (4.1) with $|\bar{b}_1| \leq \mu$ for every solution \bar{b}_1 of (P₁) or every Lagrange multiplier associated with the constraint $|b|_{W^{1,2}} \leq \mu$ is positive.*

A special case arises when z_1 is obtained from piecewise linear interpolation of discrete data. Then z_1 is not sufficiently smooth to be a candidate for a solution of (4.1); if moreover fc^{-1} is nonconstant on every subinterval of $(0, 1)$, then $\lambda > 0$.

If $u(\bar{b}_1) \neq z_1$ almost everywhere on $[0, 1]$, then by the above considerations, $u(\bar{b}_1) = \text{constant}$ on $[0, 1]$ and as a consequence of the Dirichlet boundary conditions we have $u(\bar{b}_1) = 0$ on $[0, 1]$. This leads to

LEMMA 4.4. *Let \bar{b}_1 be a solution of (P_1) with $u(b_1)(x) \neq z_1(x)$ a.e. on $[0, 1]$. Then either $f = 0$ (a.e.) on $[0, 1]$ or every Lagrange multiplier associated with the constraint $|b|_{W^{1,2}} \leq \mu$ is positive.*

In the case $\lambda > 0$, any solution \bar{b}_1 of (P_1) satisfies the Neumann problem

$$(4.14) \quad \begin{aligned} -(\bar{b}_1)_{xx} + \bar{b}_1 &= \frac{1}{\lambda} u_x(\bar{b})p \quad \text{in } (0, 1), \\ (\bar{b}_1)_x(1) &= (\bar{b}_1)_x(0) = 0, \end{aligned}$$

where $u_x(\bar{b})p \in W_0^{1,2}(0, 1)$.

LEMMA 4.5. *If $\lambda > 0$, then the solution \bar{b}_1 of (P_1) belongs to $W^{3,2}(0, 1)$ and $|\bar{b}_1|_{W^{1,2}} = \mu$.*

Lemmas 4.1 and 4.2 can obviously be used to give sufficient conditions for nontriviality of λ . A different condition guaranteeing $\lambda > 0$ is given by:

(Λ) There exists an interval $W \subset [0, 1]$ such that $fc^{-1} \neq \text{constant}$ and $\mu(b_1) \neq z$ a.e. on W .

We will show that (Λ) implies that (4.13) cannot hold and that therefore $\lambda > 0$. In fact, if (4.13) holds, then there exists an interval $V_1 \subset W$ such that $u_x(\bar{b}_1) = 0$ or $p = 0$ on V_1 . Using (4.1) and (4.8) this contradicts (Λ) and $\lambda > 0$.

We summarize some of the results for the case $i = 1$ in a theorem.

THEOREM 4.1. (a) *If fc^{-1} is nonconstant (a.e.) on every subinterval of $[0, 1]$ and z_1 is not a solution of (4.1) for any $b \in B$ then all solutions \bar{b}_1 of (P_1) belong to $W^{3,2}(0, 1)$.*

(b) *If $u(\bar{b}_1)(x) \neq z_1(x)$ a.e. for a solution \bar{b}_1 of (P_1) and $f \neq 0$ in $L^2(0, 1)$, then $\bar{b}_1 \in W^{3,2}(0, 1)$.*

(c) *If (Λ) holds for a solution \bar{b}_1 of (P_1) , then $\bar{b}_1 \in W^{3,2}(0, 1)$. In all these cases $\lambda > 0$, $(\bar{b}_1)_x(0) = (\bar{b}_1)_x(1) = 0$ and $|\bar{b}_1| = \mu$.*

For the case $i = 2$ we easily find an analogous result.

THEOREM 4.2. *Let fc^{-1} be nonconstant (a.e.) on an open interval containing x_0 . Then either $z_2 = u(b)(x_0)$ for some $b \in B$ or every solution \bar{b}_2 of (P_2) belongs to $W^{3,2}(0, 1)$, $(\bar{b}_2)_x(0) = (\bar{b}_2)_x(1) = 0$ and $|\bar{b}_2|_{W^{1,2}} = \mu$.*

Remark 4.1. Quite similar calculations can be used to discuss regularity for the case when the observation operator is given by $C_3\varphi = \dot{\varphi}$. In this case the optimization problem becomes:

$$(P_3) \quad \min_{b \in B} |C_3 u(q) - z_3|^2,$$

for a given $z_3 \in L^2(0, 1)$. One can derive the following result: If fc^{-1} is nonconstant on every subinterval of $[0, 1]$, then $z_3 \in W^{1,2}(0, 1)$ and $u_x(\bar{b}_3) = z_3$ for some (and thus all) solutions \bar{b}_3 of (P_3) or $\bar{b}_3 \in W^{3,2}(0, 1)$, $(\bar{b}_3)_x(0) = (\bar{b}_3)_x(1) = 0$ and $|\bar{b}_3|_{W^{1,2}} = \mu$ for all solutions \bar{b}_3 of (P_3) .

We next turn to discuss a certain regularity property that occurs when c is estimated. The calculations are similar to the case that we just described when estimating b and will therefore be omitted.

Let us assume that a , b and f are known, with $b = 0$ and $f \in L^2(0, 1)$, and that c has to be estimated in

$$(4.15) \quad \begin{aligned} -(au_x)_x + cu &= f \quad \text{in } [0, 1], \\ u(0) &= u(1) = 0, \end{aligned}$$

where $a(x) \geq \alpha > 0$ and $|a|_{W^{1,2}} \leq \mu$. Let $C = \{c \in L^2(0, 1): |c|_{L^2} \leq \tilde{\mu}, \text{ with } \tilde{\mu} \in (0, 2\alpha)\}$ and $C_\delta = \{c \in L^2(0, 1): |c|_{L^2} \leq \alpha + (\tilde{\mu}/2)\}$.

Note that $\tilde{\mu} < \alpha + (\tilde{\mu}/2) < 2\alpha$. Using Remark 2.3, one can easily show that there exist constants c_4 and c_5 such that for all φ and ψ in $W_0^{1,2}(0, 1)$ and $c \in C_\delta$ we have $l(\varphi, \varphi) \geq c_4 |\varphi|_{W_0^{1,2}}$ and $|l(\varphi, \psi)| \leq c_5 |\psi|_{W_0^{1,2}} |\varphi|_{W_0^{1,2}}$. As before we consider the problems:

$$(P_i^c) \quad \min J_i(c) = \min |C_i u(c) - z_i|^2 \text{ over } C \text{ subject to } u(c) \text{ satisfying (4.15)}$$

or equivalently

$$(P_i^c) \quad \min J_i(c) = \min |C_i u(c) - z_i|^2 \text{ such that } G(c) \leq 0 \text{ and } u(c) \text{ satisfies (4.15).}$$

Here $G(c) = |c|_{L^2}^2 - \tilde{\mu}^2$, $i = 1$ or 2 and $u(c)$ denotes the solution of (4.15). Note that $J_i(c)$ is defined on C_δ . Again any solution \bar{c} of (P_i^c) is seen to be a regular point of $G(c) \leq 0$ and there exists a $\lambda \geq 0$ such that \bar{c} is a stationary point of $\Lambda: C_\delta \rightarrow \mathbb{R}$ given by

$$\Lambda(c) = J_i(c) + \lambda G(c).$$

We now restrict ourselves to the case of distributed observations C_1 . For every $h \in L^2(0, 1)$ we have

$$(4.16) \quad \delta_c \Lambda(\bar{c})h = 2 \int_0^1 (u(\bar{c}_1)(x) - z_1(x))(\delta_c u(\bar{c}_1)h)(x) dx + 2\lambda (\bar{c}_1, h)_{L^2} = 0,$$

where for $c \in L^2(0, 1)$, $\delta_c u(c): L^2(0, 1) \rightarrow W^{2,2}(0, 1)$ is given by $\delta_c u(c)h = \eta(h)$, with $\eta(h) \in W^{2,2}(0, 1)$ satisfying

$$(4.17) \quad \begin{aligned} -(a\eta_x(h))_x + c\eta(h) &= -hu(c), \\ \eta(0) &= \eta(1) = 0. \end{aligned}$$

Consider the adjoint equation given by

$$(4.18) \quad \begin{aligned} -(ap_x)_x + cp &= u(c) - z_1 \quad \text{in } (0, 1), \\ p(0) &= p(1) = 0. \end{aligned}$$

From (4.17) and (4.18) we conclude that

$$(4.19) \quad \int_0^1 (u(c) - z_1)\eta(h) dx = - \int_0^1 pu(c)h dx.$$

Using this in (4.16), we obtain

$$(4.20) \quad \int_0^1 (-pu(\bar{c}_1) + \lambda \bar{c}_1)h dx = 0 \quad \text{for all } h \in L^2(0, 1).$$

Therefore we have the relationship

$$(4.21) \quad \lambda \bar{c}_1(x) - p(x)u(\bar{c}_1)(x) = 0 \quad \text{almost everywhere in } [0, 1].$$

If $\lambda = 0$, then $p(x)u(x) = 0$ for all $x \in [0, 1]$, since p and u are continuous. Suppose $p(\xi) \neq 0$ for some $\xi \in [0, 1]$; then $p(x) \neq 0$ for all x in an interval V containing ξ . This implies that $u(\bar{c}_1) = 0$ on V and consequently $f = 0$ a.e. on V . If we assume that f is not a.e. zero on any subinterval of $[0, 1]$, then from this last observation we obtain that $p = 0$ on $[0, 1]$ or $\lambda = 0$. In the case $p = 0$ on $[0, 1]$ we have $u(\bar{c}_1) = z_1$ on $[0, 1]$ by (4.19) and therefore $z_1 \in W^{2,2}(0, 1)$.

THEOREM 4.3. *If f is not almost everywhere zero on any subinterval of $[0, 1]$, then either $z_1 = u(c)$ in $[0, 1]$ for some $c \in C$ or $\lambda > 0$. If $\lambda > 0$, then $\bar{c}_1 = (1/\lambda)pu(\bar{c}_1)$, $\bar{c}_1 \in W^{2,2}(0, 1)$ and $|\bar{c}_1|_{L^2} = \mu$.*

The case of point observation C_2 is treated quite similarly. We only need to replace (4.16), (4.18) and (4.19) by

$$(4.16') \quad \delta_c \Lambda(\bar{c}_2)h = 2(u(\bar{c}_2)(x_0) - z_2)(\delta_c u(\bar{c}_2)h)(x_0) + 2\lambda(\bar{c}_2, h)_{L^2} = 0,$$

$$(4.18') \quad \begin{aligned} -(ap_x)_x + cp &= (u(c)(x_0) - z_2)\delta(x^0), \\ p(0) &= p(1) = 0 \end{aligned}$$

and

$$(4.19') \quad (u(c)(x_0) - z_2)\eta(h)(x_0) = - \int_0^1 pu(c)h \, dx.$$

We therefore obtain $\delta_c \Lambda(\bar{c}_2)h = -2 \int_0^1 pu(\bar{c}_2)h \, dx + 2\lambda(\bar{c}_2, h)_{L^2} = 0$, which is (4.20). Proceeding as in the case of distributed observations, we derive the following result:

THEOREM 4.4. *If f is not almost everywhere zero on any subinterval of $[0, 1]$, then either $z_2 = u(c)(x_0)$ for some $c \in C$ or $\lambda > 0$. If $\lambda > 0$, then $\bar{c}_2 = (1/\lambda)pu(\bar{c}_2)$, $\bar{c}_2 \in W^{1,2}(0, 1)$ and $|\bar{c}_2|_{L^2} = \mu$.*

Remark 4.2. The study of possible regularity of a in (4.1), and more generally, of pointwise and norm bounds simultaneously requires a somewhat different analysis which will be carried out elsewhere.

5. Penalization in the presence of regularity. We return to the case of estimating b that was discussed at the beginning of § 4, and make use of the regularity $\bar{b} \in W^{3,2}(0, 1)$ that was shown to hold for the solutions \bar{b} of (P_i) .

The following condition is assumed to hold for the subspaces of the approximating functions for the coefficient b . Let $B_P = \{b \in W^{1,2}: |b(x)| \leq d\mu \text{ for all } x \in [0, 1]\}$.

(H4) The sequence of finite dimensional linear subspaces $B^m \subset W^{1,2}(0, 1)$ satisfies that for each $b \in W^{3,2}(0, 1)$ with $|b|_{W^{1,2}} \leq \mu$ there exists a sequence $\tilde{b}^m \in B^m \cap B_P$ with $|\tilde{b}^m - b|_{W^{1,2}} \leq \rho_B(m)|b|_{W^{3,2}}$ and $\lim_{m \rightarrow \infty} \rho_B(m) = 0$ independently of b .

If B^m , $m = 1, 2, \dots$, is chosen as the space of linear spline functions with equidistant knots (i/m) , $i = 0, \dots, m$, then $\rho_B(m) = O(1/m^2)$. Since $|b|_\infty \leq d\mu$ for b with $|b|_{W^{1,2}} \leq \mu$, one can take $Q^m b \in B^m \cap B_P$, where Q^m denotes the quasi-interpolation operator (see [13, pp. 299, 230]); compare also Remark 3.5.

For $i = 1$ or 2 we consider the problems

$$(P_{i_h(m)}^{N,m}) \quad \min_{b \in B^m \cap B_P} (J^N(b) + h(m)\psi_B(b)).$$

By Remark 2.2 and Lemma 3.1 there exist solutions $\bar{b}_{h(m)}^{N,m}$ of $(P_{i_h(m)}^{N,m})$, if only (H1) and (H2) hold. Utilizing the regularity results for the convection coefficient given in § 4, we now improve the approximation result of Theorem 3.2.

THEOREM 5.1. *Assume that (H1), (H2) and (H4) hold, and that $\lim_{m \rightarrow \infty} h(m) = \infty$, $\lim_{m \rightarrow \infty} h(m)\rho_B(m)^2 = 0$. If b^* is any weak limit point of $\{\bar{b}_{h(m)}^{N,m}\}$, such that $\bar{b}_{h(m_k)}^{N,m_k} \rightharpoonup b^*$ in $W^{1,2}(0, 1)$ and $N_k \rightarrow \infty$, $m_k \rightarrow \infty$ as $k \rightarrow \infty$, then $b^* \in B$ is a solution of (P_i) , $u(\bar{b}_{h(m_k)}^{N,m_k}) \rightarrow u(b^*)$ in $W^{1,2}(0, 1)$ and $J^{N_k}(\bar{b}_{h(m_k)}^{N,m_k}) + h(m_k)\psi_B(\bar{b}_{h(m_k)}^{N,m_k}) \rightarrow J(b^*)$.*

Proof. Let \bar{b} be any solution of (P_i) , for $i = 1$ or 2 and determine \tilde{b}^{m_k} according to (H4). Then for all k the inequality

$$(5.1) \quad J^{N_k}(\bar{b}_{h(m_k)}^{N,m_k}) + h(m_k)\psi_B(\bar{b}_{h(m_k)}^{N,m_k}) \leq J^{N_k}(\tilde{b}^{m_k}) + h(m_k)\psi_B(\tilde{b}^{m_k})$$

holds. If $|\bar{b}| < \mu$, then $\lim_{k \rightarrow \infty} h(m_k) \psi_B(\tilde{b}^{m_k}) = 0$. Otherwise we have $h(m_k) \psi_B(\tilde{b}^{m_k}) = h(m_k)(|\tilde{b}^{m_k}|_{W^{1,2}} - \mu)^2 \leq h(m_k)|\tilde{b}^{m_k}|_{W^{1,2}} - |\bar{b}|_{W^{1,2}}|^2 \leq h(m_k)(|\tilde{b}^{m_k} - \bar{b}|_{W^{1,2}})^2 = h(m_k) \rho_B(m_k)^2 |\bar{b}|_{W^{2,2}}$, and again $\lim_{k \rightarrow \infty} h(m_k) \psi_B(\tilde{b}^{m_k}) = 0$.

Consequently the right-hand side of (5.1) converges to $J(\bar{b})$ as $k \rightarrow \infty$, and the left-hand side of (5.1) is bounded. This implies that $\lim_{k \rightarrow \infty} \psi_B(\bar{b}_{h(m_k)}^{N_k, m_k}) = 0$ and therefore $b^* \in B$. Since

$$(5.2) \quad J^{N_k}(\bar{b}_{h(m_k)}^{N_k, m_k}) \leq J^{N_k}(\bar{b}_{h(m_k)}^{N_k, m_k}) + h(m_k) \psi_B(\bar{b}_{h(m_k)}^{N_k, m_k}),$$

we have from (5.1)

$$(5.3) \quad J^{N_k}(\bar{b}_{h(m_k)}^{N_k, m_k}) \leq J^{N_k}(\tilde{b}^{m_k}) + h(m_k) \psi_B(\tilde{b}^{m_k}),$$

for all k . Corollary 2.1 implies that $u^{N_k}(\tilde{b}^{m_k}) \rightarrow u(\bar{b})$ and $u^{N_k}(\bar{b}_{h(m_k)}^{N_k, m_k}) \rightarrow u(b^*)$ in $W^{1,2}(0, 2)$ as $k \rightarrow \infty$. Therefore, taking the limit as $k \rightarrow \infty$ in (5.3), we have

$$(5.4) \quad J_i(b^*) \leq J_i(\bar{b}).$$

Since \bar{b} is an arbitrary solution of (P_i) , b^* is a solution as well.

The convergence of the states has already been discussed. The final claim concerning convergence of the penalized fit-to-data criteria follows from (5.2)–(5.4), where we note that (5.4) holds with the inequality replaced by an equality.

6. Numerical examples. In this section we give some numerical data for the identification of the coefficients b and c in (2.1) via the output least squares method; in these examples unconstrained minimization is not successful numerically, whereas realization of the norm constraint by a penalty functional leads to quite satisfactory fits of the “unknown” coefficient \bar{b} or \bar{c} . In all our examples the observation z was taken as the known analytical solution u corresponding to a parameter \bar{b} or \bar{c} as was used in the L^2 -fit-to-data criterion. It is our experience with earlier calculations that adding small amounts of noise to the data does not significantly change the numerical results and we have thus not considered this case here.

To solve the problems $(P_M^{N, m})$, we took as subspaces H^N cubic B -splines with an equidistant grid $\{i/N\}_{i=0}^N$ and modified to satisfy Dirichlet boundary conditions and for Q^m piecewise linear spline functions or again cubic B -splines, with grid $\{i/m\}_{i=0}^m$. Thus $\dim(H^N) = N + 1$ and $\dim(Q^m) = m + 1$ in the case of linear spline functions and $\dim(Q^m) = m + 3$ for cubic B -splines. The minimization was carried out by Newton's method with the derivatives of the discretized systems calculated analytically. The startup value for the unknown coefficient was chosen as the constant function 2 in all cases. In all graphs the *solid lines* represent the true solution, whereas the *dotted lines* represent the numerical approximation.

Example 6.1. Here we search for the coefficient $\bar{c} = \sqrt{2} \cos 2\pi x$ in

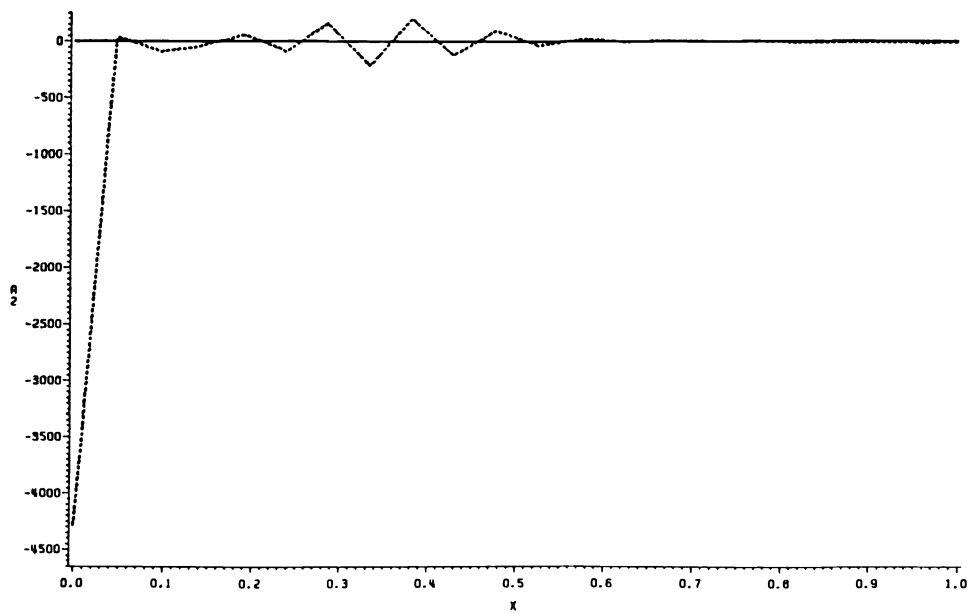
$$(6.1) \quad \begin{aligned} -4u_{xx} + \bar{c}u &= f \quad \text{in } (0, 1), \\ u(0) &= u(1) = 0, \end{aligned}$$

where $f(x) = (4\pi^2 + \sqrt{2} \cos 2\pi x) \sin \pi x$. The solution of (6.1) is given by $u(\bar{c}) = \sin \pi x$. The penalty functional, slightly different from the previous section, was taken to be

$$\psi(c) = \begin{cases} (|c|_{L^2}^2 - 1)^2 & \text{if } |c|_{L^2} > 1, \\ 0 & \text{if } |c|_{L^2} \leq 1. \end{cases}$$

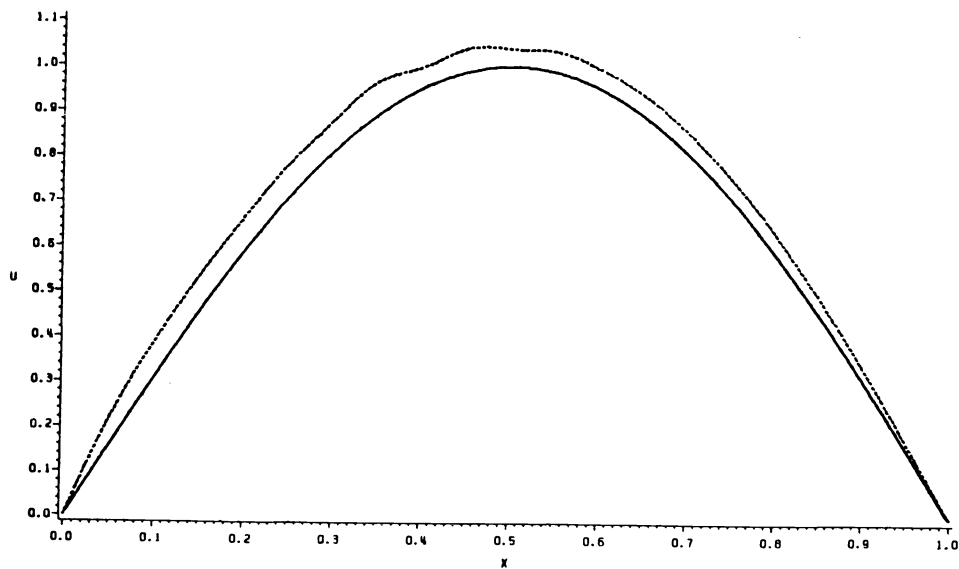
Thus the fit-to-data criterion is given by

$$|u^N(c^m) - z|^2 + M\psi(c^m)$$



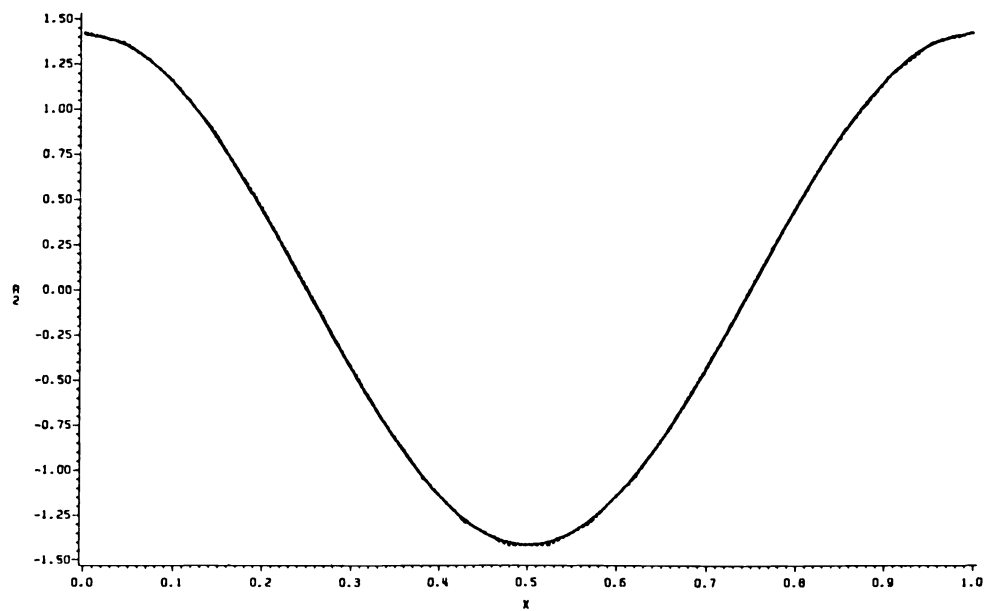
c ESTIMATE
 $c(x) = \sqrt{2} \cos 2\pi x, N = 20, m = 21, M = 0$

FIG. 1



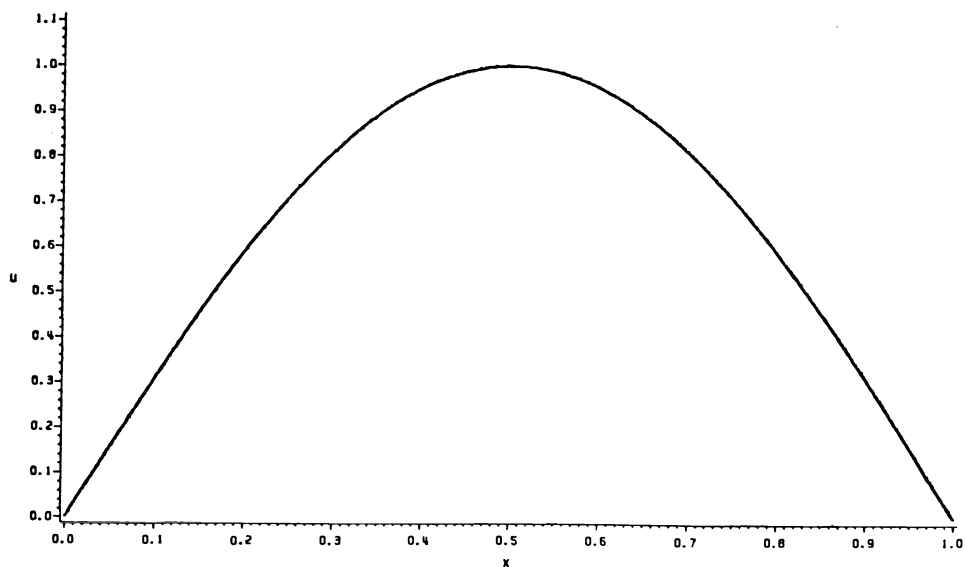
u ESTIMATE
 $u(x) = \sin \pi x, N = 20, m = 21, M = 0$

FIG. 2



c ESTIMATE
 $c(x) = \sqrt{2} \cos 2\pi x$, $N = 20$, $m = 21$, $M = .1$

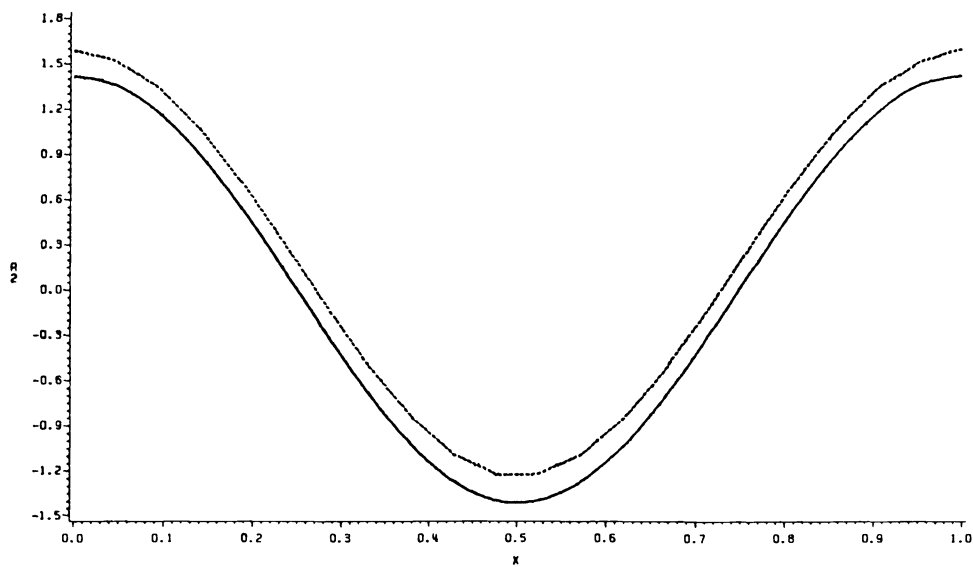
FIG. 3



u ESTIMATE
 $u(x) = \sin \pi x$, $N = 20$, $m = 21$, $M = .1$

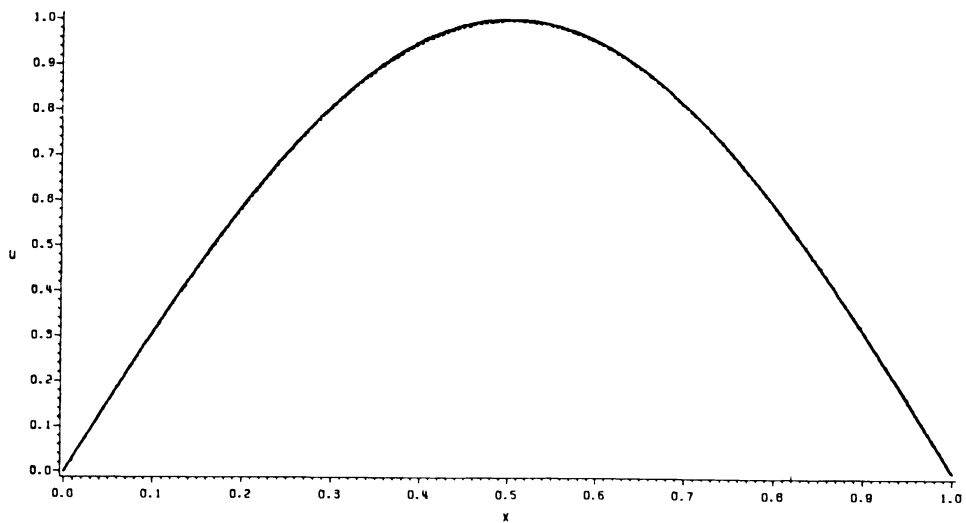
FIG. 4

with $z = u(\bar{c})$. In Figs. 1-6 we give the graphs of the converged values for $\bar{c}_M^{N,m}$ and $u^N(\bar{c}_M^{N,m})$ in the case that $N=20$, $m=21$ and for the values $M=0, .1$, and 10 . The graph for $M=1$ is indistinguishable from that for $M=.1$. The best fits are obtained for $M=.1$ and 1 ; for $M=10$ the result is less accurate, which might be attributed to the well-known numerical ill-conditioning of the penalty method for large values of



c ESTIMATE
 $c(x) = \sqrt{2} \cos 2\pi x, N = 20, m = 21, M = 10$

FIG. 5



u ESTIMATE
 $u(x) = \sin \pi x, N = 20, m = 21, M = 10$

FIG. 6

the penalty parameter M . The reader should observe that several different scales are used for the graphs.

Example 6.2. We now identify $\bar{b} = \sqrt{2} \cos \pi x$ in

$$(6.2) \quad \begin{aligned} -4u_{xx} + \bar{b}u_x + 4u &= f \quad \text{in } (0, 1), \\ u(0) &= u(1) = 0, \end{aligned}$$

where $f(x) = (4\pi^2 + 4) \sin \pi x + \sqrt{2}\pi \cos^2 \pi x$. The solution of (6.2) is given by $u(\bar{b}) = \sin \pi x$. As a penalty functional we took

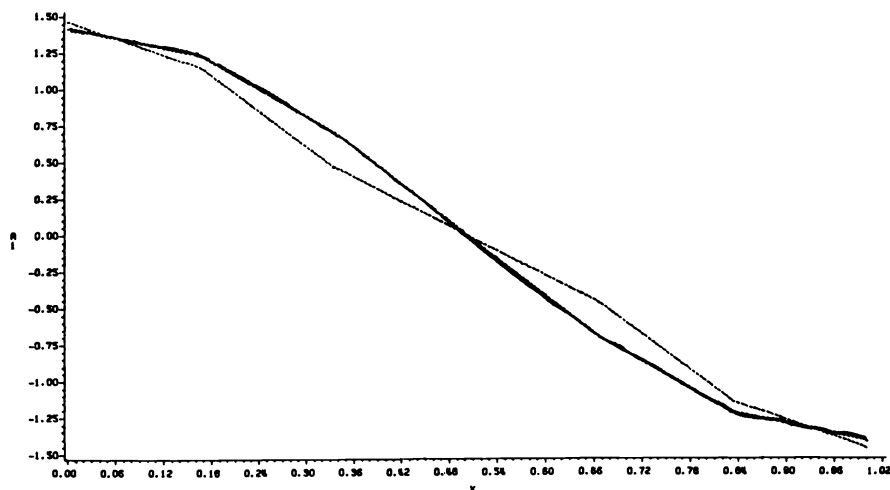
$$(6.3) \quad \psi(b) = \begin{cases} (|b|_{H^1}^2 - (1 + \pi^2))^2 & \text{if } |b|_{H^1} > 1 + \pi^2, \\ 0 & \text{if } |b|_{H^1} \leq 1 + \pi^2, \end{cases}$$

and the fit-to-data criterion is defined accordingly. In Figs. 7-9 we give the graphs of the converged values $\bar{b}_M^{N,m}$ for $N=10$, $M=.01$ and various values of m , when the approximating subspaces Q^m are taken as linear spline functions. We carried out the same calculations with M changed to be .001 and .1 and obtained almost identical results. It is remarkable to observe that a reasonable fit was obtained even when $m > N$. For $M=0$ the optimization algorithm did not converge. In Fig. 10 we show the result when instead of linear splines we take cubic B -splines for the space Q^m and choose $N=10$, $m=7$ and $M=.01$. For $N=10$, $m=4, 6$, or 9 and $M=.001$ or .1 the results we obtained were rather similar. For all the calculations shown in Figs. 7-12 the converged value $\bar{u}^N(b^m)$ gave a very good approximation to $u(\bar{b})$ and we therefore do not show the plots for these cases.

Finally we present the results that we obtained when the penalty functional (6.3) was replaced by

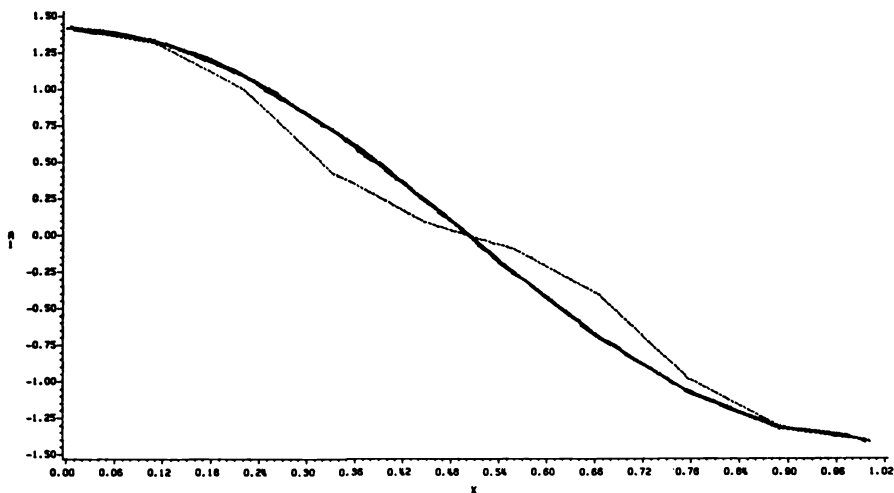
$$(6.4) \quad \psi(b) = \begin{cases} (|b|_{L^2}^2 - 1)^2 & \text{if } |b|_{L^2} > 1, \\ 0 & \text{if } |b|_{L^2} \leq 1. \end{cases}$$

Again Q^m was taken as the space of linear spline functions with knots at $\{i/m\}_{i=0}^m$.



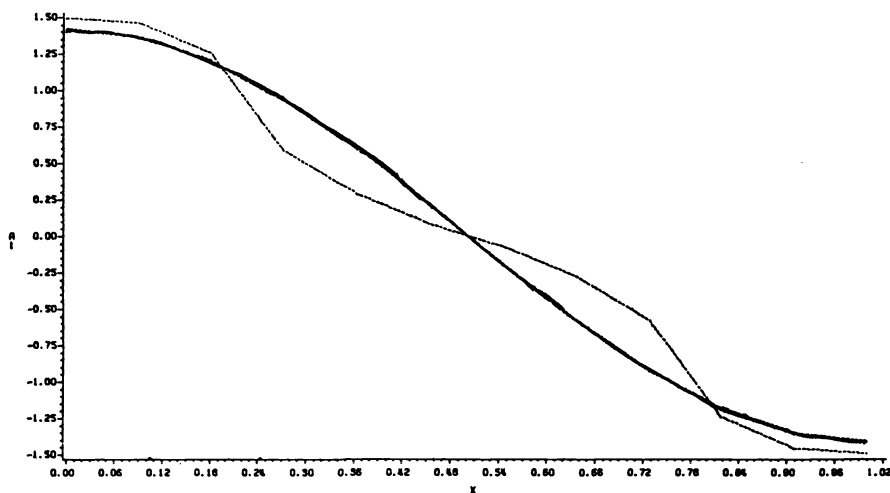
b ESTIMATE
 $b(x) = \sqrt{2} \cos(\pi x)$, $N=10$, $m=6$, $M=.01$

FIG. 7



b ESTIMATE (DASHED)
 $b(x) = \sqrt{2} \cos(\pi x)$, $N = 10$, $m = 9$, $M = .01$

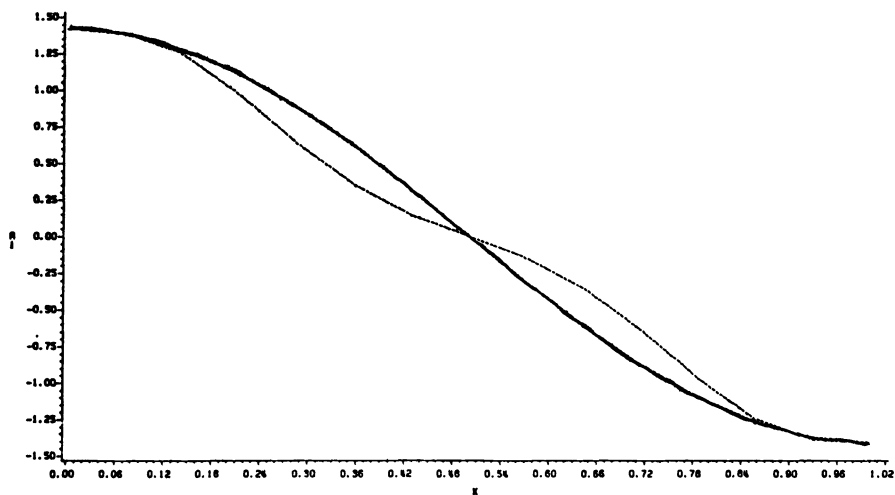
FIG. 8



b ESTIMATE
 $b(x) = \sqrt{2} \cos(\pi x)$, $N = 10$, $m = 11$, $M = .01$

FIG. 9

Figures 11 and 12 give the results for the case $N = 10$, $M = 1$ and $m = 6$ and 11. We made several other runs with different values for M and m . Throughout the penalty functional (6.3) employing the H^1 -norm produced better fits than the (6.4) criterion which uses the L^2 -norm.

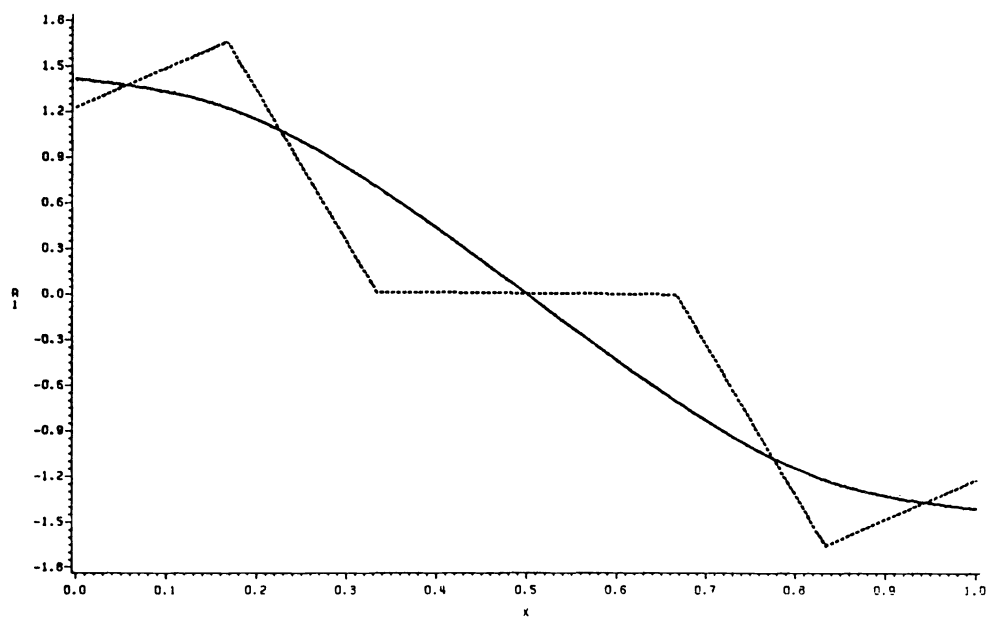


ESTIMATE

$$b(x) = \sqrt{2} \cos(\pi x), \quad N = 10, \quad m = 7, \quad M = .01$$

(cubic spline approximation)

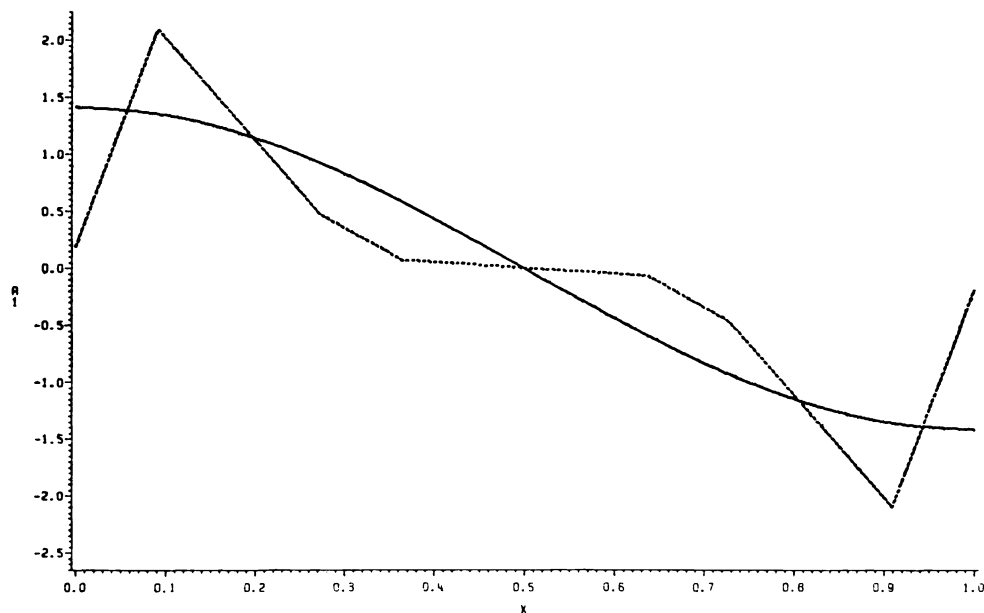
FIG. 10

 b ESTIMATE

$$b = \sqrt{2} \cos \pi x, \quad N = 10, \quad m = 6, \quad M = 1$$

(L^2 -penalty functional)

FIG. 11



ESTIMATE
 $b = \sqrt{2} \cos \pi x$, $N = 10$, $m = 9$, $M = 1$
 (L^2 -penalty functional)

FIG. 12

REFERENCES

- [1] R. A. ADAMS, *Sobolev Spaces*, Academic Press, New York, 1975.
- [2] H. T. BANKS, P. L. DANIEL AND E. S. ARMSTRONG, *Parameter estimation for static models of the Maypole hoop/column antenna surface*, Proc. IEEE International Large Scale Symposium, Virginia Beach, VA, 1982.
- [3] H. T. BANKS, J. M. CROWLEY AND K. KUNISCH, *Cubic spline approximation techniques for parameter estimation in distributed systems*, IEEE Trans. Automat. Control, 28 (1983), pp. 773-786.
- [4] F. COLONIUS AND K. KUNISCH, *Stability of parameter estimation in two point boundary value problem*, Zeitschrift für die reine und angewandte Mathematik, to appear.
- [5] R. FALK, *Error estimates for the numerical identification of a variable coefficient*, Math. Comp., 40 (1983), pp. 537-546.
- [6] H. G. HEUSER, *Functional Analysis*, John Wiley, New York, 1982.
- [7] C. KRAVARIS AND J. H. SEINFELD, *Identification of parameters in distributed parameter systems by regularization*, this Journal, 23 (1985), pp. 217-241.
- [8] K. KUNISCH AND L. W. WHITE, *Parameter estimation for elliptic equations in multidimensional domains with point and flux observations*, Nonlinear Anal. TMA, 10 (1986), pp. 121-146.
- [9] J. L. LIONS, *Optimal Control Systems Governed by Partial Differential Equations*, Springer, New York, 1971.
- [10] D. G. LUENBERGER, *Optimization by Vector Space Methods*, John Wiley, New York, 1969.
- [11] H. L. ROYDEN, *Real Analysis*, Macmillan, New York, 1963.
- [12] M. H. SCHULTZ, *Spline Analysis*, Prentice-Hall, Englewood Cliffs, NJ, 1973.
- [13] L. L. SCHUMAKER, *Spline Functions: Basic Theory*, John Wiley, New York, 1981.

THE LINEAR QUADRATIC CONTROL PROBLEM FOR INFINITE DIMENSIONAL SYSTEMS WITH UNBOUNDED INPUT AND OUTPUT OPERATORS*

A. J. PRITCHARD† AND D. SALAMON‡

Abstract. This paper establishes a general semigroup framework for solving quadratic control problems with infinite dimensional state space and unbounded input and output operators.

Key words. infinite dimensional systems, linear quadratic control, unbounded inputs and outputs, semigroups

AMS(MOS) subject classification. 93C25

1. Introduction. The object of this paper is to present a general semigroup theoretic framework for solving the linear quadratic control problem (LQCP) for systems with an infinite dimensional state space and unbounded input and output operators.

The LQCP has been one of the central research problems in the area of mathematical systems theory for more than twenty years. This is partly due to its beautiful mathematical structure. Furthermore, the LQCP provides a link between the area of optimal control and structure theory for linear control systems, and last, but not least, the infinite time quadratic cost problem leads to a numerically stable procedure for stabilizing a linear system by feedback.

For finite dimensional systems the LQCP is now well understood (see e.g. Willems [28], Wonham [29]) and a more or less complete generalization of the finite dimensional theory has been developed for infinite dimensional systems with bounded input and output operators (see e.g. Datko [6], Curtain and Pritchard [4], Lions [19], Gibson [10], Bensoussan, Delfour and Mitter [2], Zabczyk [30]).

In many dynamical systems, the control and observation processes are severely limited. For example there may be delays in the control actuators and measurement devices. Also for systems described by partial differential equations (PDE) it may not be possible to influence or sense the state at each point of the spatial domain. Instead controls and sensors are restricted to a few points or parts of the boundary. Modelling such limitations results in unbounded input and output operators. For infinite dimensional systems with unbounded input and output operators the LQCP has recently been studied by various authors. One of the first papers in this direction was by Lukes and Russell [20] and involved spectral operators. The classical reference for parabolic systems is of course the book of Lions [19]. His results have only recently been generalized to parabolic systems with a larger degree of unboundedness in the input and output operators (Pollock and Pritchard [22], Balakrishnan [1], Flandoli [9], Lasiecka and Triggiani [16], Sorine [26], [27]). The LQCP for first order hyperbolic PDE's has been studied by Russell [23]. Lasiecka and Triggiani [18] consider the higher dimensional wave equation with Dirichlet boundary control. In their paper the resulting optimal feedback operator is unbounded. For retarded systems with input delays we refer to Ichikawa [12] and Delfour [8] and for neutral systems with output delays to Datko [7] and Ito and Tarn [14].

* Received by the editors February 20, 1984, and in revised form October 7, 1985.

† Control Theory Centre, University of Warwick, England CV4 7AL.

‡ Forschungsschwerpunkt Dynamische Systeme, Universität Bremen, West Germany.

All of these papers deal with very specific classes of infinite dimensional systems—so far, no attempt has apparently been made to develop a general semigroup theoretic approach for the infinite dimensional LQCP with unbounded input and output operators which applies both to parabolic and hyperbolic PDE's as well as to retarded and neutral functional differential equations (FDE). In the present paper we fill this gap. An essential feature in our approach is that the semigroup $S(t)$ which describes the dynamics of the homogeneous equation is not assumed to have any smoothing properties. This is possible by means of the theory developed in Salamon [25, Chap. 1.3] and provides the basis for our approach to the LQCP.

In § 2 we solve the finite quadratic control problem in the general semigroup theoretic framework. In particular, we derive the existence of a unique-nonnegative solution $P(t)$ of the operator differential Riccati equation and we show that the unique optimal control is given by a time-varying feedback law involving this operator $P(t)$. We point out that the solution operator $P(t)$ of the Riccati equation has smoothing properties and that the associated feedback operator is bounded.

Section 3 is devoted to the infinite time problem and the solution is described in terms of the operator algebraic Riccati equation. The solution of the algebraic Riccati equation is derived as the limit operator of the solutions of the differential equation on the interval $[0, T]$ as T tends to infinity. Generalizing the results of Zabczyk [30], we establish relationships between the stabilizability and detectability properties of the system and existence and uniqueness results for the algebraic Riccati equation.

In § 4 we show how our general theory applies to parabolic and hyperbolic PDE's with boundary control as well as for neutral FDE's with output delays. For these special classes of infinite dimensional systems we do not derive substantial new results. We do, however, obtain a number of known results, which have not even been published, as simple straightforward consequences of our general theory. Another application of this theory to retarded FDE's with delays in control and observation will be the subject of a follow up paper.

2. Finite time control. In a formal sense our basic model is

$$(2.1) \quad \begin{aligned} \dot{x}(t) &= Ax(t) + Bu(t), & x(t_0) &= x_0, \\ y(t) &= Cx(t), & t_0 &\leq t \leq t_1, \end{aligned}$$

where $u(\cdot) \in L^2[t_0, t_1; U]$, $y(\cdot) \in L^2[t_0, t_1; Y]$, U and Y are Hilbert spaces and A is the infinitesimal generator of a strongly continuous semigroup $S(t)$ on a Hilbert space H . In order to allow for possible unboundedness of the operators B and C , we assume that $B \in \mathcal{L}(U, V)$, and $C \in \mathcal{L}(W, Y)$ where W, V are Hilbert spaces such that

$$(2.2) \quad W \subset H \subset V$$

with continuous dense injections. Of course, we interpret (2.1) in the mild form which means that its solution $x(t)$ is given by the variation-of-constants formula

$$(2.3a) \quad x(t) = S(t - t_0)x_0 + \int_{t_0}^t S(t - \sigma)Bu(\sigma) d\sigma, \quad t_0 \leq t \leq t_1.$$

In order to make this formula precise and to allow for trajectories in all three spaces W, H, V , we have to assume that $S(t)$ is also a strongly continuous semigroup on W and V and that the following hypotheses are satisfied.

(H1) There exists some constant $b > 0$ such that $\int_{t_0}^{t_1} S(t_1 - \sigma)Bu(\sigma) d\sigma \in W$ and $\|\int_{t_0}^{t_1} S(t_1 - \sigma)Bu(\sigma) d\sigma\|_W \leq b\|u(\cdot)\|_{L^2(t_0, t_1; U)}$ for every $u(\cdot) \in L^2(t_0, t_1; U)$.

(H2) There exists some constant $c > 0$ such that $\|CS(\cdot - t_0)x\|_{L^2(t_0, t_1; Y)} \leq c\|x\|_V$ for every $x \in W$.

Remarks 2.1. (i) Hypothesis (H1) implies that for every $x_0 \in W$ and every $u(\cdot) \in L^2(t_0, t_1; U)$ formula (2.3a) defines a continuous function $x(\cdot)$ on the interval $[t_0, t_1]$ with values in W . Hence the output can in this case be defined by

$$(2.3b) \quad y(t) = CS(t - t_0)x_0 + C \int_{t_0}^t S(t - \sigma)Bu(\sigma) d\sigma, \quad t_0 \leq t \leq t_1,$$

and is a continuous function on the interval $[t_0, t_1]$ with values in Y . If $x_0 \in V$, then $x(\cdot)$ is only a continuous function with values in V and (2.3b) does not make sense directly. But if (H2) is satisfied, then for any $x_0 \in V$ we will use the expression $CS(t - t_0)x_0$, $t_0 \leq t \leq t_1$, to denote the function in $L^2(t_0, t_1; Y)$ which is obtained by continuous extension to $x_0 \in V$ of the operator which maps $x_0 \in W$ into $CS(\cdot - t_0)x_0 \in L^2(t_0, t_1; Y)$. In this sense the right-hand side of (2.3b) is a well defined L^2 -function of t with values in Y .

(ii) In the above sense the expression $CS(t)Bu$ has a well defined meaning as a function of t for every $u \in U$. But the expression $CS(t - \sigma)Bu(\sigma)$ will in general not be a well defined function of σ . Therefore the operator C cannot be taken under the integral sign in (2.3b).

(iii) In the following we identify the Hilbert spaces H , U , Y with their duals. Then it follows from (2.2) by duality that

$$V^* \subset H \subset W^*$$

with continuous, dense injections. Furthermore, the adjoint semigroup $S^*(t)$ is a strongly continuous semigroup on all three spaces V^* , H , W^* .

(iv) Hypothesis (H1) is satisfied if and only if the following dual statement holds.

$$(H1^*) \quad \|B^*S^*(t_1 - \cdot)x\|_{L^2(t_0, t_1; U)} \leq b\|x\|_{W^*}, \quad x \in V^*.$$

This is a simple consequence of the identity

$$\left\langle x, \int_{t_0}^{t_1} S(t_1 - \sigma)Bu(\sigma) d\sigma \right\rangle_{V^*, V} = \int_{t_0}^{t_1} \langle B^*S^*(t_1 - \sigma)x, u(\sigma) \rangle_U d\sigma$$

for $x \in V^*$ and $u(\cdot) \in L^2(t_0, t_1; U)$ and the fact that $W^{**} = W$. Similarly, the dual statement of (H2) is the following.

(H2*) For every $y(\cdot) \in L^2(t_0, t_1; Y)$ we have

$$\left\| \int_{t_0}^{t_1} S^*(\tau - t_0)C^*y(\tau) d\tau \right\|_{V^*} \leq c\|y(\cdot)\|_{L^2(t_0, t_1; Y)}.$$

(v) In view of hypothesis (H1*) the expression $B^*S^*(t)x$ has a well defined meaning as an L^2 function of t for every $x \in W^*$, in particular when $x = C^*y$ with $y \in Y$.

Associated with the control system (2.3) is the performance index

$$(2.4) \quad J(u) = \langle x(t_1), Gx(t_1) \rangle_{V, V^*} + \int_{t_0}^{t_1} [\|Cx(t)\|_Y^2 + \langle u(t), Ru(t) \rangle_U] dt,$$

where $G \in \mathcal{L}(V, V^*)$ is a nonnegative definite operator and $R \in \mathcal{L}(U)$ satisfies

$$\langle u, Ru \rangle_U \geq \varepsilon \|u\|_U^2$$

for some $\varepsilon > 0$ and every $u \in U$.

Now let us consider system (2.3) with the feedback control

$$(2.5) \quad u_F(t) = F(t)x(t), \quad t_0 \leq t \leq t_1,$$

where $F(t) \in \mathcal{L}(V, U)$ is strongly continuous on the interval $[t_0, t_1]$. Then we may define a mild evolution operator $\Phi_F(t, s) \in \mathcal{L}(V)$, $t_0 \leq s \leq t \leq t_1$, via

$$(2.6) \quad \Phi_F(t, s)x = S(t-s)x + \int_s^t S(t-\sigma)BF(\sigma)\Phi_F(\sigma, s)x \, d\sigma$$

(see Curtain and Pritchard [4]).

Remarks 2.2. (i) It follows from (2.6) that $\Phi_F(t, s)$ satisfies the equation

$$(2.7) \quad \Phi_F(t, s)x - x = \int_s^t \Phi_F(t, \sigma)[A + BF(\sigma)]x \, d\sigma, \quad t_0 \leq s \leq t \leq t_1$$

for every $x \in \mathcal{D}_V(A)$ (the domain of A regarded as an unbounded, closed operator on V). Equivalently the function $s \rightarrow \Phi_F(t, s)x \in V$ is continuously differentiable on the interval $[t_0, t]$ for every $x \in \mathcal{D}_V(A)$ and satisfies

$$(2.8) \quad \frac{\partial \Phi_F(t, s)x}{\partial s} = -\Phi_F(t, s)[A + BF(s)]x, \quad t_0 \leq s \leq t \leq t_1$$

(see Curtain and Pritchard [4].)

(ii) It is well known that the evolution operator satisfies the equation

$$(2.9) \quad \Phi_F(t, s)x = S(t-s)x + \int_s^t \Phi_F(t, \sigma)BF(\sigma)S(\sigma-s)x \, d\sigma$$

for $t_0 \leq s \leq t \leq t_1$ and $x \in V$. (See Curtain and Pritchard [4].)

(iii) Often we will consider the feedback system with an additional forcing input $v(\cdot)$ so that

$$(2.10) \quad u(t) = F(t)x(t) + v(t)$$

in (2.3). It follows easily from (2.9) that—for this control function—the corresponding solution of (2.3) is given by

$$(2.11) \quad x(t) = \Phi_F(t, t_0)x_0 + \int_{t_0}^t \Phi_F(t, \sigma)Bv(\sigma) \, d\sigma, \quad t_0 \leq t \leq t_1.$$

(iv) Using (2.6), it is easy to see that $\Phi_F(t, s)$ is also a strongly continuous evolution operator on W and V and has the following properties.

(H1') There exists a constant $b' > 0$ such that

$$\left\| \int_{t_0}^t \Phi_F(t, \sigma)Bu(\sigma) \, d\sigma \right\|_W \leq b' \|u(\cdot)\|_{L^2(t_0, t; U)}$$

for every $u(\cdot) \in L^2(t_0, t_1; U)$ and every $t \in [t_0, t_1]$.

(H2') There exists a constant $c' > 0$ such that

$$\|C\Phi_F(\cdot, s)x\|_{L^2(s, t_1; Y)} \leq c' \|x\|_V$$

for every $x \in W$ and every $s \in [t_0, t_1]$

The dual properties are the following:

(H1')* The inequality

$$\|B^*\Phi_F^*(t, \cdot)x\|_{L^2(t_0, t; U)} \leq b' \|x\|_{W^*}$$

holds for every $x \in V^*$ and every $t \in [t_0, t_1]$.

(H2')* The inequality

$$\left\| \int_s^{t_1} \Phi_F^*(\tau, s) C^* y(\tau) d\tau \right\|_{V^*} \leq c' \|y(\cdot)\|_{L^2(s, t_1; Y)}$$

holds for every $y(\cdot) \in L^2(t_0, t_1; Y)$ and every $s \in [t_0, t_1]$.

Using the condition (H2') and its dual, we can define a strongly continuous operator $P_F(t) \in \mathcal{L}(V, V^*)$, by

$$(2.12) \quad \begin{aligned} P_F(t)x = & \Phi_F^*(t_1, t) G \Phi_F(t_1, t)x + \int_t^{t_1} \Phi_F^*(\tau, t) C^* C \Phi_F(\tau, t)x d\tau \\ & + \int_t^{t_1} \Phi_F^*(\tau, t) F^*(\tau) R F(\tau) \Phi_F(\tau, t)x d\tau \end{aligned}$$

for $t_0 \leq t \leq t_1$ and $x \in V$. Then the cost of the feedback control (2.5) corresponding to an initial state $x_0 \in V$ is given by

$$(2.13) \quad J(u_F) = \langle x_0, P_F(t_0)x_0 \rangle_{V, V^*}.$$

If the initial state is in H , then this expression can be interpreted via the inner product in H .

Remark 2.3. The adjoint operator of $P_F(t) \in \mathcal{L}(V, V^*)$ is still an operator from V to V^* and coincides with $P_F(t)$. In this sense one can say that $P_F(t)$ is self adjoint. Equivalently, the operator $i^{-1}P_F(t)$ on the Hilbert space V is self adjoint with respect to the inner product on V where $i: V \rightarrow V^*$ is the canonical isometric isomorphism. Finally, $P_F(t)$ is self adjoint in the above sense if and only if its restriction to H is a self adjoint operator on H .

A formula comparing the cost of an arbitrary control $u(\cdot) \in L^2(t_0, t_1; U)$ with the cost of the feedback control (2.5) will play an important role in our analysis. In the proof of this result we will need to interchange some integrals. At some points this becomes a delicate problem since we will have to operate with terms like $C\Phi_F(t, s)B$. In order to make the results precise, we need a third hypothesis.

(H3) Suppose that

$$Z = \mathcal{D}_V(A) \subset W$$

with a continuous, dense embedding where the Hilbert space Z is endowed with the graph norm of A , regarded as an unbounded, closed operator on V . This assumption is not very restrictive. It is satisfied by all known examples of systems which satisfy (H1) and (H2) if the spaces W and V are chosen appropriately. In the following we summarize some important consequences of (H3).

Remarks 2.4. (i) If (H3) is satisfied, then A can be regarded as a bounded operator from Z into V . Correspondingly A^* becomes a bounded operator from V^* into Z^* . On the other hand A can be restricted to a closed, densely defined operator on Z . Its adjoint in this sense coincides with the above operator $A^*: V^* \rightarrow Z^*$ (Salamon [25, Lemma 1.3.2]) and moreover

$$\mathcal{D}_{W^*}(A^*) \subset \mathcal{D}_{Z^*}(A^*) = V^*.$$

(ii) It is a well-known fact from semigroup theory that

$$T_t x = \int_0^t S(s)x ds \in \mathcal{D}_V(A) = Z$$

for every $x \in V$ and every $t \geq 0$. If (H3) is satisfied, then T_t is a strongly continuous family of bounded, linear operators from V into W . It is easy to see that the adjoint operator $T_t^* \in \mathcal{L}(W^*, V^*)$ is given by

$$T_t^* x = \int_0^t S^*(s)x \, ds \in \mathcal{D}_{W^*}(A^*) \subset V^*$$

for $x \in W^*$ and $t \geq 0$.

(iii) If (H1), (H2) and (H3) are satisfied, then the following equation holds for every $u \in U$ and every $t \geq 0$

$$C \int_0^t S(s)Bu \, ds = CT_t Bu = \int_0^t CS(s)Bu \, ds.$$

This seems like a trivial fact; however, we were not able to establish this identity without assuming (H3). Note that the LHS of the above equation has to be interpreted in terms of (H1) and the RHS in terms of (H2). For establishing the equation one must approximate $Bu \in V$ by a sequence of elements in W . Then the term on the LHS will not converge in general unless range $T_t \subset W$.

LEMMA 2.5. *Suppose that (H1), (H2), (H3) are satisfied; let $F(t) \in \mathcal{L}(V, U)$, $t_0 \leq t \leq t_1$, be strongly continuous and let $\Phi_F(t, s) \in \mathcal{L}(V) \cap \mathcal{L}(W)$ be defined by (2.6). Moreover, let $u(\cdot) \in L^2(t_0, t_1; U)$ and $y(\cdot) \in L^2(t_0, t_1; U)$ be given. Then*

$$(2.14) \quad \int_{t_0}^{t_1} \int_s^{t_1} \langle C\Phi_F(t, s)Bu(s), y(t) \rangle_Y \, dt \, ds = \int_{t_0}^{t_1} \left\langle C \int_{t_0}^t \Phi_F(t, s)Bu(s) \, ds, y(t) \right\rangle_Y \, dt$$

where the first expression must be interpreted in terms of (H2) and the second in terms of (H1).

Proof. First note that, by (2.6) and (H1), $\Phi_F(t, s) - S(t-s) \in \mathcal{L}(V, W)$. Hence it is enough to establish the desired equation with $\Phi_F(t, s)$ replaced by $S(t-s)$. Secondly it is easy to see that with $T_t \in \mathcal{L}(V, W)$ defined as in Remark 2.4(ii) the equations

$$\begin{aligned} x(t) &= \int_{t_0}^t S(t-s)Bu(s) \, ds = \int_{t_0}^t T_{t-s}Bu(s) \, ds \in W, \\ z(s) &= \int_s^{t_1} S^*(t-s)C^*y(t) \, dt = - \int_s^{t_1} T_{t-s}^*C^*y(t) \, dt \in V^* \end{aligned}$$

hold for $u(\cdot) \in \mathcal{C}^1[t_0, t_1; U]$ with $u(t_0) = 0$ and $y(\cdot) \in \mathcal{C}^1[t_0, t_1; Y]$ with $y(t_1) = 0$. Interchanging integrals, we obtain from these identities that

$$\begin{aligned} \int_{t_0}^{t_1} \langle Cx(t), y(t) \rangle_Y \, dt &= \int_{t_0}^t \langle Bu(s), z(s) \rangle_{V, V^*} \, ds \\ &= \int_{t_0}^{t_1} \int_s^{t_1} \langle CS(t-s)Bu(s), y(t) \rangle_Y \, dt \, ds. \end{aligned}$$

Now the statement of the lemma follows from the fact that both sides of this equation depend continuously on $u(\cdot) \in L^2[t_0, t_1; U]$ and $y(\cdot) \in L^2[t_0, t_1; Y]$.

Now we are in the position to prove the desired comparison formula for the feedback control (2.5).

LEMMA 2.6. *Suppose that (H1), (H2), (H3) are satisfied; let $F(t) \in \mathcal{L}(V, U)$ be strongly continuous on the interval $[t_0, t_1]$ and let $P_F(t) \in \mathcal{L}(V, V^*)$ be defined by (2.12)*

and (2.6). Then the following equation holds for every $x_0 \in V$ and every $u(\cdot) \in L^2(t_0, t_1; U)$

$$\begin{aligned}
 J(u) - \langle x_0, P_F(t_0)x_0 \rangle_{V, V^*} &= \int_{t_0}^{t_1} \langle R^{-1}B^*P_F(t)x(t) \\
 &\quad + u(t), R[R^{-1}B^*P_F(t)x(t) + u(t)] \rangle dt \\
 (2.15) \quad &- \int_{t_0}^{t_1} \langle R^{-1}B^*P_F(t)x(t) \\
 &\quad + F(t)x(t), R[R^{-1}B^*P_F(t)x(t) + F(t)x(t)] \rangle dt
 \end{aligned}$$

where $x(t)$, $t_0 \leq t \leq t_1$, is given by (2.3).

Proof. We sketch only the main steps of the proof for the case $x_0 \in W$. Let $x(t)$ be the mild solution of (2.1) given by (2.3) and define

$$\begin{aligned}
 v(t) &= u(t) - F(t)x(t), \\
 z(t) &= \int_{t_0}^t \Phi_F(t, s)Bv(s) ds = x(t) - \Phi_F(t, t_0)x_0
 \end{aligned}$$

for $t_0 \leq t \leq t_1$ (see Remark 2.2(iii)). Then applying Lemma 2.5, we can obtain

$$\begin{aligned}
 2\operatorname{Re} \int_{t_0}^{t_1} \int_s^{t_1} \langle C\Phi_F(t, s)z(s), C\Phi_F(t, s)Bv(s) \rangle dt ds \\
 = \operatorname{Re} \int_{t_0}^{t_1} \int_s^{t_1} \langle Cz(t), C\Phi_F(t, s)Bv(s) \rangle dt ds = \int_{t_0}^{t_1} \|Cz(t)\|_Y^2 dt
 \end{aligned}$$

and therefore, again using Lemma 2.5,

$$\begin{aligned}
 2\operatorname{Re} \int_{t_0}^{t_1} \int_s^{t_1} \langle C\Phi_F(t, s)x(s), C\Phi_F(t, s)Bv(s) \rangle dt ds \\
 = 2\operatorname{Re} \int_{t_0}^{t_1} \langle C\Phi_F(t_1, t_0)x_0, Cz(t) \rangle dt + \int_{t_0}^{t_1} \|Cz(t)\|_Y^2 dt \\
 = \int_{t_0}^{t_1} \|Cx(t)\|_Y^2 dt - \int_{t_0}^{t_1} \|C\Phi_F(t, t_0)x_0\|_Y^2 dt.
 \end{aligned}$$

Analogous identities can be derived in a more straightforward way when C^*C is replaced by $G \in \mathcal{L}(V, V^*)$ or $F^*(t)RF(t) \in \mathcal{L}(V, V^*)$. Using first (2.12) and then these identities, we get

$$\begin{aligned}
 2\operatorname{Re} \int_{t_0}^{t_1} \langle P_F(s)x(s), Bv(s) \rangle ds \\
 = \langle x(t_1), Gx(t_1) \rangle - \langle \Phi_F(t_1, t_0)x_0, G\Phi_F(t_1, t_0)x_0 \rangle \\
 + \int_{t_0}^{t_1} \|Cx(t)\|^2 dt - \int_{t_0}^{t_1} \|C\Phi_F(t, t_0)x_0\|^2 dt \\
 + \int_{t_0}^{t_1} \langle F(t)x(t), RF(t)x(t) \rangle dt \\
 - \int_{t_0}^{t_1} \langle F(t)\Phi_F(t, t_0)x_0, RF(t)\Phi_F(t, t_0)x_0 \rangle dt \\
 = J(u) - \langle x_0, P_F(t_0)x_0 \rangle - \int_{t_0}^{t_1} \langle u(t), Ru(t) \rangle dt
 \end{aligned}$$

$$+ \int_{t_0}^{t_1} \langle F(t)x(t), RF(t)x(t) \rangle dt.$$

It is easy to see that this equation implies (2.15).

We are now able to prove the main result of this section.

THEOREM 2.7. *Let (H1), (H2) and (H3) be satisfied. Then there exists a unique strongly continuous self adjoint, nonnegative operator $P(t) \in \mathcal{L}(V, V^*)$ $t_0 \leq t \leq t_1$, solving the integral Riccati equation.*

$$(2.16) \quad \begin{aligned} P(t)x = & \Phi^*(t_1, t)G\Phi(t_1, t)x \\ & + \int_t^{t_1} \Phi^*(s, t)[C^*C + P(s)BR^{-1}B^*P(s)]\Phi(s, t)x ds \end{aligned}$$

for $x \in W$ and $t_0 \leq t \leq t_1$ where $\Phi(s, t) = \Phi_F(s, t)$ is the evolution operator defined by (2.6) with $F(t) = -R^{-1}B^*P(t) \in \mathcal{L}(V, U)$. Furthermore there is a unique optimal control which minimizes the performance index (2.4) subject to (2.3). This optimal control is given by the feedback control law

$$(2.17) \quad u_F(t) = -R^{-1}B^*P(t)x(t)$$

and the optimal cost is

$$(2.18) \quad J(u_F) = \langle x_0, P(t_0)x_0 \rangle.$$

Proof. We regard (2.16) as a fixed point problem which is to be solved by iteration. Let us define the sequence $P_k(t) \in \mathcal{L}(V, V^*)$ recursively through

$$P_0(t) = 0, \quad P_k(t) = P_F(t), \quad F(t) = -R^{-1}B^*P_{k-1}(t)$$

for $k \in \mathbb{N}$ and $t_0 \leq t \leq t_1$, where $P_F(t)$ is given by (2.12). Let us also define

$$\Phi_k(s, t) = \Phi_F(s, t), \quad F(t) = -R^{-1}B^*P_k(t),$$

so that

$$(2.19) \quad \begin{aligned} \langle z, P_{k+1}(t)x \rangle_{V, V^*} = & \langle \Phi_k(t_1, t)z, G\Phi_k(t_1, t)x \rangle_{V, V^*} \\ & + \int_t^{t_1} \langle C\Phi_k(s, t)z, C\Phi_k(s, t)x \rangle_Y ds \\ & + \int_t^{t_1} \langle B^*P_k(s)\Phi_k(s, t)z, RB^*P_k(s)\Phi_k(s, t)x \rangle_U ds \end{aligned}$$

holds for $t_0 \leq t \leq t_1$ and $x \in W$. Applying Lemma 2.6 to $F(t) = -R^{-1}B^*P_{k-1}(t)$ and $u_k(t) = -R^{-1}B^*P_k(t)x(t)$, we obtain

$$(2.20) \quad \begin{aligned} \langle x_0, P_{k+1}(t_0)x_0 \rangle = & J(u_k) = \langle x_0, P_k(t_0)x_0 \rangle \\ & - \int_{t_0}^{t_1} \langle [P_k(\tau) - P_{k-1}(\tau)]x(\tau), BR^{-1}B^* \\ & \cdot [P_k(\tau) - P_{k-1}(\tau)]x(\tau) \rangle d\tau \\ & \leq \langle x_0, P_k(t_0)x_0 \rangle \end{aligned}$$

for $k \in \mathbb{N}$ and $x_0 \in V$. Thus the sequence $\langle x_0, P_k(t_0)x_0 \rangle_{V, V^*}$, $k \in \mathbb{N}$, is monotonically decreasing and positive. Applying Kato's result [15, p. 454, Thm. 3.3] to the monotonically decreasing sequence of nonnegative operators $i^{-1}P_k(t_0)$ on V where $i: V \rightarrow V^*$ is

the canonical isomorphism (Remark 2.3), we obtain the strong convergence of this sequence to a nonnegative limit operator on V . Hence the operators $P_k(t_0) \in \mathcal{L}(V, V^*)$ converge strongly to a nonnegative self adjoint operator $P(t_0) \in \mathcal{L}(V, V^*)$. The same conclusion is valid for every $t \in [t_0, t_1]$ since $t_0 \leq t_1$ can be chosen arbitrarily.

Moreover, (2.20) shows that the operators $P_k(t) \in \mathcal{L}(V, V^*)$, $t_0 \leq t \leq t_1$, $k \in \mathbb{N}$ are uniformly bounded. Hence the limit operator $P(t) \in \mathcal{L}(V, V^*)$ is strongly measurable and uniformly bounded on the interval $[t_0, t_1]$. Therefore we can introduce a strongly continuous evolution operator $\Phi(s, t) = \Phi_F(s, t) \in \mathcal{L}(V) \cap \mathcal{L}(W)$ which is defined by (2.6) with $F(t) = -R^{-1}B^*P(t)$.

Our next step is to show that the function $\Phi_k(\cdot, t)x - \Phi(\cdot, t)x \in \mathcal{C}[t, t_1; W]$ converges to zero in the sup-norm for every $x \in V$ and every $t \in [t_0, t_1]$. For this purpose let us consider the identity

$$\begin{aligned} \Phi(s, t)x - \Phi_k(s, t)x &= \int_t^s S(s-\tau)BR^{-1}B^*[P_k(\tau) - P(\tau)]\Phi(\tau, t)x d\tau \\ &\quad - \int_t^s S(s-\tau)BR^{-1}B^*P_k(\tau)[\Phi(\tau, t)x - \Phi_k(\tau, t)x] d\tau \end{aligned}$$

and apply Gronwall's lemma. Then the desired convergence of $\Phi_k(s, t)x$ follows from the pointwise strong convergence of $P_k(\tau)$ to $P(\tau)$ together with the dominated convergence theorem.

As a consequence of this convergence result we obtain that $\Phi_k(s, t)$ converges to $\Phi(s, t)$ both in $\mathcal{L}(V)$ and in $\mathcal{L}(W)$ and that this convergence is uniform for $t \leq s \leq t_1$ (t fixed). This allows us to apply the dominated convergence theorem to formula (2.19) and hence $P(t)$ satisfies the integral Riccati equation (2.16). Finally it follows easily from (2.16) together with the strong continuity of $\Phi(s, t)$ and $\Phi^*(s, t)$ in both variables and in both spaces V and W that the operator $P(t) \in \mathcal{L}(V, V^*)$ is strongly continuous on the interval $[t_0, t_1]$. Thus we have proved the existence of a solution to (2.16).

In order to prove the uniqueness for the solution of (2.16) together with the statements on the optimal control, let us assume that $P(t) \in \mathcal{L}(V, V^*)$ is any strongly continuous, nonnegative solution of (2.16). Moreover, let $x_0 \in V$, $u(\cdot) \in L^2(t_0, t_1; U)$ be given, let $x(t) \in V$ be the corresponding solution of (2.1) which is given by (2.3) and define $v(t) = u(t) + R^{-1}B^*P(t)x(t)$ for $t_0 \leq t \leq t_1$. Then it follows from Lemma 2.6 that

$$(2.21) \quad J(u) = \langle x_0, P(t_0)x_0 \rangle + \int_{t_0}^{t_1} \langle v(t), Rv(t) \rangle dt.$$

Hence the optimal control is unique and given by the feedback law (2.17) and the optimal cost is given by (2.18). Moreover, we conclude from (2.21) that $\langle x_0, P(t_0)x_0 \rangle = \langle x_0, \hat{P}(t_0)x_0 \rangle$ for any two nonnegative solutions $P(t)$, $\hat{P}(t) \in \mathcal{L}(V, V^*)$ of (2.16) and any $x_0 \in V$. Since $t_0 \leq t_1$ can be chosen arbitrarily, this proves the uniqueness of the solution to (2.16). \square

The following result shows that the integral Riccati equation (2.16) can be converted into a differential Riccati equation.

PROPOSITION 2.8. *Suppose that (H1), (H2) and (H3) are satisfied and let $P(t) \in \mathcal{L}(V, V^*)$ be a nonnegative, self adjoint, strongly continuous operator on the interval $[t_0, t_1]$. Moreover, let the evolution operator $\Phi(s, t) = \Phi_F(s, t) \in \mathcal{L}(V)$ be defined by (2.6) with $F(t) = -R^{-1}B^*P(t)$. Then the following statements are equivalent.*

- (i) *Equation (2.16) holds for every $x \in W$ and every $t \in [t_0, t_1]$.*

(ii) For every $x \in W$ and every $t \in [t_0, t_1]$ the following equation holds

$$(2.22) \quad P(t)x = \Phi^*(t_1, t)GS(t_1 - t)x + \int_t^{t_1} \Phi^*(s, t)C^*CS(s - t)x \, ds.$$

(iii) For every $x \in W$ and every $t \in [t_0, t_1]$ the following equation holds:

$$(2.23) \quad \begin{aligned} P(t)x &= S^*(t_1 - t)GS(t_1 - t)x \\ &+ \int_t^{t_1} S^*(s - t)[C^*C - P(s)BR^{-1}B^*P(s)]S(s - t)x \, ds. \end{aligned}$$

(iv) For every $x \in Z$ the function $P(t)x$, $t_0 \leq t \leq t_1$ is continuously differentiable with values in Z^* and satisfies the differential Riccati equation

$$(2.24a) \quad \frac{d}{dt}P(t)x + A^*P(t)x + P(t)Ax - P(t)BR^{-1}B^*P(t)x + C^*Cx = 0,$$

$$(2.24b) \quad P(t_1)x = Gx.$$

In this equation A is regarded as a bounded operator from Z into V .

Proof. The equivalence of the statements (i), (ii) and (iii) can be established in a straightforward way using the formulae (2.6) and (2.9) together with Lemma 2.5.

In order to prove that (iii) implies (iv), note that the equation

$$(2.25) \quad \langle CS(t)z, CS(t)x \rangle - \langle Cz, Cx \rangle = \int_0^t [\langle CS(s)Az, CS(s)x \rangle + \langle CS(s)z, CS(s)Ax \rangle] \, ds$$

holds for all $x, z \in \mathcal{D}_W(A)$ and every $t \geq 0$. It follows from (H3) and (H2) that both sides of this equation depend continuously on $x, z \in Z = \mathcal{D}_V(A) \subset W$ and that $\mathcal{D}_Z(A) \subset \mathcal{D}_W(A) \subset Z$. Consequently $\mathcal{D}_W(A)$ is dense in Z and hence (2.25) holds for all $x, z \in Z$.

From (2.25) we see that the function $\langle z, P(t)x \rangle$ —defined by (2.23)—is continuously differentiable on the interval $[t_0, t_1]$ for all $x, z \in Z$ and satisfies the equation

$$\begin{aligned} \frac{d}{dt} \langle z, P(t)x \rangle &= -\langle S(t_1 - t)Az, GS(t_1 - t)x \rangle - \langle S(t_1 - t)z, GS(t_1 - t)Ax \rangle \\ &\quad - \langle Cz, Cx \rangle + \langle z, P(t)BR^{-1}B^*P(t)x \rangle \\ &\quad - \int_t^{t_1} [\langle CS(s - t)Az, CS(s - t)x \rangle \\ &\quad \quad + \langle CS(s - t)z, CS(s - t)Ax \rangle] \, ds \\ &\quad + \int_t^{t_1} \langle S(s - t)Az, P(s)BR^{-1}B^*P(s)S(s - t)x \rangle \, ds \\ &\quad + \int_t^{t_1} \langle S(s - t)z, P(s)BR^{-1}B^*P(s)S(s - t)Ax \rangle \, ds \\ &= -\langle Az, P(t)x \rangle - \langle z, P(t)Ax \rangle \\ &\quad - \langle Cz, Cx \rangle + \langle z, P(t)BR^{-1}B^*P(t)x \rangle. \end{aligned}$$

This implies

$$\langle z, P(t)x \rangle_{Z, Z^*} = \left\langle z, Gx + \int_t^{t_1} [A^*P(s)x + P(s)Ax - P(s)BR^{-1}B^*P(s)x + C^*Cx] ds \right\rangle_{Z, Z^*}$$

and hence (2.24a). Thus we have proved that (iii) implies (iv).

Conversely, let us assume that $P(t)$ satisfies (2.24). Then the following equation holds for every $x \in Z$ and every $t \in [t_0, t_1]$

$$\begin{aligned} S^*(t_1 - t)GS(t_1 - t)x - P(t)x &= \int_t^{t_1} \frac{d}{ds} S^*(s - t)P(s)S(s - t)x ds \\ &= \int_t^{t_1} S^*(s - t)[\dot{P}(s) + A^*P(s) + P(s)A]S(s - t)x ds \\ &= \int_t^{t_1} S^*(s - t)[C^*C - P(s)BR^{-1}B^*P(s)]S(s - t)x ds \end{aligned}$$

where the integral has to be understood in the Hilbert space Z^* and $\dot{P}(t)$ is the strong derivative of $P(t)$, $t_0 \leq t \leq t_1$ regarded as an operator in $\mathcal{L}(Z, Z^*)$. \square

3. Infinite time control. In this section we consider the control problem of minimizing the performance index

$$(3.1) \quad J(u) = \int_0^\infty [\|y(t)\|_Y^2 + \langle u(t), Ru(t) \rangle_U] dt$$

where $y(t)$ is again the output of (2.1) with $t_0 = 0$, i.e.

$$(3.2) \quad y(t) = CS(t)x_0 + C \int_0^t S(t-s)Bu(s) ds, \quad t \geq 0.$$

For this infinite time problem it is not clear that the cost will be finite for any control input $u(\cdot) \in L^2(0, \infty; U)$. So we add this as another hypothesis.

(H4) For every $x_0 \in V$ there exists a $u_{x_0}(\cdot) \in L^2[0, \infty; U]$ such that $J(u_{x_0}) < \infty$.

We will derive the optimal control via the solution of an *algebraic Riccati equation* which is actually the stationary version of (2.24). For this sake we consider the finite time control problems of minimizing the cost functionals.

$$(3.3) \quad J_T(u) = \int_0^T [\|y(t)\|_Y^2 + \langle u(t), Ru(t) \rangle_U] dt$$

subject to the constraint (3.2). The corresponding Riccati operator will be denoted by $P_T(t) \in \mathcal{L}(V, V^*)$ and satisfies the equation

$$(3.4) \quad P_T(t)x = \int_t^T S^*(s - t)[C^*C - P_T(s)BR^{-1}B^*P_T(s)]S(s - t)x ds$$

for every $x \in W$ and every $t \in [0, T]$.

LEMMA 3.1.

$$P_{T-\alpha}(t) = P_T(t + \alpha), \quad 0 \leq t \leq T - \alpha.$$

Proof. The operator $P_T(t + \alpha)$ satisfies the equation

$$\begin{aligned} P_T(t + \alpha)x &= \int_{t+\alpha}^T S^*(s - t - \alpha)[C^*C - P_T(s)BR^{-1}B^*P_T(s)]S(s - t - \alpha)x \, ds \\ &= \int_t^{T-\alpha} S^*(s - t)[C^*C - P_T(s + \alpha)BR^{-1}B^*P_T(s + \alpha)]S(s - t)x \, ds \end{aligned}$$

for $x \in W$ and $0 \leq t \leq T - \alpha$. Thus the statement of the lemma follows from the equivalence of (2.16) and (2.23) (Proposition 2.8) together with the uniqueness result (Theorem 2.7). \square

We will derive the solution of the *algebraic Riccati equation* as the limit of the solutions to *integral Riccati equations* as T goes to infinity. For this we need the following preliminary result which is a special case of Proposition 2.8.

COROLLARY 3.2. *Suppose that the hypotheses (H1), (H2) and (H3) are satisfied and let $P \in \mathcal{L}(V, V^*)$ be a nonnegative, self adjoint operator. Moreover, let $S_p(t) \in \mathcal{L}(V) \cap \mathcal{L}(W)$ be the strongly continuous semigroup which is generated by $A - BR^{-1}B^*P: \mathcal{D}_V(A) \rightarrow V$, i.e. $S_p(t)$ satisfies the equation*

$$(3.5) \quad S_p(t)x = S(t)x - \int_0^t S(t-s)BR^{-1}B^*PS_p(s)x \, ds$$

for $x \in V$ and $t \geq 0$. Then the following statements are equivalent.

(i) For every $x \in W$ and every $t \geq 0$

$$(3.6) \quad Px = S_p^*(t)PS_p(t)x + \int_0^t S_p^*(s)[C^*C + PBR^{-1}B^*P]S_p(s)x \, ds.$$

(ii) For every $x \in W$ and every $t \geq 0$

$$(3.7) \quad Px = S_p^*(t)PS(t)x + \int_0^t S_p^*(s)C^*CS(s)x \, ds.$$

(iii) For every $x \in W$ and every $t \geq 0$

$$(3.8) \quad Px = S^*(t)PS(t)x + \int_0^t S^*(s)[C^*C - PBR^{-1}B^*P]S(s)x \, ds.$$

(iv) For every $x \in Z$ the following equation holds in Z^*

$$(3.9) \quad A^*Px + PAx - PBR^{-1}B^*Px + C^*Cx = 0.$$

Now we are in the position to prove the main result of this section.

THEOREM 3.3. *Let (H1), (H2) and (H3) be satisfied. Then the following statements hold.*

(i) *The hypothesis (H4) is satisfied if and only if there exists a nonnegative self adjoint solution $P \in \mathcal{L}(V, V^*)$ of (3.9).*

(ii) *If (H4) is satisfied, then there exists a unique optimal control $u_p(\cdot) \in L^2(0, \infty; U)$ which is given by the feedback law.*

$$(3.10) \quad u_p(t) = -R^{-1}B^*Px(t), \quad t \geq 0,$$

where $P \in \mathcal{L}(V, V^*)$ is the (unique) minimal solution of (3.9). Moreover, the optimal cost is given by

$$(3.11) \quad J(u_p) = \langle x_0, Px_0 \rangle.$$

(iii) *If (H4) is satisfied, then the minimal solution $P \in \mathcal{L}(V, V^*)$ of (3.9) is strong limit of $P_T(0) \in \mathcal{L}(V, V^*)$ as T goes to infinity where $P_T(t)$ is defined by (3.4).*

Proof. First recall that the optimal control of the finite time problem on the interval $[0, T]$ is given by $u_T(t) = -R^{-1}B^*P_T(t)x(t)$, $0 \leq t \leq T$, and the optimal cost by $J_T(u_T) = \langle x_0, P_T(0)x_0 \rangle$ (Theorem 2.7). So (H4) implies that

$$\langle x_0, P_T(0)x_0 \rangle = J_T(u_T) \leq J_T(u_{x_0}) \leq J(u_{x_0}) < \infty$$

and thus there exists a limit of the increasing function $\langle x_0, P_T(0)x_0 \rangle$, $T \geq 0$, for every $x_0 \in V$. Hence there exists a nonnegative, self adjoint operator $P \in \mathcal{L}(V, V^*)$ which is the strong limit of $P_T(0)$ (Kato [15, p. 454, Thm. 3.3], compare the proof of Theorem 2.7).

By Lemma 3.1,

$$(3.12) \quad Px = s - \lim_{T \rightarrow \infty} P_T(t)x \in V^*$$

exists uniformly in t on every compact time interval. Making use of formula (3.4), we obtain for $x \in W$ and $t \geq 0$

$$\begin{aligned} Px &= \lim_{T \rightarrow \infty} P_T(0)x \\ &= \lim_{T \rightarrow \infty} \int_0^T S^*(s)[C^*C - P_T(s)BR^{-1}B^*P_T(s)]S(s)x \, ds \\ &= \lim_{T \rightarrow \infty} \int_t^T S^*(t)S^*(s-t)[C^*C - P_T(s)BR^{-1}B^*P_T(s)]S(s-t)S(t)x \, ds \\ &\quad + \lim_{T \rightarrow \infty} \int_0^t S^*(s)[C^*C - P_T(s)BR^{-1}B^*P_T(s)]S(s)x \, ds \\ &= \lim_{T \rightarrow \infty} S^*(t)P_T(t)S(t)x + \int_0^t S^*(s)[C^*C - PBR^{-1}B^*P]S(s)x \, ds \\ &= S^*(t)PS(t)x + \int_0^t S^*(s)[C^*C - PBR^{-1}B^*P]S(s)x \, ds \end{aligned}$$

and hence $P \in \mathcal{L}(V, V^*)$ is a solution of (3.6), (3.7), (3.8) and (3.9).

Conversely, let $Q \in \mathcal{L}(V, V^*)$ be any nonnegative solution of (3.9) and let $u_Q(t) = -R^{-1}B^*Qx(t)$ be the corresponding feedback control law with the associated closed loop semigroup $S_Q(t) \in \mathcal{L}(V) \cap \mathcal{L}(W)$. Then the following inequality holds for every $x_0 \in V$

$$\begin{aligned} (3.13) \quad \langle x_0, Qx_0 \rangle &= \lim_{t \rightarrow \infty} \left\{ \langle S_Q(t)x_0, QS_Q(t)x_0 \rangle \right. \\ &\quad \left. + \int_0^t \langle S_Q(s)x_0, [C^*C + QBR^{-1}B^*Q]S_Q(s)x_0 \rangle \, ds \right\} \\ &\geq \int_0^\infty \langle S_Q(t)x_0, [C^*C + QBR^{-1}B^*Q]S_Q(s)x_0 \rangle \, ds \\ &= J(u_Q) \end{aligned}$$

and hence (H4) is satisfied. Moreover, the operator $P \in \mathcal{L}(V, V^*)$ defined by (3.12) satisfies the inequality

$$\langle x_0, Px_0 \rangle = \lim_{T \rightarrow \infty} \langle x_0, P_T(0)x_0 \rangle \leq \lim_{T \rightarrow \infty} J_T(u) = J(u)$$

for every admissible control $u(\cdot) \in L^2(0, \infty; U)$. This shows that P is the minimal positive semidefinite solution of (3.6). Finally, taking $Q = P$, we conclude that the unique optimal control is given by (3.10) with cost (3.11). \square

Although the above theorem yields a solution to the infinite time problem, in a sense it is unsatisfactory. This is because we are not sure of a unique solution to the algebraic Riccati equation and also we cannot be sure that the semigroup $S_p(t)$ is exponentially stable. In order to resolve those difficulties, we need another hypothesis.

(H5) If $x_0 \in V$ and $u(\cdot) \in L^2(0, \infty; U)$ are such that $J(u) < \infty$, then $x(\cdot) \in L^2(0, \infty; V)$ where $x(t)$, $t \geq 0$, is given by (2.3) with $t_0 = 0$.

THEOREM 3.4. *Let (H1), (H2), (H3) and (H5) be satisfied. Then the algebraic Riccati equation (3.9) has at most one nonnegative, self adjoint solution $P \in \mathcal{L}(V, V^*)$. Moreover, if P is such a solution, then the closed loop semigroup $S_p(t) \in \mathcal{L}(V)$ is exponentially stable.*

Proof. If $P \in \mathcal{L}(V, V^*)$ is a positive semidefinite solution of (3.9), then the inequality (3.13) with $Q = P$ shows that the closed loop control $u_p(t) = -R^{-1}B^*Px(t)$ has a finite cost for every initial state $x_0 \in V$. By hypothesis (H5) this means that

$$\int_0^\infty \|S_p(t)x_0\|_V^2 dt < \infty$$

for every $x_0 \in V$. Hence it follows from a result of Datko [5] that the semigroup $S_p(t) \in \mathcal{L}(V)$ is exponentially stable (see Curtain and Pritchard [4]). The stability of $S_p(t)$ shows that we have equality in (3.13) and hence

$$J(u_p) = \langle x_0, Px_0 \rangle.$$

Now let $Q \in \mathcal{L}(V, V^*)$ be another nonnegative solution of (3.9) and let us apply Lemma 2.6 to the performance index

$$J_{T,Q}(u) = \langle x(T), Qx(T) \rangle + \int_0^T [\|y(t)\|_Y^2 + \langle u(t), Ru(t) \rangle] dt$$

as well as the feedback $F(t) = -R^{-1}B^*Q$ and the control input $u_p(t)$. Then $P_F(t) \equiv Q$ and hence the inequality

$$\begin{aligned} \langle x_0, Px_0 \rangle &= J(u_p) = \lim_{T \rightarrow \infty} J_{T,Q}(u_p) \\ &= \lim_{T \rightarrow \infty} \left[\langle x_0, Qx_0 \rangle + \int_0^T \langle R^{-1}B^*Qx(t) + u_p(t), R[R^{-1}B^*Qx(t) + u_p(t)] \rangle dt \right] \\ &\equiv \langle x_0, Qx_0 \rangle \end{aligned}$$

holds for every $x_0 \in V$. Interchanging the roles of P and Q , we conclude that $P = Q$. \square

Finally, let us briefly discuss the hypotheses (H4) and (H5) which are chosen in a general sense but are difficult to check in concrete examples. In most cases it might be desirable to replace them by stronger assumptions which are easier to check.

Remarks 3.5. Let (H1) and (H2) be satisfied.

(i) Suppose that system (2.1) is *stabilizable* in the sense that there exists a feedback operator $F \in \mathcal{L}(V, U)$ such that the closed loop semigroup $S_F(t) \in \mathcal{L}(V)$ defined by

$$S_F(t)x = S(t)x + \int_0^t S(t-s)BFS_F(s)x ds$$

for $t \geq 0$ and $x \in V$ is exponentially stable. Then hypothesis (H4) is satisfied.

In fact, there is an instant $T > 0$ and a constant $c_T > 0$ such that the inequalities

$$\|S_F(T)\|_{\mathcal{L}(V)} < 1, \quad \|CS_F(\cdot)x\|_{L^2(0,T;Y)} \leq c_T \|x\|_V$$

hold for every $x \in W$. This implies that

$$\|CS_F(\cdot)x\|_{L^2(0,\infty;Y)} \leq c_T \sum_{k=0}^{\infty} \|S_F(T)\|_{\mathcal{L}(V)}^k \|x\|_V$$

for $x \in W$ and hence (H4) is satisfied.

(ii) Suppose that system (2.1) is *detectable* in the sense that there exists an operator $K \in \mathcal{L}(Y, V)$ such that the output injection semigroup $S_K(t) \in \mathcal{L}(V)$ defined by

$$S_K(t)x = S(t)x + \int_0^t S_K(t-s)KCS(s)x \, ds$$

for $t \geq 0$ and $x \in W$ (see Salamon [25, Thm. I.3.9]) is exponentially stable. Then hypothesis (H5) is satisfied.

In fact, if $x(t) \in V$ and $y(t) \in Y$ are defined by (2.3) for $x_0 \in V$ and $u(\cdot) \in L^2_{\text{loc}}(0, \infty; U)$, then it is easy to see

$$x(t) = S_K(t)x_0 + \int_0^t S_K(t-s)[Bu(s) - Ky(s)] \, ds, \quad t \geq 0.$$

Hence $J(u) < \infty$ implies that $x(\cdot) \in L^2(0, \infty; V)$.

(iii) If (H4) and (H5) are satisfied, then system (2.1) is stabilizable in the sense of (i). (Theorems 3.3 and 3.4.)

(iv) For finite dimensional systems (H5) is equivalent to detectability in the sense of (ii). It seems to be an open problem whether this equivalence extends to the infinite dimensional situation.

4. Examples.

4.1. Neutral systems with output delays. We consider the linear neutral functional differential equation (NFDE)

$$(4.1) \quad \frac{d}{dt}(x(t) - Mx_t) = Lx_t + B_0u(t), \quad y(t) = Cx_t,$$

where $x(t) \in \mathbb{R}^n$, $u(t) \in \mathbb{R}^m$, $y(t) \in \mathbb{R}^p$ and x_t is defined by $x_t(\tau) = x(t + \tau)$, $-h \leq \tau \leq 0$, $h > 0$. B_0 is an $n \times m$ matrix and L, M, C are bounded linear functionals from $\mathcal{C} = \mathcal{C}[-h, 0; \mathbb{R}^n]$ into \mathbb{R}^n and \mathbb{R}^p respectively. These can be represented by matrix-functions $\eta(\tau)$, $\mu(\tau)$, $\gamma(\tau)$ of bounded variation in the following way

$$\begin{aligned} L\phi &= \int_{-h}^0 d\eta(\tau)\phi(\tau), & M\phi &= \int_{-h}^0 d\mu(\tau)\phi(\tau), \\ C\phi &= \int_{-h}^0 d\gamma(\tau)\phi(\tau), & \phi &\in \mathcal{C}. \end{aligned}$$

In order to guarantee the existence and uniqueness of solutions of (4.1), we will always assume

$$(4.2) \quad \mu(0) = \lim_{\tau \uparrow 0} \mu(\tau).$$

Moreover, we will assume at some places that $M: \mathcal{C} \rightarrow \mathbb{R}^n$ is of the special form

$$(4.3) \quad M\phi = \sum_{j=1}^{\infty} A_{-j}\phi(-h_j) + \int_{-h}^0 A_{-\infty}(\tau)\phi(\tau) \, d\tau, \quad \phi \in \mathcal{C},$$

where $0 < h_j \leq h$, $A_{-j} \in \mathbb{R}^{n \times n}$ for $j \in \mathbb{N}$, $A_{-\infty}(\cdot) \in L^1[-h, 0; \mathbb{R}^{n \times n}]$ and $\sum_{j=1}^{\infty} \|A_{-j}\| < \infty$.

A function $x(\cdot) \in L^2_{\text{loc}}(-h, \infty; \mathbb{R}^n)$ is said to be a solution of (4.1) if the function $w(t) = x(t) - Mx_t$ is absolutely continuous with an L^2 -derivative on every compact interval $[0, T]$, $T > 0$, and if $\dot{w}(t) = Lx_t + B_0u(t)$ for almost every $t \geq 0$. It is well known (Burns, Herdman and Stech [3], Salamon [25]) that (4.1) admits a unique solution $x(t)$, $t \geq -h$, for every input $u(\cdot) \in L^2_{\text{loc}}(0, \infty; \mathbb{R}^m)$ and every initial condition

$$(4.4) \quad \lim_{t \downarrow 0} x(t) - Mx_t = \phi^0, \quad x(\tau) = \phi^1(\tau), \quad -h \leq \tau < 0,$$

where $\phi = (\phi^0, \phi^1) \in M^2 = \mathbb{R}^n \times L^2(-h, 0; \mathbb{R}^n)$. Moreover it has been shown in [3], [25] that the evolution of the state

$$(4.5) \quad \hat{x}(t) = (x(t) - Mx_t, x_t) \in M^2$$

of system (4.1), (4.4) can be described by the formula

$$(4.6) \quad \hat{x}(t) = S(t)\phi + \int_0^t S(t-s)Bu(s) ds$$

where $B \in \mathcal{L}(\mathbb{R}^m, M^2)$ maps $u \in \mathbb{R}^m$ into the pair $Bu = (B_0u, 0)$ and $S(t) \in \mathcal{L}(M^2)$ is the strongly continuous semigroup generated by A , where

$$D(A) = \{\phi \in M^2: \phi^1 \in W^{1,2}, \phi^0 = \phi^1(0) - M\phi^1\}, \quad A\phi = (L\phi^1, \dot{\phi}^1).$$

Here $W^{1,2}$ denotes the Sobolev space $W^{1,2}(-h, 0; \mathbb{R}^n)$.

Obviously, the dense subspace

$$W = \{(\phi(0) - M\phi, \phi): \phi \in W^{1,2}\} = \mathcal{D}(A)$$

of M^2 —endowed with the $W^{1,2}$ norm—is invariant under $S(t)$ and $S(t)$ can be restricted to a strongly continuous semigroup on W .

The output of the system (4.1) may be described through the operator

$$C: W \rightarrow \mathbb{R}^p, \quad C\phi = \int_{-h}^0 d\gamma(\tau)\phi^1(\tau), \quad \phi \in W.$$

Remarks 4.1. (i) The infinitesimal generator A of $S(t)$ can be interpreted as a bounded operator from W into M^2 . By duality, M^2 can be regarded as a dense subspace of W^* and A^* extends to a bounded operator from M^2 into W^* .

(ii) It has been proved in Burns, Herdman and Stech [3] and Salamon [25] that system (4.1) satisfies the hypotheses (H1) and (H2) with $H = V = M^2$ and the subspace $W \subset M^2$ as defined above. Hypothesis (H1) says that the state $\hat{x}(T) \in M^2$ of (4.1) defined by (4.5) is in W for every input $u(\cdot) \in L^2(0, T; \mathbb{R}^m)$ and zero initial condition and that $\hat{x}(T) \in W$ depends continuously on $u(\cdot) \in L^2[0, T; \mathbb{R}^m]$. Hypothesis (H2) says that the output $y(\cdot)$ of the free system (4.1) (i.e. $u(t) \equiv 0$) is in $L^2(0, T; \mathbb{R}^p)$ and depends in this space continuously on the initial state $\phi \in M^2$.

(iii) If $M: \mathcal{C} \rightarrow \mathbb{R}^n$ is given by (4.3), then it is known that the semigroup $S(t) \in \mathcal{L}(M^2)$ is exponentially stable if and only if

$$\omega_0 = \sup \{\operatorname{Re} \lambda: \det \Delta(\lambda) = 0\} < 0$$

where $\Delta(\lambda) = \lambda[I - M(e^\lambda)] - L(e^\lambda)$, $\lambda \in \mathbb{C}$, is the characteristic matrix of the NFDE (4.1). A necessary condition for the exponential stability of $S(t)$ is the stability of the difference operator which means that

$$(4.7) \quad \sup \left\{ \operatorname{Re} \lambda: \det \left[I - \sum_{j=1}^{\infty} A_{-j} e^{-\lambda h_j} \right] = 0 \right\} < 0.$$

These facts have been established by Henry [11] for $S(t) \in \mathcal{L}(W)$. They extend to $S(t) \in \mathcal{L}(M^2)$ because of the similarity of these two semigroups through the transformation $\mu I - A: W \rightarrow M^2$ with $\mu \notin \sigma(A)$.

(iv) If $M: \mathcal{C} \rightarrow \mathbb{R}^n$ is given by (4.3) and if (4.7) holds, then system (4.1) is *stabilizable* in the sense that there exists a feedback operator $F \in \mathcal{L}(M^2, \mathbb{R}^m)$ such that the closed loop semigroup $S_F(t) \in \mathcal{L}(M^2)$ generated by $A + BF$ is exponentially stable if and only if

$$(4.8) \quad \text{rank} [\Delta(\lambda), B_0] = n \quad \forall \lambda \in \mathbb{C}, \quad \text{Re } \lambda \geq 0$$

(Pandolfi [21], Salamon [25]).

(v) If $M: \mathcal{C} \rightarrow \mathbb{R}^n$ is given by (4.3) and if (4.7) holds, then system (4.1) is *detectable* in the sense that there exists an output injection operator $K \in \mathcal{L}(\mathbb{R}^p, M^2)$ such that the closed loop semigroup $S_K(t) \in \mathcal{L}(M^2)$ generated by $A + KC$ is exponentially stable if and only if

$$(4.9) \quad \text{rank} \begin{bmatrix} \Delta(\lambda) \\ C(e^{\lambda \cdot}) \end{bmatrix} = n \quad \forall \lambda \in \mathbb{C}, \quad \text{Re } \lambda \geq 0$$

(Salamon [25]).

Associated with the system (4.1) we consider the performance index

$$(4.10) \quad J(u) = \int_0^\infty [\|y(t)\|_{\mathbb{R}^p}^2 + \|u(t)\|_{\mathbb{R}^m}^2] dt.$$

Then we have the following theorem (compare Ito and Tarn [14] and Datko [7]).

THEOREM 4.2. *Assume $M: \mathcal{C} \rightarrow \mathbb{R}^n$ is given by (4.3) and (4.7) is satisfied; then the following statements hold.*

(i) *If (4.8) is satisfied, there exists, for every initial state $\phi \in M^2$, a unique optimal control which minimizes the cost functional (4.10). This optimal control is given by the feedback law*

$$(4.11) \quad u_\pi(t) = -B^* \pi \hat{x}(t)$$

where $\pi \in \mathcal{L}(M^2)$ is the minimal selfadjoint, nonnegative operator which satisfies the algebraic Riccati equation

$$(4.12) \quad A^* \pi + \pi A + C^* C - \pi B B^* \pi = 0$$

(this equation must be understood in the space $\mathcal{L}(W, W^*)$). Moreover the optimal cost is given by

$$(4.13) \quad J(u_\pi) = \langle \phi, \pi \phi \rangle_{M^2}.$$

(ii) *If (4.9) is satisfied, then there exists at most one nonnegative self adjoint solution $\pi \in \mathcal{L}(M^2)$ of (4.12). Moreover if π is such a solution, the closed loop semigroup $S_\pi(t) \in \mathcal{L}(M^2)$ generated by $A - B B^* \pi$ is exponentially stable.*

4.2. Parabolic systems. Consider the system

$$(4.14) \quad \dot{x} = Ax + Bu, \quad y = Cx$$

where A is a self adjoint operator on a real Hilbert space H . We assume that A has a compact resolvent operator and that the spectrum of A consists of a strictly decreasing sequence λ_n , $n \in \mathbb{N}$, of real eigenvalues with associated eigenvector $\phi_n \in H$, $\|\phi_n\| = 1$. Then A generates the strongly continuous semigroup $S(t)$ on H given by

$$S(t)x = \sum_{n=1}^{\infty} e^{\lambda_n t} \langle x, \phi_n \rangle \phi_n.$$

We assume that W and V are given by

$$W = \left\{ x \in H \left| \sum_{n=1}^{\infty} \gamma_n \langle x, \phi_n \rangle^2 < \infty \right. \right\},$$

$$V^* = \left\{ x \in H \left| \sum_{n=1}^{\infty} \beta_n^{-1} \langle x, \phi_n \rangle^2 < \infty \right. \right\}$$

with the obvious inner products, where β_n and γ_n are positive sequences satisfying $0 < \beta_n \leq 1 \leq \gamma_n < \infty$ for $n \in \mathbb{N}$. Then the space V can be represented as a space of sequences in the following way

$$V = \left\{ x \in \mathbb{R}^{\mathbb{N}} \left| \sum_{n=1}^{\infty} \beta_n x_n^2 < \infty \right. \right\}$$

and the injection $H \subset V$ is given by identifying $x \in H$ with the sequence $\{\langle x, \phi_n \rangle\}_{n \in \mathbb{N}} \in V$. Finally, we assume that the sequences $b_n \in U$, $c_n \in Y$ satisfy

$$(4.15) \quad \sum_{n=1}^{\infty} \beta_n \|b_n\|_U^2 < \infty, \quad \sum_{n=1}^{\infty} \gamma_n^{-1} \|c_n\|_Y^2 < \infty$$

and that the operators $B \in \mathcal{L}(U, V)$ and $C \in \mathcal{L}(W, Y)$ are given by

$$Cx = \sum_{n=1}^{\infty} c_n \langle x, \phi_n \rangle, \quad Bu = \{\langle b_n, u \rangle\}_{n \in \mathbb{N}}.$$

LEMMA 4.3. (i) Let $n_0 = \max \{n \in \mathbb{N} \mid \lambda_n \geq 0\}$ and suppose that

$$(4.16) \quad \sum_{n=n_0+1}^{\infty} \frac{\gamma_n \|b_n\|^2}{|\lambda_n|} < \infty;$$

then hypothesis (H1) is satisfied.

(ii) If

$$(4.17) \quad \sum_{n=n_0+1}^{\infty} \frac{\|c_n\|^2}{\beta_n |\lambda_n|} < \infty,$$

then hypothesis (H2) is satisfied.

Proof. Statement (ii) is the dual of (i) and statement (i) follows from the inequality

$$\begin{aligned} \left\| \int_0^T S(t-s) Bu(s) ds \right\|_W &= \sum_{n=1}^{\infty} \gamma_n \left(\int_0^T e^{\lambda_n(T-s)} b_n u(s) ds \right)^2 \\ &\leq \sum_{n=1}^{\infty} \gamma_n \int_0^T e^{2\lambda_n s} ds \|b_n\|^2 \|u(\cdot)\|_{L^2[0,T;U]}^2 \\ &\leq \left[\sum_{n=1}^{n_0} \gamma_n \int_0^T e^{2\lambda_n s} ds \|b_n\|^2 + \sum_{n=n_0+1}^{\infty} \frac{\gamma_n \|b_n\|^2}{2|\lambda_n|} \right] \|u(\cdot)\|_{L^2[0,T;U]}^2. \end{aligned}$$

In concrete examples the sequences b_n , c_n are given and the spaces W and V have to be chosen in such a way that (H1), (H2) and (H3) are satisfied. The next lemma shows under which conditions this is possible.

LEMMA 4.4. Let the sequences $b_n \in U$, $c_n \in Y$, $\lambda_n \in \mathbb{R}$ be given such that λ_n is strictly decreasing and tends to $-\infty$. Then there exist positive sequences β_n , γ_n satisfying (4.15)–(4.17) if and only if

$$(4.18) \quad \sum_{n=n_0+1}^{\infty} \frac{\|b_n\| \cdot \|c_n\|}{|\lambda_n|^{1/2}} < \infty.$$

Furthermore, if (4.18) holds, then the sequences β_n, γ_n can be chosen such that $\beta_n \leq \gamma_n \leq \beta_n |\lambda_n|$ for almost every $n \in \mathbb{N}$.

Proof. The necessity of (4.18) is obvious. Conversely if (4.18) holds, then it is easy to see that the sequences

$$\beta_n = \begin{cases} \|c_n\|/\|b_n\| |\lambda_n|^{1/2}, & b_n \neq 0, \quad c_n \neq 0, \quad \lambda_n \neq 0, \\ n^2 \|c_n\|^2 / |\lambda_n|, & b_n = 0, \quad c_n \neq 0, \quad \lambda_n \neq 0, \\ 1/n^2 \|b_n\|^2, & b_n \neq 0, \quad c_n = 0, \quad \lambda_n \neq 0, \\ 1 & \text{otherwise,} \end{cases}$$

$$\gamma_n = \begin{cases} |\lambda_n|^{1/2} \|c_n\|/\|b_n\|, & b_n \neq 0, \quad c_n \neq 0, \quad \lambda_n \neq 0, \\ n^2 \|c_n\|^2, & b_n = 0, \quad c_n \neq 0, \quad \lambda_n \neq 0, \\ |\lambda_n|/n^2 \|b_n\|^2, & b_n \neq 0, \quad c_n = 0, \quad \lambda_n \neq 0, \\ \max\{1, |\lambda_n|\} & \text{otherwise,} \end{cases}$$

satisfy the requirements of the lemma.

Remarks 4.5. (i) The condition $\gamma_n = \beta_n |\lambda_n|$ for almost every $n \in \mathbb{N}$ (with at most a finite number of exceptions) means that

$$\mathcal{D}_V(A) \subset \mathcal{D}_V((-A)^{1/2}) = \left\{ x \in V \mid \sum_{n=1}^{\infty} \beta_n |\lambda_n| x_n^2 < \infty \right\} = W \subset V$$

so that (H3) is satisfied.

(ii) We can assume without loss of generality that $W \subset H \subset V$, i.e. the sequences β_n and γ_n^{-1} are bounded. This can always be achieved by redefining b_n, c_n, β_n and γ_n .

(iii) It is well known that system (4.14) is stabilizable in the space V if and only if $b_n \neq 0$ for $n = 1, \dots, n_0$ (Curtain and Pritchard [4]).

The system is detectable through the unbounded output operator $C: W \rightarrow Y$ if and only if $c_n \neq 0$ for $n = 1, \dots, n_0$. This follows from an obvious generalization of the standard result for bounded output operators using a perturbation result in Salamon [25].

We are now in the position to apply the Theorems 3.3 and 3.4 to the Cauchy problem (4.14) with the performance index (4.10). Hence there exists a unique nonnegative operator $P \in \mathcal{L}(V, V^*)$ satisfying the algebraic Riccati equation.

$$(4.19) \quad AP + PA - PBB^*P + C^*C = 0$$

if $b_n \neq 0$ and $c_n \neq 0$ for $n = 1, \dots, n_0$. Furthermore the optimal control is given by the feedback law

$$(4.20) \quad u(t) = -B^*Px(t).$$

Example 4.6. As a specific example consider

$$(4.21a) \quad z_t = z_{\xi\xi}, \quad 0 < \xi < 1,$$

$$(4.21b) \quad z_{\xi}(t, 0) = u(t), \quad z_{\xi}(t, 1) = 0, \quad t > 0,$$

$$(4.21c) \quad y(t) = \int_0^1 c(\xi) z(t, \xi) d\xi, \quad t > 0.$$

It can be shown (see Curtain and Pritchard [4]) that this system is equivalent to a Cauchy problem of the form (4.14) with $H = L^2[0, 1]$, $\lambda_0 = 0$, $\phi_0(\xi) \equiv 1$, $\lambda_n = -n^2\pi^2$, $\phi_n(\xi) = \sqrt{2} \cos n\pi\xi$, and $Bu = -\delta u$ (δ being the Dirac delta impulse at $\xi = 0$). Hence

we get $b_0 = -1$, $b_n = -\sqrt{2}$ for $n \in \mathbb{N}$ and $c_n = \langle c, \phi_n \rangle$ for $n = 0, 1, 2, \dots$. So condition (4.18) is satisfied if and only if

$$\sum_{n=1}^{\infty} \frac{|c_n|}{n} < \infty.$$

This allows for arbitrary bounded linear output operators $C : L^2[0, 1] \rightarrow \mathbb{R}$ and even for a class of unbounded output operators. If C is bounded, then c_n is square summable and we may choose $\gamma_n = 1$, $\beta_n = n^{-2}$, which means that

$$W = L^2[0, 1], \quad V^* = H^1[0, 1].$$

Remark 4.7. Existence and uniqueness results for the differential Riccati equation associated with parabolic systems have been established by Pritchard and Pollock [22], Flandoli [9] and Sorine [26], [27] under weaker hypothesis. The assumptions in these papers are, roughly speaking, that A is a self adjoint nonpositive operator on H and that

$$W = V^* = \mathcal{D}((-A)^{1/2}).$$

In [9] and [22] it is assumed that the function $\|CS(t)B\|_{\mathcal{L}(U, V)}$ is integrable on $[0, T]$, whereas our results are only applicable if this function is square integrable. However, in [9] and [22] the Riccati operator $P(t)$ will only be in $\mathcal{L}(V, H) \cap \mathcal{L}(H, V^*)$ and correspondingly the optimal feedback operator $F(t) = -B^*P(t)$ will only be in $\mathcal{L}(H, U)$ as opposed to $\mathcal{L}(V, U)$.

4.3. Hyperbolic systems. Consider the system

$$(4.22) \quad \dot{z} = Az + Bu, \quad y = Cz,$$

where A is a self adjoint operator on a real Hilbert space H . We assume that A has a compact resolvent operator and that its (simple) negative eigenvalues $\lambda_n = -\omega_n^2$ satisfy

$$(4.23) \quad \omega_1 \geq \delta, \quad \omega_{n+1} - \omega_n \geq \delta, \quad n \in \mathbb{N},$$

for some $\delta > 0$. The corresponding eigenvectors are denoted by $\phi_n \in H$, $\|\phi_n\| = 1$. Furthermore, we assume that the spaces W_0, V_1 are given by

$$W_0 = \left\{ x \in H \left| \sum_{n=1}^{\infty} \gamma_n |\lambda_n| \langle x, \phi_n \rangle^2 < \infty \right. \right\},$$

$$V_1^* = \left\{ x \in H \left| \sum_{n=1}^{\infty} \beta_n^{-1} \langle x, \phi_n \rangle^2 < \infty \right. \right\},$$

where β_n and γ_n are positive sequences satisfying $0 \leq \beta_n \leq 1 \leq \gamma_n \leq \beta_n |\lambda_n|$ for $n \in \mathbb{N}$. Finally, we assume that the sequences $b_n \in U$, $c_n \in Y$ satisfy

$$(4.24) \quad \sum_{n=1}^{\infty} \beta_n \|b_n\|^2 < \infty, \quad \sum_{n=1}^{\infty} \frac{\|c_n\|^2}{\gamma_n |\lambda_n|} < \infty,$$

and that the operators $B \in \mathcal{L}(U, V_1)$, $C \in \mathcal{L}(W_0, Y)$ are given by

$$B^*x = \sum_{n=1}^{\infty} \langle x, \phi_n \rangle b_n, \quad Cx = \sum_{n=1}^{\infty} \langle x, \phi_n \rangle c_n.$$

Defining $V \subset H$ by

$$V = \mathcal{D}((-A)^{1/2}) = \left\{ x \in H \left| \sum_{n=1}^{\infty} |\lambda_n| \langle x, \phi_n \rangle^2 < \infty \right. \right\}$$

and identifying H with its dual, we obtain that $V \subset H \subset V^*$ and A extends to a bounded operator from V to V^* . In order to transform (2.22) into a first order system, we introduce the product space

$$\mathcal{H} = V \times H$$

with inner product

$$\langle x, \hat{x} \rangle = \sum_{n=1}^{\infty} [|\lambda_n| \langle x_0, \phi_n \rangle \langle \hat{x}, \phi_n \rangle + \langle x_1, \phi_n \rangle \langle \hat{x}_1, \phi_n \rangle].$$

Then the operator $\mathcal{A}: \mathcal{D}(\mathcal{A}) \rightarrow \mathcal{H}$ defined by

$$\mathcal{A} = \begin{bmatrix} 0 & I \\ A & 0 \end{bmatrix}, \quad \mathcal{D}(\mathcal{A}) = \mathcal{D}(A) \times V$$

is the infinitesimal generator of the strongly continuous semigroup $\mathcal{L}(t) \in \mathcal{L}(\mathcal{H})$ which is given by

$$\mathcal{L}(t)x = \begin{bmatrix} \sum_{n=1}^{\infty} [(\cos \omega_n t) \langle x_0, \phi_n \rangle + \omega_n^{-1} (\sin \omega_n t) \langle x_1, \phi_n \rangle] \phi_n \\ \sum_{n=1}^{\infty} [\omega_n (\sin \omega_n t) \langle x_0, \phi_n \rangle + (\cos \omega_n t) \langle x_1, \phi_n \rangle] \phi_n \end{bmatrix}.$$

Finally, we introduce the spaces

$$\mathcal{W} = \left\{ x \in \mathcal{H} \left| \sum_{n=1}^{\infty} \gamma_n [|\lambda_n| \langle x_0, \phi_n \rangle^2 + \langle x_1, \phi_n \rangle^2] < \infty \right. \right\},$$

$$\mathcal{V}^* = \left\{ x \in \mathcal{H} \left| \sum_{n=1}^{\infty} \beta_n^{-1} [|\lambda_n| \langle x_0, \phi_n \rangle^2 + \langle x_1, \phi_n \rangle^2] < \infty \right. \right\}$$

and the operators $\mathcal{B} \in \mathcal{L}(U, \mathcal{V})$, $\mathcal{C} \in \mathcal{L}(\mathcal{W}, Y)$ by

$$\mathcal{B} = \begin{bmatrix} 0 \\ B \end{bmatrix}, \quad \mathcal{C} = [C \quad 0].$$

Then (4.22) is equivalent to the Cauchy problem

$$(4.25) \quad \dot{x} = \mathcal{A}x + \mathcal{B}u, \quad y = \mathcal{C}x,$$

by means of the identification $x = (z, \dot{z})$. Note that we identify $\mathcal{H} = V \times H$ with its dual.

LEMMA 4.8. (i) *If*

$$(4.26) \quad \sup_{n \in \mathbb{N}} \gamma_n \|b_n\|_U^2 < \infty,$$

then the operator \mathcal{B} satisfies (H1).

(ii) *If*

$$(4.27) \quad \sup_{n \in \mathbb{N}} \frac{\|c_n\|_Y^2}{\beta_n |\lambda_n|} < \infty,$$

then the operator \mathcal{C} satisfies (H2).

Proof. Statement (ii) is the dual of (i). In order to prove statement (i), note first that

$$\int_0^T \mathcal{L}(T-s) \mathcal{B}u(s) ds = \begin{bmatrix} \sum_{n=1}^{\infty} \omega_n^{-1} \int_0^T (\sin \omega_n(T-s)) \langle b_n, u(s) \rangle ds \phi_n \\ \sum_{n=1}^{\infty} \int_0^T (\cos \omega_n(T-s)) \langle b_n, u(s) \rangle ds \phi_n \end{bmatrix}$$

for every $u(\cdot) \in L^2[0, T; U]$ and hence

$$\begin{aligned}
 & \left\| \int_0^T \mathcal{L}(T-s) \mathcal{B}u(s) ds \right\| \\
 &= \sum_{n=1}^{\infty} \gamma_n \left\{ \left[\int_0^T (\sin \omega_n(T-s)) \langle b_n, u(s) \rangle ds \right]^2 \right. \\
 &\quad \left. + \left[\int_0^T (\cos \omega_n(T-s)) \langle b_n, u(s) \rangle ds \right]^2 \right\} \\
 &\leq (\sup_{n \in \mathbb{N}} \gamma_n \|b_n\|^2) \sum_{n=1}^{\infty} \left\{ \left\| \int_0^T (\sin \omega_n(T-s)) u(s) ds \right\|^2 \right. \\
 &\quad \left. + \left\| \int_0^T \cos \omega_n(T-s) u(s) ds \right\|^2 \right\} \\
 &\leq \text{const.} (\sup_{n \in \mathbb{N}} \gamma_n \|b_n\|^2) \cdot \|u(\cdot)\|_{L^2[0, T; U]}^2.
 \end{aligned}$$

The final inequality is a consequence of (4.23) together with some properties of Fourier series (see Ingham [13] and Russell [24]).

The next lemma shows under which conditions the spaces \mathcal{W} and \mathcal{V} can be chosen in such a way that (H1), (H2), (H3) are satisfied if the sequences b_n, c_n are given.

LEMMA 4.9. *Let the sequences $b_n \in U$, $c_n \in Y$, $\lambda_n \in \mathbb{R}$ be given such that (4.23) is satisfied. Then there exist positive sequences β_n, γ_n satisfying (4.24), (4.26) and (4.27) if and only if*

$$(4.28) \quad \sum_{n=1}^{\infty} \frac{\|b_n\|^2 \cdot \|c_n\|^2}{|\lambda_n|} < \infty.$$

Furthermore, if (4.28) holds, then the sequences β_n, γ_n can be chosen such that $\mathcal{D}_{\mathcal{V}}(\mathcal{A}) \subset \mathcal{W} \subset \mathcal{V}$.

Proof. The necessity of (4.28) is obvious. Conversely, if (4.28) holds, then it is easy to see that the sequences

$$\begin{aligned}
 \beta_n &= \begin{cases} \|c_n\|^2 / |\lambda_n|, & \|b_n\|_U \|c_n\| \geq 1 \quad \text{or } b_n = 0, \quad c_n \neq 0, \\ 1 / \|b_n\|^2 |\lambda_n|, & \|b_n\| \|c_n\| \leq 1 \quad \text{and } b_n \neq 0, \\ 1 / |\lambda_n|, & b_n = 0, \quad c_n = 0, \end{cases} \\
 \gamma_n &= \begin{cases} 1 / \|b_n\|^2, & b_n \neq 0, \\ \|c_n\|^2, & b_n = 0, \quad c_n \neq 0, \\ 1, & b_n = 0, \quad c_n = 0, \end{cases}
 \end{aligned}$$

satisfy the requirements of the lemma. In particular β_n / γ_n is bounded and $\gamma_n \leq \beta_n |\lambda_n|$ for every $n \in \mathbb{N}$.

We are now in the position to apply Theorem 2.7 to the Cauchy problem (4.25) with the performance index

$$(4.29) \quad J(u) = \int_0^T [\|y(t)\|^2 + \|u(t)\|^2] dt.$$

Hence there exists a unique nonnegative strongly continuous operator $\mathcal{P}(t) \in \mathcal{L}(\mathcal{V}, \mathcal{V}^*)$ satisfying the differential Riccati equation

$$\begin{aligned}
 & \frac{d}{dt} \mathcal{P}(t)x + \mathcal{A}^* \mathcal{P}(t)x + \mathcal{P}(t)\mathcal{A}x - \mathcal{P}(t)\mathcal{R}\mathcal{B}^* \mathcal{P}(t)x + \mathcal{C}^* \mathcal{C}x = 0, \\
 & \mathcal{P}(T)x = 0, \quad x \in \mathcal{D}_{\mathcal{V}}(\mathcal{A}).
 \end{aligned}
 \tag{4.30}$$

Furthermore, the optimal control is given by the feedback law

$$(4.31) \quad u(t) = -\mathcal{B}^* \mathcal{P}(t)x(t).$$

Example 4.10. As a specific example we consider the system

$$(4.32a) \quad z_{tt} = z_{\xi\xi}, \quad 0 < \xi < 1, \quad t > 0,$$

$$(4.32b) \quad z(t, 0) = u(t), \quad z(t, 1) = 0, \quad t > 0,$$

$$(4.32c) \quad y(t) = \int_0^1 c(\xi) z(t, \xi) d\xi, \quad t > 0,$$

in the Hilbert space

$$\mathcal{H} = H_0^1[0, 1] \times L^2[0, 1]$$

which we identify with its dual. Then the operator $A = \Delta : H^2[0, 1] \cap H_0^1[0, 1] \rightarrow L^2[0, 1]$ has the eigenvalues $\lambda_n = n^2 \pi^2$ with corresponding eigenfunctions $\phi_n(\xi) = \sqrt{2} \sin n\pi\xi$. Furthermore, the input operator for (2.32) takes the form $Bu = -\delta' u$, where δ' is the distributional derivative of the Dirac delta impulse at $\xi = 0$ (see Curtain and Pritchard [4]). Hence

$$b_n = \sqrt{2} n\pi, \quad c_n = \sqrt{2} \int_0^1 c(\xi) \sin n\pi\xi d\xi$$

for $n \in \mathbb{N}$. So condition (4.28) is satisfied if c_n is square integrable. The proof of Lemma 4.9 shows that we may choose

$$\beta_n = \max \left\{ \frac{|c_n|^2}{|\lambda_n|}, \frac{1}{|\lambda_n|^2} \right\}, \quad \gamma_n = \frac{1}{|\lambda_n|}, \quad n \in \mathbb{N}.$$

In particular this means that the boundary control system (4.32) has continuous solutions in the space

$$L^2[0, 1] \times H^{-1}[0, 1]$$

for every input $u(\cdot) \in L^2[0, T]$. This result has also been established by Lasiecka and Triggiani [17]. For the output operator we can allow an arbitrary bounded linear map from $L^2[0, 1]$ into \mathbb{R} . The space \mathcal{V} depends on this map. In any case $\mathcal{W} \subset \mathcal{V}$ and hence $\mathcal{P}(t) : \mathcal{V} \rightarrow \mathcal{V}^*$ has a smoothing effect with respect to \mathcal{W} .

Remark 4.11. An analogous result has been developed by Lasiecka and Triggiani [18] for the higher dimensional wave equation. In their paper the output operator is the identity on the displacement component of the state in $\mathcal{W} = L^2 \times H^{-1}$. This case cannot be treated within our framework. However, the results in [18] are weaker than ours. The uniqueness for the solution of the Riccati equation has not been established in [18]. Furthermore, the Riccati operator in [18] is in $\mathcal{L}(\mathcal{W})$ and does not have smoothing properties with respect to \mathcal{W} . Consequently the feedback operator becomes unbounded with respect to this space. It seems that for hyperbolic PDE's our assumptions are close to the weakest possible in order to derive a bounded feedback operator.

REFERENCES

- [1] A. V. BALAKRISHNAN, *Boundary control of parabolic equations: L-Q-R theory*, in Theory of Nonlinear Operators, Proc. 5th International Summer School, Akademie-Verlag, Berlin, 1978.
- [2] A. BENSOUSSAN, M. C. DELFOUR AND S. K. MITTER, *The linear quadratic optimal control problem for infinite dimensional systems over an infinite horizon: survey and examples*, 1976 IEEE Conference on Decision and Control, IEEE Publications, New York, 1976, pp. 745-751.

- [3] J. A. BURNS, T. L. HERDMAN AND H. W. STECH, *Linear functional differential equations as semigroups in product spaces*, SIAM J. Math. Anal., 14 (1983), pp. 98–116.
- [4] R. F. CURTAIN AND A. J. PRITCHARD, *Infinite Dimensional Linear Systems Theory*, Lecture Notes in Computer and Information Science 8, Springer-Verlag, Berlin, 1978.
- [5] R. DATKO, *Uniform asymptotic stability of evolutionary processes in Banach space*, SIAM J. Math. Anal., 3 (1972), pp. 428–454.
- [6] ———, *Unconstrained control problems with quadratic cost*, this Journal, 11 (1973), pp. 32–52.
- [7] ———, *Neutral autonomous functional equations with quadratic cost*, this Journal, 12 (1974), pp. 70–82.
- [8] M. C. DELFOUR, *The linear quadratic optimal control problem with delays in the state and control variables: A state space approach*, Centre de Recherche de Mathématiques Appliquées, Université de Montréal, CRMA-1012, 1981.
- [9] F. FLANDOLI, *Riccati equation arising in a boundary control problem with distributed parameters*, this Journal, 22 (1984), pp. 76–86.
- [10] J. S. GIBSON, *The Riccati integral equations for optimal control problems on Hilbert space*, this Journal, 17 (1979), pp. 537–565.
- [11] D. HENRY, *Linear autonomous neutral functional differential equations*, J. Differential Equations, 15 (1974), pp. 106–128.
- [12] A. ICHIKAWA, *Quadratic control of evolution equations with delays in control*, this Journal, 20 (1982), pp. 645–668.
- [13] A. E. INGHAM, *Some trigonometrical inequalities on the theory of series*, Math. Z., 41 (1936), pp. 367–379.
- [14] K. ITO AND T. J. TARN, *A linear quadratic control problem for neutral systems*, J. Nonlinear Anal.—TMA, to appear.
- [15] T. KATO, *Perturbation Theory of Linear Operators*, Springer-Verlag, New York, 1966.
- [16] I. LASIECKA AND R. TRIGGIANI, *Dirichlet boundary control problem for parabolic equations with quadratic cost: analyticity and Riccati feedback synthesis*, this Journal, 21 (1983), pp. 41–67.
- [17] ———, *Riccati equations for hyperbolic partial differential equations with $L_2(0, T; L_2(\Gamma))$ —Dirichlet boundary terms*, this Journal, 24 (1986), pp. 884–925.
- [18] ———, *An L^2 theory for the quadratic optimal cost problem of hyperbolic equations with control in the Dirichlet boundary conditions*, in Control Theory for Distributed Parameter Systems and Applications, F. Kappel, K. Kunisch and W. Schappacher, eds., Lecture Notes in Computer and Information Science 54, Springer-Verlag, Berlin, 1983, pp. 138–152.
- [19] J. L. LIONS, *Optimal Control of Systems Governed by Partial Differential Equations*, Springer-Verlag, New York, 1971.
- [20] D. L. LUKES AND D. L. RUSSELL, *The quadratic criterion for distributed systems*, this Journal, 7 (1969), pp. 101–121.
- [21] L. PANDOLFI, *Stabilization of neutral functional differential equations*, J. Opt. Theory Appl., 20 (1976), pp. 191–204.
- [22] J. POLLACK AND A. J. PRITCHARD, *The infinite time quadratic cost control problem for distributed systems with unbounded control action*, J. Inst. Math. Appl., 25 (1980), pp. 287–309.
- [23] D. L. RUSSELL, *Quadratic performance criteria in boundary control of linear symmetric hyperbolic systems*, this Journal, 11 (1973), pp. 475–509.
- [24] ———, *Closed loop eigenvalue specification for infinite dimensional systems: augmented and deficient hyperbolic cases*, MRC, University of Wisconsin-Madison, TSR #2021, 1979.
- [25] D. SALAMON, *Control and Observation of Neutral Systems*, RNM 91, Pitman, London, 1984.
- [26] M. SORINE, *Une resultat d'existence et unicité pour l'équation de Riccati stationnaire*, Rapport INRIA, no. 55, 1981.
- [27] ———, *Sur le semigroupe non linéaire associé a l'équation de Riccati*, Centre de Recherche de Mathématiques Appliquées, Université de Montréal, CRMA-1055, 1981.
- [28] J. C. WILLEMS, *Least squares stationary optimal control and the algebraic Riccati equation*, IEEE Trans. Automat. Control, AC-16 (1972), pp. 621–634.
- [29] W. M. WONHAM, *Linear Multivariable Control: A Geometric Approach*, Lecture Notes in Economics and Mathematical Systems 101, Springer-Verlag, New York, 1974.
- [30] J. ZABCZYK, *Remarks on the algebraic Riccati equation in Hilbert space*, Appl. Math. Opt., 2 (1976), pp. 251–258.

THE MAXIMUM PRINCIPLE FOR AN OPTIMAL SOLUTION TO A DIFFERENTIAL INCLUSION WITH END POINTS CONSTRAINTS*

HALINA FRANKOWSKA†

Abstract. We derive Pontryagin's maximum principle for a general optimal control problem using the set-valued version of variational equation. We achieve this aim by exploiting an adequate differential calculus of set-valued maps. Furthermore, the calmness condition is replaced by a surjectivity condition involving reachable sets of the "set-valued linearization" of the initial control problem. Duality then provides both the "adjoint differential inclusion" and the maximum principle.

Key words. differential inclusion, tangent cone, derivative of a set-valued map, convex process, variational inclusion, maximum principle

AMS(MOS) subject classifications. 49B10, 49B34, 49B30, 93C15

1. Introduction. We shall derive the maximum principle for the optimization problem

$$(1) \quad \text{minimize} \left\{ g(x(0), x(1)) + \int_0^1 \phi(x(\tau)) d\tau \right\}$$

over the solutions to a differential inclusion

$$(2) \quad x'(t) \in F(x(t))$$

satisfying end points constraints

$$(3) \quad x(0) \in C_0, \quad x(1) \in C_1.$$

By regarding this problem as an abstract optimization problem

$$(4) \quad \min_{x \in \mathcal{K}} \phi(x)$$

where \mathcal{K} is the set of solutions to (2), (3) in an appropriate functional space, the maximum principle is the three-century-old Fermat rule stating that the gradient of the functional ϕ vanishes at a solution \bar{x} of (4) lying in the interior of \mathcal{K} . Since most often in infinite-dimensional space \mathcal{K} has an empty interior (i.e., \bar{x} lies in the boundary of \mathcal{K}) this Fermat rule becomes

$$(5) \quad \langle \phi'(\bar{x}), w \rangle \geq 0$$

for all w in $T_{\mathcal{K}}(\bar{x})$, where $T_{\mathcal{K}}(\bar{x})$ is the contingent cone to \mathcal{K} at \bar{x} , introduced by Bouligand in the 1930s (see Aubin-Ekeland [3, Chap. 7]).

When we wish to derive the maximum principle from this inequality, we observe that we cannot "characterize explicitly" all the elements of this cone, but only of a subcone. We already followed this abstract approach in Frankowska [12] by using an appropriate subcone: the intermediate tangent cone $I_{\mathcal{K}}(\bar{x})$, introduced by Ursescu [22]. Hence we shall deduce the maximum principle from inequality (5), when w ranges over a closed convex subcone of the intermediate tangent cone. The convexity allows us to state a dual version of the Fermat rule (5), which is the familiar form in which the maximum principle is stated.

* Received by the editors December 4, 1984, and in revised form August 26, 1985.

† CEREMADE, Université de Paris—Dauphine, 75775 Paris, France.

We may always choose the Clarke tangent cone $C_{\mathcal{K}}(\bar{x})$, introduced by Clarke in [6]. We refer also to Clarke [7], Rockafellar [17] and Aubin-Ekeland [3, Chap. 7], for an exhaustive exposition of the properties of this cone. Let us mention only a result of J. Treiman [20], generalizing to the case of Banach spaces a result of Cornet [8]. It implies in essence that when a subcone of the contingent cone is lower semicontinuous, it is actually a subcone of the Clarke tangent cone. However, the choice of the Clarke tangent cone is not always wise. First, it involves unnecessary regularity assumptions. Second, it happens to be difficult to handle and to characterize explicitly. This is the reason why we state our results for an arbitrary closed convex subcone of the intermediate tangent cone.

In this paper, we replace the “calmness” assumption introduced in Clarke [7] by a surjectivity assumption, which involves naturally the end points constraints.

The outline of the paper is as follows. In § 2 we recall some notions with which we are dealing. Section 3 is devoted to the study of the derivative of the solutions set of the differential inclusion (2). In the fourth section we derive the maximum principle. An example is provided in § 5.

2. Tangent cones. We recall first

DEFINITION 2.1. Let K be a subset of a Banach space E and $x \in \bar{K}$ (closure of K).

(a) Bouligand's contingent cone:

$$T_K(x) := \left\{ w \in E : \liminf_{h \rightarrow 0+} \text{dist} \left(w, \frac{K-x}{h} \right) = 0 \right\},$$

(b) Intermediate tangent cone:

$$I_K(x) := \left\{ w \in E : \lim_{h \rightarrow 0+} \text{dist} \left(w, \frac{K-x}{h} \right) = 0 \right\},$$

(c) Clarke's tangent cone:

$$C_K(x) := \left\{ w \in E : \lim_{\substack{h \rightarrow 0+ \\ y \rightarrow x \\ K}} \text{dist} \left(w, \frac{K-y}{h} \right) = 0 \right\},$$

(d) Dubovickii-Miljutine tangent cone:

$$D_K(x) := \{ w \in E : \exists \varepsilon > 0, x +]0, \varepsilon][w + \varepsilon B) \subset K \},$$

where B denotes the closed unit ball in E .

All the above sets are cones. (a), (b), (c) are closed, (d) is open. Moreover if K is convex then $T_K(x) = C_K(x) = I_K(x)$, $D_K(x) = \text{Int } T_K(x)$.

The cone $C_K(x)$ is always convex and

$$C_K(x) \subset I_K(x) \subset T_K(x), \quad \text{Int } C_K(x) \subset D_K(x) \subset I_K(x).$$

For further properties of $T_K(x)$ see Aubin-Ekeland [3, Chap. 7]. The cone $C_K(x)$ was studied in Clarke [7], Rockafellar [17] and Aubin-Ekeland [3]. For the cone $D_K(x)$ see Dubovickii-Miljutine [9]. Properties of $I_K(x)$ are given in Ursescu [22].

We recall

DEFINITION 2.2. For $Q: \mathbb{R} \times E \rightarrow \mathbb{R} \cup \{+\infty\}$, set

$$\limsup_{h \rightarrow 0+} \inf_{u' \rightarrow u} Q(h, u') := \sup_{\varepsilon > 0} \inf_{\delta > 0} \sup_{h \in]0, \delta[} \inf_{u' \in u + \varepsilon \tilde{B}} Q(h, u')$$

(see Rockafellar [17]).

DEFINITION 2.3. For $f: E \rightarrow \mathbb{R} \cup \{+\infty\}$, $x \in \text{Dom}(f)$, $u \in E$ set

$$i_+f(x)(u) := \limsup_{h \rightarrow 0+} \inf_{u' \rightarrow u} \frac{f(x + hu') - f(x)}{h}.$$

If $\text{Ep} f$ denotes the epigraph of f , we observe that

$$I_{\text{Ep} f}(x, f(x)) = \text{Ep } i_+f(x)$$

(see Frankowska [12]).

Here we shall use the closed convex subcones of the intermediate tangent cone $I_K(x)$. The Clarke tangent cone $C_K(x)$ is the first example of such a subcone. Another example is provided by the *asymptotic tangent cone* given by

$$I_K^a(x) = \{u \in I_K(x) : u + I_K(x) \subset I_K(x)\}.$$

$I_K^a(x)$ is a closed convex cone and $C_K(x) \subset I_K^a(x) \subset I_K(x)$.

In this paper we shall study a “linearization” of a differential inclusion given by a closed (convex) process. We recall

DEFINITION 2.4. Let E, E_1 be Banach spaces and let $A: E \rightrightarrows E_1$ be a set-valued map. A is called a closed (respectively, convex) process if $\text{graph } A$ is a closed (respectively, convex) cone. The transposed process $A^*: E_1^* \rightrightarrows E^*$ of a closed convex process $A: E \rightrightarrows E_1$ is defined by

$$v \in A^*(u) \text{ if and only if for all } (x, y) \in \text{graph } A, \langle v, x \rangle \leq \langle u, y \rangle.$$

DEFINITION 2.5 [13]. Let E, E_1 be given Banach spaces and let $F: E \rightrightarrows E_1$ be a set-valued map, Lipschitzian at x and $y \in F(x)$. The intermediate derivative $dF(x, y)$ of F at (x, y) is a set-valued map from E into E_1 defined by

$$dF(x, y)(u) = \left\{ v \in E_1 : \lim_{h \rightarrow 0+} \text{dist} \left(v, \frac{F(x + hu) - y}{h} \right) = 0 \right\}$$

for all $u \in E$, or equivalently

$$v \in dF(x, y)(u) \text{ if and only if } (u, v) \in I_{\text{graph } F}(x, y).$$

In the case of single-valued maps, this notion generalizes the Fréchet derivative.

EXAMPLE 2.6. Let $F: E \rightrightarrows E_1$ be a set-valued map, Lipschitzian at x , $y \in F(x)$. Then $dF(x, y)$ is a closed process.

The cones $I_{\text{graph } F}^a(x, y)$ and $C_{\text{graph } F}(x, y)$ define closed convex processes $d_aF(x, y)$, $CF(x, y)$, respectively, by

$$v \in d_aF(x, y)(u) \text{ if and only if } (u, v) \in I_{\text{graph } F}^a(x, y),$$

$$v \in CF(x, y)(u) \text{ if and only if } (u, v) \in C_{\text{graph } F}(x, y).$$

In the next two lemmas we study some properties of the intermediate derivative $dF(x, y)$.

LEMMA 2.7. Let $F: E \rightrightarrows \mathbb{R}^n$ be a given set-valued map, Lipschitzian at x with a constant M and $y \in F(x)$. If F is differentiable at (x, y) , in the sense that

$$T_{\text{graph } F}(x, y) = I_{\text{graph } F}(x, y),$$

then the map $dF(x, y)(\cdot)$ is an M -Lipschitzian closed process. If, moreover, it is soft in the sense that

$$T_{\text{graph } F}(x, y) = C_{\text{graph } F}(x, y),$$

then $dF(x, y)$ is a closed convex process.

Proof. The set $dF(x, y)(0) \neq \emptyset$ because it contains zero. Pick any $u \in E$ such that $dF(x, y)(u) \neq \emptyset$ and let $v \in dF(x, y)(u)$, $u_1 \in E$. For all small $h > 0$ let $v_h \in (F(x + hu) - y)/h$ be such that $\lim_{h \rightarrow 0+} v_h = v$. By Lipschitzianity, for all small $h > 0$, there exists $w_h \in (F(x + hu_1) - y)/h$ such that $\|w_h - v_h\| \leq M\|u - u_1\|$. So there exists a subsequence $\{w_{h_i}\}$ converging to some w . Thus

$$(u_1, w) \in T_{\text{graph } F}(x, y), \quad \|v - w\| \leq M\|u - u_1\|.$$

By the assumption of Lemma 2.7,

$$w \in dF(x, y)(u_1).$$

The second statement follows from the convexity of Clarke's tangent cone. \square

LEMMA 2.8. *Let $F: E \rightrightarrows E_1$ be a Lipschitzian at x set-valued map with convex images and $y \in F(x)$. Then for all $u \in \text{Dom } dF(x, y) := \{w: dF(x, y)(w) \neq \emptyset\}$, $dF(x, y)(u)$ is a closed convex set and*

$$\text{cl}(dF(x, y)(u) + T_{F(x)}(y)) \subset dF(x, y)(u).$$

Proof. Fix $u \in \text{Dom } dF(x, y)$. The set $dF(x, y)(u)$ is closed because $\text{graph } dF(x, y)$ is closed. It is convex because F has convex images. To prove the last statement we have to verify that

$$dF(x, y)(u) + \bigcup_{h>0} \frac{1}{h}(F(x) - y) \subset dF(x, y)(u).$$

Pick $h_0 > 0$ and $v \in dF(x, y)(u)$ and let w be so that $y + h_0 w \in F(x)$. For all small $h > 0$ let $v_h \in E_1$ be such that $\lim_{h \rightarrow 0+} v_h = v$ and

$$(2.9) \quad y + hv_h \in F(x + hu).$$

Since F is Lipschitzian at x there exists $M > 0$ such that for all $h > 0$ sufficiently small

$$(2.10) \quad y + h_0 w \in F(x + hu) + h|u|MB.$$

By the convexity of $F(x + hu)$ and (2.9), (2.10) for all small $h > 0$

$$\left(1 - \frac{h}{h_0}\right)(y + hv_h) + \frac{h}{h_0}(y + h_0 w) \in F(x + hu) + \frac{h^2}{h_0}|u|MB.$$

Hence

$$y + h(v_h + w) - \frac{h^2}{h_0}v_h \in F(x + hu) + \frac{h^2}{h_0}|u|MB$$

and by Definition 2.5, $v + w \in dF(x, y)(u)$. \square

In this paper we consider families of uniformly Lipschitzian closed processes. The next lemma provides a sufficient condition for the uniform Lipschitz continuity.

Let $A: E \rightrightarrows E_1$ be a given closed convex process. If $\text{Dom}(A) := \{x \in E: A(x) \neq \emptyset\} = E$ the Robinson-Ursescu theorem (see Robinson [16], Ursescu [21] or Aubin-Ekeland [3, p. 132]) tells us that A is M -Lipschitzian for some $M > 0$ and that for all $q \in \text{Dom}(A^*)$

$$\sup_{p \in A^*(q)} \|p\| \leq M\|q\|.$$

LEMMA 2.11. *Let T be a topological space and let $A_\tau: \mathbb{R}^n \rightrightarrows \mathbb{R}^q$, $\tau \in T$ be a family of closed convex processes. Assume that the map $\tau \rightarrow \text{graph } A_\tau$ is lower semicontinuous at τ_0 . Then the following statements are equivalent.*

- (i) The processes A_τ are uniformly Lipschitzian on a neighborhood V of τ_0 , in the sense that there exists a constant $M > 0$ such that

$$A_\tau \text{ is } M\text{-Lipschitzian for all } \tau \in V,$$

- (ii) $\text{Dom } A_{\tau_0} = \mathbb{R}^n$.

Proof. Clearly (i) implies (ii). Conversely assume that (ii) is verified. By the Robinson–Ursescu theorem, $A_{\tau_0}(\cdot)$ is γ -Lipschitzian for some $\gamma > 0$. Let $u_0, \dots, u_n \in \mathbb{R}^n$ be such that

$$B \subset \text{Int co } \{u_i : i = 0, \dots, n\},$$

$$m = \max_i \|u_i\|.$$

By Lipschitzianity, for all i there exists

$$v_i \in A_{\tau_0}(u_i) \cap \gamma m B.$$

By the lower semicontinuity of graph A_τ there exists a neighborhood V of τ_0 such for all $\tau \in V$ and $i = 0, \dots, n$ we can find

$$(2.12) \quad (u_i(\tau), v_i(\tau)) \in \text{graph } (A_\tau) \cap (mB \times 2\gamma mB)$$

such that

$$B \subset \text{Int co } \{u_i(\tau) : i = 0, \dots, n\}.$$

Since graph A_τ is convex, we deduce from (2.12) that for all $u \in mB$

$$A_\tau(u) \cap 2\gamma mB \neq \emptyset$$

and thus

$$\frac{1}{2\gamma} B \subset A_\tau^{-1}(B).$$

This inclusion implies that $A_\tau(\cdot)$ is Lipschitzian with a constant 2γ . \square

COROLLARY 2.13. Let $F: \mathbb{R}^n \rightrightarrows \mathbb{R}^q$ be a set-valued map. Assume that there exists a family of closed convex processes $P(x, y): \mathbb{R}^n \rightrightarrows \mathbb{R}^q$, $(x, y) \in \text{graph } F$ satisfying

$$\{u \in \mathbb{R}^n : \exists v \in \mathbb{R}^q, (u, v) \in \liminf_{\substack{(x', y') \rightarrow (x, y) \\ (x', y') \in \text{graph } F}} \text{graph } P(x, y)\} = \mathbb{R}^n.$$

Then, for all compact set $G \subset \text{graph } F$, any family of closed convex processes $A(x, y): \mathbb{R}^n \rightrightarrows \mathbb{R}^q$, $(x, y) \in G$ satisfying

$$\text{graph } P(x, y) \subset \text{graph } A(x, y) \text{ for all } (x, y) \in G$$

is uniformly Lipschitzian.

Proof. By the compactness of G it is enough to show that for all $(x, y) \in \text{graph } F$ there exist $\varepsilon > 0$ and $M > 0$ such that for all $(x', y') \in ((x, y) + \varepsilon B) \cap \text{graph } F$ $A(x', y')$ is M -Lipschitzian. Fix $(x, y) \in \text{graph } F$ and set

$$Q = \liminf_{\substack{(x', y') \rightarrow (x, y) \\ (x', y') \in \text{graph } F}} \text{graph } P(x', y').$$

For all $(x', y') \in \text{graph } F$ and $u \in \mathbb{R}^n$ set

$$S(x', y')(u) = \begin{cases} \{v \in \mathbb{R}^q : (u, v) \in Q\} & \text{if } (x', y') = (x, y), \\ P(x', y')(u) & \text{otherwise.} \end{cases}$$

By Lemma 2.11 there exist $\varepsilon > 0$, $M > 0$ such that for all $(x', y') \in ((x, y) + \varepsilon B) \cap \text{graph } F$ $S(x', y')$ is M -Lipschitzian, i.e., $(1/M)B \subset S(x'y')^{-1}(B) \subset P(x'y')^{-1}(B) \subset A(x', y')^{-1}(B)$. This implies the M -Lipschitzianity of $A(x', y')$. \square

We shall need the following

LEMMA 2.14. *Let $\phi: \mathbb{R}^n \rightarrow \mathbb{R}$ be a Lipschitzian function and $f: L^1(0, 1) \rightarrow \mathbb{R}$ be defined by*

$$f(x) = \int_0^1 \phi(x(\tau)) d\tau.$$

Further let $z \in L^1(0, 1)$ and $\psi: [0, 1] \times \mathbb{R}^n \rightarrow \mathbb{R}$ be such a function that

- (i) $\psi(\cdot, u)$ is measurable for all $u \in \mathbb{R}^n$,
- (ii) $\{\psi(\tau, \cdot): \tau \in [0, 1]\}$ is a family of uniformly Lipschitzian convex functions and $\psi(\tau, 0) = 0$ for all $\tau \in [0, 1]$,
- (iii) $\psi(\tau, \cdot) \geq i_+ \phi(z(\tau))(\cdot)$ for almost all $\tau \in [0, 1]$.

Then the function $g: L^1(0, 1) \rightarrow \mathbb{R}$ defined by

$$g(u) = \int_0^1 \psi(\tau, u(\tau)) d\tau$$

is convex, Lipschitzian. Furthermore $g \geq i_+ f(z)$ and if ξ is a subgradient of g at zero, then for almost all $\tau \in [0, 1]$

$$\xi(\tau) \in \partial \psi(\tau, 0).$$

Proof. It is enough to prove the last statement. By Fatou's lemma and Lipschitzianity of ϕ , for all $u \in L^1(0, 1)$

$$\begin{aligned} i_+ f(z)(u) &\leq \int_0^1 \limsup_{h \rightarrow 0+} \frac{\phi(z(\tau) + hu(\tau)) - \phi(z(\tau))}{h} d\tau \\ &= \int_0^1 i_+ \phi(z(\tau))(u(\tau)) d\tau \leq \int_0^1 \psi(\tau, u(\tau)) d\tau = g(u). \end{aligned}$$

If $\xi \in \partial g(0)$, then $\xi \in L^\infty(0, 1)$ and for all $u \in L^1(0, 1)$

$$\int_0^1 \langle \xi(\tau), u(\tau) \rangle d\tau \leq \int_0^1 \psi(\tau, u(\tau)) d\tau.$$

We may assume that ξ is a bounded function. Since $\psi(\tau, \cdot) - \langle \xi(\tau), \cdot \rangle$ are uniformly Lipschitzian functions equal to zero at the point zero and u is an arbitrary function of $L^1(0, 1)$ the last inequality implies that for almost all $\tau \in [0, 1]$ and all $x \in \mathbb{R}^n$, $\psi(\tau, x) - \langle \xi(\tau), x \rangle \geq 0$, equivalently,

$$\psi(\tau, x) - \psi(\tau, 0) \geq \langle \xi(\tau), x \rangle$$

which ends the proof. \square

3. The intermediate derivative of the solution set to the differential inclusion $x' \in F(x)$. Let $F: \mathbb{R}^n \rightrightarrows \mathbb{R}^n$ be a set-valued map. Recall that a function x of the Sobolev space $W^{1,1}(0, 1)$ is called a solution to the differential inclusion

$$(3.1) \quad x' \in F(x)$$

whenever $x'(t) \in F(x(t))$ almost everywhere in $[0, 1]$. We denote by S_1 the set of all solutions of (3.1) defined on the time interval $[0, 1]$. For all $\eta \in \mathbb{R}^n$ set

$$S(\eta) = \{x \in S_1: x(0) = \eta\}.$$

Let z belong to S_1 . For deriving the maximum principle we need to regard S as a set-valued map from \mathbb{R}^n to the space $C(0, 1)$ of continuous functions on $[0, 1]$ into \mathbb{R}^n and to characterize the derivative $dS(z(0), z)$ (the intermediate derivative of $S: \mathbb{R}^n \rightrightarrows C(0, 1)$).

We denote by $\text{co } F$ the set-valued map whose value at x is equal to the convex hull of $F(x)$.

Consider the convexified inclusion

$$(3.2) \quad x' \in \text{co } F(x)$$

and let S_1^{co} denote the set of solutions of (3.2) defined on the time interval $[0, 1]$. For all $\eta \in \mathbb{R}^n$ set

$$S^{\text{co}}(\eta) = \{x \in S_1^{\text{co}}: x(0) = \eta\}.$$

We wish to compare $dS(z(0), z)$ and $dS^{\text{co}}(z(0), z)$.

THEOREM 3.3. *Assume that F has compact images and is Lipschitzian on a neighborhood of $z([0, 1])$. Then*

$$dS(z(0), z) = dS^{\text{co}}(z(0), z).$$

Proof. Let $\varepsilon > 0$ be so that F is Lipschitzian on $z([0, 1]) + 2\varepsilon B$. By the Filippov-Ważewski relaxation theorem (see Aubin-Cellina [2, p. 128]) for all $x \in S^{\text{co}}(\eta)$ satisfying $x([0, 1]) \subset z([0, 1]) + \varepsilon B$ there exists a sequence $x_i \in S(\eta)$, $i \geq 1$ which converges to x in $C(0, 1)$. Hence the result. \square

In the theorem below we characterize subsets of $dS(z(0), z)$.

THEOREM 3.4. *Assume that F has compact images and is Lipschitzian on a neighborhood of $z([0, 1])$. Then the set of solutions to the differential inclusion*

$$(3.5) \quad \begin{aligned} w'(t) &\in d \text{co } F(z(t), z'(t))(w(t)) \quad \text{a.e. in } [0, 1], \\ w(0) &= \eta, \end{aligned}$$

is contained in $dS(z(0), z)(\eta)$.

Proof. From [13] follows that every solution of the inclusion (3.5) is contained in $dS^{\text{co}}(z(0), z)(\eta)$. Theorem 3.3 ends the proof. \square

4. The maximum principle. Consider the differential inclusion

$$(4.1) \quad x' \in F(x)$$

and let S_1 denote the set of all solutions of (4.1) defined on $[0, 1]$.

Let C_0, C_1 be subsets of \mathbb{R}^n and let

$$\phi: \mathbb{R}^n \rightarrow \mathbb{R}, \quad g: \mathbb{R}^n \times \mathbb{R}^n \rightarrow \mathbb{R}$$

be given Lipschitzian functions.

Consider the problem

$$(4.2) \quad \text{minimize } \left\{ g(x(0), x(1)) + \int_0^1 \phi(x(t)) dt: x \in S_1, x(0) \in C_0, x(1) \in C_1 \right\}.$$

Let z be a solution of (4.2). We wish to prove the maximum principle by using a “linearization” of differential inclusion (4.1) along z . For this we consider:

I. A family of closed convex processes $A(t): \mathbb{R}^n \rightrightarrows \mathbb{R}^n$ where $t \in [0, 1]$ satisfying

$$(4.3) \quad \begin{aligned} t \rightarrow A(t)(x) &\text{ is a measurable map,} \\ A(t)(x) &\subset d \text{co } F(z(t), z'(t))(x), \end{aligned}$$

for all $x \in \mathbb{R}^n$.

II. A family of convex functions $\psi(t): \mathbb{R}^n \rightarrow \mathbb{R}$, where $t \in [0, 1]$, such that for all $u \in \mathbb{R}^n$

$$(4.4) \quad \begin{aligned} & t \rightarrow \psi(t)(u) \text{ is measurable,} \\ & \psi(t)(u) \geq i_+ \phi(z(t))(u), \quad \psi(t)(0) = 0, \\ & \psi(t)(\cdot) \text{ is positively homogeneous.} \end{aligned}$$

III. A convex function $\psi_1: \mathbb{R}^n \times \mathbb{R}^n \rightarrow \mathbb{R}$ such that

$$(4.5) \quad \begin{aligned} & \psi_1 \geq i_+ g(z(0), z(1)), \quad \psi_1(0) = 0, \\ & \psi_1 \text{ is positively homogeneous.} \end{aligned}$$

IV. Closed convex cones $P, Q \subset \mathbb{R}^n$ satisfying

$$(4.6) \quad P \subset I_{C_0}(z(0)),$$

$$(4.7) \quad \emptyset \neq \text{Int } Q \subset D_{C_1}(z(1))$$

We denote their negative polars by

P^-, Q^- , respectively.

Let us consider the linearization

$$(4.8) \quad \begin{aligned} & w'(t) \in \text{cl}(A(t)w(t) + T_{\text{co } F(z(t))}(z'(t))) \quad \text{a.e. in } [0, 1], \\ & w(0) \in P, \end{aligned}$$

and let $R(1)$ denote the reachable set of (4.8) at time 1, i.e.,

$$R(1) = \{w(1): w \in W^{1,1}(0, 1) \text{ is a solution of (4.8)}\}.$$

Remark. Lemma 2.8 implies that $\text{cl}(A(t)w + T_{\text{co } F(z(t))}(z'(t))) \subset d \text{ co } F(z(t), z'(t))(w)$.

THEOREM 4.9. Assume that $z \in S_1$ solves problem (4.2), F has compact images and is Lipschitzian on a neighborhood of $z([0, 1])$. Let $\{A(t): t \in [0, 1]\}$ be a family of uniformly Lipschitzian closed convex processes, let $\{\psi(t): t \in [0, 1]\}$ be a family of uniformly Lipschitzian convex functions, let ψ_1 be a Lipschitzian convex function, and let P, Q be closed convex cones such that the relations (4.3)–(4.7) are satisfied.

If the following surjectivity assumption holds true

- (i) $R(1) - Q = \mathbb{R}^n$, or equivalently
 - (ii) there exists a solution \bar{w} of (4.8) such that $\bar{w}(1) \in \text{Int } Q$,
- then there exists a $q \in W^{1,\infty}(0, 1)$ satisfying

- (a) the adjoint inclusion:

$$q'(t) \in A^*(t)(-q(t)) + \partial\psi(t)(0),$$

- (b) the maximum principle:

$$\langle q(t), z'(t) \rangle = \max_{e \in F(z(t))} \langle q(t), e \rangle,$$

- (c) the transversality condition:

$$(q(0), -q(1)) \in \partial\psi_1(0) + P^- \times Q^-.$$

Remark. In particular if for a.e. $t \in [0, 1]$ the set-valued map F is soft at $(z(t), z'(t))$, then by Lemma 2.7 the theorem holds with $A(t) = dF(z(t), z'(t))$.

Proof. Fix any number $p > 1$. We introduce the following notation:

$$E = L^p(0, 1), \quad W = W^{1,p}(0, 1), \quad T = \mathbb{R}^n \times \mathbb{R}^n,$$

$$\gamma \in \mathcal{L}(W, T): \gamma(x) = (x(0), x(1)),$$

$$L \in \mathcal{L}(W, E): Lx = \dot{x},$$

$$f: E \rightarrow E: f(x) = \int_0^1 \phi(x(t)) dt,$$

$$\mathcal{F}: E \rightrightarrows E: \mathcal{F}(x) = \{y \in E: y(t) \in F(x(t)) \text{ a.e.}\}.$$

With this new notation, z solves the following problem

$$\text{minimize } \{f(x) + g \circ \gamma(x): Lx \in \mathcal{F}(x); x(0) \in C_0, x(1) \in C_1\}.$$

For all $t \in [0, 1]$, $x \in \mathbb{R}^n$ set

$$G(t, x) := \text{cl} (A(t)x + T_{\infty F(z(t))}(z'(t))).$$

The set-valued map G is measurable in t and Lipschitzian in x . Moreover, for almost all $t \in [0, 1]$ graph $G(t, \cdot)$ is a closed convex cone. Hence by the time-dependent version of the Filippov theorem (see Clarke [7, p. 115]), $W^{1,\infty}(0, 1)$ -solutions of (4.8) are dense in $W^{1,1}(0, 1)$ -solutions of (4.8). Therefore we may assume

(ii)' there exists a solution $\bar{w} \in W^{1,\infty}(0, 1)$ of (4.8) such that $\bar{w}(1) \in \text{Int } Q$.

Moreover the set reachable at time 1 by $W^{1,\infty}(0, 1)$ -solutions of (4.8) is dense in $R(1)$.

We shall proceed in several steps.

Step 1. Denote by SL the set of all solutions of (4.8) belonging to $W^{1,p}(0, 1)$ which satisfy $w(1) \in Q$. We claim that for all $w \in SL$

$$(4.10) \quad \int_0^1 \psi(t)(w(t)) dt + \psi_1(\gamma w) \geq 0.$$

Indeed by the convexity of SL and the remarks preceding Step 1

$$(4.11) \quad \{w(1): w \in SL\} - Q = \mathbb{R}^n.$$

Let $\bar{w} \in W^{1,\infty}(0, 1)$ be as in assumption (ii)'. Then for all $n \geq 1$ and $w \in SL$ the function

$$w_n = \left(1 - \frac{1}{n}\right)w + \frac{1}{n}\bar{w} \in SL.$$

Moreover,

$$(4.12) \quad \begin{aligned} w_n(1) &\subset D_{C_1}(z(1)), \\ w_n &\rightarrow w \quad \text{in } W^{1,p}(0, 1). \end{aligned}$$

Since $\{\psi(t)\}$ are uniformly Lipschitzian and ψ_1 is Lipschitzian it is enough to prove (4.10) for all w_n . So fix n . By Lemma 2.8 and Theorem 3.4 for all $h > 0$ there exist $w_n^h \in W^{1,1}(0, 1)$ satisfying $z + hw_n^h \in S_1$, $w_n^h \rightarrow w_n$ in $C(0, 1)$.

By Lipschitzianity of F and Filippov's theorem (see Aubin-Cellina [2, p. 120]) we may assume that $z(0) + hw_n^h(0) \in C_0$. Moreover by (4.12) for all h sufficiently small $z(1) + hw_n^h(1) \in C_1$. Since z is a solution and f, g are Lipschitzian, we obtain

$$\begin{aligned} 0 &\leq \limsup_{h \rightarrow 0+} \frac{f(z + hw_n^h) - f(z)}{h} + \limsup_{h \rightarrow 0+} \frac{g \circ \gamma(z + hw_n^h) - g \circ \gamma(z)}{h} \\ &= i_+ f(z)(w_n) + i_+ g(\gamma z)(\gamma w_n). \end{aligned}$$

Lemma 2.14 and assumptions (4.4), (4.5) end the proof of Step 1.

Step 2. We claim that for all $u, e \in E$ there exists a solution $w \in W^{1,p}(0, 1)$ of the differential inclusion

$$w'(t) \in G(t, w(t) + u(t)) + e(t),$$

$$w(0) \in P, \quad w(1) \in Q.$$

Indeed since $G(t, \cdot)$ are uniformly Lipschitzian there exist $w_1 \in W^{1,p}(0, 1)$ satisfying

$$w'_1(t) \in G(t, w_1(t) + u(t)) + e(t) \quad \text{a.e.},$$

$$w_1(0) = 0$$

(see [7, p. 115]). On the other hand by (4.11) there exists $w_2 \in W^{1,p}(0, 1)$ satisfying

$$w'_2(t) \in G(t, w_2(t)),$$

$$w_2(0) \in P,$$

$$-w_1(1) \in w_2(1) - Q.$$

Consider then $w = w_1 + w_2$.

By convexity, for almost all t

$$\begin{aligned} w'(t) &= w'_1(t) + w'_2(t) \in G(t, w_1(t) + u(t)) + e(t) + G(t, w_2(t)) \\ &\subset G(t, w(t) + u(t)) + e(t) \end{aligned}$$

and

$$w(0) = w_2(0) \in P,$$

$$w(1) = w_1(1) + w_2(1) \in Q,$$

which ends the proof of Step 2.

Step 3. Consider the closed convex process $Q: E \rightrightarrows E$ given by

$$Q(x) = \{y \in E: y(t) \in G(t, x(t)) \text{ a.e.}\}$$

and let $N_{\text{co } F(z(t))}(z'(t))$ denote the normal cone (of convex analysis) to $\text{co } F(z(t))$ at $z'(t)$.

We claim that the transposed process $Q^*: E^* \rightrightarrows E^*$ is given by $p \in Q^*(q)$ if and only if for almost all $t \in [0, 1]$

$$(4.13) \quad q(t) \in -N_{\text{co } F(z(t))}(z'(t)),$$

$$(4.14) \quad p(t) \in A(t)^*(q(t)).$$

Indeed fix $q \in E^*$ and $p \in Q^*(q)$. Then for all $(x, y) \in \text{graph } Q$ $\langle p, x \rangle \leq \langle q, y \rangle$. Since $0 \in G(t, 0)$ it implies that for all measurable set $A \subset [0, 1]$ and all measurable functions $x, y: A \rightrightarrows B$ satisfying $y(t) \in G(t, x(t))$ a.e. in A

$$(4.15) \quad \int_A \langle (p(t), -q(t)), (x(t), y(t)) \rangle dt \leq 0.$$

Let A' be the set of all $t \in [0, 1]$ such that

$$\sup_{(a,b) \in \text{graph } G(t, \cdot)} \langle (p(t), -q(t)), (a, b) \rangle > 0.$$

It is measurable. We check that its measure $\mu(A')$ is zero. Indeed there exists a measurable set $A \subset A'$ and $\varepsilon > 0$ such that $\mu(A') = 2\mu(A)$ and for all $t \in A$, $\sup \{ \langle (p(t), -q(t)), (a, b) \rangle: (a, b) \in (B \times B) \cap \text{graph } G(t, \cdot) \} \geq \varepsilon$. The set-valued map

$$A \ni t \rightarrow \text{graph } G(t, \cdot) \cap \{(a, b) \in B \times B: \langle (p(t), -q(t)), (a, b) \rangle \geq \varepsilon\}$$

is measurable with closed nonempty images. Thus it has a measurable selection, a function $u(t) \in \text{graph } G(t, \cdot) \cap B \times B$ such that for all $t \in A$ $\langle p(t), -q(t) \rangle, u(t) \rangle \geq \varepsilon$. Inequality (4.15) implies then that $\mu(A) = 0$. Hence for almost all $t \in [0, 1]$ and all $a \in \mathbb{R}^n$, $b \in \text{cl}(A(t)(a) + T_{\text{co}F(z(t))}(z'(t)))$, $\langle p(t), a \rangle \leq \langle q(t), b \rangle$. Setting $a = 0$ we deduce (4.13). Because zero is contained in the tangent cone we also obtain that for almost all t and all $(a, b) \in \text{graph } A(t)$ $\langle p(t), a \rangle \leq \langle q(t), b \rangle$. This implies (4.14). One can easily check that if $p, q \in E^*$ satisfy the inclusions (4.13), (4.14) then $p \in Q^*(q)$.

Step 4. For all $w \in W$ set

$$\begin{aligned}\pi(w) &:= \int_0^1 \psi(\tau)(w(\tau)) \, d\tau, \\ \omega(t) &= \begin{cases} \psi_1(t) & \text{if } t \in P \times Q, \\ +\infty & \text{otherwise,} \end{cases} \\ \Pi &= \partial\pi(0), \quad \Omega = \partial\omega(0)\end{aligned}$$

(the subdifferentials of π and ω at zero in the sense of convex analysis). Steps 1, 2 and Lemma 3.4 [12] imply that there exist $\bar{q} \in W^{1,p^*}(0, 1)$ (where $1/p + 1/p^* = 1$) and $\xi \in \Pi \subset L^\infty(0, 1)$ satisfying

$$-\bar{q}' \in \xi + Q^*(\bar{q}), \quad (-\bar{q}(0), \bar{q}(1)) \in \Omega.$$

By Lemma 2.14, $\xi(t) \in \partial\psi(t)(0)$ a.e. Set $q = -\bar{q}$. Step 3 implies that q satisfies the adjoint inclusion (a) and the maximum principle (b). Because $\{A(t)^*(-q(t)): t \in [0, 1]\}$ is a bounded set and $\xi \in L^\infty(0, 1)$ we also obtain that $q' \in L^\infty(0, 1)$.

Since $\Omega = \partial\psi_1(0) + P^- \times Q^-$ we obtain the transversality condition (c). So the proof is complete. \square

5. An example. Let U be a compact metric space and let a continuous function $f: \mathbb{R}^n \times U \rightarrow \mathbb{R}^n$ be given. Consider the control system

$$(5.1) \quad x' = f(x, u(t)), \quad u(t) \in U.$$

We denote by S_1 the set of solutions of (5.1).

Let two subsets $C_0, C_1 \subset \mathbb{R}^n$ and Lipschitzian functions $\phi: \mathbb{R}^n \rightarrow \mathbb{R}$, $g: \mathbb{R}^n \times \mathbb{R}^n \rightarrow \mathbb{R}$ be given.

We study the problem

$$\text{minimize } \left\{ g(x(0), x(1)) + \int_0^1 \phi(x(t)) \, dt : x \in S_1, x(0) \in C_0, x(1) \in C_1 \right\}.$$

Assume that a trajectory-control pair (z, \bar{u}) solves the above problem and let $N_{C_0}(z(0))$, $N_{C_1}(z(1))$ denote the Clarke normal cone to C_0 at $z(0)$ and to C_1 at $z(1)$, respectively.

Consider the linear inclusion

$$(5.2) \quad \begin{aligned} w' &\in \frac{\partial f}{\partial x}(z(t), \bar{u}(t))w + T_{\text{co}f(z(t), U)}(z'(t)), \\ w(0) &\in C_{C_0}(z(0)). \end{aligned}$$

We denote by

$$R(1) = \{w(1): w \in W^{1,1}(0, 1) \text{ is a solution of (5.2)}\}.$$

THEOREM 5.3. Assume that there exist $L > 0$ and an open neighborhood V of $z([0, 1])$ such that $f(\cdot, u)$ is L -Lipschitzian for all $u \in U$, and for almost all t , the derivatives

$(\partial f/\partial x)(z(t), \bar{u}(t))$, $\phi'(z(t))$ and the derivative $g'(z(0), z(1))$ are well-defined. Assume further that Clarke's tangent cone $C_{C_1}(z(1))$ has a nonempty interior. If the following surjectivity assumption holds true:

- (i) $R(1) - C_{C_1}(z(1)) = \mathbb{R}^n$,
 - (ii) there exists a solution \bar{w} of (5.2) such that $\bar{w}(1) \in \text{Int } C_{C_1}(z(1))$,
- then there exists $q \in W^{1,\infty}(0, 1)$ satisfying

$$q'(t) = \phi'(z(t)) - \frac{\partial f}{\partial x}(z(t), \bar{u}(t))^* q(t),$$

$$\langle q(t), f(z(t), \bar{u}(t)) \rangle = \max_{u \in U} \langle q(t), f(z(t), u) \rangle,$$

$$(q(0), -q(1)) \in g'(z(0), z(1)) + N_{C_0}(z(0)) \times N_{C_1}(z(1)).$$

Proof. One can easily verify that for $F(\cdot) = f(\cdot, U)$

$$\frac{\partial f}{\partial x}(z(t), \bar{u}(t)) \in dF(z(t), z'(t)) \subset d \text{ co } F(z(t), z'(t)).$$

Moreover, for all t the map

$$A(t): w \rightarrow \frac{\partial f}{\partial x}(z(t), \bar{u}(t))w$$

is L -Lipschitzian closed convex process and

$$A(t)^*(w) = \frac{\partial f}{\partial x}(z(t), \bar{u}(t))^* w.$$

Theorem 4.9 ends the proof. \square

REFERENCES

- [1] J. P. AUBIN, *Applied Functional Analysis*, Wiley Interscience, New York, 1979.
- [2] J. P. AUBIN AND A. CELLINA, *Differential Inclusions*, Springer-Verlag, 1984.
- [3] J. P. AUBIN AND I. EKELAND, *Applied Nonlinear Analysis*, Wiley Interscience, New York, 1984.
- [4] J. P. AUBIN, H. FRANKOWSKA AND C. OLECH, *Controllability of convex processes*, this Journal, 24 (1986), pp. 1192-1211.
- [5] H. BERLIOCCI AND J. M. LASRY, *Principe de Pontriagin pour des systèmes régis par une équation différentielle multivoque*, Note de CRAS 277 (1973), pp. 1103-1105.
- [6] F. H. CLARKE, *Generalized gradient and applications*, Trans. Amer. Math. Soc., 205 (1975), pp. 247-262.
- [7] ———, *Optimization and Non-smooth Analysis*, Wiley Interscience, New York, 1983.
- [8] B. CORNET, *Regularity properties of tangent and normal cones*, Cah. Mat. Décision, Univ. Paris-Dauphine (1981), no. 81-30.
- [9] A. I. DUBOVICKII AND A. M. MILJUTINE, *Extremum problems with constraints*, Soviet Math., 4 (1963), pp. 452-455.
- [10] A. F. FILIPPOV, *Classical solutions of differential equations with multivaried right-hand side*, this Journal, 5 (1967), pp. 609-621.
- [11] H. FRANKOWSKA AND C. OLECH, *Boundary solutions to differential inclusions*, J. Differential Equations, 44 (1982), pp. 156-165.
- [12] H. FRANKOWSKA, *The adjoint differential inclusions associated to a minimal trajectory of a differential inclusion*, Annales nonlinéaires de I.H.P., no. 2 (1985), pp. 75-99.
- [13] ———, *Local controllability and infinitesimal generators of semi-groups of set-valued maps*, this Journal, 25 (1987), to appear.
- [14] ———, *The first order necessary conditions for nonsmooth variational and control problems*, this Journal, 22 (1984), pp. 1-12.

- [15] A. IOFFE, *Nonsmooth analysis: Differential calculus of nondifferentiable mappings*, Trans. Amer. Math. Soc., 266 (1981), pp. 1–56.
- [16] S. ROBINSON, *Normed convex processes*, Trans. Amer. Math. Soc., 174 (1972), pp. 127–140.
- [17] R. T. ROCKAFELLAR, *Generalized directional derivatives and subgradients of nonconvex functions*, Canad. J. Math., 32 (1980), pp. 257–280.
- [18] ———, *Convex Analysis*, Princeton University Press, Princeton, NJ, 1970.
- [19] ———, *La théorie des sous-gradients et ses applications à l'optimization*, Les Presses de l'Université de Montréal, Montréal, 1978.
- [20] J. S. TREIMAN, *Characterization of Clarke's tangent and normal cones in finite and infinite dimensions*, Nonlinear Anal., 7 (1983), pp. 771–783.
- [21] C. URSESCU, *Multifunctions with closed convex graph*, Czech. Math. J., 25 (1975), pp. 438–441.
- [22] ———, *Tangent set's calculus and necessary conditions for extremality*, this Journal, 20 (1982), pp. 563–574.
- [23] W. H. WAGNER, *Survey of measurable selection theorems*, this Journal, 15 (1977), pp. 859–903.
- [24] T. WAZEWSKI, *On an optimal control problem*, Proc. Conference, Differential Equations and Their Applications, Prague, 1964, pp. 229–242.

A GENERAL THEOREM ON LOCAL CONTROLLABILITY*

H. J. SUSSMANN†

Abstract. We prove a general sufficient condition for local controllability of a nonlinear system at an equilibrium point. Earlier results of Brunovsky, Hermes, Jurdjevic, Crouch and Byrnes, Sussmann and Grossmann, are shown to be particular cases of this result. Also, a number of new sufficient conditions are obtained. All these results follow from one simple general principle, namely, that local controllability follows whenever brackets with certain symmetries can be “neutralized,” in a suitable way, by writing them as linear combinations of brackets of a lower degree. Both the class of symmetries and the definition of “degree” can be chosen to suit the problem.

Key words. nonlinear control, local controllability, nilpotent approximation, symmetries

AMS(MOS) subject classifications. 49B10, 93C10

Introduction. In recent years, several papers have been published giving sufficient conditions for a nonlinear control system to be locally controllable from a point. (Cf. Brunovsky [3], Crouch and Byrnes [5], Grossmann [8], Hermes [10]–[13], Jurdjevic [14], Stefani [21], [22], Sussmann [24], [25].) The purpose of this article is to prove a general theorem which contains all these results as particular cases and, in addition, gives stronger results. Our result (Theorem 2.4) shows that many known sufficient conditions can be derived in a unified way from a single general principle, namely, the combination of a nilpotent approximation with the use of input symmetries. Section 2 is devoted to the statement of the main theorem, preceded by an outline of the basic facts and definitions needed for its formulation. Section 3 reviews the basic formalism needed to set up the nilpotent approximation, and proves a number of technical lemmas needed to turn this approximation into a tool for establishing local controllability results. Sections 4 and 5 introduce the basic ingredients of our main result, namely, dilations and invariant elements. The proof of the main theorem is then given in § 6. In § 7, we review in detail the various controllability results referred to above, and explain how they all follow from our result. We also prove stronger versions of several of those theorems, and some new sufficient conditions. Finally, in § 8 we discuss some of the limitations of our method. In addition to the observations of § 8, we remark that there are recent results by R. M. Bianchini and G. Stefani, as well as work by H. Knobloch and K. Wagner, which provide new sufficient conditions that are not contained in the ones given here.

1. Preliminaries. The local controllability problem has a long history, beginning with the classical controllability theory for linear systems, and the first nonlinear local controllability result, namely, the theorem which states that if the linearization of a system at an equilibrium point p is controllable, then the system itself is locally controllable from p in small time (i.e. for every $T > 0$, the time T reachable set from p contains p in its interior; cf., for example, Lee and Markus [17]). This “small-time local controllability” property, henceforth abbreviated as STLC, is of interest to control theorists for a number of reasons, such as: (a) that a sufficient condition for STLC is obviously equivalent to a necessary condition for the constant trajectory $x(t) \equiv p$ to

* Received by the editors April 3, 1985. This research was partially supported by the National Science Foundation under grant MCS 78-02442.

† Mathematics Department, Rutgers University, New Brunswick, New Jersey 08903.

lie on the boundary of the attainable set from p ; since the simplest form of the Pontryagin Maximum Principle is precisely one such necessary condition, the STLC problem can be viewed as a particular case of the general problem of “high order optimality conditions;” (b) that STLC is equivalent to an important property of the optimal time function V , namely, continuity at p . (Here V is defined by letting $V(q)$ be the infimum of all the times T such that q can be reached from p in time T . If we wish to study the more common V function, defined in terms of the time it takes to steer q to p , then the continuity of V at p is equivalent to the STLC property for the system obtained by running the original system backwards.)

More recently, the problem has attracted the interest of “differential geometric control theorists,” i.e. of those who take the point of view that a control system is primarily a family of vector fields on a manifold, and a lot of the control-theoretically interesting information about the system should be contained in the Lie brackets of these vector fields (cf. [9], [16], [19], [23], [26]). At the early stages of the development of this “Lie theoretic approach,” attention was concentrated on proving those results that followed most naturally from the method. In particular, it was recognized right away that the Lie algebraic method yielded a complete characterization of local controllability for real analytic systems with the somewhat unnatural property of being “symmetric,” i.e. such that every trajectory run backwards is also a trajectory. (Hermann [9], Lobry [19]; the result is known as “Chow’s Theorem.”) On the other hand, for “reasonable” (i.e. not necessarily symmetric, but real-analytic) systems, the method yielded a complete characterization of a property which is related to, but not quite the same as, STLC. D. Elliott introduced the name “accessibility property” to refer to the property that the reachable set from p has an interior point. The so-called “positive form of Chow’s Theorem” (Krener [16]; cf. also [23]) characterizes this property in terms of Lie brackets. In 1974, P. Brunovsky [3] started from the observation that the “Lie theoretic” theorem about symmetric systems does not even give the most classical of all local controllability theorems, namely, the one for linear systems. He then proceeded to single out a class of systems (called “odd systems”) which could be proved to be STLC by Lie theoretic methods, and contained the class of linear systems. Since then, other local controllability results have been proved, as indicated above. The common feature of all these results is the exploitation of certain “structural symmetries” of a problem.

The traditional approach towards proving local controllability theorems has been to construct “control variations.” Heuristically, if one can construct control variations in all possible directions, then the reachable set ought to be a full neighborhood of the starting point. The argument can usually be made rigorous by some topological consideration. Ideally, the construction of variations in various directions should involve Lie bracket calculations. In practice, however, these calculations become rather cumbersome, and a different method is desirable which would construct, once and for all, a large collection of variations. One such method was used by us in [25], to prove a conjecture of H. Hermes. (Our earlier paper [24], which proved a different sufficient condition for STLC, was based on constructing variations, and it has only recently become clear to us that the result of [24] also follows using the method of [25].) The goal of the present paper is to prove the most general result that can be obtained by means of the method of [25].

A brief outline of the approach is as follows. Since a control system is primarily a family $\mathcal{V} = \{V_i: i \in I\}$ of vector fields, one can associate with it the Lie algebra $L(\mathcal{V})$ of vector fields generated by \mathcal{V} . Forgetting about rigor, one can think about “the Lie group” $G(\mathcal{V})$ with Lie algebra $L(\mathcal{V})$, and obtain an “action” of $G(\mathcal{V})$ on the state

space of the system. A $g \in G(\mathcal{V})$ is a product of exponentials $\exp(t_j V_{i_j})$, and therefore the result gp of acting on p by g is a point obtained by starting from p and following integral trajectories of the V_i , with switchings of vector fields allowed, and with motion “backwards in time” permitted as well. Those g ’s for which all the t_j are positive constitute a subsemigroup S of $G(\mathcal{V})$, which gives rise to the true trajectories of the control system. The reachable set from p is $S \cdot p$. For \mathcal{V} to be locally controllable from p , $S \cdot p$ has to have a nonempty interior, and so $G(\mathcal{V})p$ must be open, which means that, at least locally, $G(\mathcal{V})$ has to act transitively. If H is the isotropy group at p of this action, then a sufficient condition for p to be an interior point of $S \cdot p$ is that the interior of S in $G(\mathcal{V})$ should contain an element of H .

To make this rigorous, an algebraic formalism is needed to surmount the obstacles arising from the fact that $L(\mathcal{V})$ is, in general, infinite dimensional, and therefore $G(\mathcal{V})$ is not a well defined “Lie group.” Rather than work with $L(\mathcal{V})$ one works formally, with a free Lie algebra $L(\mathbf{X})$ in indeterminates X_i , and with its completion, the Lie algebra $\hat{L}(\mathbf{X})$ of formal Lie series in the X_i . Then there is a well defined group $\hat{G}(\mathbf{X})$, the group of exponentials of Lie series (cf., for example, Serre [20]). The controls can be embedded in $\hat{G}(\mathbf{X})$ as a subsemigroup S , by means of a map which assigns to each control a noncommutative formal power series, obtained by solving the differential equation of the system formally, using the indeterminates rather than the vector fields. (This map, introduced by Chen in [4], has been extensively used in control theory by M. Fliess, under the name of “Chen series,” cf. [6], [7].) Although obvious convergence and integrability difficulties arise if one tries to make $\hat{G}(\mathbf{X})$ act on the state space, the subsemigroup S does act in an obvious way, since S is identified with the set of admissible controls. And the series of a control $u(\cdot)$ contains a lot of information about the action of $u(\cdot)$. (More precisely, it is an asymptotic series, and it converges in the analytic case if $u(\cdot)$ is sufficiently small, cf. [1], [2], [7], [18], [25].) Since $\hat{G}(\mathbf{X})$ is not yet a true Lie group, one then makes a nilpotent approximation $G^N(\mathbf{X})$ of $\hat{G}(\mathbf{X})$ by killing all brackets of degree $> N$. If I is finite, $G^N(\mathbf{X})$ is now a Lie group in the usual sense. Then there is a corresponding approximating semigroup S^N . Although it is not possible in general to have $G^N(\mathbf{X})$ act on the state space, one can still define an “approximate action” and an “approximate isotropy group.” To get local controllability one must be able to prove (modulo technicalities) that the interior of S^N intersects the isotropy group. This we do by proving a general lemma that says that the interior of S^N always contains an element of a “very special form.” It then follows that, if one hypothesizes that all these “very special” elements are in the isotropy group, one gets controllability. As will be made clear in § 7, all known local controllability theorems amount to various forms of this hypothesis.

The special elements are obtained as the fixed points of the action of a finite group Λ on “input symmetries.” An input symmetry is, roughly, a linear map from $L(\mathbf{X})$ to $L(\mathbf{X})$ whose exponential maps S to S . Examples of such symmetries are: (a) multiplying a control by -1 , if its range of values permits it; (b) interchanging two controls; (c) time reversal. If a system has many symmetries, then there will be few Λ -fixed elements, and the resulting local controllability theorem will be very strong. As an example, we remark that the introduction of time-reversal, which was not used in [25], enables us here to prove a result which is considerably stronger than the Hermes conjecture proved in [25].

It turns out that the condition that certain “special elements” of the semigroups S^N be in the “isotropy group” can be rephrased, by passing to the logarithms, as the requirement that certain Lie brackets should vanish at p . It then becomes apparent that one can do slightly better. The brackets need not vanish. It suffices for them to

be “neutralized,” i.e. expressible as linear combinations of brackets of lower degree. And there is a certain amount of freedom as to the concept of “degree” to be used. One can use any one-parameter group of dilations to define “degree,” provided that certain technical conditions hold.

In order to avoid unnecessary complications, we will only work with systems that can be studied using a free Lie algebra generated by a *finite* set of indeterminates. That is, we will only study systems where the collection \mathcal{V} of associated vector fields is either finite, or a set of linear combinations of a finite set of vector fields. That is, we will only work with systems of the form

$$(1.1) \quad \dot{x} = \sum_{i=1}^k v_i g_i(x)$$

where the control $v = (v_1, \dots, v_k)$ is required to satisfy a constraint $v \in J$, where J is some subset of \mathbb{R}^k . It is then clear that we can assume that J linearly spans \mathbb{R}^k . If J does not affinely span \mathbb{R}^k , let A be the affine hull of J . By making a linear change of coordinates, we may assume that A is the set $\{1\} \times \mathbb{R}^{k-1}$. Then the system (1.1) becomes

$$(1.2) \quad \dot{x} = f_0(x) + \sum_{i=1}^m u_i f_i(x)$$

with control constraint $u = (u_1, \dots, u_m) \in K$. (Here K is such that $J = \{1\} \times K$, and $m = k - 1$.)

If J affinely spans \mathbb{R}^k , then we let $f_0 \equiv 0$, $m = k$, $g_i = f_i$ for $i = 1, \dots, m$, $K = J$, $u_i = v_i$ for $i = 1, \dots, m$. Then our system is also of the form (1.2), with a control constraint $u \in K$, where K affinely spans \mathbb{R}^m . It is in this form that, from now on, all our systems will be expressed.

2. Statement of the main theorem. In this section we will state our main local controllability theorem. In order to get to the statement as quickly as possible, we will omit a number of definitions. Detailed definitions of all the concepts occurring in the statement are given in subsequent sections.

We consider control systems of the form

$$(2.1) \quad \dot{x} = f_0(x) + \sum_{i=1}^m u_i f_i(x), \quad x \in M$$

with a control constraint

$$(2.2) \quad u = (u_1, \dots, u_m) \in K$$

where

(CS1) M is a smooth (i.e. C^∞) manifold,

(CS2) $\mathbf{f} = (f_0, \dots, f_m)$ is an $(m+1)$ -tuple of C^∞ vector fields on M ,

and

(CS3) K is a subset of \mathbb{R}^m such that

$$(2.3) \quad \text{Aff}(K) = \mathbb{R}^m.$$

Here $\text{Aff}(K)$ denotes the affine hull of K , i.e. the set of all finite linear combinations $\sum \alpha_i u^i$ with the $u^i \in K$, $\alpha_i \in \mathbb{R}$, and $\sum_i \alpha_i = 1$.

To specify a system we must give M , f and K . So we will simply refer to the triple $\Sigma = (M, f, K)$ as the control system, it being understood that M , f and K are supposed to satisfy (CS1), (CS2) and (CS3).

An *admissible control* for Σ is a Lebesgue integrable, K -valued function defined on some interval $[0, T]$. If $u(\cdot): [0, T] \rightarrow K$ is an admissible control, a *trajectory* for $u(\cdot)$ is an absolutely continuous curve $x(\cdot): [0, T] \rightarrow M$ such that

$$(2.4) \quad \dot{x}(t) = f_0(x(t)) + \sum_{i=1}^m u_i(t) f_i(x(t))$$

for almost all $t \in [0, T]$.

If $q \in M$ is of the form $x(T)$ for some trajectory such that $x(0) = p$, then q will be said to be *reachable from p in time T* . The set of all q that are reachable from p in time T for the system $\Sigma = (M, f, K)$ is the *time T reachable set from p* , and will be denoted by $\text{Reach}(\Sigma, T, p)$. Also we write

$$(2.5) \quad \text{Reach}(\Sigma, \leq T, p) = \bigcup_{0 \leq t \leq T} \text{Reach}(\Sigma, t, p)$$

for $T \geq 0$.

The system Σ is *small-time locally controllable* (STLC) from p if p is an interior point of $\text{Reach}(\Sigma, \leq T, p)$ for all $T > 0$. An equivalent characterization of this condition involves the optimal time function $V_{\Sigma, p}$. We define $V_{\Sigma, p}(q)$ to be the infimum of those T such that q is reachable from p in time T . (If no such T exists, then $V_{\Sigma, p}(q) = +\infty$.) Then Σ is STLC from p if and only if $V_{\Sigma, p}$ is continuous at p .

One can also consider the reachable sets obtained by restricting the class of admissible controls. For instance, we let $\text{Reach}_{pc}(\Sigma, T, p)$, $\text{Reach}_{pc}(\Sigma, \leq T, p)$ be the reachable sets obtained by using *piecewise constant* controls, and we say that Σ is STLC_{pc} from p if p is an interior point of $\text{Reach}_{pc}(\Sigma, \leq T, p)$ for all $T > 0$. The sufficient condition stated below in our main theorem is for STLC. However, under the hypotheses of the theorem, STLC and STLC_{pc} are equivalent, as will be observed below (cf. Proposition 2.3), so that the distinction between these two types of controllability need not worry us here.

If K is compact and convex, then one can also consider the sets reachable by bang-bang controls. (A bang-bang control is a piecewise constant control with values in the set of extreme points of K .) The corresponding small-time local controllability property is denoted by STLC_{bb} . Again, Proposition 2.3 will show that STLC and STLC_{bb} are equivalent under the hypotheses of our main theorem.

If \mathcal{F} is a family of C^∞ vector fields on a manifold M , then $L(\mathcal{F})$ denotes the Lie algebra of vector fields generated by the elements of \mathcal{F} . If \mathcal{V} is any set of vector fields on M , and $p \in M$, then we write

$$(2.6) \quad \mathcal{V}(p) = \{V(p) : V \in \mathcal{V}\}.$$

The family \mathcal{F} is said to satisfy the *Lie algebra rank condition* (LARC) at p if $L(\mathcal{F})(p)$ is the whole tangent space of M at p . An \mathcal{F} -trajectory is a curve $x(\cdot)$ which is a finite concatenation of integral arcs of members of \mathcal{F} . (Note: if an integral arc γ of a member f of \mathcal{F} is reparametrized by reversing the sense of time, then the resulting curve is an integral arc of $-f$, and need not be an \mathcal{F} -trajectory, since $-f$ need not belong to \mathcal{F} .) The family \mathcal{F} has the *accessibility property* (AP) from p if, for every $T > 0$, the set of points that can be reached from p by \mathcal{F} -trajectories in time $\leq T$ has a nonempty

interior. The following is a standard result from accessibility theory (the “positive form of Chow’s Theorem,” cf. Krener [16], Sussmann and Jurdjevic [23]).

PROPOSITION 2.1. *Let \mathcal{F} be a family of C^∞ vector fields on a C^∞ manifold M . Then the LARC at p implies the AP from p . Conversely, the AP from p implies the LARC at p if M is a real-analytic manifold and the members of \mathcal{F} are real-analytic.*

To a system Σ of the form (2.1), with a control constraint (2.2), we associate the family \mathcal{F}_Σ whose members are all the vector fields $f_0 + \sum_{i=1}^m u_i f_i$, for $(u_1, \dots, u_m) \in K$. The hypothesis that $\text{Aff}(K) = \mathbb{R}^m$ implies that the linear span of the members of \mathcal{F}_Σ is precisely the same as the linear span of f_0, \dots, f_m . Therefore $L(\mathcal{F}_\Sigma) = L(\mathbf{f})$, so that \mathcal{F}_Σ satisfies the LARC at p if and only if \mathbf{f} does. On the other hand, it is easy to see that an \mathcal{F}_Σ trajectory is precisely the same as a trajectory of Σ for a piecewise constant control. Hence Σ cannot be STLC_{pc} from p unless \mathcal{F}_Σ satisfies the AP from p . On the other hand, if f_0, \dots, f_m are real analytic vector fields, \mathcal{F}_Σ satisfies the AP from p if and only if \mathbf{f} satisfies the LARC from p . Therefore, in the analytic case, it is no restriction to assume that \mathbf{f} satisfies the LARC from p , if we seek to characterize the STLC_{pc} property. Actually, it is easy to prove:

PROPOSITION 2.2. *A system Σ of the form (2.1), with a control constraint (2.2), and f_0, \dots, f_m real analytic, cannot be STLC from a point p unless \mathbf{f} satisfies the LARC from p .*

Moreover, when the LARC from p holds, the distinction between STLC , STLC_{pc} and STLC_{bb} disappears, as shown by the following result, whose proof is given in the Appendix.

PROPOSITION 2.3. *Let Σ be a system of the form (2.1), with a control constraint (2.2) that satisfies (2.3). Assume that \mathbf{f} satisfies the LARC at p .*

Let \tilde{K} be the closure of the convex hull of K , and let $\tilde{\Sigma}$ be the system $(M, \mathbf{f}, \tilde{K})$. Then $\tilde{\Sigma}$ is SLTC from p if and only if Σ is SLTC_{pc} from p .

In particular, Proposition 2.3 implies that, for an arbitrary Σ , the STLC and STLC_{pc} properties from p are equivalent, if \mathbf{f} satisfies the LARC at p . Also, if K is compact and convex, STLC and STLC_{bb} are equivalent.

The sufficient condition for STLC to be proved here involves two main ingredients, namely, a finite group of symmetries and a one-parameter group of dilations. The symmetries considered will be mappings of a Lie algebra which is naturally associated to our problem. Precisely, we consider $L(\mathbf{X})$, the free Lie algebra in the indeterminates $\mathbf{X} = (X_0, \dots, X_m)$. We will be interested in linear maps $\lambda : L(\mathbf{X}) \rightarrow L(\mathbf{X})$ which are not necessarily Lie algebra automorphisms, but have a weaker property which we now define.

Let L be a Lie algebra over \mathbb{R} . We define $[L]^k$ for $k = 1, 2, \dots$ by $[L]^1 = L$, $[L]^{k+1} = [L, L^k]$. Clearly, any Lie algebra automorphism of L maps each $[L]^k$ into itself. A linear mapping $\lambda : L \rightarrow L$ which is a linear isomorphism and satisfies $\lambda([L]^k) \subseteq [L]^k$ for each k will be called a *pseudoautomorphism* of L .

In the particular case when L is $L(\mathbf{X})$, the $[L]^k$ are the ideals $L_k(\mathbf{X})$, where $L_k(\mathbf{X})$ is the sum of all the homogeneous components $L^{j, \text{hom}}(\mathbf{X})$ of degree j , for $j \geq k$. If $\lambda : L(\mathbf{X}) \rightarrow L(\mathbf{X})$ is a pseudoautomorphism, then λ gives rise to a linear map $\hat{\lambda}$ from $\hat{L}(\mathbf{X})$ to $\hat{L}(\mathbf{X})$, where $\hat{L}(\mathbf{X})$ is the Lie algebra of formal Lie series in X_0, \dots, X_m . (If $S \in \hat{L}(\mathbf{X})$, and $S = \sum_{j=1}^{\infty} S_j$, where S_j is homogeneous of degree j , then $\hat{\lambda}$ is defined by $\hat{\lambda}(S) = \sum_j \lambda(S_j)$. The sum is well defined because, for each k , the only terms that may contribute to the homogeneous component of degree k are the $\lambda(S_j)$ for $j \leq k$.) It is clear that $\hat{\lambda}\hat{\mu} = \hat{\lambda}\hat{\mu}$ if λ, μ are pseudoautomorphisms.

The class of controls is embedded as a subsemigroup $\hat{S}(\mathbf{X}, K)$ of the group $\hat{G}(\mathbf{X}) = \{\exp(Z) : Z \in \hat{L}(\mathbf{X})\}$. A pseudoautomorphism λ of $L(\mathbf{X})$ gives rise to a mapping

$\lambda^\#$ from $\hat{G}(\mathbf{X})$ to $\hat{G}(\mathbf{X})$ by letting

$$(2.7) \quad \lambda^\#(\exp(Z)) = \exp(\hat{\lambda}(Z)) \quad \text{for } Z \in \hat{L}(\mathbf{X}).$$

An *input symmetry* for Σ is a pseudoautomorphism λ of $L(\mathbf{X})$ such that the corresponding map $\lambda^\#$ maps $\hat{S}(\mathbf{X}, K)$ to $\hat{S}(\mathbf{X}, K)$. (Actually, the definition of input symmetry only depends on m and K , and not on the particular choice of M, f_0, \dots, f_m .)

The second important ingredient is a one parameter group of dilations $\{\Delta(\rho): 0 < \rho < \infty\}$ of the linear space $V = L^{1,\text{hom}}(\mathbf{X})$. Then Δ gives rise to groups of dilations Δ^Λ, Δ^L of the free associative algebra $A(\mathbf{X})$ in the indeterminates X_0, \dots, X_m , and of $L(\mathbf{X})$, respectively. Also, one obtains a one-parameter group $\hat{\Delta}^\Lambda$ of automorphisms of the algebra $\hat{A}(\mathbf{X})$ of formal power series in X_0, \dots, X_m . We call Δ *compatible* with $\hat{S}(\mathbf{X}, K)$ if the $\hat{\Delta}^\Lambda(\rho)$ map $\hat{S}(\mathbf{X}, K)$ into itself for $0 < \rho \leq 1$. (Equivalently, Δ is compatible with $\hat{S}(\mathbf{X}, K)$ if and only if, for every $u = (u_1, \dots, u_m) \in K$ and every ρ such that $0 < \rho \leq 1$, $\Delta(\rho)(X_0 + \sum_{i=1}^m u_i X_i)$ is of the form $T(X_0 + \sum_{i=1}^m v_i X_i)$ for some $T > 0$, $v = (v_1, \dots, v_m) \in K$.)

A group of dilations Δ as above can be used to define the Δ -degree of an element Z of $A(\mathbf{X})$. We call Z Δ -homogeneous of degree r if $\Delta^\Lambda(\rho)(Z) = \rho^r Z$ for every ρ . If Z is arbitrary, then Z is a finite sum of homogeneous elements, and the Δ -degree of Z (denoted by $\deg_\Delta(Z)$) is the largest of the degrees of the homogeneous components of Z .

If $\mathbf{f} = (f_0, \dots, f_m)$ is an $(m+1)$ -tuple of C^∞ vector fields on a C^∞ manifold M , then we can consider the map $\text{Ev}(\mathbf{f})$ which assigns to every element P of $L(\mathbf{X})$ the vector field obtained by plugging in each f_i for the corresponding indeterminate X_i . If $p \in M$, then we also define the map $\text{Ev}_p(\mathbf{f})$, from $L(\mathbf{X})$ to the tangent space $T_p M$, given by $\text{Ev}_p(\mathbf{f})(P) = \text{Ev}(\mathbf{f})(P)(p)$.

We now define what it means for a $Z \in L(\mathbf{X})$ to be Δ -neutralized for \mathbf{f} at p . If Z is Δ -homogeneous, we say that Z is Δ -neutralized for \mathbf{f} at p if $\text{Ev}_p(\mathbf{f})(Z)$ can be expressed as a sum of vectors $\text{Ev}_p(\mathbf{f})(Q_j)$, where the Q_j are elements of $L(\mathbf{X})$ such that $\deg_\Delta(Q_j) < \deg_\Delta(Z)$. (Clearly, the Q_j can always be chosen to be Δ -homogeneous.) If Z is not necessarily homogeneous, then we write Z as a sum of homogeneous components, and we say that Z is Δ -neutralized for \mathbf{f} at p if each homogeneous component is.

With these definitions, our main result is the following.

THEOREM 2.4. *Let $\Sigma = (M, \mathbf{f}, K)$ be a control system, and let $p \in M$. Assume that:*

- (i) Σ satisfies the Lie algebra rank condition at p ,
- (ii) there exist (a) a finite group Λ of input symmetries and (b) a one-parameter group of dilations $\Delta = \{\Delta(\rho): \rho > 0\}$ of $L^{1,\text{hom}}(\mathbf{X})$ which is compatible with $\hat{S}(\mathbf{X}, K)$, such that every Λ -fixed element of $L(\mathbf{X})$ is Δ -neutralized for \mathbf{f} at p .

Then Σ is small-time locally controllable at p .

3. Exponential Lie series and the nilpotent approximation. We review the basic facts about the formalism of noncommutative power series and nilpotent approximation (cf. [4], [6], [25]). The idea is to solve (2.1) formally, by using indeterminates X_0, \dots, X_m rather than the vector fields f_0, \dots, f_m , and then regard a given control system as an action of a "Lie group" $\hat{G}(\mathbf{X})$ of exponential Lie series, together with the specification of a subsemigroup $\hat{S}(\mathbf{X}, K)$ which is identified with the class of controls. We now make this precise.

Let $\mathbf{X} = (X_0, \dots, X_m)$ be a finite sequence of indeterminates. We let $A(\mathbf{X})$ denote the free associative algebra over \mathbb{R} generated by the X_j . For any multiindex $I = (i_1, \dots, i_k)$, with $i_j \in \{0, \dots, m\}$ for $j = 1, \dots, k$, we let $X_I = X_{i_1} \cdots X_{i_k}$. Then $A(\mathbf{X})$ is the set of all sums $\sum_I a_I X_I$, where the coefficients a_I are real numbers, the summation

runs over all possible multiindices I , and all but finitely many a_I vanish. (It is understood that $X_\phi = 1$.)

We also let $\hat{A}(\mathbf{X})$ denote the set of all *formal power series* in the noncommuting indeterminates X_j , i.e. the set of all sums $\sum_I a_I X_I$ as above, except that the a_I are no longer required to vanish for all but finitely many I . In both $A(\mathbf{X})$ and $\hat{A}(\mathbf{X})$, addition is done componentwise, and multiplication is carried out using the formula $X_I X_J = X_{I*J}$, where $I * J$ is the concatenation of I and J (i.e. the multiindex obtained by writing, in order, first the components of I and then those of J).

For any nonnegative integer N , we use $A^{N,\text{hom}}(\mathbf{X})$ to denote the homogeneous component of degree N of $A(\mathbf{X})$, and $A^N(\mathbf{X})$ to denote the sum of the $A^{j,\text{hom}}(\mathbf{X})$ for $j = 0, \dots, N$. The space $A^N(\mathbf{X})$ is embedded as a linear subspace of $A(\mathbf{X})$ but, naturally, it is not a subalgebra. On the other hand, $A^N(\mathbf{X})$ is an algebra if one defines multiplication as in $A(\mathbf{X})$, with the extra proviso that monomials of degree greater than N are set equal to zero. Thus regarded, $A^N(\mathbf{X})$ is the *free nilpotent associative algebra of step $N+1$ in the indeterminates X_0, \dots, X_m* . Then $A^N(\mathbf{X})$ can be identified with the quotient of $A(\mathbf{X})$ by the ideal of all linear combinations of monomials of degree strictly larger than N . The canonical projection from $A(\mathbf{X})$ onto $A^N(\mathbf{X})$ is the *truncation map* τ_X^N . We will write τ^N rather than τ_X^N whenever the context makes it clear which \mathbf{X} is being referred to. Clearly, one can also think of $A^N(\mathbf{X})$ as a quotient of $\hat{A}(\mathbf{X})$. The corresponding truncation map from $\hat{A}(\mathbf{X})$ onto $A^N(\mathbf{X})$ will be denoted by $\hat{\tau}_X^N$ or $\hat{\tau}^N$. The kernels of $\tau^N, \hat{\tau}^N$ are denoted by $A_N(\mathbf{X}), \hat{A}_N(\mathbf{X})$, respectively. In particular, $\hat{A}_0(\mathbf{X})$ is the set of formal power series $\sum_I a_I X_I$ for which $a_\phi = 0$. The *exponential map* is a well defined bijection

$$(3.1) \quad \exp: \hat{A}_0(\mathbf{X}) \rightarrow 1 + \hat{A}_0(\mathbf{X})$$

whose inverse is a map from $1 + \hat{A}_0(\mathbf{X})$ to $\hat{A}_0(\mathbf{X})$ denoted by “log.” If $S \in \hat{A}_0(\mathbf{X})$, then $\exp(S)$ and $\log(1 + S)$ are given by the usual power series.

One can also define $A_k^N(\mathbf{X})$ to be the set of all elements of $A^N(\mathbf{X})$ that are linear combinations of monomials of degree $> k$. Then

$$(3.2) \quad A_k^N(\mathbf{X}) = \tau^N(A_k(\mathbf{X})) = \hat{\tau}^N(\hat{A}_k(\mathbf{X})).$$

The exponential map

$$(3.3) \quad \exp_N: A_0^N(\mathbf{X}) \rightarrow 1 + A_0^N(\mathbf{X})$$

and its inverse \log_N are given, in this case, by power series that are actually finite sums, due to the nilpotency of $A^N(\mathbf{X})$.

The algebras $A(\mathbf{X}), \hat{A}(\mathbf{X}), A^N(\mathbf{X})$ are Lie algebras in the usual way. We let $L(\mathbf{X})$ denote the Lie subalgebra of $A(\mathbf{X})$ generated by X_0, \dots, X_m . An element S of $A(\mathbf{X})$ will be said to be a *Lie element* iff $S \in L(\mathbf{X})$. It is clear that S is a Lie element iff all the homogeneous components of S are Lie elements. Therefore, if we let

$$(3.4) \quad L^{N,\text{hom}}(\mathbf{X}) = L(\mathbf{X}) \cap A^{N,\text{hom}}(\mathbf{X}),$$

we see that $L(\mathbf{X})$ is the direct sum of the $L^{N,\text{hom}}(\mathbf{X})$, $N = 1, 2, 3, \dots$.

The Lie algebra $L(\mathbf{X})$ is spanned by the *formal brackets* of X_0, \dots, X_m . Precisely, we define $\text{Br}(\mathbf{X})$ to be the smallest subset of $L(\mathbf{X})$ that contains X_0, X_1, \dots, X_m and is closed under bracketing. The elements of $\text{Br}(\mathbf{X})$ will be referred to as *brackets* of \mathbf{X} . It is clear that every $B \in \text{Br}(\mathbf{X})$ is homogeneous. (Notice that we have chosen not to define a “bracket” as a formal expression but as an element of $L(\mathbf{X})$ so that, for example, $[[X_0, X_1], [X_0, X_2]]$ and $[[X_1, X_0], [X_2, X_0]]$ are the same element of $\text{Br}(\mathbf{X})$.) Naturally, the elements of $\text{Br}(\mathbf{X})$ are not linearly independent. Several systematic

procedures for singling out subsets of $\text{Br}(\mathbf{X})$ that form bases of $L(\mathbf{X})$ can be found in the literature, cf., for example, [20], [27], but we shall not need those results here.)

We can also define $\hat{L}(\mathbf{X})$ to be the set of all formal sums $\sum_{N=1}^{\infty} S_N$ such that each S_N is in $L^{N,\text{hom}}(\mathbf{X})$, i.e. the set of those elements of $\hat{A}(\mathbf{X})$ all of whose homogeneous components are Lie. The members of $\hat{L}(\mathbf{X})$ will be referred to as *Lie elements* of $\hat{A}(\mathbf{X})$, and they clearly form a Lie subalgebra of $\hat{A}(\mathbf{X})$. The Lie algebras $L(\mathbf{X})$, $\hat{L}(\mathbf{X})$ are known, respectively, as the *free Lie algebra in the indeterminates* X_0, \dots, X_m and the *algebra of Lie series in* X_0, \dots, X_m .

Since $\hat{L}(\mathbf{X}) \subseteq \hat{A}_0(\mathbf{X})$, the exponential map is well defined on $\hat{L}(\mathbf{X})$. The elements of $\hat{A}(\mathbf{X})$ that are of the form $\exp(S)$ for some $S \in \hat{L}(\mathbf{X})$ are the *exponential Lie series* in X_0, \dots, X_m . The set of all such series is denoted by $\hat{G}(\mathbf{X})$. It follows from the Campbell-Hausdorff formula that $\hat{G}(\mathbf{X})$ is a group under multiplication. The exponential map, restricted to $\hat{L}(\mathbf{X})$, is a bijection from $\hat{L}(\mathbf{X})$ onto $\hat{G}(\mathbf{X})$, which will also be denoted by “exp,” while we will use “log” to denote the inverse map.

The group $\hat{G}(\mathbf{X})$ is almost “a Lie group whose Lie algebra is $\hat{L}(\mathbf{X})$,” but it fails to be a true Lie group, since it is infinite-dimensional. However, its truncated versions

$$(3.5) \quad G^N(\mathbf{X}) = \hat{\tau}^N(\hat{G}(\mathbf{X}))$$

are true Lie groups. (As for $\hat{G}(\mathbf{X})$ itself, it is a projective limit of the $G^N(\mathbf{X})$, but we will not make use of this fact.) Each $G^N(\mathbf{X})$ is a connected, simply connected, nilpotent Lie group, with Lie algebra $L^N(\mathbf{X})$, where

$$(3.6) \quad L^N(\mathbf{X}) = \tau^N(L(\mathbf{X})) = \hat{\tau}^N(\hat{L}(\mathbf{X})).$$

The exponential map from $L^N(\mathbf{X})$ to $G^N(\mathbf{X})$ is none other than the restriction of \exp_N to $L^N(\mathbf{X})$ (which is a subset of $A_0^N(\mathbf{X})$). We will therefore also use \exp_N to denote this map. Then \exp_N is a bijection from $L^N(\mathbf{X})$ onto $G^N(\mathbf{X})$, whose inverse map will, as expected, be denoted by \log_N . Then $L^N(\mathbf{X})$ is the *free nilpotent Lie algebra of step $N+1$ in* X_0, \dots, X_m , and we shall refer to the group $G^N(\mathbf{X})$ as the *free nilpotent Lie group of step $N+1$ infinitesimally generated by* X_0, \dots, X_m .

Now suppose that we are given a C^∞ manifold M and an $(m+1)$ -tuple $\mathbf{f} = (f_0, \dots, f_m)$ of C^∞ vector fields on M . Each f_j is therefore a member of $D(M)$, the algebra of all partial differential operators $P: C^\infty(M) \rightarrow C^\infty(M)$. (Here $C^\infty(M)$ denotes the space of C^∞ real-valued functions on M .) There is therefore a well defined *evaluation map*

$$(3.7) \quad \text{Ev}(\mathbf{f}): A(\mathbf{X}) \rightarrow D(M)$$

obtained by “plugging in the f_j for the X_j ,” so that

$$(3.8) \quad \text{Ev}(\mathbf{f}) \left(\sum_I a_I X_I \right) = \sum_I a_I f_I,$$

where, if $I = (i_1, \dots, i_k)$, we write

$$(3.9) \quad f_I = f_{i_1} f_{i_2} \cdots f_{i_k}.$$

The image $\text{Ev}(\mathbf{f})(A(\mathbf{X}))$ will be denoted by $A(\mathbf{f})$. Then $A(\mathbf{f})$ is the subalgebra of $D(M)$ generated by f_0, \dots, f_m . The evaluation map $\text{Ev}(\mathbf{f})$ can be restricted to $L(\mathbf{X})$. The corresponding map, which we will also denote by $\text{Ev}(\mathbf{f})$, is a surjective homomorphism from $L(\mathbf{X})$ onto $L(\mathbf{f})$, where $L(\mathbf{f})$ is the Lie algebra of vector fields generated by f_0, \dots, f_m .

The kernel of $\text{Ev}(\mathbf{f}): A(\mathbf{X}) \rightarrow A(\mathbf{f})$ is the set of all *algebraic identities satisfied by* f_0, \dots, f_m , and we will denote it by $\text{AI}(\mathbf{f})$. Similarly, the kernel of $\text{Ev}(\mathbf{f}): L(\mathbf{X}) \rightarrow L(\mathbf{f})$ is the set of *Lie algebraic identities satisfied by* f_0, \dots, f_m , and we denote it by $\text{LI}(\mathbf{f})$.

If p is a point in M , then we use $D_p(M)$ to denote the set of all partial differential operators at p , i.e. the quotient of $D(M)$ modulo the set of $P \in D(M)$ such that $(P\phi)(p) = 0$ for every $\phi \in C^\infty(M)$. Also, we let $T_p(M)$ denote the tangent space of M at p . We then have the *evaluation at p map* $\text{Ev}_p(\mathbf{f}): A(\mathbf{X}) \rightarrow D_p(M)$ given by $\text{Ev}_p(\mathbf{f})(S) = (\text{Ev}(\mathbf{f})(S))(p)$. The kernel of this map is the set of *algebraic relations among the f_j at p* , and will be denoted by $\text{AR}(\mathbf{f}, p)$. Similarly, $\text{Ev}_p(\mathbf{f})$ maps $L(\mathbf{X})$ to $T_p(M)$. The kernel of this map, denoted by $\text{LR}(\mathbf{f}, p)$, is the set of *Lie algebraic relations* (or, simply, *Lie relations*) among the f_j at p . (For instance, $[X_0, X_1] + X_2$ is a Lie identity satisfied by f_0, f_1, f_2 iff the vector field $[f_0, f_1] + f_2$ vanishes identically. Similarly, $[X_0, X_1] + X_2$ is a Lie relation among the f_j at p iff $[f_0, f_1] + f_2$ vanishes at p .)

The image $\text{Ev}_p(\mathbf{f})(L(\mathbf{X}))$ is precisely the subspace $L(\mathbf{f})(p)$ of $T_p(M)$, where

$$(3.10) \quad L(\mathbf{f})(p) = \{V(p) : V \in L(\mathbf{f})\}.$$

The system \mathbf{f} satisfies the *Lie algebra rank condition* (LARC) at p if $L(\mathbf{f})(p) = T_p(M)$, i.e. if $\text{Ev}_p(\mathbf{f})$ maps $L(\mathbf{X})$ onto $T_p(M)$.

The evaluation maps $\text{Ev}(\mathbf{f}), \text{Ev}_p(\mathbf{f})$ can formally be applied to series S in $\hat{A}(\mathbf{X})$, giving rise to formal infinite sums of partial differential operators (which, if $S \in \hat{L}(\mathbf{X})$, are vector fields). However, if one wishes to make sense of $\text{Ev}(\mathbf{f})(S)$ as a mathematical object in a rigorous way, technical difficulties arise. (For instance, suppose that f_0, f_1 are C^∞ vector fields that satisfy $[f_0, f_1] = f_1$, and S is the Lie series $\sum_{k=0}^{\infty} (-1)^k (\text{ad } X_0)^k(X_1)$. Should the general definition of $\text{Ev}(\mathbf{f})(S)$ be such that, in this particular case, $\text{Ev}(\mathbf{f})(S)$ is the zero series?) Rather than attempt to overcome these difficulties, we shall avoid them, by agreeing to refer to the series $\text{Ev}(\mathbf{f})(S)$ (or $\text{Ev}_p(\mathbf{f})(S)$) *only* as part of purely heuristic discussions which are not expected to be rigorous anyhow, *or* as part of statements that are given a precise mathematical translation. (For instance, the phrase “ $\text{Ev}_p(\mathbf{f})(S)$, applied to a function ϕ , is asymptotic to . . .” will be translated into a collection of inequalities involving only the truncations $\text{Ev}_p(\mathbf{f})(\hat{\tau}^N(S))$, in which only finite sums occur.)

We can also define truncated evaluation maps $\text{Ev}^N(\mathbf{f}), \text{Ev}_p^N(\mathbf{f})$ by restricting $\text{Ev}(\mathbf{f})$ and $\text{Ev}_p(\mathbf{f})$ to $A^N(\mathbf{X})$ or to $L^N(\mathbf{X})$. However, the algebra structure of $A^N(\mathbf{X})$ and the Lie algebra structure of $L^N(\mathbf{X})$ do not turn $A^N(\mathbf{X}), L^N(\mathbf{X})$ into subalgebras of $A(\mathbf{X}), L(\mathbf{X})$. This implies that $\text{Ev}^N(\mathbf{f})$ need not be an algebra homomorphism from $A^N(\mathbf{X})$ to $A(\mathbf{f})$ or from $L^N(\mathbf{X})$ to $L(\mathbf{f})$. Also, the point evaluation maps $\text{Ev}_p^N(\mathbf{f})$ are defined in an obvious way as maps from $A^N(\mathbf{X})$ to $D_p(M)$ and from $L^N(\mathbf{X})$ to $T_p(M)$.

If $\mathcal{E}_p: D(M) \rightarrow D_p(M)$ is the map $Q \rightarrow Q(p)$, then $\text{Ev}_p^N(\mathbf{f}) = \mathcal{E}_p \circ \text{Ev}^N(\mathbf{f})$. We use $\text{AI}^N(\mathbf{f}), \text{LI}^N(\mathbf{f}), \text{AR}^N(\mathbf{f}, p), \text{LR}^N(\mathbf{f}, p)$ to denote, respectively, the kernels of the maps $\text{Ev}^N(\mathbf{f}): A^N(\mathbf{X}) \rightarrow D(M)$, $\text{Ev}^N(\mathbf{f}): L^N(\mathbf{X}) \rightarrow L(\mathbf{f})$, $\text{Ev}_p^N(\mathbf{f}): A^N(\mathbf{X}) \rightarrow D_p(M)$ and $\text{Ev}_p^N(\mathbf{f}): L^N(\mathbf{X}) \rightarrow T_p(M)$. Then $\text{AI}^N(\mathbf{f})$ is the set of *algebraic identities of degree $\leq N$ among the f_i* , and similar self-explanatory names will be used for the other sets $\text{LI}^N(\mathbf{f}), \text{AR}^N(\mathbf{f}, p), \text{LR}^N(\mathbf{f}, p)$. Since, as indicated earlier, Ev^N need not be a homomorphism, the sets $\text{AI}^N(\mathbf{f})$ may fail to be ideals of $A^N(\mathbf{X})$, and the $\text{LI}^N(\mathbf{f})$ need not be ideals of $L^N(\mathbf{X})$. Also, $\text{AR}^N(\mathbf{f}, p)$ can fail to be a subalgebra of $A^N(\mathbf{X})$, and $\text{LR}^N(\mathbf{f}, p)$ may fail to be a Lie subalgebra of $L^N(\mathbf{X})$. (For instance, let $\mathbf{f} = (f_0, f_1, f_2)$, and suppose that $[f_0, f_1](p) = f_1(p)$, and $f_2(p) = 0$. Then $[X_0, X_1] - X_1 \in \text{LR}^2(\mathbf{f}, p)$ and $X_2 \in \text{LR}^2(\mathbf{f}, p)$. If $\text{LR}^2(\mathbf{f}, p)$ were a Lie subalgebra of $L^2(X_0, X_1, X_2)$, it would follow that $[[X_0, X_1], X_2] - [X_1, X_2]$ is in $\text{LR}^2(\mathbf{f}, p)$, i.e. that $[X_1, X_2]$ is in $\text{LR}^2(\mathbf{f}, p)$, since $[[X_0, X_1], X_2] = 0$ in $L^2(X_0, X_1, X_2)$. So $[f_1, f_2](p) = 0$. However, it is easy to construct f_0, f_1, f_2 that satisfy the conditions stated above as well as $[f_1, f_2](p) \neq 0$.)

As in [25], \mathcal{U}_m will denote the set of all functions $u(\cdot)$ whose domain $\text{Dom}(u(\cdot))$ is a compact interval of the form $[0, T]$, such that $u(\cdot)$ takes values in \mathbb{R}^m and is

Lebesgue integrable on $[0, T]$. The time T is the *terminal time* of $u(\cdot)$ and is denoted by $T(u(\cdot))$. If $0 \leq t \leq T(u(\cdot))$, then the restriction of $u(\cdot)$ to $[0, t]$ is denoted by $u^t(\cdot)$. The components of $u(\cdot)$ are $u_1(\cdot), \dots, u_m(\cdot)$, and we write $u_0(t) \equiv 1$.

If we consider the differential equation

$$(3.11) \quad \dot{S} = S \left(X_0 + \sum_{i=1}^{\infty} u_i X_i \right)$$

for an $\hat{A}(\mathbf{X})$ -valued function $t \rightarrow S(t)$, $0 \leq t \leq T(u(\cdot))$, with the initial condition $S(0) = 1$, then the solution is

$$(3.12) \quad S(t) = \sum_I \left(\int_0^t u_I \right) X_I,$$

where $\int_0^t u_I$ is the iterated integral

$$(3.13) \quad \int_0^t u_I = \int_0^t \int_0^{\tau_k} \int_0^{\tau_{k-1}} \cdots \int_0^{\tau_2} u_{i_k}(\tau_k) u_{i_{k-1}}(\tau_{k-1}) \cdots u_{i_1}(\tau_1) d\tau_1 \cdots d\tau_k$$

if $\phi \neq I = (i_1, \dots, i_k)$. (We let $\int_0^t u_\phi = 1$.)

The series $S(T(u(\cdot)))$, with $t \rightarrow S(t)$ given as above, is the *formal power series associated with the control* $u(\cdot)$, and will be denoted by $\text{Ser}(u(\cdot))$. The mapping $\text{Ser}: \mathcal{U}_m \rightarrow \hat{A}(\mathbf{X})$ is injective and, if \mathcal{U}_m is regarded as a semigroup under the operation of concatenation, and $\hat{A}(\mathbf{X})$ is equipped with multiplication, then Ser is a semigroup homomorphism (cf. [25, Lemma 3.1]). Moreover, $\text{Ser}(u(\cdot))$ is always an exponential Lie series (cf. [25, Prop. 3.1]), so that Ser actually takes values in $\hat{G}(\mathbf{X})$. The subsemigroup $\text{Ser}(\mathcal{U}_m)$ of $\hat{G}(\mathbf{X})$ will be denoted by $\hat{S}(\mathbf{X})$. Since Ser is injective, one should think of $\hat{S}(\mathbf{X})$ as being just another way of realizing the control semigroup \mathcal{U}_m , which has the particular advantage of exhibiting \mathcal{U}_m as embedded in a group.

If K is an arbitrary subset of \mathbb{R}^m , then we can consider $\mathcal{U}_m(K)$, the subsemigroup of \mathcal{U}_m whose elements are the K -valued controls. The image of $\mathcal{U}_m(K)$ under Ser will be denoted by $\hat{S}(\mathbf{X}, K)$.

One can also consider the truncated versions of the map Ser and the semigroups $\hat{S}(\mathbf{X})$, $\hat{S}(\mathbf{X}, K)$. The truncation map $\hat{\tau}^N$ maps solutions of (3.11) to solutions of the same equation, regarded now as evolving in $A^N(\mathbf{X})$. Hence, if we let

$$(3.14) \quad \text{Ser}_N(u(\cdot)) = \hat{\tau}^N(\text{Ser}(u(\cdot)))$$

we find that

$$(3.15) \quad \text{Ser}_N(u(\cdot)) = \sum_{|I| \leq N} \left(\int_0^{T(u(\cdot))} u_I \right) X_I.$$

Moreover, $\text{Ser}_N(u(\cdot)) \in G^N(\mathbf{X})$. The sets $\text{Ser}_N(\mathcal{U}_m)$, $\text{Ser}_N(\mathcal{U}_m(K))$ will be denoted by $S^N(\mathbf{X})$, $S^N(\mathbf{X}, K)$, respectively. Clearly, these subsets are subsemigroups of $G^N(\mathbf{X})$. Moreover, $S^N(\mathbf{X})$ is the set of points that can be reached from the identity element of $G^N(\mathbf{X})$ by trajectories of the system

$$(3.16) \quad \dot{S} = \tilde{F}_0^N(S) + \sum_{i=1}^m u_i \tilde{F}_i^N(S),$$

where \tilde{F}_i^N is the restriction to $G^N(\mathbf{X})$ of the linear vector field F_i^N on $A^N(\mathbf{X})$, given by $F_i^N(S) = S X_i$. (It is clear that \tilde{F}_i^N is tangent to $G^N(\mathbf{X})$, and therefore \tilde{F}_i^N is well defined.) The Lie algebra of vector fields generated by F_0^N, \dots, F_m^N is isomorphic to $L^N(\mathbf{X})$ in an obvious way, and therefore acts transitively on $G^N(\mathbf{X})$. From this it

follows, using general results from accessibility theory, that $S^N(\mathbf{X})$ has a nonempty interior relative to $G^N(\mathbf{X})$ and, moreover, this interior is dense in $S^N(\mathbf{X})$. More generally, $S^N(\mathbf{X}, K)$ is the reachable set from the identity corresponding to the system (3.16) with the additional control constraint $(u_1, \dots, u_m) \in K$. The Lie algebra associated with this system is the Lie algebra $\Lambda^N(\mathbf{X}, K)$ generated by the vector fields $u \cdot F^N$, for $u \in K$, where we use the abbreviation $u \cdot F^N$ for $F_0^N + \sum_{i=1}^m u_i F_i^N$. (Recall that $u_0 = 1$.) Then $\Lambda^N(\mathbf{X}, K)$ acts transitively iff

$$(3.17) \quad \text{Aff}(K) = \mathbb{R}^m.$$

Since we are assuming that (3.17) holds, we can conclude that $\Lambda^N(\mathbf{X}, K)$ is indeed transitive. We then have:

LEMMA 3.1. *For every N ,*

$$\emptyset \neq \mathring{S}^N(\mathbf{X}, K) \subseteq S^N(\mathbf{X}, K) \subseteq \text{Clos } \mathring{S}^N(\mathbf{X}, K).$$

(Here “ $\mathring{}$,” and “Clos” mean interior and closure relative to $G^N(\mathbf{X})$.)

The semigroup $S^N(\mathbf{X}, K)$ is the image of $\mathcal{U}_m(K)$ under the map Ser_N . We need nice inverses of this map, i.e. ways of selecting, for $S \in S^N(\mathbf{X}, K)$, a control $u_S(\cdot) \in \mathcal{U}_m(\mathbf{X}, K)$ which “depends smoothly on S ” and is such that $\text{Ser}_N(u_S(\cdot)) = S$. The construction of such inverses was already done in [25]. However, we shall need a slightly stronger result, which we now state.

As in [25], we let Γ be any finite sequence $(\gamma^1, \dots, \gamma^r)$ of points of \mathbb{R}^m , such that $\text{Aff}(\gamma^1, \dots, \gamma^r) = \mathbb{R}^m$. We let \mathbb{R}_+^k denote the set of k -tuples of nonnegative numbers. If $\mathbf{t} = (t_1, \dots, t_k)$ is in \mathbb{R}_+^k , then we define $\{\Gamma, \mathbf{t}\}$ to be the piecewise constant control which is equal to γ^1 during the first t_1 units of time, then to γ^2 during time t_2 , and so on. (This control is well defined even if $k > r$, because we extend the definition of γ^j to all positive integers j , by making $j \rightarrow \gamma^j$ periodic with period r , i.e., we let $\gamma^{r+1} = \gamma^1$, $\gamma^{r+2} = \gamma^2$, and so on.) Any control of the form $\{\Gamma, \mathbf{t}\}$ for some k and some $\mathbf{t} \in \mathbb{R}_+^k$ will be called a Γ -control. If $K \subseteq \mathbb{R}^m$ and Γ consists of elements of K , then Γ will be said to be a K -sequence.

The map $\nu_{k,\Gamma}^N$, defined by $\nu_{k,\Gamma}^N(\mathbf{t}) = \text{Ser}_N(\{\Gamma, \mathbf{t}\})$, takes \mathbb{R}_+^k to $S^N(\mathbf{X})$. Moreover, if Γ is a K -sequence, then $\nu_{k,\Gamma}^N$ maps \mathbb{R}_+^k to $S^N(\mathbf{X}, K)$. If $\mathbf{t}^0 \in \mathbb{R}_+^k$ is such that the differential $d\nu_{k,\Gamma}^N(\mathbf{t}_0)$ has rank equal to the dimension of $G^N(\mathbf{X})$, then the Γ -control $\{\Gamma, \mathbf{t}_0\}$ is said to be N -normal. Clearly, if Γ is a K -sequence and $\{\Gamma, \mathbf{t}_0\}$ is N -normal, then $\nu_{k,\Gamma}^N(\mathbf{t}_0) \in \mathring{S}^N(\mathbf{X}, K)$. Conversely, suppose that $S \in \mathring{S}^N(\mathbf{X}, K)$. We claim that $S = \nu_{k,\Gamma}^N(\mathbf{t}_0)$ for some K -sequence Γ and some N -normal Γ -control $\{\Gamma, \mathbf{t}_0\}$. To see this, observe first that the system (3.16), with the restriction $u \in K$, necessarily has the accessibility property from S , and the same is therefore true for the “backward system” whose trajectories are those of (3.16) run in reverse. It then follows from standard accessibility theory that, if U is any open subset of $G^N(\mathbf{X})$ containing S , then U contains a nonempty open set V such that, for the reverse system, every $S' \in V$ can be reached from S by means of a piecewise constant control. If we apply this with $U = \mathring{S}^N(\mathbf{X}, K)$, we get an open subset V of $\mathring{S}^N(\mathbf{X}, K)$ such that every $S' \in V$ can be steered to S by means of a trajectory of (3.16) that corresponds to a piecewise constant K -valued control. On the other hand, if $S' \in V$ then S' can be reached from the identity element \mathbb{I} of $G^N(\mathbf{X})$ by means of some K -valued control. This control can be approximated by piecewise constant ones. Since V is open, we conclude that some $S' \in V$ is reachable from \mathbb{I} by means of some piecewise constant control. This control is then necessarily of the form $\{\Gamma, \mathbf{t}^0\}$ for some sequence $\Gamma = (\gamma^1, \dots, \gamma^k)$ and some $\mathbf{t}^0 = (t_1^0, \dots, t_k^0) \in \mathbb{R}_+^k$ such that $t_j^0 > 0$ for all j . Since $\text{Aff}(K) = \mathbb{R}^m$, the sequence Γ can be assumed to be such that

$\text{Aff}(\gamma^1, \dots, \gamma^k) = \mathbb{R}^m$. (This may require that some new γ 's be added at the end of Γ , and then the control $\{\Gamma, t^0\}$ has to be continued by assigning positive times t_j to the new γ^j 's. However, the t^j can be taken to be arbitrarily small, and then the new S' will still be in V , since V is open.) We then get a Γ -control $\{\Gamma, t^0\}$ that steers $\mathbb{1}$ to an $S' \in V$, and is such that Γ is a K -sequence and the affine hull of the elements of Γ is \mathbb{R}^m . The proof of [25, Prop. 3.3] then implies that V contains a point S'' which is of the form $\nu_{l\Gamma(t)}^N$ for some l and some N -normal Γ -control $\{\Gamma, t\}$. (The proof of [25, Prop. 3.3] shows that, if $\Gamma = (\gamma^1, \dots, \gamma^r)$ is such that $\text{Aff}(\gamma^1, \dots, \gamma^r) = \mathbb{R}^m$, then an N -normal Γ -control exists. This was shown by choosing an l and a t such that $d\nu_{l\Gamma(t)}^N$ had the largest possible rank $\bar{\rho}$, and then constructing a submanifold M of $G^N(X)$ such that $\dim M = \bar{\rho}$, with the property that all the vector fields in the Lie algebra generated by the \tilde{F}_j^N are tangent to M , from which it follows that $\bar{\rho} = \dim G^N(X)$. The same proof applies if we now choose l, t to be such that $d\nu_{l\Gamma(t)}^N$ has the largest possible rank $\bar{\rho}$ among all l, t such that $\nu_{l\Gamma(t)}^N \in V$. Such an l, t exists because there is some l, t such that $\nu_{l\Gamma(t)}^N \in V$, namely, $l = k$ and $t = t^0$. The conclusion that $\bar{\rho} = \dim G^N(X)$ follows exactly as in [25].) If we now concatenate this N -normal control $\{\Gamma, t\}$ with a piecewise constant control that steers S'' to S , it follows easily that the resulting control is a $\tilde{\Gamma}$ -control for some $\tilde{\Gamma}$, and is N -normal. So, we have shown:

LEMMA 3.2. *Let $K \subseteq \mathbb{R}^m$, and let $S \in G^N(X)$. Then $S \in \hat{S}^N(X, K)$ if and only if there exist*

- (a) *a K -sequence $\Gamma = (\gamma^1, \dots, \gamma^r)$ such that $\text{Aff}(\gamma^1, \dots, \gamma^r) = \mathbb{R}^m$,*
- (b) *a k and a $t \in \mathbb{R}_+^k$ such that $\{\Gamma, t\}$ is N -normal and $\nu_{k\Gamma(t)}^N = S$.*

The existence of “nice local inverses” to the map Ser_N follows easily.

COROLLARY 3.3. *Let $K \subseteq \mathbb{R}^m$, and let $S \in \hat{S}^N(X, K)$. Then there exist:*

- (a) *a K -sequence $\Gamma = (\gamma^1, \dots, \gamma^r)$ such that $\text{Aff}(\gamma^1, \dots, \gamma^r) = \mathbb{R}^m$,*
- (b) *a positive integer k ,*
- (c) *an open subset W of $G^N(X)$ such that $S \in W$,*
- (d) *a real analytic map $\psi: W \rightarrow \mathbb{R}_+^k$, such that*

$$(3.18) \quad \text{Ser}_N(\{\Gamma, \psi(S')\}) = S' \quad \text{for all } S' \in W.$$

The proof is just a straightforward application of the Implicit Function theorem.

The group $\hat{G}(X)$ is the “Lie group” described at the beginning of this section. Formally, an element S of $\hat{G}(X)$ is an exponential of a Lie series in the indeterminates X_0, \dots, X_m , and therefore $\text{Ev}(\mathbf{f})(S)$ is the exponential of a vector field on M , i.e. a map from M to M . If $S \in \hat{S}(X, K)$, then S can be thought of as a control, and $\text{Ev}_p(\mathbf{f})(S)$ is the point of M to which p is steered by this control. Then $\hat{L}(X)$ is the “Lie algebra” of the Lie group $\hat{G}(X)$. Those elements $Z \in \hat{L}(X)$ such that $\text{Ev}_p(\mathbf{f})(Z) = 0$ constitute the “isotropy subalgebra,” and their exponentials are the “isotropy subgroup.” The reachable set from p is $\text{Ev}_p(\mathbf{f})(\hat{S}(X, K))$. The Lie algebra rank condition says that $\hat{G}(X)$ “acts transitively on M near p .” Hence p will be an interior point of the reachable set if the interior of $\hat{S}(X, K)$ intersects the isotropy subgroup.

The preceding formal considerations are not rigorous, because $\hat{G}(X)$ is not a true Lie group and, as explained above, $\text{Ev}(\mathbf{f})$ is not well defined on $\hat{G}(X)$. In order to obtain a rigorous local controllability theorem one has to consider the nilpotent approximations $G^N(X)$ to $\hat{G}(X)$. The $G^N(X)$ are true Lie groups, with Lie algebra $L^N(X)$, and the subsemigroups $S^N(X, K)$ represent the nilpotent approximations to $\hat{S}(X, K)$. Pursuing the analogy with our earlier discussion, we may think of $\text{LR}^N(\mathbf{f}, p)$ as the “isotropy subalgebra” corresponding to the “action” of $G^N(X)$, and of $H^N(\mathbf{f}, p) = \exp_N(\text{LR}^N(\mathbf{f}, p))$ as the “isotropy group.” If N is large enough (so that $\text{Ev}_p^N(\mathbf{f})(L^N(X))$ is the whole tangent space $T_p M$), then the “action” of $G^N(X)$ on M

is transitive. So we might expect to be able to prove that, if the interior of $S^N(\mathbf{X}, K)$ intersects H^N , then p is in the interior of the reachable set from p . Also, it should follow that, if $\dot{S}^N(\mathbf{X}, K) \cap H^N$ contains points reachable from the identity in arbitrarily small time, then (M, \mathbf{f}, K) is STLC from p . However, this reasoning is not valid, since $\text{Ev}^N(\mathbf{f})$ need not be a true Lie algebra homomorphism, $\text{LR}^N(\mathbf{f}, p)$ need not be a Lie subalgebra of $L^N(\mathbf{X})$, and $G^N(\mathbf{X})$ does not really act on M . If $S \in \dot{S}^N(\mathbf{X}, K) \cap H^N(\mathbf{f}, p)$ and we write $S = \exp_N(Z)$, then $\text{Ev}_p^N(\mathbf{f})(Z) = 0$, and so $\text{Ev}_p(\mathbf{f})(Z) = 0$. Therefore $\exp(Z)$ is equal to the identity map plus a series of differential operators that vanish at p . However, there is no reason for $\exp(Z)$ to be the series of a control $u(\cdot)$. What can be said is that $\exp_N(Z) = \text{Ser}_N(u(\cdot))$ for some $u(\cdot)$. But then $\text{Ser}(u(\cdot))$ will not necessarily be equal to $\exp(Z)$, although it will be equal to $\exp(Z)$ up to terms of degree N . So $u(\cdot)$ will not necessarily steer p to p . However, it will steer p to a point q which is close to p . If U is a neighborhood of Z in $L^N(\mathbf{X})$, and $\exp(U)$ is small enough so that $\exp(U) \subseteq S^N(\mathbf{X}, K)$, then one can choose a $u'(\cdot)$ such that $\text{Ser}_N(u'(\cdot)) = \exp_N(Z')$ for each $Z' \in U$. Then the controls $u'(\cdot)$ will steer p to a neighborhood V of q . If U is large enough, then we may expect V to be such that $p \in V$. To make all this rigorous, we have to be able to choose $u'(\cdot)$ in a continuous fashion as a function of Z' . This requires that we confine ourselves to neighborhoods U such that, if $W = \exp(U)$, then there is a map ψ that satisfies the conditions of Corollary 3.3. So we define a *normal neighborhood* of a point $S \in \dot{S}^N(\mathbf{X}, K)$ to be an open subset W of $G^N(\mathbf{X})$ such that there exist Γ, k, ψ for which the conditions of Corollary 3.3 hold. Then Corollary 3.3 simply says that every point of $\dot{S}^N(\mathbf{X}, K)$ has a normal neighborhood. The sufficient condition for STLC from p will then say that, if $\dot{S}^N(\mathbf{X}, K) \cap H^N(\mathbf{f}, p)$ contains points S_i reachable from the identity in arbitrarily small time t_i , then (M, \mathbf{f}, K) is STLC from p , provided that N is sufficiently large, and that these points have normal neighborhoods whose size does not decrease too fast as $t \rightarrow 0$. It will be clear from the proof that it is not necessary to have a lower bound for the size of the neighborhood in all directions, but only in directions transversal to $H^N(\mathbf{f}, p)$. To make this precise, let $\mathcal{E} = (E_1, \dots, E_k)$ be a finite sequence of elements of $L^N(\mathbf{X})$, and let $Z \in L^N(\mathbf{X})$. We define, for $r > 0$

$$(3.19) \quad B_{\mathcal{E}}(Z, r) = \left\{ Z + \sum_{i=1}^k x_i E_i : \sum_{i=1}^k x_i^2 \leq r^2 \right\}.$$

($B_{\mathcal{E}}(Z, r)$ is the \mathcal{E} -ball of radius r and center Z . We will only use this definition for sequences \mathcal{E} such that E_1, \dots, E_k are linearly independent.)

Also, we define a function $T^N : A^N(\mathbf{X}) \rightarrow \mathbb{R}$ by letting $T^N(S)$ be the coefficient of X_0 in S . (In particular, if $S = \text{Ser}_N(u(\cdot))$ for some control $u(\cdot)$, then $T^N(S)$ is the terminal time of $u(\cdot)$.) We then have:

THEOREM 3.4. *Let (M, \mathbf{f}, K) be a control system, and let $p \in M$. Assume that K is a bounded set. Let N be a positive integer, and let $\mathcal{E} = (E_1, \dots, E_n)$ be a sequence of elements of $L^N(\mathbf{X})$ such that $(\text{Ev}_p(\mathbf{f})(E_1), \dots, \text{Ev}_p(\mathbf{f})(E_n))$ is a basis of the tangent space $T_p M$. Assume that there is a sequence of points $S_j, j = 1, 2, \dots$, such that:*

- (i) $S_j \in \dot{S}^N(\mathbf{X}, K) \cap H^N(\mathbf{f}, p)$ for all j ,
- (ii) $T^N(S_j) \rightarrow 0$ as $j \rightarrow \infty$,
- (iii) *If $S_j = \exp_N(Z_j)$, then there are normal neighborhoods W_j of S_j , and a constant $\alpha > 0$, such that*

$$(3.20) \quad \log_N W_j \supseteq B_{\mathcal{E}}(Z_j, \alpha [T^N(S_j)]^N)$$

for all j . Then (M, \mathbf{f}, K) is STLC from p .

Proof. Let $\rho_j = T^N(S_j)$. For each j , choose a K -sequence Γ_j , a positive integer k_j , and a real-analytic map $\psi_j: W_j \rightarrow \mathbb{R}_+^{k_j}$ such that

$$(3.21) \quad \text{Ser}_N(\{\Gamma_j, \psi_j(S)\}) = S$$

whenever $S \in W_j$.

Choose coordinates on a neighborhood \mathcal{N} of p such that p becomes $(0, \dots, 0)$ and the vectors $\text{Ev}_p(\mathbf{f})(E_i)$ are the members e_i of the canonical basis of \mathbb{R}^n . For each control $u(\cdot)$, let $\pi_p(u(\cdot))$ be the point to which $u(\cdot)$ steers p (i.e. $\pi_p(u(\cdot)) = x(T)$, if $u(\cdot)$ is defined on $[0, T]$, and $x(\cdot)$ is the trajectory for $u(\cdot)$ such that $x(0) = p$). Proposition 4.1 of [25] implies that the series $\text{Ev}_p(\mathbf{f})(\text{Ser } u(\cdot))$ gives an asymptotic expansion for $\pi_p(u(\cdot))$ in the following sense: if ϕ is an arbitrary C^∞ function on \mathcal{N} , then there are constants β_ν and times τ_ν such that

$$(3.22) \quad \|\phi(\pi_p(u(\cdot)) - \text{Ev}_p(\mathbf{f})(\text{Ser}_\nu(u(\cdot))))\| < \beta_\nu T(u(\cdot))^{\nu+1}$$

for all ν and all controls $u(\cdot)$ such that $T(u(\cdot)) \leq \tau_\nu$. (Here $\text{Ev}_p(\mathbf{f})(\text{Ser}_\nu(u(\cdot)))$ is a finite sum of partial differential operators evaluated at p , and so $\text{Ev}_p(\mathbf{f})(\text{Ser}_\nu(u(\cdot)))\phi$ is a finite sum of numbers, namely, the results of applying those partial differential operators to ϕ . The result from [25] gives constants β_ν that also depend on a bound A for the controls, but here we are assuming that K is bounded, so that β_ν only depends on ν .)

Inequality (3.22) clearly holds for vector functions as well, so we can apply it to the identity map $\phi: \mathcal{N} \rightarrow \mathcal{N}$. From now on, ϕ denotes this map. Therefore $\phi(\pi_p(u(\cdot))) = \pi_p(u(\cdot))$, and so (3.22) becomes

$$(3.23) \quad \|\pi_p(u(\cdot)) - \text{Ev}_p(\mathbf{f})(\text{Ser}_\nu(u(\cdot)))\| \leq \beta_\nu T(u(\cdot))^{\nu+1}.$$

Now define maps μ_j from the closed unit ball \mathbb{B} of \mathbb{R}^n into M , by

$$(3.24) \quad \mu_j(x_1, \dots, x_n) = \pi_p\left(\left\{\Gamma_j, \psi_j\left(\exp_N\left(Z_j + \alpha\rho_j^N \sum_{i=1}^n x_i E_i\right)\right)\right\}\right).$$

The definition is possible because $Z_j + \alpha\rho_j^N \sum_{i=1}^n x_i E_i$ is in $B_g(Z_j, \alpha\rho_j^N)$, and so its exponential in $A^n(X)$ is in W_j . By construction, $\mu_j(x_1, \dots, x_n)$ is reachable from p , by means of the control

$$(3.25) \quad u_{j,x_1,\dots,x_n}(\cdot) = \left\{\Gamma_j, \psi_j\left(\exp_N\left(Z_j + \alpha\rho_j^N \sum_{i=1}^n x_i E_i\right)\right)\right\}.$$

The truncated series $\text{Ser}_N(u_{j,x_1,\dots,x_n}(\cdot))$ is then

$$\exp_N\left(Z_j + \alpha\rho_j^N \sum_{i=1}^n x_i E_i\right),$$

and so the terminal time $T(u_{j,x_1,\dots,x_n}(\cdot))$ is equal to $T^N(\text{Ser}_N(u_{j,x_1,\dots,x_n}(\cdot)))$, i.e. to

$$\rho_j + \alpha\rho_j^N \sum_{i=1}^n x_i \theta_i,$$

where θ_i is the coefficient of X_0 in E_i . In particular, all the points $\mu_j(x_1, \dots, x_n)$, for $(x_1, \dots, x_n) \in \mathbb{B}$, and fixed j , are reachable from p in time not greater than $\hat{\alpha}\rho_j$, where $\hat{\alpha}$ is some fixed constant which does not depend on j . Since $\rho_j \rightarrow 0$ as $j \rightarrow \infty$, our theorem will be proved if we show that $\mu_j(\mathbb{B})$ contains a neighborhood of p for sufficiently large j .

In view of (3.23), we have

$$(3.26) \quad \left\| \mu_j(x_1, \dots, x_n) - \text{Ev}_p(\mathbf{f}) \left(\exp_N \left(Z_j + \alpha \rho_j^N \sum_{i=1}^n x_i E_i \right) \right) \phi \right\| \leq \beta_N \hat{\alpha}^{N+1} \rho_j^{N+1},$$

provided that j is large enough, so that $\hat{\alpha} \rho_j \leq \tau_N$.

For $Q \in L^N(\mathbf{X})$, define

$$(3.27) \quad \widehat{\text{exp}}_N(Q) = \sum_{k=0}^N \frac{1}{k!} Q^k.$$

(Here we identify $L^N(\mathbf{X})$ with the subspace $\sum_{k=1}^N L^{k, \text{hom}}(\mathbf{X})$, but the powers Q^k are computed in $L(\mathbf{X})$, so that $\widehat{\text{exp}}_N(Q)$ is allowed to contain terms of degree greater than N .) We claim that all the coefficients of the finite series

$$\widehat{\text{exp}}_N \left(Z_j + \alpha \pi_j^N \sum_{i=1}^n x_i E_i \right) - \exp_N \left(Z_j + \alpha \rho_j^N \sum_{i=1}^n x_i E_i \right)$$

are bounded by a fixed constant times ρ_j^{N+1} . To see this, observe first that, if we write

$$(3.28) \quad Z_j = \sum_I z_I^j X_I,$$

then $z_I^j = 0$ for $|I| > N$, and there is a constant c such that $|z_I^j| \leq C \rho_j^{|I|}$ for all j, I . (Here $|I|$ is the length of the multiindex I , i.e. the degree of the monomial X_I . The first assertion follows because $Z_j \in L^N(\mathbf{X})$. The second one holds because $\exp_N(Z_j) = \text{Ser}_N(u_j(\cdot))$ for some K -valued control $u_j(\cdot)$ with terminal time ρ_j . Since the coefficients σ_I of $\text{Ser}(u_j(\cdot))$ are iterated integrals, as shown in (3.12) and (3.13), they satisfy bounds $|\sigma_I^j| \leq \text{constant} \times \rho_j^{|I|}$. Similar bounds then hold for the coefficients of the series $\log(\text{Ser}(u_j(\cdot)))$, and for those of its truncation $Z_j = \hat{\tau}^N(\log(\text{Ser}(u_j(\cdot))))$. The coefficients of $\alpha \rho_j^N \sum_{i=1}^n x_i E_i$ also satisfy a similar bound, since they all contain a factor ρ_j^N , and those of degree $> N$ vanish. So the coefficients of

$$\exp_N \left(Z_j + \alpha \rho_j^N \sum_{i=1}^n x_i E_i \right) \quad \text{and} \quad \widehat{\text{exp}}_N \left(Z_j + \alpha \rho_j^N \sum_{i=1}^n x_i E_i \right)$$

also satisfy these bounds, and then the same is true for those of the difference of these two series. However, the coefficients of this difference vanish whenever $|I| \leq N$. Hence they are bounded by a constant times ρ_j^{N+1} .

It then follows that (3.26) remains valid (possibly with a different constant in the right side) if “ \exp_N ” is replaced by “ $\widehat{\text{exp}}_N$.” Now $\widehat{\text{exp}}_N(Z_j + \alpha \rho_j^N \sum_{i=1}^n x_i E_i)$ can be written out by applying (3.27) and then expanding the powers of $Z_j + \alpha \rho_j^N \sum_{i=1}^n x_i E_i$. This leads to a finite sum of terms of the following five kinds: (i) the term $\alpha \rho_j^N \sum_{i=1}^n x_i E_i$, (ii) powers of Z_j , (iii) products of at least one Z_j factor and at least one $\alpha \rho_j^N \sum_{i=1}^n x_i E_i$, (iv) powers of $\alpha \rho_j^N \sum_{i=1}^n x_i E_i$ other than the first power, (v) the identity.

When evaluated at p , all the terms $\text{Ev}(\mathbf{f})(Z_j^k)$ vanish, because $Z_j \in \text{LR}^N(\mathbf{f}, p)$. The terms of type (iii) are $O(\rho_j^{N+1})$, because Z_j is $O(\rho_j)$. The terms of type (iv) are also $O(\rho_j^{N+1})$. Hence, modulo $O(\rho_j^{N+1})$, only the terms of types (i) and (v) count. So we get the bound

$$(3.29) \quad \left\| \mu_j(x_1, \dots, x_n) - \text{Ev}_p(\mathbf{f}) \left(\mathbb{1} + \alpha \rho_j^N \sum_{i=1}^n x_i E_i \right) \phi \right\| \leq \gamma \rho_j^{N+1}$$

for some constant γ . Then (3.29) implies (using the facts that $\text{Ev}_p(\mathbf{f})(1)\phi = \phi(p) = p = 0$, and $\text{Ev}_p(\mathbf{f})(E_i)\phi = e_i\phi = e_i$)

$$(3.30) \quad \left\| \mu_j(x_1, \dots, x_n) - \alpha \rho_j^N \sum_{i=1}^n x_i e_i \right\| \leq \gamma \rho_j^{N+1}.$$

Let

$$(3.31) \quad \nu_j(x_1, \dots, x_n) = \frac{1}{\alpha \rho_j^N} \mu_j(x_1, \dots, x_n).$$

Then $\mu_j(\mathbb{B})$ contains a neighborhood of 0 if $\nu_j(\mathbb{B})$ does. It follows from (3.30) that

$$(3.32) \quad \|\nu_j(x_1, \dots, x_n) - (x_1, \dots, x_n)\| \leq \text{constant} \times \rho_j.$$

Therefore the ν_j are continuous maps from \mathbb{B} to \mathbb{R}^n that converge uniformly to the identity map of \mathbb{B} as $j \rightarrow \infty$. This implies that $\nu_j(\mathbb{B})$ contains a neighborhood of 0 for large enough j . The proof is then complete.

Theorem 3.4 gives a sufficient condition for local controllability, but not one that is easy to check in practice. The next two sections will be devoted to providing more easily checkable conditions. Here we will just give a simple example that follows directly from Theorem 3.4.

THEOREM 3.5. *Let (M, \mathbf{f}, K) be a control system, and let $p \in M$. Let N be a positive integer such that $\text{Ev}_p(\mathbf{f})(L^N(\mathbf{X})) = T_p M$. Assume that $\hat{S}^N(\mathbf{X}, K)$ contains an element $S = \exp_N(Z)$ such that all the homogeneous components of Z are in $\text{LR}(\mathbf{f}, p)$. Then (M, \mathbf{f}, K) is STLC from p .*

Proof. For each $p > 0$, let $\Delta(\rho)$ be the automorphism of $A(\mathbf{X})$ which sends X_i to ρX_i for $i = 0, \dots, M$. Then $\Delta(\rho)$ gives rise to an automorphism $\hat{\Delta}(\rho)$ of $\hat{A}(\mathbf{X})$, defined by sending a series $S = \sum_{j=0}^{\infty} S_j$, with $S_j \in A^{j, \text{hom}}(\mathbf{X})$ to the series

$$(3.33) \quad \hat{\Delta}(\rho)(S) = \sum_{j=0}^{\infty} \rho^j S_j.$$

Clearly, $\hat{\Delta}(\rho)$ induces an automorphism of $L(\mathbf{X})$, $\hat{L}(\mathbf{X})$, and $\hat{G}(\mathbf{X})$. Moreover, since $\hat{\Delta}(\rho)$ maps $\hat{A}_N(\mathbf{X})$ to $\hat{A}_N(\mathbf{X})$, it induces an automorphism $\Delta^N(\rho)$ of $A^N(\mathbf{X})$ and, in particular, an automorphism of $G^N(\mathbf{X})$ and one of $L^N(\mathbf{X})$.

Moreover, if $0 < \rho \leq 1$, then $\hat{\Delta}(\rho)$ maps $S(\mathbf{X}, K)$ into $S(\mathbf{X}, K)$, and $\Delta^N(\rho)$ maps $S^N(\mathbf{X}, K)$ into $S^N(\mathbf{X}, K)$. (Indeed, if $t \rightarrow S(t)$, $0 \leq t \leq T$, is a solution of (3.11) corresponding to a K -valued control $t \rightarrow u(t) = (u_1(t), \dots, u_m(t))$, then $\tau \rightarrow \hat{\Delta}(\rho)(S(\tau/\rho))$ is a solution of (3.11) on the interval $[0, \rho T]$, corresponding to the control $\tau \rightarrow u(\tau/\rho)$.)

Now suppose that S is an element of $\hat{S}^N(\mathbf{X}, K)$ such that $S = \exp_N(Z)$, where

$$(3.34) \quad Z = \sum_{j=1}^N Z^j, \quad Z^j \in L^{j, \text{hom}}(\mathbf{X}),$$

and $Z \in \text{LR}(\mathbf{f}, p)$. Pick $\mathcal{E} = (E_1, \dots, E_n)$ such that

- (a) the E_i are members of $L^N(\mathbf{X})$,
- (b) each E_i is homogeneous of degree θ_i (with $\theta_i \leq N$),
- (c) $\text{Ev}_p(\mathbf{f})(E_1), \dots, \text{Ev}_p(\mathbf{f})(E_n)$ form a basis of $T_p M$.

Since $S \in \hat{S}^N(\mathbf{X}, K)$, there exists a normal neighborhood W of S . If we let $S_\rho = \Delta^N(\rho)(S)$, $W_\rho = \Delta^N(\rho)(W)$, then W_ρ is a normal neighborhood of S_ρ . Let $\tilde{\alpha} > 0$ be such that $\exp_N(Z + \sum_{i=1}^n y_i E_i) \in W_\rho$ whenever $|y_i| \leq \tilde{\alpha}$. Let $Z_\rho = \Delta(\rho)Z$. Then $\exp_N(Z_\rho + \sum_{i=1}^n \rho^{\theta_i} y_i E_i) \in W_\rho$ whenever $|y_i| \leq \tilde{\alpha}$. Since $\theta_i \leq N$, it follows that

$$(3.35) \quad B_{\mathcal{E}}(Z_\rho, \tilde{\alpha} \rho^N) \subseteq \log_N(W_\rho)$$

whenever $0 < \rho \leq 1$.

Since $Z^j \in L^{j,\text{hom}}(\mathbf{X}) \cap \text{LR}(\mathbf{f}, p)$, it follows that $Z_\rho \in \text{LR}(\mathbf{f}, p)$ for all ρ . Finally, it is clear that $T^N(S_\rho) = \rho c$, where $c = T^N(S)$. Therefore

$$(3.36) \quad B_{\mathcal{G}}(Z_\rho, \alpha[T^N(S)]^N) \subseteq \log_N(W_\rho)$$

for $0 \leq \rho \leq 1$, if $\alpha = \tilde{\alpha}c^{-N}$. So the conditions of Theorems 3.4 hold, and our desired conclusion follows.

The preceding result is too weak for applications. In the following section we will strengthen it in two ways. First, the requirement that each homogeneous component Z^j of Z be a Lie relation at p will be replaced by the weaker condition that Z^j be equal to a Lie relation plus an element of lower degree. Second, the “degree” will be allowed to be a more general one, arising from a one-parameter group of dilations which is not necessarily the family $\{\Delta(\rho): \rho > 0\}$ considered in the proof of Theorem 3.5.

4. Dilations. We now define the concept of a “group of dilations,” and prove a generalization of Theorem 3.5.

If V is a linear space over the reals, a *group of dilations* of V is a mapping $\rho \rightarrow \Delta(\rho)$ that assigns to every real $\rho > 0$ a linear endomorphism $\Delta(\rho): V \rightarrow V$, in such a way that

(DIL1) $\Delta(1) = \text{identity}$,

(DIL2) $\Delta(\rho_1)\Delta(\rho_2) = \Delta(\rho_1\rho_2)$ for all ρ_1, ρ_2 ,

(DIL3) V has a direct sum decomposition

$$(4.1) \quad V = \bigoplus_j V_j$$

such that the subspaces V_j are invariant under the $\Delta(\rho)$, and the action of $\Delta(\rho)$ on each V_j is given by multiplication by ρ_j^α for some $\alpha_j \geq 0$.

The decomposition (4.1) is clearly unique if, in addition, we require that $\alpha_j \neq \alpha_k$ whenever $j \neq k$. In this case, the V_j are referred to as *the homogeneous components of V with respect to Δ* . If $v \in V$ is such that $v \in V_j$ for some j , then v is said to be Δ -homogeneous. If $v \neq 0$, then V_j is uniquely determined by v , and the corresponding α_j is the Δ -degree of v . More generally, any $v \in V$ can be expressed in a unique way as a sum $\sum_j v_j$, $v_j \in V_j$. The Δ -degree of v is the largest α_j such that $v_j \neq 0$, and is denoted by $\deg_\Delta(v)$.

If Δ is a group of dilations of V , then Δ gives rise to groups of dilations Δ^A of $A(V)$, the free associative \mathbb{R} -algebra generated by V (i.e. the tensor algebra over V) and Δ^L of $L(V)$, the free Lie algebra generated by V . In both cases, the new group of dilations consists of automorphism of the algebraic structure, which in addition leave invariant the usual homogeneous components of $A(V)$, $L(V)$ (i.e. the homogeneous components with respect to the groups of dilations induced by $\{\Delta_0(\rho): \rho > 0\}$, where $\Delta_0(\rho): V \rightarrow V$ is multiplication by ρ).

A group of dilations Δ of V will be called *strict* if it has no component of degree zero. If Δ is strict, then Δ^L is also strict, and Δ^A is strict on $A_0(V)$, the set of elements of $A(V)$ with no constant term. (But $\Delta^A(\rho)(1) = 1$ so Δ^A is not strict on $A(V)$.)

We will use $\nu(\Delta)$ to denote the infimum of the degrees of the homogeneous components of Δ . Then $\nu(\Delta) = \nu(\Delta^L)$ for every Δ . If V is finite-dimensional, then the infimum considered above is actually a minimum, and Δ is strict if and only if $\nu(\Delta) > 0$.

In the particular case when $V = L^{1,\text{hom}}(\mathbf{X})$ (i.e. the linear span of X_0, \dots, X_m), we let $\Delta_{1,m}(\rho): V \rightarrow V$ be multiplication by ρ as above. Let Δ be any group of dilations of V . Then Δ gives rise to groups of dilations Δ^A, Δ^L of $A(\mathbf{X})$ and $L(\mathbf{X})$. Any group of dilations of $A(\mathbf{X})$ or of $L(\mathbf{X})$ which arises in this fashion from a strict group of dilations of V will be called an *admissible group of dilations*. (Clearly, a group of dilations $\Delta^\#$ of $L(\mathbf{X})$, is admissible iff $\Delta^\#$ is strict and the $\Delta(\rho)$ are automorphisms which leave the

usual homogeneous components invariant. If $\Delta^\#$ is a group of dilations of $A(X)$, then $\Delta^\#$ is admissible if and only if $\Delta^\#$ consists of automorphisms which leave the usual homogeneous components invariant, and the only elements of $\Delta^\#$ -degree zero are the constants.)

If Δ is a strict group of dilations of $L^{1,\text{hom}}(X)$ —so that Δ gives rise to admissible groups of dilations Δ^A, Δ^L —we will also refer to Δ itself as an admissible group of dilations.

In particular, the groups that arise from $\Delta_{1,m}$, denoted by $\Delta_{1,m}^A, \Delta_{1,m}^L$, are clearly admissible.

If Δ^A is any admissible group of dilations of $A(X)$ as above, arising from a group of dilations Δ of $L^{1,\text{hom}}(X)$, then every $\Delta^A(\rho)$ gives rise in an obvious way to an automorphism $\hat{\Delta}^A(\rho)$ of $\hat{A}(X)$. The $\hat{\Delta}^A(\rho)$ map $\hat{L}(X)$ to $\hat{L}(X)$ and therefore $\hat{G}(X)$ to $\hat{G}(X)$. Since $\Delta^A(\rho)$ maps $A_N(X)$ to $A_N(X)$ for each N , there are induced automorphisms $\Delta^{A,N}(\rho)$ of the algebra $A^N(X)$, which gives rise to automorphisms of the Lie algebra $L^N(X)$ and of the Lie group $G^N(X)$.

We will say that Δ is *compatible with the semigroup $\hat{S}(X, K)$* if

$$(4.2) \quad \hat{\Delta}^A(\rho)\hat{S}(X, K) \subseteq \hat{S}(X, K) \quad \text{for every } \rho \leq 1.$$

Compatibility can be described more directly as follows. The map $\Delta(\rho)$ takes $L^{1,\text{hom}}(X)$ into itself. For $u \in \mathbb{R}^m$, $u = (u_1, \dots, u_m)$, let $X(u) = X_0 + \sum_{i=1}^m u_i X_i$. If $u \in K$, then $\exp(X(u)) \in \hat{S}(X, K)$. Therefore $\hat{\Delta}^A(\rho)(\exp(X(u)))$ must belong to $\hat{S}(X, K)$, if Δ is compatible with $\hat{S}(X, K)$ and $0 < \rho \leq 1$. That is, $\exp(\hat{\Delta}^A(\rho)(X(u))) \in \hat{S}(X, K)$ and so $\exp(\hat{\Delta}^A(\rho)(X(u))) = \text{Ser}(v(\cdot))$ for some $v(\cdot) \in \mathcal{U}_m(K)$. Let T be the terminal time of $v(\cdot)$. Then T cannot equal zero for, if $T = 0$, then we would have $\text{Ser}(v(\cdot)) = 1$, and so $\hat{\Delta}^A(\rho)(X(u)) = 0$, contradicting the fact that $X(u) \neq 0$ and $\hat{\Delta}^A(\rho)$ is an automorphism. It follows from the construction of $\text{Ser}(v(\cdot))$ that the coefficient of X_0 in $\text{Ser}(v(\cdot))$ is precisely T . Moreover, the coefficient of X_0 in $\exp Z$ is the same as the coefficient of X_0 in Z . So

$$(4.3) \quad \hat{\Delta}^A(\rho)(X(u)) = TX_0 + \sum_{i=1}^m \alpha_i X_i,$$

for some choice of $(\alpha_1, \dots, \alpha_m) \in \mathbb{R}^m$. If we let $\beta = \alpha/T$, we see that

$$(4.4) \quad \hat{\Delta}^A(\rho)(X(u)) = TX(\beta).$$

Let $v': [0, T] \rightarrow \mathbb{R}^m$ be such that $v'(t) = \beta$ for $0 \leq t \leq T$. Then $\text{Ser}(v'(\cdot)) = \exp(TX(\beta))$. Since Ser is injective as a map from \mathcal{U}_m into $\hat{G}(X)$, we conclude that $v(\cdot) = v'(\cdot)$. Since $v(\cdot)$ is K -valued, we conclude that $\beta \in K$. Hence $\hat{\Delta}^A(\rho)(X(u))$ is of the form $TX(\beta)$ for some $T > 0$ and some $\beta \in K$. Conversely, if $\{\Delta(\rho)\}$ has the property that $\Delta(\rho)(X(u))$ is of the form $TX(\beta)$ for some $T > 0, \beta \in K$, whenever $0 < \rho \leq 1$ and $u \in K$, then it is easy to see that $\hat{\Delta}^A(\rho)\hat{S}(X, K) \subseteq \hat{S}(X, K)$ whenever $0 < \rho \leq 1$.

To see this, write

$$(4.5) \quad \Delta(\rho)(X(u)) = \sum_{i=0}^m \theta_i(u) X_i$$

for $u \in \mathbb{R}^m$. Then

$$\Delta(\rho)(X_0) = \sum_{i=0}^m \theta_i(0) X_i \quad \text{and} \quad \Delta(\rho)\left(\sum_{j=1}^m u_j X_j\right) = \sum_{i=0}^m [\theta_i(u) - \theta_i(0)] X_i,$$

so that each of the functions $u \rightarrow \theta_i(u) - \theta_i(0)$ is linear. Moreover, $\theta_0(u) > 0$ for every $u \in K$ and, if we define $\beta_i(u) = \theta_i(u)/\theta_0(u)$ for $i = 1, \dots, m$, $u \in \mathbb{R}^m$, $\theta_0(u) \neq 0$, then we have $\beta(u) \in K$ whenever $u \in K$, if $\beta(u) = (\beta_1(u), \dots, \beta_m(u))$. Let $S \in \hat{S}(X, K)$. Then $S = S(T)$, for some $\hat{A}(X)$ -valued function $t \rightarrow S(t)$, $0 \leq t \leq T$, that satisfies

$$(4.6) \quad \dot{S}(t) = S(t) \left(X_0 + \sum_{i=1}^m u_i(t) X_i \right),$$

where the u_i are Lebesgue integrable functions such that $(u_1(t), \dots, u_m(t)) \in K$ for all t . Let $S^\#(t) = \hat{\Delta}^A(\rho) S(t)$. Then

$$(4.7) \quad \begin{aligned} \dot{S}^\#(t) &= S^\#(t) \left(\theta_0(u(t)) X_0 + \sum_{i=1}^m \theta_i(u(t)) X_i \right) \\ &= S^\#(t) \left(\theta_0(u(t)) \left[X_0 + \sum_{i=1}^m \beta_i(u(t)) X_i \right] \right). \end{aligned}$$

Let $\tau(t) = \int_0^t \theta_0(u(s)) ds$. (The integral exists because $u(\cdot)$ is Lebesgue integrable and θ_0 is affine linear.) Since $\theta_0(u(t)) > 0$ for $0 \leq t \leq T$, τ is a strictly increasing function of t . Let $S^*(\tau) = S^\#(t)$, if $\tau = \tau(t)$. Then

$$(4.8) \quad \dot{S}^*(\tau) = S^*(\tau) \left(X_0 + \sum_{i=1}^m v_i(\tau) X_i \right),$$

where the dot now denotes differentiation with respect to τ , and $v_i(\tau) = \beta_i(u(t))$ whenever $\tau = \tau(t)$. The vector-valued function $v(\cdot)$ is Lebesgue integrable. (Notice that $\beta(u(\cdot))$ might fail to be integrable, since we do not know that $\theta_0(\cdot)$ is bounded away from zero. However, $v(\cdot)$ is necessarily integrable because, whenever ϕ is a strictly positive integrable function on $[0, T]$, and $\tau(t) = \int_0^t \phi(s) ds$ for $0 \leq t \leq T$, then the function $g: [0, \tau(T)] \rightarrow \mathbb{R}$ defined by $g(\tau(t)) = f(t)/\phi(t)$ is integrable whenever f is integrable.) Since $(v_1(\tau), \dots, v_m(\tau)) \in K$ for every τ , we see that $S^*(\tau) \in \hat{S}(X, K)$. In particular, since $S^*(\tau(T)) = S^\#(T) = \hat{\Delta}^A(\rho) S$, we see that $\hat{\Delta}^A(\rho) S \in \hat{S}(X, K)$. Therefore $\hat{\Delta}^A(\rho)$ maps $\hat{S}(X, K)$ to $\hat{K}(X, K)$. So we have shown:

LEMMA 4.1. *Let $\Delta = \{\Delta(\rho): 0 < \rho < \infty\}$ be a one-parameter group of dilations of $V = L^{1, \text{hom}}(X)$, and let $\Delta^A, \hat{\Delta}^A$ be the corresponding groups of automorphisms of $A(X), \hat{A}(X)$. Then Δ is compatible with the semigroup $\hat{S}(X, K)$ if and only if $\Delta(\rho) (X_0 + \sum_{i=1}^m u_i X_i)$ is of the form $T(X_0 + \sum_{i=1}^m v_i X_i)$ for some $T > 0$, $(v_1, \dots, v_m) \in K$, whenever $0 < \rho \leq 1$ and $(u_1, \dots, u_m) \in K$.*

If $\hat{\Delta}^A$ is compatible with $\hat{S}(X, K)$ then $\hat{\Delta}^A(\rho)$ gives rise to a map $\Delta_K^{\hat{S}}(\rho): \hat{S}(X, K) \rightarrow \hat{S}(X, K)$ whenever $0 < \rho \leq 1$, and hence to a map $\Delta_K^{\mathcal{U}}(\rho): \mathcal{U}_m(K) \rightarrow \mathcal{U}_m(K)$, since $\mathcal{U}_m(K)$ is identified with $\hat{S}(X, K)$ by means of the bijection Ser. An explicit description of this map follows from the reasoning preceding the statement of Lemma 4.1. If we write

$$(4.9) \quad \Delta(\rho) \left(X_0 + \sum u_i X_i \right) = \sum_{i=0}^m \theta_i^\rho(u) X_i,$$

for $u \in \mathbb{R}^m$, then the control $v(\cdot) = \Delta_K^{\mathcal{U}}(\rho)(u(\cdot))$ that corresponds to a given $u(\cdot) \in \mathcal{U}_m(K)$, defined on an interval $[0, T]$, is obtained from the K -valued map $t \rightarrow \beta^\rho(u(t))$, $0 \leq t \leq T$, by reparametrizing time, using $\tau = \tau(t) = \int_0^t \theta_0^\rho(u(s)) ds$ as the new time parameter. (Here $\beta^\rho(u) = (\beta_1^\rho(u), \dots, \beta_m^\rho(u))$, $\beta_i^\rho(u) = \theta_i^\rho(u)/\theta_0^\rho(u)$.) This explicit description implies, in particular, that $\Delta_K^{\mathcal{U}}(\rho)$ is continuous with respect to some natural topologies on $\mathcal{U}(T)$ (for example, L^1 , pointwise convergence), and that $\Delta_K^{\mathcal{U}}(\rho)$ maps piecewise constant controls to piecewise constant controls. More precisely, if $u(\cdot)$ is a piecewise constant control whose values are u^1, \dots, u^k , on intervals of length

t^1, \dots, t^k , then $\Delta_K^u(\rho)(u(\cdot))$ is piecewise constant with values v^1, \dots, v^k on intervals of length τ^1, \dots, τ^k , where $v^i = \beta^\rho(u^i)$, and $\tau^i = \theta_0^\rho(u^i)t^i$.

This implies, in particular:

LEMMA 4.2. *If Δ is an admissible group of dilations of $L^{1,\text{hom}}(X)$, which is compatible with K , and W is a normal neighborhood of an $S \in \hat{S}^N(X, K)$, then $\Delta^{A,N}(\rho)(W)$ is a normal neighborhood $\Delta^{A,N}(\rho)(S)$ for every $\rho \in (0, 1]$.*

Now suppose that an $(m+1)$ -tuple $\mathbf{f} = (f_0, \dots, f_m)$ of smooth vector fields on a manifold M is given, as well as a point $p \in M$. We can then define $N_0(\mathbf{f}, p)$ to be the smallest integer N such that

$$(4.10) \quad \text{Ev}_p(\mathbf{f})(L^N(\mathbf{X})) = T_p M.$$

If, in addition, an admissible group of dilations Δ on $L^{1,\text{hom}}(\mathbf{X})$ is given, we can also define $\nu_0(\mathbf{f}, p, \Delta)$ to be the largest of the Δ -degrees of all the elements of $L^{N_0(\mathbf{f}, p)}(\mathbf{X})$.

An element Z of $L(\mathbf{X})$ is said to be Δ -neutralized for \mathbf{f} at p if each Δ -homogeneous component Z_j of Z is the sum of an $R_j \in L(\mathbf{X})$ which belongs to $\text{LR}(\mathbf{f}, p)$ and a $Q_j \in L(\mathbf{X})$ such that

$$(4.11) \quad \deg_\Delta(Q_j) < \deg_\Delta(Z_j).$$

Our generalization of Theorem 3.5 is then the following

THEOREM 4.3. *Let (M, \mathbf{f}, K) be a control system, and let $p \in M$. Let Δ be an admissible group of dilations of $L^{1,\text{hom}}(\mathbf{X})$ which is compatible with $\hat{S}(\mathbf{X}, K)$. Let N be a positive integer that satisfies*

$$(4.12) \quad N \geq N_0(\mathbf{f}, p),$$

and

$$(4.13) \quad N\nu(\Delta) \geq \nu_0(\mathbf{f}, p, \Delta).$$

Assume that there exists an element Z of $L^N(\mathbf{X})$ which is Δ -neutralized for \mathbf{f} at p and satisfies $\exp_N(Z) \in \hat{S}^N(X, K)$. Then (M, \mathbf{f}, K) is STLC from p .

Remark. Theorem 3.5 is a particular case of this result. Indeed, to get Theorem 3.5 it suffices to let Δ be the group of dilations defined by $\Delta(\rho)(P) = \rho P$ for $P \in L^{1,\text{hom}}(\mathbf{X})$. A Z that satisfies the condition of Theorem 3.5 is clearly Δ -neutralized for \mathbf{f} at p .

Proof. Let $S \in \hat{S}^N(X, K)$ be such that $S = \exp_N(Z)$, $Z \in L^N(\mathbf{X})$, and Z is Δ -neutralized for \mathbf{f} at p . Let $\mathcal{E} = (E_1, \dots, E_n)$ consist of elements of $L^{N_0(\mathbf{f}, p)}(\mathbf{X})$ which are Δ -homogeneous of degrees $\sigma_1, \dots, \sigma_n$ and are such that the vectors $\text{Ev}_p(\mathbf{f})(E_i)$, $i = 1, \dots, n$ span $T_p M$. (Then, in particular, $\sigma_j \leq \nu_0(\mathbf{f}, p, \Delta)$ for $j = 1, \dots, n$.) Let W be a normal neighborhood of S . Then we can pick a neighborhood W_0 of S and a $\beta > 0$ such that $\exp_N(Z' + \sum_{j=1}^n y_j E_j) \in W$ whenever $\exp_N(Z') \in W_0$ and $|y_j| \leq \beta$ for $j = 1, \dots, n$.

Since Z is Δ -neutralized for \mathbf{f} at p , we can write

$$(4.14) \quad Z = \sum_i Z_i,$$

where the Z_i are elements of $L(\mathbf{X})$, are Δ -homogeneous of degree θ_i (with $\theta_i \neq \theta_j$ if $i \neq j$), and satisfy

$$(4.15) \quad Z_i = R_i + \sum_k Q_{ik},$$

where the Q_{ik} are Δ -homogeneous of degree η_{ik} , the R_i belong to $\text{LR}(\mathbf{f}, p)$, and the η_{ik} satisfy $\eta_{ik} < \theta_i$.

It then follows that the Z_i belong to $L^N(\mathbf{X})$, for we can write $Z = \sum_{j=1}^N Z^j$, with $Z^j \in L^{j,\text{hom}}(\mathbf{X})$, and then each Z^j is a sum of Δ -homogeneous components Z_k^j , which must necessarily belong to $L^{j,\text{hom}}(\mathbf{X})$. The Z_i are then obtained by grouping together all the Z_k^j that have the same Δ -degree, and therefore belong to $L^N(\mathbf{X})$, as stated.

The Q_{ik} can also be assumed to belong to $L^N(\mathbf{X})$. Indeed, suppose that one Q_{ik} was not in $L^N(\mathbf{X})$. Since Q_{ik} is Δ -homogeneous, we must have

$$(4.16) \quad \deg_{\Delta}(Q_{ik}) \geq (N+1) \times \nu(\Delta).$$

On the other hand, we can write

$$(4.17) \quad \text{Ev}_p(\mathbf{f})(Q_{ik}) = \sum_l q_{ikl} \text{Ev}_p(\mathbf{f})(E_l)$$

for appropriate coefficients q_{ikl} . Therefore

$$(4.18) \quad Q_{ik} = R_{ik} + \sum_l q_{ikl} E_l$$

where $R_{ik} \in \text{LR}(\mathbf{f}, p)$. The $q_{ikl} E_l$ are Δ -homogeneous of degree σ_l . Since

$$(4.19) \quad \sigma_l \leq \nu_0(f, p, \Delta) \leq (N+1)\nu(\Delta) \leq \eta_{ik} < \theta_i,$$

we can replace each Q_{ik} that occurs in (4.15) but does not belong to $L^N(\mathbf{X})$ by the sum of the $q_{ikl} E_l$, and add R_{ik} to R_i . This leads to an expression for Z_i for the form (4.15), with all the Q_{ik} in $L^N(\mathbf{X})$.

It then follows that the R_i are in $L^N(\mathbf{X})$ as well. Define

$$(4.20) \quad \tilde{Z}_{\rho} = Z - \sum_{ik} \rho^{\theta_i - \eta_{ik}} Q_{ik}.$$

Then $\exp_N(\tilde{Z}_{\rho}) \in W_0$ if ρ is small enough. Therefore, if ρ is small, W is a normal neighborhood of $\exp_N(\tilde{Z}_{\rho})$ such that

$$(4.21) \quad \exp_N\left(\tilde{Z}_{\rho} + \sum_{i=1}^n y_i E_i\right) \in W$$

whenever $|y_i| \leq \beta$ for $i = 1, \dots, n$. Let $Z_{\rho} = \Delta(\rho)(\tilde{Z}_{\rho})$. Let $W_{\rho} = \Delta(\rho)W$. Then, if ρ is sufficiently small, W_{ρ} is a normal neighborhood of $\exp_N(Z_{\rho})$ such that

$$(4.22) \quad \exp_N\left(Z_{\rho} + \sum_{i=1}^n y_i \rho^{\sigma_i} E_i\right) \in W_{\rho}$$

whenever $|y_i| \leq \beta$ for $i = 1, \dots, n$. Let

$$(4.23) \quad S_{\rho} = \exp_N(Z_{\rho}).$$

Then

$$(4.24) \quad \log_N(W_{\rho}) \supseteq B_{\mathcal{G}}(Z_{\rho}, \beta \rho^{\nu_0(\mathbf{f}, p, \Delta)}).$$

On the other hand, S_{ρ} satisfies

$$(4.25) \quad T^N(S_{\rho}) \leq c \rho^{\nu(\Delta)} \quad \text{for } 0 < \rho \leq 1,$$

for some $c > 0$. (This is because \tilde{Z}_{ρ} is a sum of Δ -homogeneous components, each of which has Δ -degree at least equal to $\nu(\Delta)$, and coefficients that are bounded as $\rho \rightarrow 0$. Therefore all the coefficients of $\Delta(\rho)\tilde{Z}_{\rho}$ are bounded by a constant times $\rho^{\nu(\Delta)}$. In particular, this is true for the coefficient of X_0 , and so (4.25) follows.)

From (4.25) we conclude that

$$(4.26) \quad \rho^{\nu_0(\mathbf{f}, p, \Delta)} \geq \rho^{N\nu(\Delta)} \geq c^{-N} [T^N(S_{\rho})]^N$$

and so

$$(4.27) \quad \log_N (W_\rho) \supseteq B_{\mathcal{G}}(Z_\rho, \alpha [T^N(S_\rho)]^N)$$

if ρ is sufficiently small, and $\alpha = \beta c^{-N}$.

Finally, we have

$$\begin{aligned} Z_\rho &= \Delta(\rho) \tilde{Z}_\rho = \Delta(\rho) \left(\sum_i \left(Z_i - \sum_k \rho^{\theta_i - \eta_{ik}} Q_{ik} \right) \right) \\ &= \sum_i \left[\rho^{\theta_i} Z_i - \sum_k \rho^{\theta_i - \eta_{ik}} \rho^{\eta_{ik}} Q_{ik} \right] \\ &= \sum_i \rho^{\theta_i} \left[Z_i - \sum_k Q_{ik} \right] \\ &= \sum_i \rho^{\theta_i} R_i. \end{aligned}$$

Therefore $Z_\rho \in \text{LR}(\mathbf{f}, p)$. Hence all the hypotheses of Theorem 3.4 are satisfied, and the desired conclusion follows.

5. Invariant elements. Theorem 4.3 says that (M, \mathbf{f}, K) is STLC from p if, for some sufficiently large N , $\hat{S}^N(\mathbf{X}, K)$ contains an element S such that $\log_N(S)$ is Δ -neutralized for \mathbf{f} at p . In order to be able to use this result, we need to know that $\hat{S}^N(\mathbf{X}, K)$ necessarily will contain elements of some very special kind, for then STLC will follow if we hypothesize that these special elements are exponentials of Δ -neutralized members of $L^N(\mathbf{X})$.

To get these “special elements” we exploit a general result about existence of points that are invariant under certain finite groups of pseudoautomorphisms (cf. § 2 for the definition of “pseudoautomorphism”).

Let L be a finite-dimensional, nilpotent Lie algebra over \mathbb{R} , and let G_L be its corresponding connected, simply connected Lie group. Then the exponential map $\exp: L \rightarrow G_L$ is a diffeomorphism onto. Therefore, if $\lambda: L \rightarrow L$ is an arbitray map, then λ gives rise to a map $\tilde{\lambda}: G_L \rightarrow G_L$, defined by letting

$$(5.1) \quad \tilde{\lambda}(\exp(z)) = \exp(\lambda(z)).$$

PROPOSITION 5.1. *Let L be a finite-dimensional, nilpotent Lie algebra over \mathbb{R} , and let G_L be the corresponding connected, simply connected Lie group. Let Λ be a finite group of pseudoautomorphisms of L , and let $\tilde{\Lambda} = \{\tilde{\lambda}: \lambda \in \Lambda\}$ be the group of bijections of G_L induced by Λ . Let S be a nonempty subset of G_L which is closed under multiplication. Suppose that every $\tilde{\lambda} \in \tilde{\Lambda}$ maps S into S . Then S contains an element s such that $\tilde{\lambda}(s) = s$ for all $\lambda \in \Lambda$.*

Proof. Start with an element $s_1 \in S$, and write $s_1 = \exp(b_1)$, where $b_1 \in L$. Let $\Lambda = \{\lambda_1, \dots, \lambda_n\}$, with $\lambda_i \neq \lambda_j$ whenever $i \neq j$. Define s_2 by

$$(5.2) \quad s_2 = \tilde{\lambda}_1(s_1) \tilde{\lambda}_2(s_1) \cdots \tilde{\lambda}_n(s_1).$$

Then $s_2 \in S$, because $\tilde{\lambda}_j(s_1) \in S$ for each j , and S is closed under multiplication. On the other hand, we have

$$(5.3) \quad \tilde{\lambda}_j(s_1) = \exp(\lambda_j(b_1)) \quad \text{for } j = 1, \dots, n.$$

Therefore the Campbell–Hausdorff formula gives

$$(5.4) \quad s_2 = \exp(z_2 + b_2)$$

where

$$(5.5) \quad z_2 = \lambda_1(b_1) + \cdots + \lambda_n(b_1)$$

and $b_2 \in [L]^2$. In view of (5.5), z_2 satisfies $\lambda(z_2) = z_2$ for every $\lambda \in \Lambda$. Assume we have proved, for some k , that there exists an $s_k \in S$ which is of the form $\exp(z_k + b_k)$, with $\lambda(z_k) = z_k$ for all $\lambda \in \Lambda$, and $b_k \in [L]^k$. Then we can define s_{k+1} by

$$(5.6) \quad s_{k+1} = \lambda_1(s_k) \lambda_2(s_k) \cdots \lambda_n(s_k)$$

and conclude from the Campbell-Hausdorff formula that

$$(5.7) \quad s_{k+1} = \exp(z_{k+1} + b_{k+1}),$$

where

$$(5.8) \quad z_{k+1} = \sum_{i=1}^n \lambda_i(z_k + b_k)$$

and b_{k+1} is a linear combination of terms, each of which is a Lie bracket of two or more elements of L of the form $\lambda_j(z_k + b_k)$ for some j . But $\lambda_j(z_k) = z_k$ and, if we let $b_{jk} = \lambda_j(b_k)$, we have $b_{jk} \in [L]^k$, because $b_k \in [L]^k$ and λ is a pseudoautomorphism. So the brackets that appear in b_{k+1} are brackets of two or more terms of the form $z_k + b_{jk}$. Now $[z_k + b_{ik}, z_k - b_{jk}] = [z_k, b_{jk}] + [b_{ik}, z_k] + [b_{ik}, b_{jk}]$, and so $[z_k + b_{ik}, z_k + b_{jk}] \in [L]^{k+1}$. So $b_{k+1} \in [L]^{k+1}$. This proves, by induction, that an $s_k \in S$ of the desired form exists for every k . Since L is nilpotent, we can take k such that $[L]^k = \{0\}$. Then $b_k = 0$, and so, if we let $s = s_k$, the condition that $\lambda(s) = s$ holds for all $\lambda \in \Lambda$.

6. End of the proof of Theorem 2.4. Assume that the conditions of Theorem 2.4 hold. Pick N so large that (4.12) and (4.13) hold. The group Λ obviously induces a group Λ_N of pseudoautomorphisms of the Lie algebra $L^N(\mathbf{X})$. The maps $\tilde{\lambda}$, for $\lambda \in \Lambda_N$, clearly map $\hat{S}^N(\mathbf{X}, K)$ into itself. The set $\hat{S}^N(\mathbf{X}, K)$ is nonempty and closed under multiplication. Proposition 5.1 then implies that $\hat{S}^N(\mathbf{X}, K)$ contains an element $S = \exp_N(Z)$, where $z \in L^N(\mathbf{X})$ is Λ_N -fixed. Then Z is Λ -fixed, and therefore Z is Δ -neutralized for \mathbf{f} at p . Theorem 4.3 then says that (M, \mathbf{f}, K) is STLC from p .

7. Applications. In all the applications discussed here, Λ will be a group obtained from a group of automorphisms Λ_0 of $L(\mathbf{X})$, by adding to it the "time reversal" map. Precisely, let $\mathbb{T}^A: A(\mathbf{X}) \rightarrow A(\mathbf{X})$ be the linear map which sends each monomial $X_{i_1} X_{i_2} \cdots X_{i_k}$ to the "reversed" monomial $X_{i_k} \cdots X_{i_2} X_{i_1}$. Then \mathbb{T}^A is an antiautomorphism of $A(\mathbf{X})$ (i.e. $\mathbb{T}^A(PQ) = \mathbb{T}^A(Q) \mathbb{T}^A(P)$ for all P, Q in $A(\mathbf{X})$). It then follows easily that $\mathbb{T}^A([P, Q]) = [\mathbb{T}^A(Q), \mathbb{T}^A(P)]$, i.e.

$$(7.1) \quad \mathbb{T}^A([P, Q]) = -[\mathbb{T}^A(P), \mathbb{T}^A(Q)],$$

for P, Q in $A(\mathbf{X})$. Then \mathbb{T}^A maps $L(\mathbf{X})$ to $L(\mathbf{X})$, and

$$(7.2) \quad \mathbb{T}(P) = (-1)^{1+k} P \quad \text{for } P \in L^{k, \text{hom}}(\mathbf{X}),$$

where \mathbb{T} denotes the restriction of \mathbb{T}^A to $L(\mathbf{X})$.

It is clear that \mathbb{T} is a pseudoautomorphism of $L(\mathbf{X})$. On the other hand \mathbb{T}^A gives rise in an obvious way to a map $\hat{\mathbb{T}}^A: \hat{A}(\mathbf{X}) \rightarrow \hat{A}(\mathbf{X})$. Clearly, $\hat{\mathbb{T}}^A(P^k) = P^k$ if $P \in A^{1, \text{hom}}(\mathbf{X})$. Therefore

$$(7.3) \quad \hat{\mathbb{T}}^A(\exp P) = \exp P$$

if $P \in A^{1, \text{hom}}(\mathbf{X})$. So, if P_1, \dots, P_k are elements of $L^{1, \text{hom}}(\mathbf{X})$, we have

$$(7.4) \quad \hat{\mathbb{T}}^A(\exp(P_1) \cdots \exp(P_k)) = \exp(P_k) \cdots \exp(P_1).$$

This implies that, if $u(\cdot)$ is a piecewise constant K -valued control, defined on $[0, T]$, then

$$(7.5) \quad \hat{\mathbb{T}}^A(\text{Ser}(u(\cdot))) = \text{Ser}(u^{\text{rev}}(\cdot))$$

where $u^{\text{rev}}(t) = u(T-t)$ for $0 \leq t \leq T$. By an elementary continuity argument, (7.5) holds for all controls $u(\cdot)$. Therefore

$$(7.6) \quad \hat{\mathbb{T}}^A(\hat{S}(\mathbf{X}, K)) = \hat{S}(\mathbf{X}, K).$$

On the other hand, \mathbb{T} gives rise to a map $\hat{\mathbb{T}}: \hat{L}(\mathbf{X}) \rightarrow \hat{L}(\mathbf{X})$, which is obviously equal to the restriction of $\hat{\mathbb{T}}^A$ to $\hat{L}(\mathbf{X})$. If P is any element of $\hat{A}_0(\mathbf{X})$, then $\hat{\mathbb{T}}^A(P^k) = [\hat{\mathbb{T}}^A(P)]^k$ for every k , and therefore

$$(7.7) \quad \hat{\mathbb{T}}^A(\exp(P)) = \exp(\hat{\mathbb{T}}^A(P)).$$

In particular, if $P \in \hat{L}(\mathbf{X})$, we get the equality

$$(7.8) \quad \hat{\mathbb{T}}^A(\exp(P)) = \exp(\hat{\mathbb{T}}(P)),$$

which implies

$$(7.9) \quad \hat{\mathbb{T}}^A(\hat{G}(\mathbf{X})) = \hat{G}(\mathbf{X}).$$

In the terminology of § 2 (cf. especially (2.7)), (7.8) shows that the restriction of $\hat{\mathbb{T}}^A$ to $\hat{G}(\mathbf{X})$ is precisely the map $\mathbb{T}^\# : \hat{G}(\mathbf{X}) \rightarrow \hat{G}(\mathbf{X})$. Hence (7.6) says that $\mathbb{T}^\#$ maps $\hat{S}(\mathbf{X}, K)$ to $\hat{S}(\mathbf{X}, K)$. So we have proved:

LEMMA 7.1. \mathbb{T} is an input symmetry.

Now suppose that Λ_0 is a finite group of *graded* linear maps from $L(\mathbf{X})$ to $L(\mathbf{X})$. (A linear map $\lambda : L(\mathbf{X}) \rightarrow L(\mathbf{X})$ is *graded* if λ maps $L^{j,\text{hom}}(\mathbf{X})$ into $L^{j,\text{hom}}(\mathbf{X})$ for each j .) Then every $\lambda \in \Lambda_0$ commutes with \mathbb{T} . Since \mathbb{T}^2 is the identity map, the set

$$(7.10) \quad \Lambda = \Lambda_0 \cup \{\lambda \mathbb{T} : \lambda \in \Lambda_0\},$$

is a finite group of pseudoautomorphisms. If Λ_0 is a group of input symmetries, then Λ is a group of input symmetries as well. We shall refer to the input symmetry \mathbb{T} as “time reversal,” and to the group Λ defined by (7.10) as “the augmentation of Λ_0 by time reversal.”

Let us call an element of $L(\mathbf{X})$ *totally odd* if all its homogeneous components have odd degree. Then it is clear that the totally odd elements of $L(\mathbf{X})$ are precisely those $P \in L(\mathbf{X})$ that satisfy $\mathbb{T}(P) = -P$. If Λ_0 is a finite group of graded linear maps of $L(\mathbf{X})$, and Λ is its augmentation by time reversal, then the Λ -fixed elements of $L(\mathbf{X})$ are precisely those $P \in L(\mathbf{X})$ that are Λ_0 -fixed and totally odd. So we can conclude from Theorem 2.4 the following:

COROLLARY 7.2. Let (M, \mathbf{f}, K) be a control system, and let $p \in M$. Assume that \mathbf{f} satisfies the LARC at p , and that there exist (a) an admissible group of dilations Δ of $L^{1,\text{hom}}(\mathbf{X})$ which is compatible with $\hat{S}(\mathbf{X}, K)$, (b) a finite group Λ_0 of graded linear maps from $L(\mathbf{X})$ to $L(\mathbf{X})$ that are input symmetries, such that every totally odd Λ_0 -fixed element of $L(\mathbf{X})$ is Δ -neutralized for \mathbf{f} at p . Then (M, \mathbf{f}, K) is STLC from p .

7.1. Symmetric systems. A symmetric system is a family $\mathcal{V} = \{V_i : i \in I\}$ of vector fields on a manifold M , such that for every $i \in I$ there is a $j \in I$ such that $V_j = -V_i$. It is well known that, if a symmetric system satisfies the LARC at p , then the system is STLC from p . For completeness, we show that our theorem implies this result. First, it is clear that we can pick vector fields f_1, \dots, f_m in this family such that the m -tuple (f_1, \dots, f_m) satisfies the LARC at p . Then we can let $f_0 = 0$. Also, we take K to be the set of all points of \mathbb{R}^m of the form $(0, 0, \dots, 0, \pm 1, 0, \dots, 0)$. We let Δ be the group

of dilations such that $\Delta(\rho)(P) = \rho P$ for $P \in L^{1,\text{hom}}(X)$, so that the Δ -degree is just the ordinary degree. We let Λ_0 be the group of automorphisms of $L(X)$ generated by $\lambda_1, \dots, \lambda_m$, where λ_i is the automorphism that takes X_j to X_j for $j \neq i$, and X_i to $-X_i$. Since the λ_i commute, Λ_0 is finite. The Λ_0 -fixed elements of $L(X)$ are those that are linear combinations of brackets where each X_i , $i = 1, \dots, m$, occurs an even number of times. Such a bracket cannot be totally odd unless it contains X_0 . But then, when the bracket is evaluated by plugging in the f_j for the X_j , the result must be zero, because $f_0 = 0$. Hence every totally odd Λ_0 -fixed element of $L(X)$ is actually in $\text{LR}(\mathbf{f}, p)$, and therefore is Δ -neutralized for \mathbf{f} at p . So we can apply Corollary 7.2 and conclude that $\Sigma = (M, \mathbf{f}, K)$ is STLC from p . Since every trajectory of Σ is a trajectory of \mathcal{V} , the small-time local controllability of \mathcal{V} follows.

7.2. The results of Brunovsky, Crouch and Byrnes. In [3], Brunovsky defined an *odd family* $\mathcal{V} = \{V_i: i \in I\}$ of vector fields on a symmetric neighborhood M of 0 to be a family such that for every $i \in I$ there is a $j \in I$ such that $V_j(-x) = -V_i(x)$ for $x \in M$. He then proved that, if \mathcal{V} is odd and satisfies the LARC, then \mathcal{V} is STLC from 0. Crouch and Byrnes [5] provided a coordinate-free generalization of this result. Suppose that $\mathcal{V} = \{V_i: i \in I\}$ is a collection of vector fields on a manifold M , and $p \in M$. Suppose that \mathcal{V} satisfies the LARC at p . Assume that there is a finite group Λ_0 of diffeomorphisms of M such that

- (i) each $\lambda \in \Lambda_0$ maps p to p ,
- (ii) each $\lambda \in \Lambda_0$ maps each V_i to some V_j in the family,
- (iii) the differentials at p of the maps $\lambda \in \Lambda_0$ have no common invariant half space.

The result of [5] then says that \mathcal{V} is small-time locally controllable from p . Brunovsky's theorem is a particular case of this, obtained by letting Λ_0 consist of the identity and the map $\lambda: x \rightarrow -x$. (If $V_j(-x) = -V_i(x)$, and λ_* denotes the differential of λ , so that $\lambda_*(v) = -v$, then λ_* maps V_i to V_j .)

We show that the result of [5] is a particular case of our Corollary 7.2. Let f_1, \dots, f_m be members of the family \mathcal{V} , chosen so that: (i) (f_1, \dots, f_m) satisfies the LARC, (ii) $f_i \neq f_j$ whenever $i \neq j$, (iii) the set $\{f_1, \dots, f_m\}$ is mapped to itself by the maps λ_* , $\lambda \in \Lambda_0$. Let $f_0 \equiv 0$, $\mathbf{f} = (f_0, \dots, f_m)$. Then consider the system (M, \mathbf{f}, K) , where $K = \{0\} \cup \tilde{K}$, and \tilde{K} is the set of all vectors of \mathbb{R}^m of the form $(0, 0, \dots, 0, 1, 0, \dots, 0)$. If $\Sigma = (M, \mathbf{f}, K)$ is STLC from p , then \mathcal{V} is. (Indeed, let q be reachable from p by a trajectory of Σ that corresponds to a piecewise constant $u(\cdot): [0, T] \rightarrow K$. Then q can also be reached by a trajectory that corresponds to a \tilde{K} -valued control, in time $T' \leq T$, by simply eliminating from $u(\cdot)$ all the pieces for which $u(\cdot)$ has the value 0.) The group Λ_0 acts on $L(X)$ for, if $\lambda \in \Lambda_0$, then λ_* permutes the elements of $\{f_1, \dots, f_m\}$, and so we can define an automorphism $g(\lambda)$ of $L(X)$ by

$$(7.11) \quad g(\lambda)(X_0) = X_0$$

and

$$(7.12) \quad g(\lambda)(X_i) = X_j \quad \text{if } \lambda_*(f_i) = f_j.$$

If \mathcal{V} is any element of $L(X)$, it follows from (7.11) and (7.12) that

$$(7.13) \quad \text{Ev}(\mathbf{f})(g(\lambda)(V)) = \lambda_*(\text{Ev}(\mathbf{f})(V)).$$

In particular, this implies that, if V is Λ_0 -fixed, then the vector $\text{Ev}_p(\mathbf{f})(V)$ is invariant under the differentials at p of all the maps $\lambda \in \Lambda_0$. Therefore $\text{Ev}_p(\mathbf{f})(V) = 0$, and so $V \in \text{LR}(\mathbf{f}, p)$.

So, if Δ is any group of dilations whatsoever, all the Λ_0 -fixed elements of $L(X)$ are Δ -neutralized for f at p , and so Σ is STLC from p .

7.3. The Hermes condition and some generalizations. Consider a system

$$(7.14) \quad \dot{x} = f_0(x) + \sum_{i=1}^m u_i f_i(x), \quad |u_i| \leq 1,$$

and assume that p is an equilibrium point of f_0 , i.e. that $f_0(p) = 0$. We let Λ_0 be the group of automorphisms of $L(X)$ generated by $\sigma_1, \dots, \sigma_m$ and all the $\tilde{\pi}$, $\pi \in S_m$, where: (a) S_m is the group of permutations of $\{1, \dots, m\}$, (b) for $\pi \in S_m$, $\tilde{\pi}$ is the automorphism of $L(X)$ which maps X_0 to X_0 and X_i to $X_{\pi(i)}$ for $i = 1, \dots, m$, (c) σ_i is the automorphism that sends X_j to X_j for $j \neq i$, and X_i to $-X_i$. It is clear that Λ_0 is finite. The Λ_0 -fixed elements are those that are linear combinations of elements of the form $\alpha(B)$, where B is a bracket of X_0, \dots, X_m , and α is the Λ_0 -symmetrization operator, i.e.

$$(7.15) \quad \alpha(V) = \sum_{\lambda \in \Lambda_0} \lambda(V).$$

It is clear that $\alpha(V) = 0$, if $\sigma_i(V) = -V$ for some i . Therefore, $\alpha(B) = 0$ if B is a bracket in which one of the X_i , $i > 0$, appears an odd number of times. Hence, in order to find the Λ_0 -fixed elements, we may limit ourselves to considering the symmetrizations of brackets B where, for $i = 1, \dots, m$, X_i appears an even number of times. For such a B , one may use the symmetrization operator β given by

$$(7.16) \quad \beta(V) = \sum_{\pi \in S_m} \tilde{\pi}(V).$$

Next, let $\theta_1, \dots, \theta_m$ be arbitrary real numbers such that $\theta_i \geq 1$ for $i = 1, \dots, m$. Define $\Delta(\rho)$ by

$$(7.17) \quad \Delta(\rho): (X_0, \dots, X_m) \rightarrow (\rho X_0, \rho^{\theta_1} X_1, \rho^{\theta_2} X_2, \dots, \rho^{\theta_m} X_m).$$

Then Δ is compatible with $\tilde{S}(X, K)$, where $K = \{(u_1, \dots, u_m) : |u_i| \leq 1 \text{ for } i = 1, \dots, m\}$. Then Corollary 7.2 implies that the system is STLC if, whenever B is a bracket with an odd number of X_0 's, and an even number of X_i 's for each $i \in \{1, \dots, m\}$, it follows that every Δ -homogeneous component of $\beta(B)$ is equal, when evaluated at p , to a linear combination of brackets of lower Δ -degree. When the θ_i are different, this condition requires too much, for $\beta(B)$ will in general fail to be homogeneous. So the most interesting case obtains when all the θ_i are equal. Let $1 \leq \theta < \infty$. Define the θ -degree δ_θ of a bracket $B \in \text{Br}(X)$ to be the sum

$$(7.18) \quad \delta_\theta(B) = \delta^0(B) + \theta \sum_{i=1}^m \delta^i(B)$$

where $\delta^i(B)$ is the number of times that X_i occurs in B . Then δ_θ is the degree that arises, in an obvious way, from a group of dilations Δ_θ . We can then apply Corollary 7.2 to get a local controllability theorem involving the group Δ_θ . However, Δ_θ only enters the theorem via the concept of Δ -neutralization, and this concept is unchanged if we multiply all the degrees by a fixed number $\nu > 0$. Hence we can use, instead of δ_θ , the degree $\hat{\delta}_\theta$ defined by $\hat{\delta}_\theta(B) = (1/\theta)\delta_\theta(B)$, i.e.

$$(7.19) \quad \hat{\delta}_\theta(B) = \frac{1}{\theta} \delta^0(B) + \sum_{i=1}^m \delta^i(B).$$

The new definition now has the advantage that $\hat{\delta}_\theta$ also makes sense for $\theta = \infty$, in which case $\hat{\delta}_\infty(B)$ is, simply, the total number of occurrences in B of the X_i for $i = 1, \dots, m$. (But $\hat{\delta}_\infty$ does not arise from an admissible group of dilations.) We can then state the following

THEOREM 7.3. *Consider a system*

$$(7.20) \quad \dot{x} = f_0(x) + \sum_{i=1}^m u_i f_i(x), \quad x \in M, \quad |u_i| \leq 1,$$

and a point $p \in M$ such that $f_0(p) = 0$. Assume that (f_0, \dots, f_m) satisfies the LARC at p . Assume that there is a $\theta \in [1, \infty]$ such that, whenever $B \in \text{Br}(X)$ is a bracket for which $\delta^0(B)$ is odd and $\delta^1(B), \dots, \delta^m(B)$ are even, then there are brackets C_1, \dots, C_k in $\text{Br}(X)$ such that

$$(7.21) \quad \text{Ev}_p(f)(\beta(B)) = \sum_{i=1}^k \xi_i \text{Ev}_p(f)(C_i)$$

for some $\xi_1, \dots, \xi_k \in \mathbb{R}$, and

$$(7.22) \quad \hat{\delta}_\theta(C_i) < \hat{\delta}_\theta(B) \quad \text{for } i = 1, \dots, k.$$

Then the system (7.20) is STLC from p .

For $\theta < \infty$, this theorem is just the result of applying Corollary 7.2 to our situation, using the group of dilations Δ_θ . To prove the theorem for $\theta = \infty$ we just “take the limit as $\theta \rightarrow \infty$.” Rigorously, this means that, if the hypotheses are satisfied for $\theta = \infty$, then they are also satisfied for some large finite θ . To see this, let us use $S_{k,l}$ to denote the linear span of all the brackets B such that $\delta^0(B) \leq k$ and $\delta^1(B) + \dots + \delta^m(B) \leq l$. If \mathbf{l} is an m -tuple (l_1, \dots, l_m) of nonnegative integers, we define $\tilde{S}_{k,\mathbf{l}}$ to be the linear span of those B 's for which $\delta^0(B) \leq k$, $\delta^1(B) = l_1, \dots, \delta^m(B) = l_m$. Also, we define

$$(7.23) \quad S_{\infty,l} = \bigcup_{k=0}^{\infty} S_{k,l},$$

$$(7.24) \quad \tilde{S}_{\infty,\mathbf{l}} = \bigcup_{k=0}^{\infty} \tilde{S}_{k,\mathbf{l}},$$

and we define spaces $S_{\infty,\mathbf{l}}^{\text{odd}}, \tilde{S}_{\infty,\mathbf{l}}^{\text{odd}}$ in exactly the same way, except that the unions are only taken over odd values of k . We call the m -tuple \mathbf{l} *even* if all its components l_1, \dots, l_m are even. Then the hypothesis of Theorem 7.3 for $\theta = \infty$ says that

$$(7.25) \quad \text{Ev}_p(f)\beta(\tilde{S}_{\infty,\mathbf{l}}^{\text{odd}}) \subseteq \text{Ev}_p(f)(S_{\infty,\|\mathbf{l}\|-1})$$

for all even \mathbf{l} . Pick an \bar{l} such that

$$(7.26) \quad \text{Ev}_p(f)(S_{\infty,\bar{l}}) = T_p M.$$

For $l = 0, 1, \dots, \bar{l}$, pick $k(l)$ such that

$$(7.27) \quad \text{Ev}_p(f)(S_{\infty,l}) = \text{Ev}_p(f)(S_{k(l),l}).$$

Then pick $\theta \in [1, \infty)$ such that $k(l) < \theta$ for $l = 0, \dots, \bar{l}$. We claim that, with this choice of θ , the hypothesis of Theorem 7.3 holds. To see this, let $B \in \tilde{S}_{\infty,\mathbf{l}}^{\text{odd}}$. Assume first that $\|\mathbf{l}\| > \bar{l}$. Then $\text{Ev}_p(f)(B) \in \text{Ev}_p(f)(S_{\infty,\bar{l}})$, so that $\text{Ev}_p(f)(B) \in \text{Ev}_p(f)(S_{k(\bar{l}),\bar{l}})$. Therefore $\text{Ev}_p(f)(B)$ is a linear combination of vectors $\text{Ev}_p(f)(C_i)$, where the C_i are in $S_{k(\bar{l}),\bar{l}}$.

But then

$$(7.28) \quad \hat{\delta}_\theta(C_i) \leq \frac{k(\bar{l})}{\theta} + \bar{l} < \bar{l} + 1 \leq |\mathbf{l}| \leq \hat{\delta}_\theta(B).$$

Next assume that $|\mathbf{l}| \leq \bar{l}$ and \mathbf{l} is even. By (7.25) and (7.27), $\text{Ev}_p(\mathbf{f})(\beta(B))$ is a linear combination of vectors $\text{Ev}_p(\mathbf{f})(C_i)$ with $C_i \in S_{k(\lambda), \lambda}$, where $\lambda = |\mathbf{l}| - 1$. But then

$$(7.29) \quad \hat{\delta}_\theta(C_i) \leq \frac{k(\lambda)}{\theta} + \lambda < \lambda + 1 = |\mathbf{l}| < \hat{\delta}_\theta(B).$$

The proof of Theorem 7.3 is now complete.

Theorem 7.3 contains as a particular case a result for single-input systems was conjectured by H. Hermes and proved by us in [25]. Precisely, the Hermes condition (HC) for a system

$$(7.30) \quad \dot{x} = f(x) + ug(x), \quad |u| \leq 1,$$

at a point p , is the condition that, if B is an arbitrary bracket of f 's and g 's with an even number of g 's, then $B(p)$ is a linear combination of values at p of brackets with fewer g 's. (In particular, by taking $B = f$, we see that the HC implies that $f(p) = 0$.) The result proved in [25] says that, if the system (7.14) satisfies the LARC and the HC at p , then it is STLC from p . If we apply Theorem 7.3 with $m = 1$ (in which case, of course, the symmetrization operator β is just the identity), we obtain a strengthened version of the theorem of [25]. The HC corresponds to $\theta = \infty$ whereas Theorem 7.3 allows other values of θ . Moreover, even if we apply Theorem 7.3 with $\theta = \infty$, the condition that has to be satisfied to get controllability is weaker than the HC, and therefore the resulting controllability theorem is stronger. (The HC demands that *every* bracket with an even number of g 's be neutralized, whereas our result only requires this for brackets with an even number of g 's *and an odd number of f 's*. As will be shown in examples below, these refinements make it possible to handle cases where the HC is insufficient.)

For general m , R. Grossmann [8] states a sufficient condition for controllability, namely, that every bracket where each of the f_i for $i = 1, \dots, m$ occurs an even number of times be expressible, at p , as a linear combination of brackets of lower total degree. This condition amounts to a weaker form of the case $\theta = 1$ of our theorem. (Theorem 7.3 only requires that the symmetrized brackets, which in addition have an odd number of f_0 's, be expressible as linear combination of lower-degree elements.)

7.4. Low order sufficient conditions for systems with a cubic control set. We now illustrate the use of Theorem 7.3 by deriving some sufficient conditions for systems of the form (7.14), in terms of brackets of low degree. Assume that p is an equilibrium point of (7.14), i.e. that $f_0(p) = 0$. Also, assume that (7.14) satisfies the LARC at p . The simplest sufficient condition for STLC is the one obtained from the Pontryagin Maximum Principle, which says that (7.14) is STLC from p if, for every $\varepsilon > 0$, the adjoint equation along the trajectory $t \rightarrow x(t) = p$, $0 \leq t \leq \varepsilon$, has no nontrivial solution $t \rightarrow \lambda(t)$ such that $\langle \lambda(t), f_i(p) \rangle = 0$ for $0 \leq t \leq \varepsilon$. The adjoint equation in this case, written in coordinates, is simply the equation

$$(7.31) \quad \dot{\lambda} = -\lambda A,$$

where A is the Jacobian matrix of f_0 at 0. If $\lambda(\cdot)$ is a solution of (7.31) such that $\langle \lambda(t), f_i(p) \rangle \equiv 0$, then $\langle \lambda(0), A^k f_i(p) \rangle = 0$ for all k . Hence, if the vectors $A^k f_i(p)$ $i = 1, \dots, m$, $k = 0, 1, \dots$ span $T_p M$, there will not exist a $\lambda(\cdot)$ with the desired properties, and so

(7.14) will be STLC from p . Clearly, $A^k f_i(p) = (\text{ad } f_0)^k(f_i)(p)$. Therefore the sufficient condition obtained from the Maximum Principle simply says that (7.14) will be STLC from p if the vectors $(\text{ad } f_0)^k(f_i)(p)$, $i = 1, \dots, m$, $k = 0, 1, \dots$ span $T_p M$. Theorem 7.3 implies a stronger result, namely

PROPOSITION 7.4. *Assume that $f_0(p) = 0$, and the vectors $(\text{ad } f_0)^k(f_i)(p)$, $i = 1, \dots, m$, $k = 0, 1, \dots$, together with the vectors $[f_i, f_j](p)$, $i, j \in \{1, \dots, m\}$, span $T_p M$. Then (7.14) is STLC from p .*

Proof. Our hypotheses imply in particular that (7.14) satisfies the LARC from p . Let $\mu > 0$ be such that the span of the vectors $(\text{ad } f_0)^k(f_i)(p)$, $i \in \{1, \dots, m\}$, $k = 0, 1, \dots$, is actually spanned by vectors of this same form with $k \leq \mu$. Pick θ such that $\mu \leq \theta < \infty$. Then $T_p M$ is spanned by vectors $\text{Ev}_p(\mathbf{f})(B)$, where the B 's are brackets such that $\hat{\delta}_\theta(B) \leq 2$. On the other hand, if C is any bracket with an odd number of X_0 's and an even number of X_i 's for each $i \in \{1, \dots, m\}$, then either $B = X_0$, in which case $\text{Ev}_p(\mathbf{f})(B) = 0$, or $\hat{\delta}_\theta(B) > 2$, in which case $\text{Ev}_p(\mathbf{f})(C)$ is certainly a linear combination of vectors $\text{Ev}_p(\mathbf{f})(B)$ with $\hat{\delta}_\theta(B) < \hat{\delta}_\theta(C)$. Hence the conditions of Theorem 7.3 are satisfied, and (7.14) is STLC from p .

If the sufficient condition of Proposition 7.4 is not satisfied, then it will be necessary to "neutralize" some brackets in order to be able to apply Theorem 7.3. The lowest total degree d where there may exist brackets to be neutralized is $d = 3$. (The case $d = 1$ is disposed of by the assumption that $f_0(p) = 0$.) The only brackets of total degree 3 where X_0 occurs an odd number of times, and each of the other X_i 's an even number of times, are the expressions $[X_i, [X_i, X_0]]$. Symmetrization yields the element

$$(7.32) \quad H = \sum_{i=1}^m [X_i, [X_i, X_0]].$$

We write $h = \text{Ev}(\mathbf{f})(H)$.

If h is "neutralized," in the sense that $h(p)$ is a linear combination of vectors $g_j(p)$, where the g_j are brackets of "lower degree," then that "releases" a whole collection of new brackets. If these brackets now span $T_p M$, then we get controllability again. Exactly which brackets are released by the neutralization of h will depend on how h is neutralized. Suppose that

$$(7.33) \quad h(p) = \sum_{i=1}^m \sum_{k=0}^{\nu} \alpha_{ik} (\text{ad } f_0)^k(f_i)(p) + \sum_{i=1}^m \sum_{j=1}^m \beta_{ij} [f_i, f_j](p)$$

for some choice of coefficients α_{ik}, β_{ij} .

Then, if we choose any θ such that $\theta \geq 1$, $\theta > \nu - 1$, we see that $\text{Ev}_p(\mathbf{f})(H)$ is a linear combination of vectors $\text{Ev}_p(\mathbf{f})(B)$ with $\hat{\delta}_\theta(B) < \hat{\delta}_\theta(H)$. The next value of the total degree d for which there may be brackets B to be neutralized is $d = 5$. And the lowest possible value $\hat{\delta}_\theta(B)$ for such brackets is $2 + (3/\theta)$. If the brackets for which $\hat{\delta}_\theta < 2 + (3/\theta)$ span $T_p M$, the system will be STLC from p . So we get

PROPOSITION 7.5. *Assume that (i) $f_0(p) = 0$, (ii) (7.33) holds for some ν and some choice of coefficients α_{ik}, β_{ij} . Assume that there is a number $\theta \in [1, \infty]$ such that $\theta > \nu - 1$, with the property that the brackets B with $k_1 f_0$'s and $k_2 f_i$'s with $i > 0$ for all k_1, k_2 such that $k_1 + \theta k_2 < 2\theta + 3$, span $T_p M$. Then (7.14) is STLC from p .*

As a simple example, suppose that $h(p) = 0$ or, more generally, that (7.33) holds with $\nu = 1$. Then θ can be chosen to be an arbitrary number in $[1, \infty]$. In particular, we can conclude that the system (7.14) is STLC from p if either (i) $T_p M$ is spanned by all the brackets of total degree ≤ 4 or (ii) $T_p M$ is spanned by all the brackets with $\delta^+ = 1$, $\delta^0 \leq 4$, together with those with $\delta^+ = 2$, $\delta^0 \leq 2$, those with $\delta^+ = 3$ and $\delta^0 \leq 1$,

and those with $\delta^+ = 4$ and $\delta^0 = 0$, or (iii) $T_p M$ is spanned by the brackets with $\delta^+ = 1$, $\delta^0 \leq 5$, together with those with $\delta^+ = 2$, $\delta^0 \leq 2$, and those with $\delta^+ = 3$, $\delta^0 = 0$, or (iv) $T_p M$ is spanned by the brackets with $\delta^+ = 1$, δ^0 arbitrary, together with those with $\delta^+ = 2$, $\delta^0 \leq 2$. (Here, for a bracket B , $\delta^i(B)$ is the number of occurrences of f^i in B , and $\delta^+(B) = \sum_{i=1}^m \delta^i(B)$. The four results stated above are obtained by taking, respectively, $\theta = 1$, $\theta = 1.1$, $\theta = 2.2$, and θ very large.) If (7.33) holds with $\nu = 2$, then we have to choose $\theta > 1$, and so we can conclude that (7.14) is STLC from p if (ii), (iii) or (iv) above hold. If (7.33) holds with $\nu = 3$, then we must choose $\theta > 2$, and we get that (7.14) is STLC from p if (iii) or (iv) hold. Finally, if (7.33) holds with some $\nu \geq 4$, we get small-time local controllability if (iv) holds.

Notice, in particular, that if H is neutralized, in the sense that (7.33) holds for some ν , then this has the effect of unconditionally releasing a number of brackets, namely, all the brackets with $\delta^+ = 2$, $\delta^0 \leq 2$.

Finally, we illustrate the result of Theorem 7.3 in the case $m = 2$, by giving some simple sufficient conditions in terms of brackets up to degree 6. The algebra to be considered here is $L(X_0, X_1, X_2)$. The homogeneous components $L^{j,\text{hom}}(X_0, X_1, X_2)$ have dimensions 3, 3, 8, 18, 48, 116, for $j = 1, 2, 3, 4, 5, 6$, respectively. So there is a total of 196 potentially linearly independent brackets of degree ≤ 6 . After we eliminate those brackets that are totally even, or odd in either X_1 or X_2 , and symmetrize, we are left with exactly eight linearly independent elements to be neutralized, namely, (a) X_0 , (b) H , and (c) six elements $B_1, B_2, B_3, B_4, B_5, B_6$ of degree five, given by

$$(7.34) \quad B_1 = [[X_1, X_2], [[X_1, X_2], X_0]],$$

$$(7.35) \quad B_2 = \sum_{i=1}^2 [X_i, (\text{ad } X_0)^3(X_i)],$$

$$(7.36) \quad B_3 = \sum_{i=1}^2 [[X_0, X_i], [X_0, [X_0, X_i]]],$$

$$(7.37) \quad B_4 = \sum_{i=1}^2 (\text{ad } X_i)^4(X_0),$$

$$(7.38) \quad B_5 = [X_1, [X_1, [X_2, [X_0, X_2]]]] + [X_2, [X_2, [X_1, [X_0, X_1]]]],$$

$$(7.39) \quad B_6 = [[X_0, X_1], [X_2, [X_1, X_2]]] + [[X_0, X_2], [X_1, [X_2, X_1]]].$$

If we apply Theorem 7.3 with $\theta = 1$, we can conclude that our system is STLC from p if the brackets of total degree ≤ 6 span $T_p M$, provided that (i) $f_0(p) = 0$, (ii) $h(p)$ is a linear combination of values at p of brackets of degree < 3 , (iii) each vector $\text{Ev}_p(f)(B_i)$, $i = 1, \dots, 6$, is a linear combination of values at p of brackets of degree < 5 .

7.5. Two single-input examples. We now analyze from the point of view of Theorem 7.3 two examples where the Hermes condition fails to hold but the system is STLC from p .

In [22], G. Stefani discusses an example of a system $\dot{x} = f(x) + ug(x)$ which is STLC from 0 even though (a) the Hermes condition is not satisfied, (b) the first bracket B needed to span the whole tangent space is one where g occurs four times. (This shows that not all brackets that are even in g are obstructions to local controllability.) We will show that Stefani's example fits the framework of our Theorem 7.3, since B is also even in f , and therefore there is no need for it to be neutralized. Stefani's example is the system $\dot{x} = u$, $\dot{y} = x$, $\dot{z} = x^3 y$, in \mathbb{R}^3 , with control constraint $|u| \leq 1$. Then

$g = (1, 0, 0)$, $f = (0, x, x^3y)$. The relevant Lie brackets are as follows:

$$\begin{aligned} [g, f] &= (0, 1, 3x^2y), [f, [g, f]] = (0, 0, 2x^3), \\ [g, [g, f]] &= (0, 0, 6xy), [f, [f, [g, f]]] = 0, \\ [g, [f, [g, f]]] &= [f, [g, [g, f]]] = (0, 0, 6x^2), \\ [g, [g, [g, f]]] &= (0, 0, 6y), \\ [g, [g, [f, [g, f]]]] &= (0, 0, 12x), \\ [g, [g, [g, [f, [g, f]]]] &= (0, 0, 12). \end{aligned}$$

(We omit brackets that vanish or that are trivially expressed in terms of the ones listed here.) In particular, the vectors $g(0)$ and $[g, f](0)$ span a two-dimensional space S . If B is any bracket of f 's and g 's of degree ≤ 5 , then $B(0) \in S$. In particular, if we take $\theta = 1$ in Theorem 7.3, we see that every bracket of degree 3 or 5 is equal, when evaluated at 0, to a linear combination of brackets of lower degree. If we now add the bracket $[g, [g, [g, [f, [g, f]]]]$, which has degree 6, we span the whole space. Notice that, since this bracket is of even total degree, Theorem 7.3 does not require that it be neutralized. Hence Theorem 7.3 implies the fact—proved by Stefani—that this system is STLC from 0.

The preceding example shows that it is possible for controllability to be achieved thanks to the effect of some brackets that are even in g , so that not all such brackets are “obstructions.” We now briefly review another example, already discussed in [25], which shows that a bracket which is even in g may be “neutralized” by a bracket with more g 's but lower total degree. Consider the system $\dot{x} = u$, $\dot{y} = x$, $\dot{z} = x^3 + y^2$, $|u| < 1$. Here $f = (0, x, x^3 + y^2)$ and $g = (1, 0, 0)$. The vectors $g(0)$ and $[f, g](0)$ span a two-dimensional space S , and all the $(\text{ad } f)^k g(0)$, $k \geq 2$, are in S . The vector $[g, [f, g]](0)$ belongs to S . However, $[g, [f, [f, [g, f]]]](0)$ is not in S , so that the Hermes condition fails to hold. On the other hand, $[g, [g, [g, f]]](0)$ is not in S either, and therefore we can apply Theorem 7.3 with $\theta = 1$ and conclude that our system is STLC from 0.

7.6. Polynomial control systems. In [14], V. Jurdjevic studied control problems of the form

$$(7.40) \quad \dot{x} = P(x) + \sum_{i=1}^m u_i b_i, \quad u = (u_1, \dots, u_m) \in K,$$

where the state variable x takes values in \mathbb{R}^n , $P: \mathbb{R}^n \rightarrow \mathbb{R}^n$ is a polynomial map each of whose components is homogeneous of degree d , and K is either a cube centered at 0 (the “restricted controls” case) or the whole space \mathbb{R}^m (the “unrestricted case”).

Jurdjevic proved that, if d is odd, then (7.40) is STLC from 0 if and only if $S = \mathbb{R}^n$, where S is the smallest linear subspace of \mathbb{R}^n which is invariant under the map P and contains b_1, \dots, b_m . Moreover, in the unrestricted case it follows that every $x \in \mathbb{R}^n$ can be reached from 0 in time T , for every $T > 0$. We show that this result follows from our general theorem. Actually, we show that it follows from Brunovsky's theorem on odd systems. First we observe that, if d is odd, then (7.40) is an “odd system” in Brunovsky's sense. Therefore, the characterization of small-time local controllability from 0 will follow if we show that the Lie algebra L generated by the vector fields $x \rightarrow P(x)$, $x \rightarrow b_i$, $x \rightarrow -b_i$ satisfies $L(0) = S$.

It follows from the definition of S that, if $x \in S$, then all the vectors b_i belong to S , and so does $P(x)$. Therefore all the members of L are tangent to S , and so $L(0) \subseteq S$. On the other hand, if $Q: \mathbb{R}^n \rightarrow \mathbb{R}^n$ is a polynomial map and v is a vector in \mathbb{R}^n , then

the Lie bracket $[v, Q]$, of the constant vector field $x \rightarrow v$ and the vector field $x \rightarrow Q(x)$, is the vector field $x \rightarrow D_v Q(x)$, where $D_v Q(x)$ is the directional derivative of Q at x in the direction of v . In particular, this implies that $(\text{ad } v)^d(P)$ is the constant vector field $x \rightarrow P(v)$. Hence, if we let Σ be set of all vectors v such that the constant vector field $x \rightarrow v$ belongs to L , we see that Σ is invariant under the map P . Therefore $S \subseteq \Sigma$. This implies that $S \subseteq L(0)$, and so $L(0) = S$.

To complete the proof, we must show that, in the unrestricted case, the condition $S = \mathbb{R}^n$ implies that the time T reachable sets are equal to \mathbb{R}^n for all $T > 0$. Assume that $S = \mathbb{R}^n$. Let $T > 0$. Then we already know that there is a neighborhood U of 0 that can be reached in time T . Let $p \in \mathbb{R}^n$. Pick r such that $0 < r < 1$ and $rp \in U$. Let $t \rightarrow x(t)$ be a trajectory such that $x(0) = 0$, $x(r^{1-d}T) = rp$. (Since rp is reachable in time T , $r^{1-d} \geq 1$, and 0 is an equilibrium, it follows that rp is also reachable in time $r^{1-d}T$.) Let $y(t) = r^{-1}x(r^{1-d}t)$ for $0 \leq t \leq T$. Then $y(0) = 0$, $y(T) = p$. Let u_1, \dots, u_m be functions on $[0, r^{d-1}T]$ such that

$$\dot{x}(s) = P(x(s)) + \sum_{i=1}^m u_i(s)b_i \quad \text{for } 0 \leq s \leq r^{d-1}T.$$

Then

$$y(t) = r^d P(x(r^{1-d}t)) + \sum_{i=1}^m r^{-d} u_i(t) b_i,$$

i.e.

$$\dot{y}(t) = P(y(t)) + \sum_{i=1}^m r^{-d} u(t) b_i.$$

Since $t \rightarrow (r^{-d}u_1(t), \dots, r^{-d}u_m(t))$ is an admissible control, it follows that p is reachable from 0 in time T .

7.7. Low order conditions with a general polyhedral control set. Consider a finite sequence $\mathcal{V} = (V_1, \dots, V_m)$ of vector fields on a manifold M . We want conditions for \mathcal{V} to be STLC from a point p . Equivalently, we want to know when the system

$$(7.41) \quad \dot{x} = \sum_{i=1}^m w_i V_i(x), \quad w = (w_1, \dots, w_m) \in J$$

is STLC from p , where J is the set of vectors $(0, 0, \dots, 0, 1, 0, \dots, 0) \in \mathbb{R}^m$.

Let $S_0(\mathcal{V}, p)$ denote the convex hull of the vectors $V_1(p), \dots, V_m(p)$. Let $I_0(\mathcal{V}, p)$ denote the largest subset I of the index set $\{1, \dots, m\}$, such that 0 is a convex combination of the vectors $V_i(p)$, $i \in I$, with strictly positive coefficients. In [24], we proved that, if (7.41) is STLC from p , then $I_0(\mathcal{V}, p)$ has to be nonempty and, moreover, there have to exist indices $i \in I_0(\mathcal{V}, p)$ such that $V_i(p) \neq 0$. The main result of [24] was a sufficient condition for (7.41) to be STLC from p . Let $S_1(\mathcal{V}, p)$ denote the convex hull of the vectors $V_i(p)$, $i \in \{1, \dots, m\}$, $[V_i, V_j](p)$, $i, j \in I_0(\mathcal{V}, p)$. The result of [24] says that, if $S_1(\mathcal{V}, p)$ contains a neighborhood of the origin in the full tangent space $T_p M$, then (7.14) is STLC from p . We now show that this result, as well as some stronger conditions, can be derived from Corollary 7.2.

First, we observe that, instead of (7.41), we can consider the system

$$(7.42) \quad \dot{x} = \sum_{i=1}^m w_i V_i(x), \quad w = (w_1, \dots, w_m) \in K_m,$$

where K_m is the convex hull of J and the vector $(0, 0, \dots, 0)$. (Indeed, let \tilde{J} be the union of J and $\{(0, \dots, 0)\}$. It is clear that, if small-time local controllability holds with J replaced by \tilde{J} , then it holds for the system (7.41). On the other hand, Proposition 2.3 says that \tilde{J} can be replaced by its convex hull as well.)

Let us assume that the origin is an interior point of $S_1(\mathcal{V}, p)$. Also, let us relabel the indices so that $I_0(\mathcal{V}, p) = \{1, \dots, \mu\}$, where $2 \leq \mu \leq m$.

Then we can express 0 as a convex combination

$$(7.43) \quad 0 = \sum_{i=1}^{\mu} \lambda_i V_i(p), \quad \lambda_i > 0, \quad \sum_{i=1}^{\mu} \lambda_i = 1.$$

On the other hand, the hypothesis that 0 is an interior point of $S_1(\mathcal{V}, p)$ implies that: (i) *the vectors $V_1(p), \dots, V_m(p)$, together with the $[V_i, V_j](p)$, for $i = 1, \dots, \mu$; $j = 1, \dots, \mu$ span the tangent space $T_p M$* , (ii) *it is possible to express 0 as a convex combination*

$$(7.44) \quad 0 = \sum_{i=1}^m \alpha_i V_i(p) + \sum_{1 \leq i < j \leq \mu} \beta_{ij} [V_i, V_j](p)$$

where the α_i, β_{ij} are strictly positive, and $\sum \alpha_i + \sum \beta_{ij} = 1$.

(To see that (ii) follows, pick $\delta > 0$ so small that $-\delta Z(p) \in S_1(\mathcal{V}, p)$ whenever $Z = V_i$ for some i or $Z = [V_i, V_j]$ for some $i, j \in \{1, \dots, \mu\}$. Then each $-\delta V_i(p)$, $i \in \{1, \dots, m\}$, can be written as a convex combination of the $V_j(p)$, $j \in \{1, \dots, m\}$, and the $[V_j, V_k](p)$, $j, k \in \{1, \dots, \mu\}$. So 0 can be written as a linear combination of these same vectors in which all the coefficients are nonnegative and the coefficient of $V_i(p)$ is strictly positive. The same is true for $[V_i, V_j](p)$, if $i, j \in \{1, \dots, \mu\}$. If we then add all these expressions and divide by the sum of the coefficients, we obtain an expression of the desired form.)

Now define $f_0 = 0, f_i = \lambda_i V_i$ for $i = 1, \dots, \mu, f_i = \alpha_i V_i$ for $i = \mu + 1, \dots, m$. Consider the system

$$(7.45) \quad \dot{x} = f_0(x) + \sum_{i=1}^m u_i f_i(x), \quad u = (u_1, \dots, u_m) \in K_m.$$

It is clear that every trajectory of (7.45) is a trajectory of (7.42). Hence it suffices to prove that (7.45) is STLC from p .

To prove that (7.45) is STLC from p , we work with the free Lie algebra $L(X_0, \dots, X_m)$. We let Λ_0 be the group of all automorphisms g_π of $L(X_0, \dots, X_m)$ that are induced by a permutation π of the indices $\{0, \dots, m\}$ that satisfies $\pi(0) = 0, \pi(\{1, \dots, \mu\}) = \{1, \dots, \mu\}$. It is clear that all the g_π are input symmetries. We define dilations $\Delta(p)$ by assigning Δ -degree one to X_0, \dots, X_μ , and Δ -degree δ to $X_{\mu+1}, \dots, X_m$, where δ is some number such that $2 < \delta < 3$.

We now show that all the totally odd Λ_0 -fixed elements of $L(X_0, \dots, X_m)$ are Δ -neutralized for \mathbf{f} at p . Since $f_0(p), \dots, f_m(p)$, and the $[f_i, f_j](p)$ with $i, j \in \{1, \dots, \mu\}$ span $T_p M$, it is clear that $T_p M$ is spanned by the evaluations at \mathbf{f}, p of elements of $L(X_0, \dots, X_m)$ of Δ -degree not greater than δ . Among these, the Λ_0 -fixed elements of Δ -degree one are spanned by X_0 and $X_1 + \dots + X_\mu$. But

$$(7.46) \quad f_0(p) = (f_1 + \dots + f_\mu)(p) = 0,$$

and so these elements are neutralized. The elements of Δ -degree 2 are totally odd and therefore need not be considered. The Λ_0 -fixed elements of Δ -degree δ are spanned by $X_{\mu+1} + \dots + X_m$, and (7.44) shows that $f_{\mu+1}(p) + \dots + f_m(p)$ is equal to the value at p of an element of Δ -degree $< \delta$. Hence Corollary 7.2 can be applied. This completes the proof that the result of [24] is a particular case of Corollary 7.2.

It should be clear from the preceding proof that one can get more sophisticated results by just applying the same method. A detailed analysis of what can be so obtained will be the subject of a future paper. At the moment, we limit ourselves to two examples. In these examples, if $\lambda_1, \dots, \lambda_\mu$ are such that (7.43) holds, we let g_1, g_2 denote the vector fields

$$(7.47) \quad g_1 = \lambda_1 V_1 + \dots + \lambda_\mu V_\mu,$$

$$(7.48) \quad g_2 = \sum_{i=1}^{\mu} \lambda_i^2 (\text{ad } V_i)^2(g_1).$$

(That is, $g_1 = f_1 + \dots + f_\mu$, $g_2 = \sum_{i=1}^{\mu} [f_i, [f_i, g_1]]$.)

Then, if (7.43) holds, the system (7.41) is STLC from p if one of the following conditions holds:

- (I) (a) $g_2(p)$ is a linear combination of the vectors $V_i(p)$, $i \in \{1, \dots, \mu\}$ and the $[V_i, V_j](p)$, $i, j \in \{1, \dots, \mu\}$,
- (b) 0 is a convex combination, with strictly positive coefficients, of the $V_i(p)$, $i \in \{1, \dots, m\}$, the $[V_i, V_j](p)$, $i, j \in \{1, \dots, \mu\}$, and the $[V_i, [V_j, V_k]](p)$, $i, j, k \in \{1, \dots, \mu\}$,
- (c) $T_p M$ is spanned by the $V_i(p)$, $i \in \{1, \dots, m\}$, the $[V_i, V_j](p)$, $i \in \{1, \dots, m\}$, $j \in \{1, \dots, \mu\}$, the $[V_i, [V_j, V_k]](p)$, $i, j, k \in \{1, \dots, \mu\}$, and the $[V_i, [V_j, [V_k, V_l]]](p)$, $i, j, k, l \in \{1, \dots, \mu\}$;
- (II) (a) 0 is a convex combination with strictly positive coefficients of the $V_i(p)$, $i \in \{1, \dots, m\}$ and the $[V_i, V_j](p)$, $i, j \in \{1, \dots, \mu\}$,
- (b) $g_2(p)$ is a linear combination of the $V_i(p)$, $i \in \{1, \dots, m\}$ and the $[V_i, V_j](p)$, $i, j \in \{1, \dots, \mu\}$,
- (c) $T_p M$ is spanned by the $V_i(p)$, $i \in \{1, \dots, m\}$, the $[V_i, V_j](p)$, $i, j \in \{1, \dots, m\}$, the $[V_i, [V_j, V_k]](p)$, where i, j, k are in $\{1, \dots, \mu\}$, and the $[V_i, [V_j, [V_k, V_l]]](p)$, $i, j, k, l \in \{1, \dots, \mu\}$.

To see that (I) implies small-time local controllability from p , we reason as before, but with $3 < \delta < 4$. Condition (c) of (I) says that $T_p M$ is spanned by the brackets of Δ -degree not greater than $1 + \delta$. The Λ_0 -fixed elements of Δ -degree one are spanned by X_0 and $X_1 + \dots + X_\mu$, which are obviously neutralized, since $f_0 \equiv 0$ and (7.43) holds. The Λ_0 -fixed elements of Δ -degree 2 do not matter, because they are totally even. In Δ -degree 3 there is only one Λ_0 -fixed element, namely G_2 , where we let $G_1 = X_1 + \dots + X_\mu$, $G_2 = \sum_{i=1}^{\mu} [X_i, [X_i, G_1]]$. Condition (a) then says that this element is neutralized. In Δ -degree δ there is one Λ_0 -fixed element, namely, $X_{\mu+1} + \dots + X_m$. Condition (b) then says that this element is neutralized. Finally, all the elements of Δ -degree 4 or $1 + \delta$ are totally even, and therefore need not be considered.

To see the sufficiency of (II) we again use the same argument, with δ such that $2 < \delta < 2.5$. Then condition (c) says that $T_p M$ is spanned by evaluations of brackets of Δ -degree not greater than 2δ . The possible Δ -degrees of such brackets are 1, 2, δ , 3, $1 + \delta$, 4, $2 + \delta$ and 2δ . The only Δ -degrees where totally odd brackets occur are 1, δ , 3 and $2 + \delta$. In Δ -degree 1, the Λ_0 -fixed elements are X_0 and $X_1 + \dots + X_\mu$, which are neutralized. In degree δ , the Λ_0 -fixed element is $X_{\mu+1} + \dots + X_m$, which is neutralized by condition (a). In Δ -degree 3, the Λ_0 -fixed element is G_2 , which is neutralized by condition (b). So Corollary 7.2 applies, and small-time local controllability follows.

8. Conclusion. The main implication of the results proven here is that, so far, one method appears to suffice to prove most known small-time local controllability results. It seems to us that this method is still very special, and it should be possible to obtain better results by making a more detailed analysis of the semigroups $S^N(X, K)$.

One important application of the theory developed in this paper is to the problem of High Order Optimality Conditions. Small-time local controllability is a particular instance of this general problem, in which we are concerned with finding sufficient conditions for a particular trajectory (given by $x(t) \equiv p = \text{constant}$) to lie in the interior of the attainable set from p , which is the same as finding necessary conditions for the trajectory to lie on the boundary of the reachable set. The methods of this paper can be used to prove results on the construction of control variations for more general trajectories, and to obtain necessary conditions for optimality. The results will be reported in subsequent papers.

Appendix.

Proof of Proposition 2.3. Clearly, all that needs to be shown is that, if $\tilde{\Sigma}$ is STLC from p , it follows that Σ is STLC_{pc} from p . Suppose that $\tilde{\Sigma}$ is STLC from p . Let $T > 0$. Pick T' such that $0 < T' < T$, and let U be an open set such that $p \in U$ and $U \subseteq \text{Reach}(\tilde{\Sigma}, \leq T', p)$. Shrink U , if necessary, so that $L(f)(q)$ is the full tangent space at q for every $q \in U$. Let $q \in U$. Let \mathcal{F}_{Σ} be the family of vector fields associated with Σ , and let $-\mathcal{F}_{\Sigma} = \{-V : V \in \mathcal{F}_{\Sigma}\}$. Then $L(f) = L(\mathcal{F}_{\Sigma}) = L(-\mathcal{F}_{\Sigma})$. Let \mathcal{H} be the family of restrictions to U of the members of $-\mathcal{F}_{\Sigma}$. Then \mathcal{H} has the AP from q . So there is a nonempty open subset W of U such that every $r \in W$ is reachable from q by an \mathcal{H} -trajectory in time not greater than $T - T'$, so that q is reachable from every $r \in W$ by an \mathcal{F}_{Σ} -trajectory in time not greater than $T - T'$.

Now pick an $r \in W$. Since $W \subseteq U$, r is reachable from p in time τ , for some $\tau \in [0, T']$ by means of a trajectory of $\tilde{\Sigma}$ that corresponds to a control $u(\cdot) : [0, \tau] \rightarrow \tilde{K}$. Then $u(\cdot)$ can be approximated in $L^1([0, \tau], \mathbb{R}^m)$ by a sequence $\{u_n(\cdot)\}$ of piecewise constant K -valued controls. If $x_n(\cdot)$ is the trajectory for $u_n(\cdot)$ such that $x_n(0) = p$, and we let $r_n = x_n(\tau)$, then $r_n \in W$ for sufficiently large n . Therefore W contains a point r' which is reachable from p in time τ by means of a piecewise constant \tilde{K} -valued control. Let $\text{co}(K)$ denote the convex hull of K . Since \tilde{K} is the closure of $\text{co}(K)$, every piecewise constant \tilde{K} -valued control can be approximated in $L^1([0, \tau], \mathbb{R}^m)$ by piecewise constant $\text{co}(K)$ -valued controls. Therefore W must contain a point r'' which is reachable from p in time τ by means of a piecewise constant $\text{co}(K)$ -valued control. Finally, a piecewise constant $\text{co}(K)$ -valued control can be approximated weakly by piecewise constant K -valued controls. Hence W contains a point r''' which is reachable from p in time τ by a piecewise constant K -valued control. So $r''' \in \text{Reach}_{pc}(\Sigma, \leq T', p)$. Since $q \in \text{Reach}_{pc}(\Sigma, \leq (T - T'), r''')$, we conclude that $q \in \text{Reach}_{pc}(\Sigma, \leq T, p)$. Since q was an arbitrary point of U , we see that $U \subseteq \text{Reach}_{pc}(\Sigma, \leq T, p)$. Since U is open, $p \in U$, and T was an arbitrary positive number, it follows that Σ is STLC_{pc} from p .

Acknowledgments. We thank an anonymous referee for very helpful suggestions and remarks. Also, we are especially grateful to Klaus Wagner who read the paper with great care and pinpointed a large number of misprints as well as some mathematical inaccuracies.

REFERENCES

- [1] A. AGRACHEV AND R. GAMKRELIDZE, *The exponential representation of flows and the chronological calculus*, Math. USSR Sbornik, 35 (1979), pp. 727-785.
- [2] R. BROCKETT, *Volterra series and geometric control theory*, Automatica, 12 (1976), pp. 167-176.
- [3] P. BRUNOVSKY, *Local controllability of odd systems*, Banach Center Publications, Warsaw, Poland, 1 (1974), pp. 39-45.
- [4] K. J. CHEN, *Integration of paths, geometric invariants and a generalized Baker-Hausdorff formula*, Ann. Math., 65 (1957), pp. 163-178.

- [5] P. CROUCH AND C. BYRNES, *Symmetries and local controllability*, to appear.
- [6] M. FLIESS, *Fonctionnelles causales non linéaires et indéterminées non commutatives*, Bull. Soc. Math. France, 109 (1981), pp. 3–40.
- [7] ———, *Développements fonctionnels en indéterminées non commutatives des solutions d'équations différentielles non linéaires forcées*, C. R. Acad. Sci. Paris, Ser. A, 287 (1978), pp. 1133–1135.
- [8] R. GROSSMANN, Unpublished doctoral dissertation, Princeton Univ., Princeton, NJ, 1985.
- [9] R. HERMANN, *On the accessibility problem in control theory*, in International Symposium on Nonlinear Differential Equations and Nonlinear Mechanics, Academic Press, New York, 1963, pp. 325–332.
- [10] H. HERMES, *Lie algebras of vector fields and local approximation of attainable sets*, this Journal, 16 (1978), pp. 715–727.
- [11] ———, *On local controllability*, this Journal, 20 (1982), pp. 211–220.
- [12] ———, *Local controllability and sufficient conditions in singular problems*, J. Differential Equations, 20 (1976), pp. 213–232.
- [13] ———, *Control systems which generate decomposable Lie algebras*, J. Differential Equations, 44 (1982), pp. 166–187.
- [14] V. JURDJEVIC, *Polynomial control systems*, Proc. 22nd IEEE CDC, Vol. 2, 1983, pp. 904–906.
- [15] A. KRENER, *Local approximation of control systems*, J. Differential Equations, 19 (1975), pp. 125–133.
- [16] ———, *A generalization of Chow's theorem and the bang-bang theorem to nonlinear control systems*, this Journal, 12 (1974), pp. 43–52.
- [17] E. B. LEE AND L. MARKUS, *Foundations of Optimal Control Theory*, John Wiley, New York, 1976.
- [18] C. LESIAK AND A. KRENER, *The existence and uniqueness of Volterra series for nonlinear systems*, IEEE Trans. Automat. Control, 23 (1978), pp. 1090–1095.
- [19] C. LOBRY, *Controlabilité des systèmes non linéaires*, this Journal, 8 (1970), pp. 573–605.
- [20] J. P. SERRE, *Lie Algebras and Lie Groups*, W. A. Benjamin, New York, 1965.
- [21] G. STEFANI, *On local controllability and related topics*, preprint, Facoltà di Ingegneria, Università di Firenze.
- [22] ———, *Local properties of nonlinear control systems*, preprint, Facoltà di Ingegneria, Università di Firenze.
- [23] H. J. SUSSMANN AND V. JURDJEVIC, *Controllability of nonlinear systems*, J. Differential Equations, 12 (1972), pp. 95–116.
- [24] H. J. SUSSMANN, *A sufficient condition for local controllability*, this Journal, 16 (1978), pp. 790–802.
- [25] ———, *Lie brackets and local controllability: a sufficient condition for scalar-input systems*, this Journal, 21 (1983), pp. 686–713.
- [26] ———, *Lie brackets, real analyticity and geometric control*, in Differential Geometric Control Theory, R. W. Brockett, R. S. Millman and H. J. Sussmann, eds., Birkhauser, Boston, 1983, pp. 1–116.
- [27] G. VIENNOT, *Algèbres de Lie Libres et Monoides Libres*, Lecture Notes in Mathematics 691, Springer-Verlag, New York, 1978.

MINIMIZING OR MAXIMIZING THE EXPECTED TIME TO REACH ZERO*

D. HEATH†, S. OREY‡, V. PESTIEN§ AND W. SUDDERTH¶

Abstract. We treat the following control problems: the process $X_1(t)$ with values in the interval $(-\infty, 0]$ (or $[0, \infty)$) is given by the stochastic differential equation

$$dX_1(t) = \mu(t) dt + \sigma(t) dW_t, \quad X_1(0) = x_1$$

where the nonanticipative controls μ and σ are to be chosen so that $(\mu(t), \sigma(t))$ remains in a given set \mathcal{S} and the object is to minimize (or maximize) the expected time to reach the origin. The minimization problem had been discussed earlier by Heath, Pestien and Sudderth under various restrictions on the set \mathcal{S} . Here an improved verification lemma is established which is used to solve the minimization and maximization problems for any \mathcal{S} . An application to a portfolio problem is discussed.

Key words. stochastic control, portfolio selection, gambling theory

AMS(MOS) subject classifications. 60G40, 60J60, 93E20

1. Introduction. Consider a real-valued process $\{X_1(t)\}$ given by a stochastic differential equation

$$dX_1(t) = \mu(t) dt + \sigma(t) dW_t, \quad X_1(0) = x_1$$

where $\{W_t\}$ is standard Brownian motion and $\mu(t)$ and $\sigma(t)$ are nonanticipative controls to be chosen so that $(\mu(t), \sigma(t))$ remains in a specified set \mathcal{S} . The problems of minimizing or maximizing the expected time to reach the origin are treated in § 3. The minimization problem has been studied in [9] and [3], though with an exponential change of variables putting the problem on $(0, 1]$. For a more detailed discussion, see Remark 2 in § 3.

The solution of these control problems uses a new refinement of the verification lemma of [9], which is proved in § 2. This result should be of independent interest.

Section 4 deals with a portfolio planning problem which turns out to be a special case of the minimization problem. This portfolio problem was originally solved in [3].

2. Continuous-time stochastic control. The formulation of stochastic control problems given here is adapted from Pestien and Sudderth [9]. Our notation and terminology is the same as theirs, but we consider a more general class of processes and establish a verification lemma more suited to the present applications.

A *continuous-time gambling problem* is a triple (F, Σ, u) where

- (2.1) the *state space* F is Polish (we shall use a Borel subset of ordinary Euclidean space),
- (2.2) the *gambling house* Σ is a mapping which assigns to each $x \in F$ a nonempty collection $\Sigma(x)$ of processes $X = \{X_t, t \geq 0\}$ with state space F such that $X_0 = x$ and X has right-continuous paths with left-limits,
- (2.3) the *utility function* u is a Borel function from F to the real line.

* Received by the editors March 18, 1985, and in revised form October 15, 1985.

† School of Operations Research and Industrial Engineering, Cornell University, Ithaca, New York 14853.

‡ School of Mathematics, University of Minnesota, Minneapolis, Minnesota 55455. The research of this author was supported by National Science Foundation grant MCS 83-01080.

§ Department of Mathematics and Computer Science, University of Miami, Coral Gables, Florida 33124.

¶ School of Statistics, University of Minnesota, Minneapolis, Minnesota 55455. The research of this author was supported by National Science Foundation grant DMS-8421208.

A process $X \in \Sigma(x)$ is said to be *available* at x . Each available X is defined on some probability space (Ω, \mathcal{F}, P) and is adapted to an increasing filtration $(\mathcal{F}_t, t \geq 0)$ of complete sub-sigma fields of \mathcal{F} . The probability space and filtration may depend on X .

A player, starting at position $x \in F$, selects a process $X \in \Sigma(x)$ and receives payoff $u(X)$ defined by

$$(2.4) \quad u(X) = E[\limsup_{t \rightarrow \infty} u(X_t)].$$

The expectation occurring on the right is assumed to be well-defined for every available process X .

The *value function* V is defined by

$$V(x) = \sup \{u(X) : X \in \Sigma(x)\}$$

for every $x \in F$. A process $X \in \Sigma(x)$ is *optimal* at x if

$$u(X) = V(x).$$

From now on we shall require that F be a Borel subset of the Euclidean space \mathbb{R}^d having nonempty interior, and each process $X = \{X_t\}$ under consideration will be an *Ito process* of the form

$$(2.5) \quad X_t = x + \int_0^t \alpha(s) ds + \int_0^t \beta(s) dW_s$$

where $W = \{W_t\}$ is a standard m -dimensional Brownian motion process on (Ω, \mathcal{F}, P) adapted to increasing, right-continuous σ -fields $\{\mathcal{F}_t\}$, and \mathcal{F}_t is independent of $\{W_{t+s} - W_t, s \geq 0\}$. The function $\alpha = \alpha(t, \omega)$ is to be \mathbb{R}^d -valued, progressively measurable, adapted to $\{\mathcal{F}_t\}$ and such that

$$(2.6) \quad \int_0^t |\alpha(s)| ds < \infty \quad \text{a.s. for all } t.$$

The function $\beta = \beta(t, \omega)$ has as values real $d \times m$ matrices, is progressively measurable, adapted to $\{\mathcal{F}_t\}$, and satisfies

$$(2.7) \quad \int_0^t |\beta(s)|^2 ds < \infty \quad \text{a.s. for all } t.$$

For each pair (a, b) , where $a \in \mathbb{R}^d$ is a $d \times 1$ vector and b is a $d \times m$ real-valued matrix, define the differential operator $D(a, b)$ for sufficiently smooth functions $Q : \mathbb{R}^d \rightarrow \mathbb{R}$ by

$$D(a, b)Q(y) = Q_x(y)a + \frac{1}{2} \sum_{i=1}^d \sum_{j=1}^m Q_{x_i x_j}(y)(bb')_{ij}$$

where

$$Q_x(y) = \left(\frac{\partial Q}{\partial x_1}, \dots, \frac{\partial Q}{\partial x_d} \right), \quad Q_{x_i x_j} = \frac{\partial^2 Q}{\partial x_i \partial x_j},$$

and b' is the transpose of b .

We now specify $\Sigma(x)$ by specifying the possible values of α and β . To this end, for each $x \in F$ let $C(x)$ be a nonempty set of pairs (a, b) , where $a \in \mathbb{R}^d$ and b is a real $d \times m$ matrix. (The idea is that $C(x)$ is the set from which a player at state x may choose the value of (α, β) .) Assume also that every available process X is absorbed at the time T_X of its first exit from F° , the interior of F . These conditions define a

function Σ_C on F where $\Sigma_C(x)$ is the collection of all processes X having paths in F and satisfying (2.5), (2.6) and (2.7) together with

$$(2.8) \quad (\alpha(t, \omega), \beta(t, \omega)) \in C(X_t(\omega)) \quad \text{for all } (t, \omega),$$

$$(2.9) \quad (\alpha(t, \omega), \beta(t, \omega)) = (0, 0) \quad \text{for } t \geq T_X(\omega),$$

$$(2.10) \quad C(x) = \{(0, 0)\} \quad \text{for } x \in F - F^\circ.$$

Let Σ be a gambling house such that $\Sigma(x) \subset \Sigma_C(x)$ for every $x \in F$.

The following proposition, which is related to Lemmas 2 and 3 of [9], will be applied in the next two sections.

PROPOSITION. *Let G be an open subset of \mathbb{R}^d which contains F . Suppose $Q: G \rightarrow \mathbb{R}$ and $Q_n: G \rightarrow \mathbb{R}$ for $n = 1, 2, \dots$. Suppose also that each Q_n has continuous second-order derivatives on G and that*

$$(i) \quad \lim_{n \rightarrow \infty} Q_n(x) = Q(x) \quad \text{for every } x \in F.$$

Assume the following conditions for every $x \in F^\circ$ and every $X \in \Sigma(x)$:

(ii) $Q(X) \geq u(X)$ where $Q(X) = E[\limsup_{t \rightarrow \infty} Q(X_t)]$ is assumed to be well-defined.

(iii) There exists a sequence $\{k_n\}$ of nonnegative constants such that $\lim_{n \rightarrow \infty} k_n = 0$ and with probability one, for all n and all $t \geq 0$,

$$D(\alpha(t), \beta(t))Q_n(X_t) \leq k_n.$$

(Here α and β are related to X by (2.5).)

(iv) There exist integrable random variables Z, Y_1, Y_2, \dots such that, for all n and all $t \geq 0$,

$$Z \leq Q_n(X_t) \leq Y_n.$$

Then $Q \geq V$.

The following lemma is the chief tool for the proof of the proposition.

LEMMA. *Suppose $Q: G \rightarrow \mathbb{R}$ has continuous second-order derivatives, $x_0 \in F^\circ$, $X \in \Sigma(x_0)$, and τ is an almost surely finite $\{\mathcal{F}_t\}$ -stopping time. Also assume*

(i) there is a nonnegative constant k such that with probability one, for all $s \geq 0$,

$$D(\alpha(s), \beta(s))Q(X_s) \leq k,$$

(ii) there exist integrable random variables Y and Z such that for all $t \geq 0$,

$$Z \leq Q(X_t) \leq Y.$$

Then

$$EQ(X_\tau) \leq Q(x_0) + kE\tau.$$

Proof. Apply Ito's lemma to write

$$(2.11) \quad Q(X_t) = Q(x_0) - A_t + M_t = Q(x_0) - (kt + A_t) + kt + M_t$$

where

$$A_t = - \int_0^t D(\alpha(s), \beta(s))Q(X_s) ds,$$

$$M_t = \int_0^t Q_x(X_s)\beta(s) dW_s.$$

(Here α and β are related to X by (2.5) and satisfy (2.6) and (2.7).)

Assume without loss of generality that $E\tau < \infty$. Hence,

$$(2.12) \quad EQ(X_\tau) = Q(x_0) + kE\tau + E[M_\tau - (k\tau + A_\tau)].$$

It suffices to show that the final expectation in (2.12) is less than or equal to zero. By condition (i), $-(k\tau + A_\tau) \leq 0$. We will show that $EM_\tau \leq 0$. (Notice that EM_τ is well-defined by the first equality in (2.11) and condition (ii).)

Let T_j be a sequence of stopping times such that $\{M_{t \wedge T_j}, \mathcal{F}_t\}$ is a uniformly integrable martingale for every j and $T_j \rightarrow \infty$ a.s. Let $B_j = [\tau > T_j]$. Then $M_\tau = M_{\tau \wedge T_j}$ on B_j^c , and

$$\begin{aligned} \int_{B_j^c} M_\tau &= \int_{B_j^c} M_{\tau \wedge T_j} = EM_{\tau \wedge T_j} - \int_{B_j} M_{T_j} = 0 - \int_{B_j} [Q(X_{T_j}) + A_{T_j} - Q(x_0)] \\ &\leq \int_{B_j} [-Z + k\tau + Q(x_0)]. \end{aligned}$$

That is,

$$\int_{B_j^c} M_\tau^+ - \int_{B_j^c} M_\tau^- \leq \int_{B_j} [-Z + k\tau + Q(x_0)].$$

Let $j \rightarrow \infty$, and conclude

$$EM_\tau = \int_{\Omega} M_\tau^+ - \int_{\Omega} M_\tau^- \leq \int_{\emptyset} [-Z + k\tau + Q(x_0)] = 0. \quad \square$$

Proof of the proposition. Let $x_0 \in F$ and $X \in \Sigma(x_0)$. By condition (ii) and [9, Lemma 1], it suffices to show

$$(2.13) \quad EQ(X_\tau) \leq Q(x_0)$$

for every almost surely finite stopping time τ .

Assume first that τ is bounded. Then, by the lemma and Fatou's inequality,

$$EQ(X_\tau) \leq \liminf_{n \rightarrow \infty} EQ_n(X_\tau) \leq \lim_{n \rightarrow \infty} Q_n(x_0) + (\lim_{n \rightarrow \infty} k_n)E\tau = Q(x_0).$$

If τ is unbounded, use Fatou's inequality again:

$$EQ(X_\tau) \leq \liminf_{n \rightarrow \infty} EQ(X_{\tau \wedge n}) \leq Q(x_0). \quad \square$$

3. Minimizing or maximizing the expected time to reach zero. The problems described in the Introduction will now be formulated as continuous-time gambling problems in \mathbb{R}^2 . Consider first the problem of minimizing expected time. The first coordinate, x_1 , of the state vector x will correspond to the player's position on $(-\infty, 0]$, while the second coordinate, x_2 , will represent time.

It is convenient to allow negative as well as positive times and define

$$F = \{x \in \mathbb{R}^2: -\infty < x_1 \leq 0\}.$$

Because the object is to minimize expected time, let

$$u(x) = -x_2.$$

Recall the notation from § 2. The interior of F is $F^\circ = (-\infty, 0) \times (-\infty, \infty)$ and by our conventions each available process X will be absorbed at time

$$T = T_X = \inf\{t: X_1(t) = 0\}.$$

In the present example the set $C(x)$ will not depend on x for $x \in F^\circ$. Let $\mathcal{S} \subset \mathbb{R} \times [0, \infty)$,

$$(3.0) \quad C_0 = \left\{ \left(\begin{pmatrix} \mu \\ 1 \end{pmatrix}, \begin{pmatrix} \sigma \\ 0 \end{pmatrix} \right) : (\mu, \sigma) \in \mathcal{S} \right\}$$

and let $C(x) = C_0$ for $x \in F^\circ$. Every $X \in \Sigma_C(x)$ can be specified by stochastic differential equations

$$\begin{aligned} dX_1(t) &= \mu(t) dt + \sigma(t) dW_t, \\ dX_2(t) &= dt, \\ X_1(0) &= x_1, X_2(0) = x_2, \end{aligned} \quad (3.1)$$

where μ and σ are progressively measurable and $(\mu(t), \sigma(t)) \in \mathcal{S}$, $t < T$; and $X_t = X_T$ for $t \geq T$. Note that for every $X \in \Sigma_C(x)$ the second coordinate process $\{X_2(t)\}$ increases deterministically at rate 1 up to time T , and by (2.4) and the definition of u

$$u(X) = -x_2 - ET. \quad (3.2)$$

Now let

$$\begin{aligned} \Sigma(x) &= \{X \in \Sigma_C(x) : u(X) > -\infty\} \\ &= \{X \in \Sigma_C(x) : ET < \infty\}. \end{aligned} \quad (3.3)$$

From (3.2) and (3.3) one sees that

$$V(x_1, x_2) = V(x_1) - x_2 \quad (3.4)$$

where $V(x_1) = V(x_1, 0)$. Furthermore, for $x_1 < y_1 < 0$, a strategy starting at x_1 and minimizing the time to 0 must first minimize the time to y_1 and, having gotten there, minimize the time to 0. This argument leads to $V(x_1) = V(x_1 - y_1) + V(y_1)$. Since V is also continuous and vanishes at the origin, one may conclude

$$V(x_1) = \lambda x_1$$

where $\lambda \geq 0$ depends on \mathcal{S} . (We omit a formal proof because we will not rely on this formula below.)

If in (3.1) $\mu(t) = \mu(X_1(t))$ and $\sigma(t) = \sigma(X_1(t))$, where μ and σ are measurable real-valued functions on $(-\infty, 0)$, we say that X is given by a *stationary Markovian strategy*. For given functions μ and σ then X as defined by (3.1) depends only on the initial conditions, so we may write $u(X) = v(x_1, x_2)$ and from (3.2)

$$v(x_1, x_2) = v(x_1) - x_2 \quad (3.5)$$

where $v(x_1) = v(x_1, 0)$. Now $u(X)$ can be obtained explicitly. Assume for simplicity that μ and σ are piecewise continuous functions, and $\sigma(x_1) \geq \sigma_0 > 0$ for all x_1 . Note that if $\sigma(x_1) \geq \sigma_0 > 0$ and σ and μ are measurable and bounded on compact sets then (3.1) will have a unique weak solution: for μ vanishing this is shown by a time-change argument ([4, Chap. IV, Ex. 4.2]), and for μ not vanishing one uses transformation of drift ([4, Chap. IV, Thm. 4.2]).

By definition $v(x_1)$ is simply the negative of the expected time it takes the diffusion to reach the origin if it is started at x_1 . If $X \in \Sigma(x)$, then T is finite with probability one and $v(x_1)$ is the limit as $M \rightarrow \infty$ of $-v_M(x_1)$, where $v_M(x_1)$ is the expected time to exit the interval $[-M, 0]$. Let us set

$$a(x) = \sigma^2(x).$$

Then v_M is determined by

$$\mu v'_M + \frac{1}{2} a v''_M + 1 = 0, \quad v_M(0) = v_M(-M) = 0.$$

Solving for v'_M and letting $M \rightarrow \infty$ gives

$$v'(x_1) = e^{-B(x_1)} \int_{-\infty}^{x_1} e^{B(z)} \frac{2}{a(z)} dz \quad (3.6)$$

where

$$(3.7) \quad B(x_1) = \int_r^{x_1} \frac{2\mu(y)}{a(y)} dy$$

and r is an arbitrary point in $(-\infty, 0]$. Of course

$$(3.8) \quad v(x_1) = \int_0^{x_1} v'(y) dy.$$

Recall (see, for example, [5]) that the diffusion determined by μ and a has a scale function and speed measure determined respectively by

$$(3.9) \quad dp(x_1) = e^{-B(x_1)} dx_1, \quad dm(x_1) = \frac{2}{a(x_1)} e^{B(x_1)} dx_1.$$

The formulas (3.9) are correct under the same conditions mentioned above for the existence of a unique weak solution. Note first that if p is as in (3.9), $p(X_1(t))$ is a martingale, as can be seen by applying the Ito formula as extended by Krylov [7]. To check the speed measure first consider the case where μ vanishes. Then as remarked above the diffusion can be obtained by a time change of a Brownian motion \tilde{W} , the time change being the inverse of $\phi_t = \int_0^t \sigma^{-2}(\tilde{W}(s)) ds$. Comparing this with the Ito-McKean construction of diffusion we find that the speed measure satisfies $m(dy) = 2\sigma^{-2}(y) dy$, see [2, Thm. 2.123]. When there is drift present consider the drift-free diffusion $Y_t = p(X_1(t))$. From the Ito formula we can read off the diffusion coefficient of Y , hence also the speed measure \tilde{m} for Y and since the speed measure for X_1 is given by $m(dx) = \tilde{m}(p(dx))$ this is now determined and indeed given as in (3.9).

Consider now $\mu(t) \equiv \mu_0$, $\sigma(t) \equiv \sigma_0$, where μ_0 and σ_0 are constants. This will determine a diffusion with $ET < \infty$ if and only if $\mu_0 > 0$, and then

$$(3.10) \quad v(x_1) = \frac{x_1}{\mu_0}$$

which is a special case of (3.6) if $\sigma_0 > 0$ and obvious if $\sigma_0 = 0$.

It is natural, especially in the light of (3.10), to conjecture that an optimal strategy is to choose the drift μ to achieve the supremum

$$M = \sup \{ \mu : (\mu, \sigma) \in \mathcal{S} \text{ for some } \sigma \}.$$

As is explained in Remark 2 below, a similar strategy was proposed by Kelly [6] for certain discrete-time problems. However, these "Kelly strategies" need not be optimal if the set of the possible σ 's is unbounded. The exact criterion for our continuous-time problem involves another quantity

$$I = \inf_{\varepsilon > 0} \sup \{ \mu + \varepsilon \sigma^2 : (\mu, \sigma) \in \mathcal{S} \}.$$

THEOREM 1. *Let $x \in F^\circ$.*

(a) *If $0 < M < \infty$ and $I < \infty$ then $V(x) = x_1/M - x_2$. If in addition $(M, \sigma_0) \in \mathcal{S}$, then the process $X \in \Sigma(x)$ with $\mu(t) \equiv M$ and $\sigma(t) \equiv \sigma_0$ is optimal.*

(b) *If $M \leq 0$ and $I < \infty$ then $V(x) = -\infty$.*

(c) *If $M = \infty$ or $I = \infty$ then $V(x) = -x_2$ (i.e., the origin can be reached in an arbitrarily small expected time.)*

Proof. (a) Let $Q(x) = x_1/M - x_2$. It is clear from (3.5) and (3.10) that $Q \leq V$. It remains to verify that $Q \geq V$. (Once this is done, the final assertion of (a) will follow from (3.5) and (3.10).) This inequality will be proved by applying the proposition of § 2.

For $\delta > 0$, let

$$I(\delta) = \sup \left\{ \mu + \frac{\delta}{2} \sigma^2 : (\mu, \sigma) \in \mathcal{S} \right\},$$

and notice that by condition (a), $I(\delta) < \infty$ for δ sufficiently small. For such δ , let

$$Q^\delta(x) = \frac{e^{x_1 \delta} - 1}{\delta I(\delta)} - x_2.$$

Now verify the conditions of the proposition, with $Q = Q^\delta$ and $Q_n = Q^\delta$ for each n . Condition (i) is automatic and (ii) follows easily. With $k_n = 0$ for each n , it is routine to check that (iii) holds. As to condition (iv), observe that in the formula for $Q^\delta(x)$ the first term on the right is bounded uniformly in x_1 for each fixed δ . So $Q^\delta(X_t)$ is bounded above and below by a constant plus $X_2(t)$, and since $x_2 \leq X_2(t) \leq x_2 + T$ and $X \in \Sigma(x)$ implies that T is integrable, the proposition gives $Q^\delta \geq V$. Finally, because $I(\delta) \rightarrow M$ and $Q^\delta \rightarrow Q$ as $\delta \rightarrow 0$, we have $Q \geq V$.

(b) We reduce the result to (a). Let $\varepsilon > 0$ and consider a new problem based on the set

$$\mathcal{S}_\varepsilon = \mathcal{S} \cup \{(\varepsilon, 0)\}.$$

The quantity corresponding to M for the new problem is $M_\varepsilon = \varepsilon$. Thus part (a) can be applied to obtain the value function

$$V_\varepsilon(x) = \frac{x_1}{\varepsilon} - x_2.$$

Clearly $V(x) \leq V_\varepsilon(x) \rightarrow -\infty$ as $\varepsilon \rightarrow 0$.

(c) If $M = \infty$ the desired conclusion $V(x) = -x_2$ follows easily from (3.10). So assume now that $M < \infty$ and $I = \infty$. Then there exists a sequence (μ_i, σ_i) , with $(\mu_i, \sigma_i) \in \mathcal{S}$, $\sigma_i > 0$, and $\sigma_i \uparrow \infty$ and

$$(3.11) \quad \mu_i \geq -h(a_i)a_i, \quad i = 1, 2, \dots,$$

where $a_i = \sigma_i^2$ and $h(s)$ is a nonnegative function on $[0, \infty)$ which decreases to zero as $s \rightarrow \infty$. Let

$$\sigma(x_1) = \sigma_{i(x_1)}, \quad \mu(x_1) = \mu_{i(x_1)}$$

where i is a function from $(-\infty, 0)$ to the positive integers with $i(x_1)$ increasing rapidly to ∞ as x_1 decreases to $-\infty$. Use (3.6), (3.7) and (3.11) to obtain

$$\begin{aligned} v'(x_1) &= \int_{-\infty}^{x_1} \frac{2}{a(z)} \exp \left(- \int_z^{x_1} \frac{2\mu(y)}{a(y)} dy \right) dz \\ &\geq \int_{-\infty}^{x_1} \frac{2}{a(z)} \exp \left(\int_z^{x_1} h(a(y)) dy \right) dz \\ &\geq \left[\exp \left(\int_{-\infty}^0 h(a(y)) dy \right) \right] \left[\int_{-\infty}^{x_1} \frac{2}{a(z)} dz \right]. \end{aligned}$$

For any $\varepsilon > 0$ we can choose i so that $a(y) = a_{i(y)}$ increases sufficiently fast so that both integrals occurring in the final expression are arbitrarily small. It follows that i can be chosen to make v as small as desired, and then (3.8) gives the desired conclusion. (Notice $T < \infty$ with probability one because $p(-\infty) = -\infty$ by (3.9) and σ is bounded below by σ_1 .) \square

For the maximization problem it seems natural to work on $[0, \infty)$ rather than $(-\infty, 0]$ and to think of maximizing the expected time until bankruptcy occurs. Here is the formal definition of the gambling problem:

$$F = \{x \in \mathbb{R}^2: 0 < x_1 < \infty\},$$

$$u(x) = x_2,$$

$$C(x) = C_0 \quad \text{for } x \in F^\circ$$

where C_0 is given by (3.0),

$$\Sigma(x) = \Sigma_C(x).$$

Then, for $x \in F$ and $X \in \Sigma(x)$,

$$(3.12) \quad u(X) = x_2 + ET$$

where $T = \inf \{t: X_1(t) = 0\}$. As before

$$V(x_1, x_2) = V(x_1) + x_2$$

where

$$V(x_1) = V(x_1, 0).$$

It is natural, as it was for the minimization problem, to conjecture that an optimal strategy will choose μ to achieve

$$M = \sup \{ \mu: (\mu, \sigma) \in \mathcal{S} \text{ for some } \sigma \}.$$

This time the conjecture is essentially correct.

THEOREM 2. *Let $x \in F^\circ$.*

(a) *If $M < 0$, then $V(x) = -x_1/M + x_2$. If in addition $(M, \sigma_0) \in \mathcal{S}$, then the process $X \in \Sigma(x)$ with $\mu(t) \equiv M$ and $\sigma(t) \equiv \sigma_0$ is optimal.*

(b) *If $M \geq 0$, then $V(x) = \infty$.*

Proof. Suppose X is given by a stationary Markov strategy $\mu(t) \equiv \mu_0$, $\sigma(t) \equiv \sigma_0$ where μ_0 and σ_0 are constants. Because we have changed from $(-\infty, 0]$ to $[0, \infty)$, formulas (3.5), (3.10) and (3.12) now imply

$$(3.13) \quad u(X) = \begin{cases} -\frac{x_1}{\mu_0} + x_2 & \text{if } \mu_0 < 0, \\ \infty & \text{if } \mu_0 \geq 0. \end{cases}$$

Part (b) of the theorem is immediate. For (a), let $Q(x) = -x_1/M + x_2$. By (3.13), $Q \leq V$. The reverse inequality will be proved by another application of the proposition of § 2.

Let $\{\beta_n\}$ be a sequence of numbers in the interval $(0, 1)$ which increase up to 1. Define

$$Q_n(x) = \frac{e^{\lambda(\beta_n)x_1} - 1}{\log \beta_n} + x_2 \beta_n^{x_2}$$

where

$$\lambda(\beta) = \frac{-M - \sqrt{M^2 - 2\sigma_0^2 \log \beta}}{\sigma_0^2}$$

and $\sigma_0 > 0$. (The first term on the right-hand side in the definition of $Q_n(x)$ is equal to the expectation of $\int_0^T (\beta_n)^s ds$ for a process $\mu(t) \equiv M$, $\sigma(t) \equiv \sigma_0$ and thus corresponds to a discounted payoff.)

Condition (i) of the proposition is easily verified, and (ii) is obvious because $Q \geq u$. For (iii) let $(a, b) \in C(x)$ where $a = \binom{\mu}{1}$, $b = \binom{\sigma}{0}$ and calculate (with $\beta = \beta_n$)

$$\begin{aligned} D(a, b)Q_n(x) &= \frac{\lambda(\beta) e^{\lambda(\beta)x_1}}{\log \beta} \left(\mu + \frac{1}{2} \lambda(\beta) \sigma^2 \right) + \beta^{x_2} (1 + x_2 \log \beta) \\ &\leq \frac{\lambda(\beta)M}{\log \beta} + \beta^{x_2} (1 + x_2 \log \beta). \end{aligned}$$

The inequality holds because $\mu \leq M$ and $\lambda(\beta) < 0$. Now recall that $X_2(s)$ is an increasing process and (iii) will follow after some calculus. Condition (iv) is an easy consequence of the definition of Q_n together with the facts that $X_1(s) \geq 0$ and $X_2(s) \geq X_2(0)$. \square

Remark 1. Since the set \mathcal{S} is not assumed to be bounded, and the σ with $(\mu, \sigma) \in \mathcal{S}$ are not bounded away from zero, the usual approach via Bellman's equation for the value function V could not be used above. For a continuous time gambling problem as defined in § 2 the Bellman equation can be written in the form

$$(3.14) \quad \sup D(a, b) V(x) = 0$$

where the supremum is taken over all $(a, b) \in C(x)$. For the minimization problem of this section, (3.4) applies and (3.14) becomes

$$(3.15) \quad \sup_{(\mu, \sigma) \in \mathcal{S}} \left[\mu V'(x_1) + \frac{1}{2} \sigma^2 V''(x_1) - 1 \right] = 0.$$

Under condition (c) of Theorem 1 the value function $V(x_1) \equiv 0$ does not satisfy (3.15). Furthermore if $I = \infty$ and $M < \infty$ with $(M, \sigma_0) \in \mathcal{S}$, the function x_1/M does solve (3.15) but does not represent the value function. Under condition (a) of the theorem the value function $V(x_1) = x_1/M$ is a solution of (3.15) but this fact does not follow from standard theorems.

Remark 2. Consider the problem of a process on the interval $0 < \tilde{x}_1 \leq 1$ determined by the equation

$$(3.16) \quad \tilde{X}_1(0) = \tilde{x}_1, d\tilde{X}_1(t) = \tilde{X}_1(t) [\tilde{\mu}(t) dt + \tilde{\sigma}(t) dW_t]$$

where $\tilde{\mu}(t)$, $\tilde{\sigma}(t)$ are nonanticipating controls required to satisfy $(\tilde{\mu}(t), \tilde{\sigma}(t)) \in \tilde{\mathcal{S}}$ and the object is to minimize the expectation of $\tilde{T} = \inf \{t: \tilde{X}_1(t) = 1\}$. This problem reduces to that of Theorem 1 by the change of variables $X_1(t) = \log \tilde{X}_1(t)$. This follows from Ito's formula, and one finds $\mu(t) = \tilde{\mu}(t) - \sigma^2(t)/2$, $\sigma(t) = \tilde{\sigma}(t)$. So one can formulate the theorem to apply to the \tilde{X}_1 process. Note that the role of M is assumed by

$$\tilde{M} = \sup \{ \tilde{\mu} - \tilde{\sigma}^2/2 : (\tilde{\mu}, \tilde{\sigma}) \in \tilde{\mathcal{S}} \}$$

and the role of I is taken by

$$\tilde{I} = \inf_{\varepsilon > 0} \sup \left\{ \tilde{\mu} - \left(\frac{1}{2} - \varepsilon \right) \tilde{\sigma}^2 : (\tilde{\mu}, \tilde{\sigma}) \in \tilde{\mathcal{S}} \right\}.$$

The problem for the \tilde{X}_1 process was considered in [9] and [3] and solved under some restrictions on $\tilde{\mathcal{S}}$. In [9] it was assumed that $\lambda \tilde{\mathcal{S}} \subseteq \tilde{\mathcal{S}}$ for all $\lambda \geq 0$, while in [3] this assumption was needed only for $0 \leq \lambda \leq 1$.

As discussed in [9] and [3] various models lead to the problem on $[0, 1]$. One of these, the "portfolio problem," will be explained in § 4. In [6] Kelly introduced a plan in discrete time based on the criterion of maximizing, at each stage, the expected value of the logarithm. This "Kelly criterion" was further studied by Breiman [1] who

established certain asymptotic optimality properties. Theorem 1 may be interpreted to imply that a continuous time Kelly criterion is in fact optimal under the hypotheses of (a), but not under those of (c).

4. A portfolio problem. Consider the problem of managing a portfolio of stocks, bonds and cash so as to minimize the expected time to reach a given total worth. For a simple model suppose that there is one bond whose price B_t at time t satisfies

$$dB_t = r_B B_t dt,$$

and one stock whose price S_t at time t satisfies

$$dS_t = r_S S_t dt + \sigma_S S_t dW_t$$

where r_B , r_S and σ_S are positive constants and $\{W_t\}$ is a standard Brownian motion. A recent paper by Malliaris [8] explains the use of stochastic differential models in finance and has numerous references to the financial literature. Let $\tilde{X}_1(t)$ be the total fortune of an investor at time t , let $f_S(t)$ be the fraction of that fortune invested in the stock, and let $f_B(t)$ be the fraction invested in the bond. Then \tilde{X}_1 satisfies

$$(4.1) \quad d\tilde{X}_1(t) = \tilde{X}_1(t)[r_S f_S(t) + r_B f_B(t)] dt + \sigma_S f_S(t) dW_t.$$

Let

$$\tilde{S} = \{(\tilde{\mu}, \tilde{\sigma}): \tilde{\mu} = r_S f_S + r_B f_B, \tilde{\sigma} = \sigma_S f_S, f_B \geq 0, f_S \geq 0, f_B + f_S \leq 1\}.$$

Then (4.1) and $\tilde{X}_1(0) = \tilde{x}_1$ are equivalent to (3.16) and $(\tilde{\mu}(t), \tilde{\sigma}(t)) \in \tilde{S}$. We are in the situation of Remark 2 of §3. Theorem 1 applies, and one is in case (a). If $r_B > r_S$ one should obviously take $f_B(t) \equiv 1$. If $r_B \leq r_S$ one finds

$$\tilde{M} = \begin{cases} r_B + \frac{(r_S - r_B)^2}{2\sigma_S^2} & \text{if } r_S \leq r_B + \sigma_S^2, \\ r_S - \frac{\sigma_S^2}{2} & \text{otherwise.} \end{cases}$$

The corresponding optimal policies are given by $f_B = 1 - f_S$ and

$$f_S = \begin{cases} \frac{r_S - r_B}{\sigma_S^2} & \text{if this is less than 1,} \\ 1 & \text{otherwise.} \end{cases}$$

In particular, the Kelly strategy is optimal.

REFERENCES

- [1] LEO BREIMAN, *Optimal gambling systems for favorable games*, Proc. the Fourth Berkeley Symposium on Mathematical Statistics and Probability, Univ. California Press, Berkeley, CA, 1 (1961), pp. 65-78.
- [2] DAVID FREEDMAN, *Brownian Motion and Diffusion*. Holden-Day, San Francisco, CA, 1971.
- [3] D. HEATH AND W. SUDDERTH, *Continuous time portfolio management: minimizing the expected time to reach a goal*, Preprint, Institute for Mathematics and Applications, Univ. Minnesota, Minneapolis, MN, 1984.
- [4] NOBUYUKI IKEDA AND SHINZO WATANABE, *Stochastic Differential Equations and Diffusion Processes*. Kodansha Ltd., Tokyo, 1981.

- [5] K. ITO AND H. MCKEAN, JR., *Diffusion Processes and their Sample Paths*. Academic Press, New York, 1965.
- [6] J. L. KELLY, JR., *A new interpretation of information rate*, Bell System Tech. J., 35 (1956), pp. 917-926.
- [7] N. V. KRYLOV, *Controlled Diffusion Processes*. Springer-Verlag, New York, 1980.
- [8] A. G. MALLIARIS, *Ito's calculus in financial decision making*, SIAM Rev., 25 (1983), pp. 481-496.
- [9] VICTOR C. PESTIEN AND WILLIAM D. SUDDERTH, *Continuous-time Red and Black: How to control a diffusion to a goal*, Math. Oper. Res., 10 (1983), pp. 599-611.

SUPERVISORY CONTROL OF A CLASS OF DISCRETE EVENT PROCESSES*

P. J. RAMADGE†‡ AND W. M. WONHAM†

Abstract. The paper studies the control of a class of discrete event processes, i.e., processes that are discrete, asynchronous and possibly nondeterministic. The controlled process is described as the generator of a formal language, while the controller, or *supervisor*, is constructed from a recognizer for a specified target language that incorporates the desired closed-loop system behavior. The existence problem for a supervisor is reduced to finding the largest *controllable* language contained in a given *legal* language. Two examples are provided.

Key words. discrete event systems, control, automata

AMS(MOS) subject classifications. 93C10, 93B50, 93C30

1. Introduction. In this paper we study the control of a class of systems broadly known as discrete event processes. The principal features of such processes are that they are discrete, asynchronous and (possibly) nondeterministic. Typical instances include computer networks, flexible manufacturing systems, and the start-up and shut-down procedures of industrial plants.

While numerous practical examples are described in the literature on simulation (see especially Fishman [1978] and Zeigler [1984]), there is at the present time apparently no unifying theory for the control of discrete event processes. Nor is it entirely clear what such a theory ought to encompass. Numerous approaches to the modeling of discrete event processes have appeared in the literature. A general sampling of these could include boolean models (Aveyard [1974]); Petri nets (Peterson [1981]); formal languages (Beauquier and Nivat [1980], Park [1981]); temporal logic (Pnueli [1979], Hailpern and Owicki [1983]); and port automata and flow networks (Milne and Milner [1979], Steenstrup, Arbib and Manes [1981]). All of this work is concerned, in one way or another, with the problem of how to achieve or verify the orderly flow of events; and to this end how to bring together ideas from logic, language and automaton theory. However, while control problems are implicit in much of the work just cited, control-theoretic ideas as such have found little application there. The variety of approaches reflects the diversity of areas in which discrete event processes play an important role. It also indicates that to date no dominant paradigm has emerged upon which a theory of control might be based.

In this article we investigate a simple abstract model of a controlled discrete event process, our main objective being to determine qualitative structural features of the relevant basic control problems. Specifically we take the controlled process to be the generator of a formal language, and study how the recognizer of a specified (target) language may be employed as a controller. In this regard we found suggestive the work of Shaw [1978] and Shields [1979] on flow expressions and path expressions

* Received by the editors August 17, 1984, and in final revised form March 21, 1985.

† Systems Control Group, Department of Electrical Engineering, University of Toronto, Toronto, Ontario, Canada M5S 1A4. This research was partially supported by the Natural Sciences and Engineering Research Council of Canada under grant A-7399.

‡ Present address, Department of Electrical Engineering and Computer Science, Princeton University, Princeton, New Jersey 08544.

respectively; while C. A. R. Hoare has recently brought to our attention certain points of similarity with his linguistic approach to concurrent processes in Hoare [1983, Chap. 2]. Nevertheless our definition of “controllable language,” and our main results. (Theorems 7.1 and 10.1) on the existence and structure of controllers are believed to be quite new. Our approach is similar in spirit to some qualitative theories of multivariable control synthesis that have emerged over the last decade in the context of standard dynamic systems (for example, Wonham [1979], Nijmeijer [1983]). The present article is based on Ramadge [1983], and is summarized in Ramadge and Wonham [1984], while earlier versions appeared as Ramadge and Wonham [1982a, b].

The paper is organized as follows. In § 2 we define the class of controlled processes and controllers (*supervisors*) of interest; and in § 3 we discuss various associated formal languages. Sections 4 and 5 develop criteria for the existence of a supervisor for which the corresponding closed-loop controlled system satisfies given linguistic requirements; the main new idea here is that of a *controllable language*. Section 6 introduces the notion of a supervisor that is *proper*, namely *nonblocking* and *nonrejecting*. In § 7 we pose two problems of supervisor synthesis: the Supervisory Marking Problem (SMP) and the Supervisory Control Problem (SCP). Each of these is then shown to be solvable in a *minimally restrictive*, or “optimal,” fashion in the class of proper supervisors, the “optimality” depending on a semilattice property of the relevant classes of languages. Section 8 defines a *projection* (or simplification) of supervisors. The latter, combined with some notions of reduction of languages and recognizers in § 9, leads to our second main result in § 10, the *Quotient Structure Theorem*. According to this, every efficiently constructed supervisor is structurally equivalent to a quotient (i.e., high-level, or lumped, model) of a recognizer of the desired closed-loop generated language. We conclude in §§ 11 and 12 with two simple but practical illustrations.

2. Controlled discrete-event processes.

2.1. Generators. To establish notation we first recall various standard ideas from automaton and language theory (cf. Hopcroft and Ullman [1979]). We define a *generator* to be a 5-tuple

$$\mathcal{G} = (Q, \Sigma, \delta, q_0, Q_m)$$

where Q is the set of *states* q , Σ is the *alphabet* or set of output symbols σ , $\delta: \Sigma \times Q \rightarrow Q$ is the *transition function*, $q_0 \in Q$ is the *initial state* and $Q_m \subset Q$ is a subset of states to be called *marker states*.¹ We always assume that Σ , but not necessarily Q or Q_m , is finite. In general, δ is only a partial function (pfn), meaning that, for each fixed $q \in Q$, $\delta(\sigma, q)$ is defined only for some subset $\Sigma(q) \subset \Sigma$ that depends on q . Formally \mathcal{G} is equivalent to a directed graph with node set Q and an edge $q \rightarrow q'$ labeled σ for each triple (σ, q, q') such that $q' = \delta(\sigma, q)$. Such an edge, or state transition, will be called an *event*.

We interpret \mathcal{G} as a device that starts in q_0 and executes state transitions, i.e., generates a sequence of events, by following its graph. Events are considered to occur spontaneously (no auxiliary forcing mechanism is postulated), asynchronously (i.e., without reference to a clock) and instantaneously. An event is thought of as signaled (to an outside observer, say) by its *label* σ . \mathcal{G} may be nondeterministic in the sense that more than one event may be available for selection at a given node of its graph; however, distinct events at a given node always carry distinct labels.

¹ The terms *generator* and *marker state* are nonstandard, but better suited to our interpretation than, for example, “automaton” and “final state.” Our “generator” is a special case of Harrison’s “transition system” (Harrison [1965]); it will play the role of “plant” in the sense of control theory.

Let Σ^* denote the set of all finite strings s of elements of Σ , including the empty string, 1.² In standard fashion we construct the extended transition function

$$\delta: \Sigma^* \times Q \rightarrow Q \quad (\text{pfn})$$

according to

$$\delta(1, q) = q, \quad q \in Q,$$

and

$$\delta(s\sigma, q) = \delta(\sigma, \delta(s, q))$$

whenever $q' = \delta(s, q)$ and $\delta(\sigma, q')$ are both defined. Any subset of Σ^* is a *language* over Σ . The strings of a language are often called *words*. The language *generated* by \mathcal{G} is

$$L(\mathcal{G}) = \{w: w \in \Sigma^* \text{ and } \delta(w, q_0) \text{ is defined}\}.$$

The language *marked* by \mathcal{G} is

$$L_m(\mathcal{G}) = \{w: w \in L(\mathcal{G}) \text{ and } \delta(w, q_0) \in Q_m\}.$$

We interpret $L(\mathcal{G})$ as the set of all possible finite sequences of events that can occur; while $L_m(\mathcal{G}) \subset L(\mathcal{G})$ is a distinguished subset of these sequences that may be “marked,” or recorded, perhaps representing completed “tasks” (or sequences of tasks) carried out by the physical process that \mathcal{G} is intended to model.³

To conclude this subsection we remark that it is usually convenient to eliminate states of \mathcal{G} that can never be reached (or “accessed”) from q_0 . Namely let

$$Q_{ac} = \{q: \exists w \in \Sigma^*, \delta(w, q_0) = q\},$$

$$Q_{ac,m} = Q_{ac} \cap Q_m,$$

$$\delta_{ac} = \delta|_{(\Sigma \times Q_{ac})}.$$

The *accessible component* of \mathcal{G} , denoted by $Ac(\mathcal{G})$, is then defined to be

$$Ac(\mathcal{G}) = (Q_{ac}, \Sigma, \delta_{ac}, q_0, Q_{ac,m}).$$

A generator \mathcal{G} is *accessible* if $\mathcal{G} = Ac(\mathcal{G})$.

We say that \mathcal{G} is *co-accessible* if every string in $L(\mathcal{G})$ can be completed to a string in $L_m(\mathcal{G})$, i.e.,

$$(\forall w) w \in L(\mathcal{G}) \Rightarrow (\exists s) s \in \Sigma^* \quad \text{and} \quad ws \in L_m(\mathcal{G}).$$

If \mathcal{G} is both accessible and co-accessible it is said to be *trim* (Eilenberg [1974]). It is well known (cf. Eilenberg [1974, § III.5]) that to every language (i.e., subset of Σ^*) there corresponds a trim generator that is essentially unique.

2.2. Controlled discrete event processes. To a generator $\mathcal{G} = (Q, \Sigma, \delta, q_0, Q_m)$ we now adjoin a means of control. That is, \mathcal{G} will play the role of the “plant” (object to be controlled) of standard control theory. For this let $\Sigma_c \subset \Sigma$ be a distinguished subset of the alphabet; we say that an event (σ, q, q') is a *controlled event* if $\sigma \in \Sigma_c$. Let

$$\Gamma = \{0, 1\}^{\Sigma_c}$$

² 1 plays the role of identity of string concatenation, i.e., $1s = s1 = s$.

³ Here there is no implication that generating action halts after the completion of some marked sequence; marked states of \mathcal{G} need not be “final” states.

be the set of all binary assignments to the elements of Σ_c . Each assignment $\gamma \in \Gamma$, i.e., each function

$$\gamma: \Sigma_c \rightarrow \{0, 1\},$$

is a *control pattern*. An event (with label) σ is said to be *enabled by γ* if $\gamma(\sigma) = 1$, or *disabled by γ* if $\gamma(\sigma) = 0$. It is convenient to extend each $\gamma \in \Gamma$ to a map $\gamma: \Sigma \rightarrow \{0, 1\}$ by defining $\gamma(\sigma) = 1$ for each $\sigma \in \Sigma - \Sigma_c$. If $\delta: \Sigma \times Q \rightarrow Q$ is the transition function of \mathcal{G} , we define an augmented transition function

$$\delta_c: \Gamma \times \Sigma \times Q \rightarrow Q \quad (\text{pfn})$$

according to

$$\delta_c(\gamma, \sigma, q) = \begin{cases} \delta(\sigma, q) & \text{if } \delta(\sigma, q) \text{ is defined and } \gamma(\sigma) = 1, \\ \text{undefined} & \text{otherwise.} \end{cases}$$

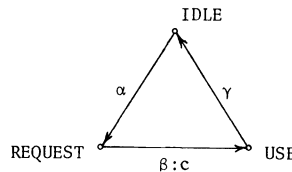
Formally, the object

$$\mathcal{G}_c = (Q, \Gamma \times \Sigma, \delta_c, q_0, Q_m)$$

is just another generator, constructed from \mathcal{G} by a specification of Σ_c . However, we interpret \mathcal{G}_c as a version of \mathcal{G} that admits external control, as follows. For brevity call “an event labeled σ ” simply “an event σ .” For each fixed $\gamma \in \Gamma$ there is a generator $\mathcal{G}(\gamma)$ formed by deleting from the graph of \mathcal{G} those events σ with $\gamma(\sigma) = 0$, i.e., those events that the control pattern γ disables. Then external control action would consist simply in switching the control pattern through a sequence of elements $\gamma, \gamma', \gamma'', \dots$ in Γ , like switching the pattern of red and green lights in a traffic network. Observe that such control is “permissive” (cf. Peterson [1981]): while disabled events are certainly prevented from occurring, enabled events are not necessarily forced to occur.

A structure \mathcal{G}_c as described above will be called a *controlled discrete event process* (CDEP).

2.3. Example—a primitive CDEP. A user of a resource may be modeled as a deterministic CDEP with three states I (IDLE), R (REQUEST) and U (USE), and with transitions as shown. Here we take (with some change of notation)



$\Sigma = \{\alpha, \beta, \gamma\}$ and $\Sigma_c = \{\beta\}$. The (two) control patterns correspond to evaluations $c = 0$ or $c = 1$ of the control variable c . A transition $R \rightarrow U$ may occur only when $c = 1$.

More interesting examples arise with the concurrent control of several CDEPs; these may then be combined into a single nondeterministic CDEP. At this stage the reader may skip ahead to §§ 11 and 12 for a glance at examples of this type.

2.4. Supervisors. Our objective will be to design a controller that switches control patterns in such a way that a given CDEP, \mathcal{G}_c , as described in § 2.2, behaves in obedience to various constraints. Such a controller will be called a *supervisor*. Formally a supervisor \mathcal{S} is a pair

$$\mathcal{S} = (S, \phi).$$

Here

$$S = (X, \Sigma, \xi, x_0, X_m)$$

is a deterministic automaton with (possibly infinite) state set X , input alphabet Σ , transition (partial) function $\xi: \Sigma \times X \rightarrow X$, initial state x_0 and marker subset $X_m \subset X$; while

$$\phi: X \rightarrow \Gamma$$

is a (total) function that maps supervisor states x into control patterns γ . Thus for each $x \in X$,

$$\gamma := \phi(x) \in \{0, 1\}^{\Sigma_c}.$$

(As before we extend $\phi(x)$ to a map $\phi(x): \Sigma \rightarrow \{0, 1\}$ with $\phi(x)(\sigma) = 1$ for each $\sigma \in \Sigma - \Sigma_c$.) S will always be assumed to be accessible. We call ϕ the *state feedback map*.

In many applications it will be the case that $X_m = X$, i.e., the supervisor plays no auxiliary “marking” role; but the extra generality with $X_m \neq X$ is obtained with little effort.

We interpret S conventionally, as a device that executes a sequence of state transitions (according to ξ) in response to an appropriate input string $w \in \Sigma^*$. Thus we may couple \mathcal{S} to \mathcal{G}_c in a feedback loop by allowing the state transitions of S to be forced by \mathcal{G}_c , and requiring \mathcal{G}_c to be constrained by the successive control patterns determined by the states of S . Formally define the partial function

$$\xi \times \delta_c: \Sigma \times X \times Q \rightarrow X \times Q \quad (\text{pfn})$$

according to

$$(\sigma, x, q) \mapsto (\xi(\sigma, x), \delta_c(\phi(x), \sigma, q)).$$

Thus $(\xi \times \delta_c)(\sigma, x, q)$ is defined iff $\delta(\sigma, q)$ is defined, $\phi(x)(\sigma) = 1$, and $\xi(\sigma, x)$ is defined. This yields the generator

$$(X \times Q, \Sigma, \xi \times \delta_c, (x_0, q_0), X_m \times Q_m).$$

We define the *supervised discrete event process* (SDEP), denoted by $\mathcal{S}/\mathcal{G}_c$, to be the accessible generator⁴

$$(2.1) \quad \mathcal{S}/\mathcal{G}_c = \text{Ac}(X \times Q, \Sigma, \xi \times \delta_c, (x_0, q_0), X_m \times Q_m).$$

From now on we shall assume that $\xi \times \delta_c$ has been extended to strings of Σ^* in the way described in § 2.1 for δ . Of course, so far there is nothing to guarantee that $(X \times Q)_{ac}$ is anything more than the singleton $\{(x_0, q_0)\}$, or that $L(\mathcal{S}/\mathcal{G}_c)$ is any larger than the singleton $\{1\}$ consisting of the empty string alone.

In analogy to the case of \mathcal{G} itself, we wish to interpret the language $L(\mathcal{S}/\mathcal{G}_c)$ generated by $\mathcal{S}/\mathcal{G}_c$ as the set of all possible finite sequences of events that can occur when \mathcal{S} is coupled to \mathcal{G}_c as just described. For this it is necessary to ensure that transitions of S are actually defined whenever they can occur in \mathcal{G} and are enabled by ϕ . To formalize this relationship we shall say that \mathcal{S} is *complete with respect to \mathcal{G}_c* provided the following is true: for all $s \in \Sigma^*$, $\sigma \in \Sigma$ the three conditions

- (i) $s \in L(\mathcal{S}/\mathcal{G}_c)$,
- (ii) $s\sigma \in L(\mathcal{G})$ (i.e., $\delta(s\sigma, q_0)$ is defined),
- (iii) $[\phi \circ \xi(s, x_0)](\sigma) = 1$ (i.e., σ is enabled at $\xi(s, x_0)$), together imply that
- (iv) $s\sigma \in L(\mathcal{S}/\mathcal{G}_c)$ (i.e., $\xi(s\sigma, x_0)$ is defined).

While the definition (2.1) is logically acceptable as it stands, it will be of real value only when it is physically interesting, namely when \mathcal{S} is complete with respect to \mathcal{G}_c .

⁴ The notation is intended to suggest simply that “ \mathcal{G}_c is under supervision by \mathcal{S} ,” no quotient structure is implied!

Before continuing with the general development, the reader might wish to glance at the opening paragraphs of §§ 11 and 12, where two concrete examples of supervisory control problems are provided.

3. Languages of $\mathcal{S}/\mathcal{G}_c$.

3.1. Definitions. Let $L \subset \Sigma^*$. The *closure* of L , denoted by \bar{L} , is the set of all strings that are prefixes of words of L , i.e.,

$$\bar{L} = \{s : s \in \Sigma^* \text{ and } (\exists t) t \in \Sigma^*, st \in L\}.$$

For instance, if $L = \emptyset$ then $\bar{L} = \emptyset$ and if $L \neq \emptyset$ then $1 \in \bar{L}$. A language L is *closed* if $L = \bar{L}$. If \mathcal{G} is any generator then $L(\mathcal{G})$ is closed; if in addition \mathcal{G} is trim then

$$L(\mathcal{G}) = \bar{L}_m(\mathcal{G}).$$

Let \mathcal{G}_c be a CDEP constructed from a generator \mathcal{G} . For simplicity we shall denote \mathcal{G}_c simply by its underlying generator \mathcal{G} . The notation $L(\mathcal{G})$ will henceforth denote the language generated by \mathcal{G} if disabling control action were absent, i.e., all events $\sigma \in \Sigma_c$ were permanently enabled. Similarly we refer to $L_m(\mathcal{G})$ as the *uncontrolled (discrete-event) process language*. Let \mathcal{S} be a supervisor for \mathcal{G} , $L(\mathcal{S}/\mathcal{G})$ the language generated by \mathcal{S}/\mathcal{G} and $L_m(\mathcal{S}/\mathcal{G})$ the language marked by \mathcal{S}/\mathcal{G} . Define the *language controlled by \mathcal{S} in \mathcal{G}* to be

$$(3.1) \quad L_c(\mathcal{S}/\mathcal{G}) := L(\mathcal{S}/\mathcal{G}) \cap L_m(\mathcal{G}).$$

In other words, $L_c(\mathcal{S}/\mathcal{G})$ consists of those (marked) strings of the uncontrolled process language that “survive” in the presence of supervision.

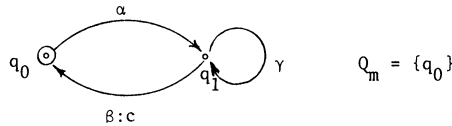
It is clear from the definitions that

$$(3.2) \quad L_m(\mathcal{S}/\mathcal{G}) \subset L_c(\mathcal{S}/\mathcal{G}) \subset L_m(\mathcal{G})$$

and, if \mathcal{G} is trim,

$$(3.3) \quad \bar{L}_m(\mathcal{S}/\mathcal{G}) \subset \bar{L}_c(\mathcal{S}/\mathcal{G}) \subset \bar{L}(\mathcal{S}/\mathcal{G}) [= L(\mathcal{S}/\mathcal{G})] \subset L(\mathcal{G}) = \bar{L}_m(\mathcal{G}).$$

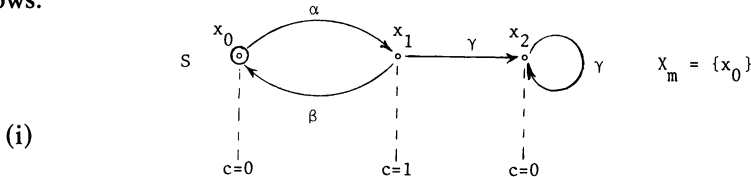
3.2. Examples. Let \mathcal{G} be the generator over $\Sigma = \{\alpha, \beta, \gamma\}$ displayed below.



Then⁵

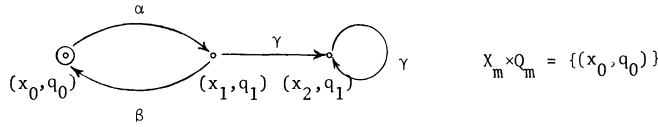
$$L_m(\mathcal{G}) = (\alpha\gamma^*\beta)^*.$$

We shall consider two different supervisors, each specified by its transition graph, as follows.



⁵ For the notation of regular expressions used here and below see, for example, Hopcroft and Ullman [1979].

This gives for \mathcal{S}/\mathcal{G} the transition graph



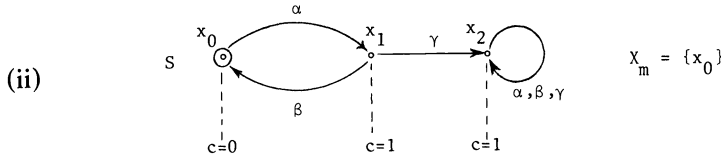
It is seen that

$$L(\mathcal{S}/\mathcal{G}) = (\alpha\beta)^*(1 + \alpha\gamma^*),$$

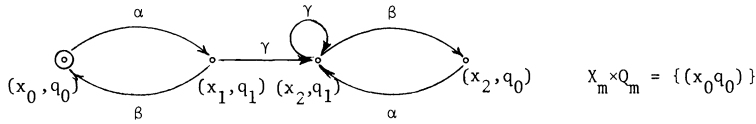
$$L_c(\mathcal{S}/\mathcal{G}) = L(\mathcal{S}/\mathcal{G}) \cap L_m(\mathcal{G}) = (\alpha\beta)^* = L_m(\mathcal{S}/\mathcal{G}).$$

For future reference (§ 6) we note, however, that

$$\bar{L}_c(\mathcal{S}/\mathcal{G}) = (\alpha\beta)^*(1 + \alpha) \subsetneq L(\mathcal{S}/\mathcal{G}).$$



This gives for \mathcal{S}/\mathcal{G} the transition graph



It is seen that

$$L_c(\mathcal{S}/\mathcal{G}) = L_m(\mathcal{G}) = (\alpha\gamma^*\beta)^*.$$

Again for future reference (§ 6), we note that

$$L_m(\mathcal{S}/\mathcal{G}) = (\alpha\beta)^* \subsetneq L_c(\mathcal{S}/\mathcal{G})$$

and

$$\bar{L}_m(\mathcal{S}/\mathcal{G}) \subsetneq \bar{L}_c(\mathcal{S}/\mathcal{G}).$$

4. Marking and control. A supervisor \mathcal{S} performs two essentially independent tasks: marking (as described by $L_m(\mathcal{S}/\mathcal{G})$) and control (as described by $L_c(\mathcal{S}/\mathcal{G})$, $L(\mathcal{S}/\mathcal{G})$). If a given controlled behavior is achievable, then any marking task is simultaneously achievable that is consistent with the controlled behavior.

Without essential loss of generality we assume that the generator \mathcal{G} is trim, namely

$$L(\mathcal{G}) = \bar{L}_m(\mathcal{G}).$$

PROPOSITION 4.1. (i) *For each sublanguage $K \subset L_m(\mathcal{G})$ there exists a complete supervisor \mathcal{S} such that, for \mathcal{S}/\mathcal{G} , we have*

$$L(\mathcal{S}/\mathcal{G}) = L(\mathcal{G}), \quad L_m(\mathcal{S}/\mathcal{G}) = K.$$

(ii) *Let L be a closed sublanguage of $L(\mathcal{G})$. If there exists a complete supervisor \mathcal{S} for which $L(\mathcal{S}/\mathcal{G}) = L$ then for every sublanguage K of $L \cap L_m(\mathcal{G})$ there exists a complete supervisor \mathcal{S}_K such that, correspondingly,*

$$L(\mathcal{S}_K/\mathcal{G}) = L, \quad L_m(\mathcal{S}_K/\mathcal{G}) = K.$$

Before proving Proposition 4.1 we make the (more or less standard) definition: if $L \subset \Sigma^*$, a recognizer \mathcal{M} for L is an accessible generator \mathcal{G} such that $L_m(\mathcal{G}) = L$. While

a recognizer and an accessible generator are formally no different, we interpret a recognizer $\mathcal{M} = (Q, \Sigma, \delta, q_0, Q_m)$ as a device which, like a supervisor, is forced externally by strings in Σ^* ; its action is thus to “recognize” precisely the words of L , regarded as input strings to \mathcal{M} .

Proof of Proposition 4.1. (i) Let $S = (X, \Sigma, \xi, x_0, X_m)$ be a recognizer for K . By adjoining a “dump” state to X , if necessary, we can arrange that $\xi(\sigma, x)$ is defined for all $(\sigma, x) \in \Sigma \times X$. Define

$$\phi : X \rightarrow \{0, 1\}^\Sigma$$

according to

$$\phi(x)(\sigma) = 1, \quad x \in X, \quad \sigma \in \Sigma.$$

It is clear that $\mathcal{S} = (S, \phi)$ has the required properties.

(ii) Let

$$T = (Y, \Sigma, \eta, y_0, Y_m)$$

be a recognizer for K . By adjoining a “dump” state to Y , if necessary, it can be arranged that $\eta(\sigma, y)$ is defined for all $(\sigma, y) \in \Sigma \times Y$. Let

$$\mathcal{S} = (S, \phi), \quad S = (X, \Sigma, \xi, x_0, X)$$

be a complete supervisor for which

$$L(\mathcal{S} / \mathcal{G}) = L.$$

Define the supervisor

$$\mathcal{S}_K = (S', \phi'), \quad S' = (X', \Sigma, \xi', x'_0, X'_m)$$

according to

$$\begin{aligned} X' &= X \times Y, & \xi'(\sigma, x, y) &= (\xi(\sigma, x), \eta(\sigma, y)), \\ x'_0 &= (x_0, y_0), & X'_m &= X \times Y_m, & \phi'(x, y) &= \phi(x). \end{aligned}$$

Since the control action of \mathcal{S}_K is the same as that of \mathcal{S} , it is clear that $L(\mathcal{S}_K / \mathcal{G}) = L$, and obviously $L_m(\mathcal{S}_K / \mathcal{G}) = K$. Also, $\xi'(\sigma, x, y)$ is defined just when $\xi(\sigma, x)$ is defined, so that S_k is complete with respect to \mathcal{G}_c . \square

In this proof our construction merely installs a recognizer that acts as a marking device, either alone in part (i), or “in parallel” with the original supervisor \mathcal{S} in part (ii). This does nothing to change the control action, but might be thought of as a means of recording when words in K have been completed.

5. Controllability. In this section we introduce a definition of controllability that will play a key role in characterizing those languages that can be generated by closed-loop structures $\mathcal{S} / \mathcal{G}$ with a given CDEP \mathcal{G} and a suitable choice of complete supervisor \mathcal{S} .

Let $\mathcal{G} = (Q, \Sigma, \delta, q_0, Q_m)$ be a fixed CDEP. We assume that \mathcal{G} is trim, i.e., $L(\mathcal{G}) = \bar{L}_m(\mathcal{G})$. Write $\Sigma_u = \Sigma - \Sigma_c$, i.e., Σ_u is the set of (labels of) events that cannot be disabled. Let $K \subset \Sigma^*$, $L \subset \Sigma^*$ be arbitrary languages. We say that K is

- (i) *L-closed* if $K = \bar{K} \cap L$,
- (ii) (Σ_u, L) -*invariant* if $\bar{K} \Sigma_u \cap L \subset \bar{K}$,
- (iii) *controllable* if $K \subset L(\mathcal{G})$ and K is $(\Sigma_u, L(\mathcal{G}))$ -invariant i.e., $\bar{K} \Sigma_u \cap L(\mathcal{G}) \subset \bar{K}$.

Recall that \bar{K} is the language consisting of K together with all the prefixes (including the empty word) of words in K . Thus a sublanguage K of L is *L-closed* iff any prefix of K that is a word of L is also a word of K .

The language $\bar{K}\Sigma_u \cap L$ consists of all strings $s' = s\sigma$ where $s' \in L$, $s \in \bar{K}$ and $\sigma \in \Sigma_u$. If we think of L as representing “physically possible behavior,” and \bar{K} as “legally admissible behavior,” then the string $s\sigma$ is a legally admissible string s followed by an uncontrolled symbol σ such that $s\sigma$ is physically possible. K is (Σ_u, L) -invariant precisely when all such strings are legally admissible, i.e., certain instances of uncontrolled behavior are nonetheless legal.

Finally, thinking of $L(\mathcal{G})$ as the uncontrolled process language, i.e., the physically possible uncontrolled behavior of our CDEP, we have that K is controllable if every prefix $s \in \bar{K}$ is physically possible, and every physically possible string $s\sigma$, with $s \in \bar{K}$ and σ uncontrolled, is again in \bar{K} .

The following technical proposition will support our main results (Theorems 6.1, 7.1) on existence of supervisors.

PROPOSITION 5.1. *Let $K_1 \subset L_m(\mathcal{G})$, $K_2 \subset L_m(\mathcal{G})$ and $K_3 \subset L(\mathcal{G})$ with $K_3 \neq \emptyset$. There exists a complete supervisor \mathcal{S} such that for the closed-loop system \mathcal{S}/\mathcal{G} ,*

$$(5.1) \quad L_m(\mathcal{S}/\mathcal{G}) = K_1, \quad L_c(\mathcal{S}/\mathcal{G}) = K_2, \quad L(\mathcal{S}/\mathcal{G}) = K_3,$$

iff

- (i) $K_1 \subset K_2$,
- (ii) $K_2 = K_3 \cap L_m(\mathcal{G})$,
- (iii) K_3 is closed and controllable.

Proof. (Only if). Let the complete supervisor \mathcal{S} satisfy (5.1). Condition (i) follows by (3.2) and (ii) by the definition (3.1) of $L_c(\mathcal{S}/\mathcal{G})$. We have already noted that $L(\mathcal{S}/\mathcal{G}) (= K_3)$ is closed. To show that K_3 is controllable, suppose that $s\sigma \in L(\mathcal{G})$, with $s \in \bar{K}_3 (= K_3 = L(\mathcal{S}/\mathcal{G}))$ and $\sigma \in \Sigma_u$. If $\mathcal{S} = (S, \phi)$ with $S = (X, \Sigma, \xi, x_0, X_m)$ then, in the notation of § 2.4,

$$(x, q) := (\xi \times \delta_c)(s, x_0, q_0)$$

is defined. Since $\sigma \in \Sigma_u$ we have $\phi(x)(\sigma) = 1$, and as $s\sigma \in L(\mathcal{G})$ it follows that $q' := \delta(\sigma, q)$ is defined. Therefore

$$\delta_c(\phi(x), \sigma, q) = q';$$

and because \mathcal{S} is complete with respect to \mathcal{G}_c ,

$$x' := \xi(\sigma, x)$$

is defined. Therefore

$$(\xi \times \delta_c)(\sigma, x, q) = (\xi(\sigma, x), \delta_c(\phi(x), \sigma, q)) = (x', q')$$

is defined, namely

$$s\sigma \in L(\mathcal{S}/\mathcal{G}) = K_3 = \bar{K}_3,$$

so K_3 is controllable.

(If). By Proposition 4.1 it is enough to construct a complete supervisor \mathcal{S} such that $L(\mathcal{S}/\mathcal{G}) = K_3$. For this let $S = (X, \Sigma, \xi, x_0, X)$ be a trim recognizer for K_3 . Since K_3 is closed and S is trim, the marker set of S is X itself, as indicated; and we have that $\xi(s, x_0)$ is defined iff $s \in K_3$. For $x \in X$ let

$$\Sigma_x^0 = \{\sigma : (\exists s) s \in K_3 \text{ and } \xi(s, x_0) = x \text{ and } s\sigma \in L(\mathcal{G}) \text{ and } s\sigma \notin K_3\},$$

$$\Sigma_x^1 = \{\sigma : (\exists s) s \in K_3 \text{ and } \xi(s, x_0) = x \text{ and } s\sigma \in K_3\}.$$

We claim that $\Sigma_x^0 \cap \Sigma_x^1 = \emptyset$. In essence this follows by the fact that S is a trim recognizer for K_3 . Indeed suppose that $\sigma \in \Sigma_x^0 \cap \Sigma_x^1$ with

$$\begin{aligned} s^0 \in K_3, \xi(s^0, x_0) = x, & \quad s^0 \sigma \notin K_3, \\ s^1 \in K_3, \xi(s^1, x_0) = x, & \quad s^1 \sigma \in K_3. \end{aligned}$$

Then

$$\xi(\sigma, x) = \xi(\sigma, \xi(s^0, x_0)) = \xi(s^0 \sigma, x_0)$$

fails to be defined (since $s^0 \sigma \notin K_3$); whereas

$$\xi(\sigma, x) = \xi(\sigma, \xi(s^1, x_0)) = \xi(s^1 \sigma, x_0)$$

must be defined (since $s^1 \sigma \in K_3$): a contradiction.

By controllability of K_3 , $\Sigma_x^0 \subset \Sigma_c$. Let

$$\phi : X \rightarrow \{0, 1\}^\Sigma$$

be any function such that, if $\phi(x) =: \gamma$, then

$$\gamma(\Sigma_x^0) = 0, \quad \gamma(\Sigma_x^1) = 1, \quad \gamma(\Sigma_u - \Sigma_x^1) = 1.$$

It is clear from the claim just proved that such a function ϕ exists.

Now let $\mathcal{S} = (S, \phi)$. It will be shown that $L(\mathcal{S}/\mathcal{G}) = K_3$. By the definition of S , it is clear that $L(\mathcal{S}/\mathcal{G}) \subset K_3$. For the reverse inclusion, we use induction on the length $|s|$ of strings $s \in L(\mathcal{G})$. If $|s| = 1$, i.e., $s = \sigma$ for some $\sigma \in \Sigma$, then

$$\sigma \in \Sigma_{x_0}^1 \quad \text{if } \sigma \in K_3$$

so $\sigma \in L(\mathcal{S}/\mathcal{G})$ if $\sigma \in K_3$. For a language $L \subset \Sigma^*$ write

$$L^{(j)} := \{s : s \in L \text{ and } |s| = j\}, \quad j = 0, 1, 2, \dots,$$

and note that $L = \bigcup_{j=0}^{\infty} L^{(j)}$. Assume for the induction step that

$$L^{(i)}(\mathcal{S}/\mathcal{G}) = K_3^{(i)}, \quad i = 0, 1, \dots, j.$$

Let $s \in L^{(j)}(\mathcal{S}/\mathcal{G})$ and consider the string $s\sigma \in L(\mathcal{G})$. Now $x := \xi(s, x_0)$ is defined, and so $\sigma \in \Sigma_x^0 \cup \Sigma_x^1$; therefore $s\sigma \in K_3$ implies

$$\sigma \in \Sigma_x^1 \text{ and } \xi(\sigma, x) \text{ is defined}$$

implies

$$\phi(x)(\sigma) = 1 \text{ and } \xi(\sigma, x) \text{ is defined}$$

implies

$$s\sigma \in L(\mathcal{S}/\mathcal{G}).$$

Therefore

$$L^{(j+1)}(\mathcal{S}/\mathcal{G}) = K_3^{(j+1)}.$$

It only remains to show that \mathcal{S} is complete with respect to \mathcal{G} . For this let

$$s \in L(\mathcal{S}/\mathcal{G}) = K_3, \quad s\sigma \in L(\mathcal{G}),$$

and

$$[\phi \circ \xi(s, x_0)](\sigma) = 1.$$

Now if $s\sigma \notin K_3$, then we must have

$$\sigma \in \Sigma_{\xi(s, x_0)}^0.$$

But this implies that

$$[\phi \circ \xi(s, x_0)](\sigma) = 0,$$

a contradiction. Hence $s\sigma \in K_3$ and \mathcal{S} is complete. \square

6. Proper supervisors. To specify controlled behavior in a way that is intuitively satisfying, more stringent conditions must be placed on the three languages

$$L_m(\mathcal{S}/\mathcal{G}), \quad L_c(\mathcal{S}/\mathcal{G}), \quad L(\mathcal{S}/\mathcal{G})$$

that describe the closed-loop system \mathcal{S}/\mathcal{G} . We shall say that \mathcal{S} is *nonblocking* if

$$\bar{L}_c(\mathcal{S}/\mathcal{G})[\bar{L}(\mathcal{S}/\mathcal{G}) \cap \bar{L}_m(\mathcal{G})] = L(\mathcal{S}/\mathcal{G})$$

and that \mathcal{S} is *nonrejecting* if

$$\bar{L}_c(\mathcal{S}/\mathcal{G}) = \bar{L}_m(\mathcal{S}/\mathcal{G}).$$

By definition we always have $\bar{L}_c(\mathcal{S}/\mathcal{G}) \subset L(\mathcal{S}/\mathcal{G})$. If \mathcal{S} blocks, i.e., fails to be nonblocking, then there exists a string s generated by \mathcal{S}/\mathcal{G} (i.e., $s \in L(\mathcal{S}/\mathcal{G})$) that can never be completed to a word $st \in L_c(\mathcal{S}/\mathcal{G})$, i.e., $s \notin \bar{L}_c(\mathcal{S}/\mathcal{G})$. In this sense the CDEP may be blocked from ever completing a “task.” This undesirable situation is illustrated by supervisor (i) of § 3.2. Here, for instance, the string $\alpha\gamma \in L(\mathcal{S}/\mathcal{G}) - \bar{L}_c(\mathcal{S}/\mathcal{G})$.

If \mathcal{S} rejects, i.e., fails to be nonrejecting, then there exists a string $s \in \bar{L}_c(\mathcal{S}/\mathcal{G})$ that can be completed to a “task” in $L_c(\mathcal{S}/\mathcal{G})$ but never to a task that is marked, i.e., (say) recorded. By contrast, if $\bar{L}_c(\mathcal{S}/\mathcal{G}) = \bar{L}_m(\mathcal{S}/\mathcal{G})$, so that

$$L_m(\mathcal{S}/\mathcal{G}) \subset L_c(\mathcal{S}/\mathcal{G}) \subset \bar{L}_m(\mathcal{S}/\mathcal{G}),$$

then for every $s \in L_c(\mathcal{S}/\mathcal{G})$ there is some t such that $st \in L_m(\mathcal{S}/\mathcal{G})$, and then $st \in L_c(\mathcal{S}/\mathcal{G})$ as well. In § 3.2 the supervisor (ii) rejects: only strings of the form $(\alpha\beta)^*$ are marked, while $L_c(\mathcal{S}/\mathcal{G}) = (\alpha\gamma^*\beta)^*$ represents the complete set of tasks that may be performed.

A supervisor \mathcal{S} will be said to be *proper* if it is complete, nonblocking and nonrejecting; namely \mathcal{S} is complete and

$$\bar{L}_m(\mathcal{S}/\mathcal{G}) = \bar{L}_c(\mathcal{S}/\mathcal{G}) = L(\mathcal{S}/\mathcal{G}).$$

THEOREM 6.1. *Let $K \subset L_m(\mathcal{G})$, $K \neq \emptyset$.*

(i) *There exists a proper supervisor \mathcal{S} such that $L_m(\mathcal{S}/\mathcal{G}) = K$ iff K is controllable. In that case,*

$$L_c(\mathcal{S}/\mathcal{G}) = L_m(\mathcal{G}) \cap \bar{K}.$$

(ii) *There exists a proper supervisor \mathcal{S} such that $L_c(\mathcal{S}/\mathcal{G}) = K$ iff K is controllable and $L_m(\mathcal{G})$ -closed.*

Proof. (i) K is controllable iff \bar{K} is closed and controllable, iff the triple

$$(K_1, K_2, K_3) := (K, \bar{K} \cap L_m(\mathcal{G}), \bar{K})$$

satisfies conditions (i)–(iii) of Proposition 5.1, iff there exists a complete supervisor \mathcal{S} such that

$$(L_m(\mathcal{S}/\mathcal{G}), L_c(\mathcal{S}/\mathcal{G}), L(\mathcal{S}/\mathcal{G})) = (K, \bar{K} \cap L_m(\mathcal{G}), \bar{K})$$

and this condition means that \mathcal{S} is proper.

(ii) K is controllable and $L_m(\mathcal{G})$ -closed iff the triple

$$(K_1, K_2, K_3) := (K, K, \bar{K})$$

satisfies the conditions of Proposition 5.1, iff there exists a complete supervisor \mathcal{S} such that

$$(L_m(\mathcal{S}/\mathcal{G}), L_c(\mathcal{S}/\mathcal{G}), L(\mathcal{S}/\mathcal{G})) = (K, K, \bar{K}),$$

and again this means that \mathcal{S} is proper. \square

7. Supervisor synthesis problems. Let languages $L_a, L_g \in \Sigma^*$ be given, with

$$\emptyset \neq L_a \subset L_g \subset L_m(\mathcal{G}).$$

We interpret L_g as “legal behavior,” i.e., each word of L_g is a “legal task;” and L_a as “minimal acceptable behavior,” i.e., control of the CDEP \mathcal{G} in such a way that a language smaller than L_a is generated is considered inadequate. We now introduce the

Supervisory Marking Problem (SMP). Construct a proper supervisor \mathcal{S} for \mathcal{G} such that

$$L_a \subset L_m(\mathcal{S}/\mathcal{G}) \subset L_g.$$

Similarly we define the

Supervisory Control Problem (SCP). Construct a proper supervisor \mathcal{S} for \mathcal{G} such that

$$L_a \subset L_c(\mathcal{S}/\mathcal{G}) \subset L_g.$$

If SCP is solvable then by the proof of Theorem 6.1(ii) we can always arrange that $L_m(\mathcal{S}/\mathcal{G}) = L_c(\mathcal{S}/\mathcal{G})$, so that automatically SMP is solvable as well. For a converse to this statement, consider the special but interesting case where L_g is $L_m(\mathcal{G})$ -closed, i.e.,

$$L_g = \bar{L}_g \cap L_m(\mathcal{G}).$$

Then L_g is a sublanguage of $L_m(\mathcal{G})$ with the property that if a string $st \in L_g$ and $s \in L_m(\mathcal{G})$ then also $t \in L_g$. Now if SMP is solvable, the language $L_m(\mathcal{S}/\mathcal{G})$ satisfies

$$L_a \subset L_m(\mathcal{S}/\mathcal{G}) \subset L_g,$$

so that

$$L_a \subset L_m(\mathcal{S}/\mathcal{G}) \subset L_c(\mathcal{S}/\mathcal{G}).$$

Also, since \mathcal{S} is proper,

$$\bar{L}_m(\mathcal{S}/\mathcal{G}) = \bar{L}_c(\mathcal{S}/\mathcal{G})$$

so that

$$L_c(\mathcal{S}/\mathcal{G}) \subset \bar{L}_c(\mathcal{S}/\mathcal{G}) = \bar{L}_m(\mathcal{S}/\mathcal{G}) \subset \bar{L}_g.$$

But $L_c(\mathcal{S}/\mathcal{G}) \subset L_m(\mathcal{G})$ by definition, i.e.,

$$L_c(\mathcal{S}/\mathcal{G}) \subset \bar{L}_g \cap L_m(\mathcal{G}) = L_g.$$

Hence

$$L_a \subset L_c(\mathcal{S}/\mathcal{G}) \subset L_g$$

and so SCP is solvable as well.

When SMP or SCP is solvable, it may be considered desirable that the solution be *minimally restrictive* in the sense that $L_m(\mathcal{S}/\mathcal{G})$ or $L_c(\mathcal{S}/\mathcal{G})$, considered as a sublanguage of $L_m(\mathcal{G})$, be as large as possible, subject to the constraint that it is a sublanguage of L_g . The fact that minimally restrictive solutions are possible in principle is due to a certain semilattice property that we now describe. For this, let $L \subset L(\mathcal{G})$ be an arbitrary sublanguage of $L(\mathcal{G})$. Let

$$C_{\mathcal{G}}(L) := \{K : K \subset L \text{ and } K \text{ is controllable}\},$$

$$F_{\mathcal{G}}(L) := \{K : K \subset L \text{ and } K = \bar{K} \cap L_m(\mathcal{G})\}.$$

Thus $C_{\mathcal{G}}(L)$ (respectively, $F_{\mathcal{G}}(L)$) are the controllable (respectively, $L_m(\mathcal{G})$ -closed) sublanguages of L .

PROPOSITION 7.1. *$C_{\mathcal{G}}(L)$ and $F_{\mathcal{G}}(L)$ are nonempty classes of languages that are closed under arbitrary unions.*

Proof. Let \emptyset be the empty language (i.e., the empty set in Σ^*). Clearly

$$\emptyset \in C_{\mathcal{G}}(L) \quad \text{and} \quad \emptyset \in F_{\mathcal{G}}(L)$$

so $C_{\mathcal{G}}(L)$ and $F_{\mathcal{G}}(L)$ are nonempty classes. If $K_{\alpha} \in C_{\mathcal{G}}(L)$ for α in some index set A , then

$$\bar{K}_{\alpha} \Sigma_u \cap L(\mathcal{G}) \subset \bar{K}_{\alpha}, \quad \alpha \in A.$$

By the definition of closure it follows immediately that

$$\overline{\bigcup_{\alpha} K_{\alpha}} = \bigcup_{\alpha} \bar{K}_{\alpha}.$$

Therefore

$$\begin{aligned} \left(\overline{\bigcup_{\alpha} K_{\alpha}} \right) \Sigma_u \cap L(\mathcal{G}) &= \left(\bigcup_{\alpha} \bar{K}_{\alpha} \right) \Sigma_u \cap L(\mathcal{G}) = \bigcup_{\alpha} (\bar{K}_{\alpha} \Sigma_u) \cap L(\mathcal{G}) \\ &= \bigcup_{\alpha} [\bar{K}_{\alpha} \Sigma_u \cap L(\mathcal{G})] \subset \bigcup_{\alpha} \bar{K}_{\alpha} = \overline{\bigcup_{\alpha} K_{\alpha}}, \end{aligned}$$

and so

$$\bigcup_{\alpha} K_{\alpha} \in C_{\mathcal{G}}(L)$$

as claimed. The proof for $F_{\mathcal{G}}(L)$ is similar. \square

By Proposition 7.1 each of $C_{\mathcal{G}}(L)$, $F_{\mathcal{G}}(L)$ contains a unique supremal element with respect to inclusion, which we denote by

$$\sup C_{\mathcal{G}}(L), \quad \sup F_{\mathcal{G}}(L),$$

respectively. In fact, $C_{\mathcal{G}}(L)$ and $F_{\mathcal{G}}(L)$ are complete subsemilattices of the semilattice of all sublanguages of L , partially ordered by inclusion, and with join operation the union of languages.

On the basis of Theorem 6.1 and Proposition 7.1 we immediately obtain our first main result.

THEOREM 7.1. (i) *SMP is solvable iff*

$$\sup C_{\mathcal{G}}(L_g) \supset L_a.$$

(ii) *SCP is solvable iff*

$$\sup \{C_{\mathcal{G}}(L_g) \cap F_{\mathcal{G}}(L_g)\} \supset L_a.$$

In each case the corresponding supervisor is minimally restrictive. \square

8. Projections of supervisors. Let $\mathcal{S} = (S, \phi)$ and $\hat{\mathcal{S}} = (\hat{S}, \hat{\phi})$ each be supervisors for \mathcal{G} , where as usual

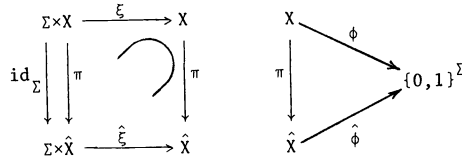
$$S = (X, \Sigma, \xi, x_0, X_m), \quad \phi : X \rightarrow \{0, 1\}^\Sigma,$$

$$\hat{S} = (\hat{X}, \Sigma, \hat{\xi}, \hat{x}_0, \hat{X}_m), \quad \hat{\phi} : \hat{X} \rightarrow \{0, 1\}^\Sigma.$$

We shall say that a (total) function $\pi : X \rightarrow \hat{X}$ is a *projection* from \mathcal{S} to $\hat{\mathcal{S}}$, and write $\pi : \mathcal{S} \rightarrow \hat{\mathcal{S}}$, provided

- (i) $\pi : X \rightarrow \hat{X}$ is surjective,
- (ii) $\pi(x_0) = \hat{x}_0$ and $X_m = \pi^{-1}(\hat{X}_m)$,
- (iii)⁶ $\hat{\xi} \circ (\text{id}_\Sigma \times \pi)(\sigma, x) = \pi \circ \xi(\sigma, x)$ for all (σ, x) where $\xi(\sigma, x)$ is defined,
- (iv) $\hat{\phi} \circ \pi = \phi$.

Under these conditions we shall refer to \hat{S} as the *quotient* of S under π . The situation is displayed in the diagrams⁷ below.



Projections represent very close relationships between supervisors, as expressed in the following. We assume that $\text{id}_\Sigma \times \pi$ is extended to a map $\text{id}_\Sigma \times \pi : \Sigma^* \times X \rightarrow \Sigma^* \times \hat{X}$ in the natural way.

PROPOSITION 8.1. *Let \mathcal{S} be complete with respect to \mathcal{G} , and let $\pi : \mathcal{S} \rightarrow \hat{\mathcal{S}}$ be a projection. Then*

- (i) π is unique,
- (ii) $(L_m, L_c, L)(\mathcal{S}/\mathcal{G}) = (L_m, L_c, L)(\hat{\mathcal{S}}/\mathcal{G})$,
- (iii) $\hat{\mathcal{S}}$ is complete with respect to \mathcal{G} ,
- (iv) \mathcal{S} is nonblocking (respectively, nonrejecting, proper) iff $\hat{\mathcal{S}}$ is nonblocking (respectively, nonrejecting, proper).

Proof. As usual we write

$$\mathcal{S} = (S, \phi), \quad S = (X, \Sigma, \xi, x_0, X_m)$$

and similarly for $\hat{\mathcal{S}}$. Recall that, by definition, both S and \hat{S} are accessible.

- (i) We have $\pi(x_0) = \hat{x}_0$. If $x \in X$ then, since \mathcal{S} is accessible, there is a string $s \in \Sigma^*$ such that $x = \xi(s, x_0)$, so

$$\pi(x) = \pi \circ \xi(s, x_0) = \hat{\xi} \circ (\text{id}_\Sigma \times \pi)(s, x_0) = \hat{\xi}(s, \hat{x}_0)$$

and this formula determines $\pi(x)$ uniquely.

- (ii) Write

$$L = L(\mathcal{S}/\mathcal{G}), \quad \hat{L} = L(\hat{\mathcal{S}}/\mathcal{G}).$$

If $s \in L$ with $|s| = 1$, i.e., $s = \sigma$, then $\phi(x_0)(\sigma) = 1$, and

$$(\xi \times \delta_c)(\sigma, x_0, q_0) = (\xi(\sigma, x_0), \delta_c(\phi(x_0), \sigma, q_0)) = (\xi(\sigma, x_0), \delta(\sigma, q_0))$$

⁶ By definition $\text{id}_\Sigma \times \pi : \Sigma \times X \rightarrow \Sigma \times \hat{X} : (\sigma, x) \mapsto (\sigma, \pi(x))$.

⁷ The symbol \supset means that the left-hand diagram is only “partially commutative,” in the sense of (iii).

is defined; hence $\hat{\phi}(\hat{x}_0)(\sigma) = 1$ and

$$\begin{aligned} (\hat{\xi} \times \delta_c)(\sigma, \hat{x}_0, q_0) &= (\hat{\xi}(\sigma, \hat{x}_0), \delta(\sigma, q_0)) = (\pi \circ \xi(\sigma, x_0), \delta(\sigma, q_0)) \\ &= (\pi \times \text{id}_Q) \circ (\xi \times \delta_c)(\sigma, x_0, q_0) \end{aligned}$$

is defined. This shows that

$$L^{(1)} \subset \hat{L}^{(1)}.$$

By induction on $|s|$ it is readily seen that

$$L^{(j)} \subset \hat{L}^{(j)}, \quad j = 0, 1, \dots,$$

hence $L \subset \hat{L}$.

For the reverse inclusion suppose first that $(\hat{\xi} \times \delta_c)(\sigma, \hat{x}_0, q_0)$ is defined. Then $\sigma \in \hat{L}$ and $\hat{\phi}(\hat{x}_0)(\sigma) = 1$. So

$$\phi(x_0)(\sigma) = (\hat{\phi} \circ \pi(x_0))(\sigma) = \hat{\phi}(\hat{x}_0)(\sigma) = 1.$$

Also

$$\delta(\sigma, q_0) = \delta_c(\hat{\phi}(\hat{x}_0), \sigma, q_0)$$

is defined, so by virtue of completeness $\xi(\sigma, x_0)$ is defined. Therefore

$$(\xi \times \delta_c)(\sigma, x_0, q_0) = (\xi(\sigma, x_0), \delta(\sigma, q_0))$$

is defined, and so $\hat{L}^{(1)} \subset L^{(1)}$. Assuming $\hat{L}^{(i)} \subset L^{(i)}$ ($i = 0, 1, \dots, j$), let $s \in \hat{L}^{(j)}$ and consider $s\sigma \in \hat{L}^{(j+1)}$. Then $s \in L^{(j)}$, so

$$x := (\xi \times \delta_c)(s, x_0, q_0), \quad \hat{x} := (\hat{\xi} \times \delta_c)(s, \hat{x}_0, q_0)$$

are defined, and $\pi(x) = \hat{x}$. By exactly the same argument as before, applied to (x, \hat{x}) in place of (x_0, \hat{x}_0) , we conclude that $s\sigma \in L^{(j+1)}$. So $\hat{L} \subset L$ and $\hat{L} = L$. It is now immediate that

$$L_c(\mathcal{S}/\mathcal{G}) = L(\mathcal{S}/\mathcal{G}) \cap L_m(\mathcal{G}) = L(\hat{\mathcal{S}}/\mathcal{G}) \cap L_m(\mathcal{G}) = L_c(\hat{\mathcal{S}}/\mathcal{G}).$$

Finally,

$$\begin{aligned} s \in L_m(\mathcal{S}/\mathcal{G}) & \\ \text{iff } (\xi \times \delta_c)(s, x_0, q_0) &\in X_m \times Q_m, \\ \text{iff } \xi(s, x_0) \in X_m &\text{ and } s \in L_c(\mathcal{S}/\mathcal{G}), \\ \text{iff } \xi(s, x_0) \in \pi^{-1}(\hat{X}_m) &\text{ and } s \in L_c(\mathcal{S}/\mathcal{G}), \\ \text{iff } \pi \circ \xi(s, x_0) \in \hat{X}_m &\text{ and } s \in L_c(\hat{\mathcal{S}}/\mathcal{G}), \\ \text{iff } \hat{\xi}(s, \hat{x}_0) \in \hat{X}_m &\text{ and } s \in L_c(\hat{\mathcal{S}}/\mathcal{G}), \\ \text{iff } (\hat{\xi} \times \delta_c)(s, \hat{x}_0, q_0) &\in \hat{X}_m \times Q_m, \\ \text{iff } s \in L_m(\hat{\mathcal{S}}/\mathcal{G}). & \end{aligned}$$

(iii) To verify that $\hat{\mathcal{S}}$ is complete with respect to \mathcal{G} , let

$$s \in L(\hat{\mathcal{S}}/\mathcal{G}), \quad s\sigma \in L(\mathcal{G})$$

and

$$(\hat{\phi} \circ \hat{\xi}(s, \hat{x}_0))(\sigma) = 1.$$

Then $s \in L(\mathcal{S}/\mathcal{G})$ by (ii), and

$$\phi \circ \xi(s, x_0)(\sigma) = (\hat{\phi} \circ \pi \circ \xi(s, x_0))(\sigma) = (\hat{\phi} \circ \hat{\xi}(s, \hat{x}_0))(\sigma) = 1.$$

Since \mathcal{S} is complete it follows that $\xi(s\sigma, x_0)$ is defined, hence (because π is a projection) $\hat{\xi}(s\sigma, \hat{x}_0)$ is defined as well, namely $\hat{\mathcal{S}}$ is complete.

(iv) Immediate from (ii) and (iii). \square

9. Efficient supervisor. In this section we give a simple abstract characterization of an “efficiently constructed” supervisor for a given nonempty, controllable and $L_m(\mathcal{G})$ -closed language $K \subset \Sigma^*$. By Theorem 6.1(ii) we know that a proper supervisor $\mathcal{S} = (S, \phi)$ exists such that $K = L_m(\mathcal{S}/\mathcal{G}) = L_c(\mathcal{S}/\mathcal{G})$, so that

$$\bar{K} = L(\mathcal{S}/\mathcal{G}).$$

Furthermore, by the construction used in the proof of Proposition 5.1 (“if” statement), we can arrange that, for a string $s \in \bar{K}$, the state x reached by S is such that, for all $\sigma \in \Sigma_c$,

$$(9.1) \quad \phi(x)(\sigma) = \begin{cases} 0, & s\sigma \notin \bar{K}, \\ 1, & s\sigma \in \bar{K}. \end{cases}$$

On the basis of (9.1) we define an equivalence relation on Σ^* as follows. Strings $s, s' \in \Sigma^*$ are *control-equivalent*, written $s \sim s'$, if for all $\sigma \in \Sigma_c$, $s\sigma \in \bar{K}$ iff $s'\sigma \in \bar{K}$. Thus two strings are control-equivalent if the control action (9.1) immediately following either one is the same for every $\sigma \in \Sigma$.

Recall from automaton theory (for example, Harrison [1965]) that an equivalence relation \mathbf{e} on Σ^* is a *right-congruence* if, whenever $s, s' \in \Sigma^*$ and $s \equiv s' \pmod{\mathbf{e}}$, then for all $t \in \Sigma^*$, $st \equiv s't \pmod{\mathbf{e}}$. Now let $\{\mathbf{e}_\alpha : \alpha \in A\}$ be an arbitrary nonempty family of equivalence relations on Σ^* . Their lattice-theoretic join, written

$$(9.2) \quad \mathbf{e} = \sup \{\mathbf{e}_\alpha : \alpha \in A\},$$

is defined as follows (cf. Szász [1963]): $s \equiv s' \pmod{\mathbf{e}}$ if there exists an integer $k \geq 1$, elements $\alpha_0, \dots, \alpha_k \in A$, and strings $s_1, \dots, s_k \in \Sigma^*$ such that

$$\begin{aligned} s &\equiv s_1 \pmod{\mathbf{e}_{\alpha_0}} \\ s_1 &\equiv s_2 \pmod{\mathbf{e}_{\alpha_1}} \\ &\vdots \\ s_{k-1} &\equiv s_k \pmod{\mathbf{e}_{\alpha_{k-1}}} \\ s_k &\equiv s' \pmod{\mathbf{e}_{\alpha_k}}. \end{aligned}$$

It is easy to check that if, in particular, the \mathbf{e}_α are right-congruences, then so is \mathbf{e} .

The lattice-theoretic ordering of equivalence relations on Σ^* is defined as follows: $\mathbf{e}_1 \leq \mathbf{e}_2$ if, for all $s, s' \in \Sigma^*$, $s \equiv s' \pmod{\mathbf{e}_1}$ implies $s \equiv s' \pmod{\mathbf{e}_2}$; \mathbf{e}_1 is said to be *finer than* \mathbf{e}_2 (or \mathbf{e}_2 is *coarser than* \mathbf{e}_1). Then \mathbf{e} in (9.2) is the finest equivalence relation on Σ^* that is coarser than each \mathbf{e}_α , $\alpha \in A$.

In general, control-equivalence \sim is not a right-congruence. However, if we define $s \equiv s' \pmod{\mathbf{o}}$ if $s = s'$, then trivially \mathbf{o} is a right-congruence and $\mathbf{o} \leq \sim$. It follows from the preceding that the equivalence

$$\approx := \sup \{\mathbf{e} : \mathbf{e} \text{ a right-congruence on } \Sigma^* \text{ and } \mathbf{e} \leq \sim\}$$

exists, and is the coarsest right-congruence on Σ^* that is finer than \sim .

For $s \in \Sigma^*$ let $[s]$ be the equivalence class of $s \pmod{\approx}$. In standard fashion we construct the corresponding automaton, defined on strings in \bar{K} . Let

$$\bar{S} = (\bar{X}, \Sigma, \bar{\xi}, \bar{x}_0, \bar{X}).$$

Here

$$\bar{X} = \{[s] : s \in \bar{K}\}, \quad \bar{x}_0 = [1]$$

(note that $1 \in \bar{K}$ as \bar{K} is nonempty and closed); finally $\bar{\xi}(\sigma, \bar{x}) = \bar{x}'$ if $\bar{x} = [s]$, $s \in \bar{K}$, $s\sigma \in \bar{K}$ and $[s\sigma] = \bar{x}'$; otherwise $\bar{\xi}$ is not defined. Next we define a control law

$$\bar{\phi}: \bar{X} \rightarrow \{0, 1\}^{\Sigma}$$

according to

$$\bar{\phi}(\bar{x})(\sigma) = 0;$$

if $\sigma \in \Sigma_c$, there exists $s \in \bar{K}$ with $[s] = \bar{x}$ and $s\sigma \notin \bar{K}$; otherwise $\bar{\phi}(\bar{x})(\sigma) = 1$. By our construction of \bar{X} , $\bar{\xi}$ and $\bar{\phi}$ are unambiguously determined. We can now define the efficient supervisor

$$\bar{\mathcal{P}} = (\bar{S}, \bar{\phi}).$$

Evidently $\bar{\mathcal{P}}$ is “efficient” in the sense that any automaton \hat{S} , that supports the defined control action (9.1) on each string $s \in \bar{K}$, must have a state structure (right-congruence on Σ^*) at least as fine as that of \bar{S} .

It is easy to see that $\bar{\mathcal{P}}$ is complete, since

$$s \in \bar{K}, [s] = \bar{x}, \bar{\phi}(\bar{x})(\sigma) = 1, s\sigma \in L(\mathcal{G}),$$

implies $s\sigma \in \bar{K}$ and therefore $\bar{\xi}(\sigma, \bar{x})$ is defined. Much as in the proof of Proposition 8.1 it is straightforward to verify that $L(\bar{\mathcal{P}}/\mathcal{G}) = \bar{K} = L(\mathcal{P}/\mathcal{G})$. Finally, as $\bar{X}_m = \bar{X}$,

$$L_m(\bar{\mathcal{P}}/\mathcal{G}) = L(\bar{\mathcal{P}}/\mathcal{G}) \cap L_m(\mathcal{G}) = \bar{K} \cap L_m(\mathcal{G}) = K$$

since K is $L_m(\mathcal{G})$ -closed; so that

$$K = L_m(\bar{\mathcal{P}}/\mathcal{G}) = L_c(\bar{\mathcal{P}}/\mathcal{G})$$

and $\bar{\mathcal{P}}$ is proper. Thus $\bar{\mathcal{P}}$ performs the same control action on \mathcal{G} as the supervisor \mathcal{P} with which we started.

Let $\equiv (\text{mod } \bar{K})$ denote \bar{K} -equivalence on Σ^* : $s \equiv s' (\text{mod } \bar{K})$ if for all $t \in \Sigma^*$, $st \in \bar{K}$ iff $s't \in \bar{K}$. Clearly $\equiv (\text{mod } \bar{K})$ is a right-congruence. Also, if $s \equiv s' (\text{mod } \bar{K})$ then in particular for all $\sigma \in \Sigma_c$, $s\sigma \in \bar{K}$ iff $s'\sigma \in \bar{K}$. It follows that

$$\equiv (\text{mod } \bar{K}) \leq \approx;$$

i.e., for all $s, s' \in \Sigma^*$, $s \equiv s' (\text{mod } \bar{K})$ implies $s \approx s'$. In particular, if $s, s' \in \bar{K}$ and $s \equiv s' (\text{mod } \bar{K})$ then

$$\bar{\xi}(s, \bar{x}_0) = \bar{\xi}(s', \bar{x}_0).$$

We shall refer to the latter property by saying that the automaton \bar{S} is \bar{K} -reduced. By its construction, \bar{S} is also \bar{K} -trim, namely every state of \bar{S} is visited by a word of \bar{K} ; that is, for every $\bar{x} \in \bar{X}$ there is $s \in \bar{K}$ such that $\bar{\xi}(s, \bar{x}_0) = \bar{x}$. In the next section it is shown that any supervisor with these two properties can be projected from a supervisor based on a recognizer for \bar{K} .

10. Quotient structure theorem. We can now prove the second main result of this paper. It states, roughly, that “every efficiently constructed supervisor is a quotient (high-level, or lumped, model) of the desired closed-loop behavior.”

Let $\mathcal{P} = (S, \phi)$ be a complete supervisor for \mathcal{G} . Write $K_1 := L_m(\mathcal{P}/\mathcal{G})$, $K_3 := L(\mathcal{P}/\mathcal{G})$ and assume that S is K_3 -reduced and K_3 -trim. These properties hold for the “efficient” supervisor of the previous section. Finally let

$$\hat{S}^0 = (X^0, \Sigma, \xi^0, x_0^0, X^0)$$

be a trim recognizer for K_3 .

THEOREM 10.1. *Subject to the foregoing hypotheses, there exist a subset $X_m^0 \subset X^0$ and a state feedback map $\phi^0: X^0 \rightarrow \{0, 1\}^\Sigma$ with the following properties:*

(i) *The supervisor*

$$\mathcal{S}^0 := (S^0, \phi^0), \quad S^0 := (X^0, \Sigma, \xi^0, x_0^0, X_m^0)$$

is a complete supervisor for \mathcal{G} with

$$L_m(\mathcal{S}^0/\mathcal{G}) = K_1, \quad L(\mathcal{S}^0/\mathcal{G}) = K_3.$$

(ii) *There is a projection $\pi: \mathcal{S}^0 \rightarrow \mathcal{S}$.*

(iii) *If \mathcal{S} is proper then so is \mathcal{S}^0 .*

Proof. Write

$$S = (X, \Sigma, \xi, x_0, X_m).$$

Let $x^0 \in X^0$. Since \hat{S}^0 is trim there exists $s \in K_3$ such that

$$\xi^0(s, x_0^0) = x^0.$$

Let $\xi(s, x_0) =: x \in X$ and define $\pi: X^0 \rightarrow X$ according to $\pi(x^0) = x$.

To show that π is well-defined, let $t \in K_3$, $\xi^0(t, x_0^0) = x^0$, and let $\xi(t, x_0) =: y \in X$. Since \hat{S}^0 is a recognizer for K_3 and $\xi^0(s, x_0^0) = \xi^0(t, x_0^0)$, we have $s \equiv t \pmod{K_3}$. Since S is K_3 -reduced, $\xi(s, x_0) = \xi(t, x_0)$, i.e., $x = y$.

We claim that π is a projection. First let $x \in X$. Since S is K_3 -trim, there exists $s \in K_3$ such that $\xi(s, x_0) = x$. Let $\xi^0(s, x_0^0) =: x^0$. Then as already shown, $\pi(x^0) = x$, so π is surjective. To verify that π respects ξ^0 , let $\xi^0(\sigma, x^0) = y^0$. We have $x^0 = \xi^0(s, x_0^0)$ for some $s \in K_3$, and then $s\sigma \in K_3$. Since $K_3 = L(\mathcal{S}/\mathcal{G})$, $\xi(s\sigma, \pi(x^0))$ is defined and, as shown already, coincides with $\pi \circ \xi^0(\sigma, x^0)$. To establish that π is a projection it only remains to define $X_m^0 = \pi^{-1}(X_m)$ together with $\phi^0 := \phi \circ \pi$.

It must be shown that $L(\mathcal{S}^0/\mathcal{G}) = K_3$. By the argument used in the proof of Proposition 5.1 it is enough to show that

(i) $(\forall \sigma, x^0)(\exists s)\xi^0(s, x_0^0) = x^0$ and $s\sigma \in L(\mathcal{G})$ and $s\sigma \notin K_3 \Rightarrow \phi^0(x^0)(\sigma) = 0$,

(ii) $(\forall \sigma, x^0)(\exists s)\xi^0(s, x_0^0) = x^0$ and $s\sigma \in K_3 \Rightarrow \phi^0(x^0)(\sigma) = 1$.

For (i), let $\xi^0(s, x_0^0) = x^0$, $s\sigma \in L(\mathcal{G})$, $s\sigma \notin K_3$. Clearly $s \in K_3$. If $\xi(s, x_0) = x$ then, as in the proof of Proposition 5.1, it follows necessarily that $\phi(x)(\sigma) = 0$ and therefore

$$\phi^0(x^0)(\sigma) = (\phi \circ \pi(x^0))(\sigma) = \phi(x)(\sigma) = 0.$$

The proof of (ii) is similar.

To show that \mathcal{S}^0 is complete with respect to \mathcal{G} we note that

$$s \in L(\mathcal{S}^0/\mathcal{G}), \quad s\sigma \in L(\mathcal{G}) \quad \text{and} \quad [\phi^0 \circ \xi^0(s, x_0^0)](\sigma) = 1,$$

iff

$$s \in L(\mathcal{S}/\mathcal{G}), \quad s\sigma \in L(\mathcal{G}) \quad \text{and} \quad [\phi \circ \xi(s, x_0)](\sigma) = 1.$$

But since \mathcal{S} is complete the latter condition implies that $s\sigma \in L(\mathcal{S}/\mathcal{G}) = L(\mathcal{S}^0/\mathcal{G})$.

Finally let $s \in K_3$. Then

$$s \in K_1$$

iff $\xi(s, x_0) \in X_m$ and $\delta(s, q_0) \in Q_m$,

iff $\pi \circ \xi^0(s, x_0^0) \in X_m$ and $\delta(s, q_0) \in Q_m$,

iff $\xi^0(s, x_0^0) \in X_m^0$ and $\delta(s, q_0) \in Q_m$,

iff $s \in L_m(\mathcal{S}^0/\mathcal{G})$.

So $L_m(\mathcal{S}^0/\mathcal{G}) = L_m(\mathcal{S}/\mathcal{G})$. In particular if \mathcal{S} is proper, so is \mathcal{S}^0 . \square

11. Example 1. We consider two users of a single resource, each modeled as in § 2.3, giving the state transition graphs $\mathcal{G}_1, \mathcal{G}_2$ of Fig. 11.1. For \mathcal{G} we take the “shuffle” of $\mathcal{G}_1, \mathcal{G}_2$, namely the process determined by the concurrent actions of \mathcal{G}_1 and \mathcal{G}_2 under the assumption that these actions are asynchronous and independent. This assumption rules out the simultaneous occurrence of an event in \mathcal{G}_1 with an event in \mathcal{G}_2 , but otherwise places no constraint on their joint behavior. The graph of \mathcal{G} is thus as shown in Fig. 11.2. Here the state \odot is both q_0 and (as a singleton) Q_m , while $L_m(\mathcal{G})$ consists of all words over the alphabet

$$\Sigma = \{\alpha_1, \beta_1, \gamma_1, \alpha_2, \beta_2, \gamma_2\}$$

corresponding to paths in the graph that begin and end at \odot .

The objective of supervisory control is to manipulate the binary controls c_1, c_2 in order to satisfy the following synchronization requirements.

(i) *Mutual exclusion:* $\mathcal{G}_1, \mathcal{G}_2$ never simultaneously occupy their respective USE states.

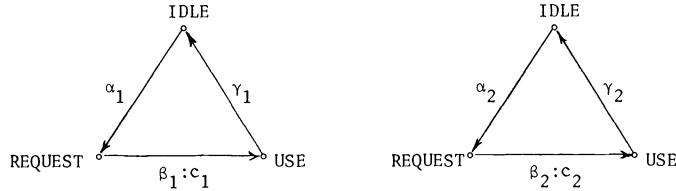


FIG. 11.1. Example 1: Independent CDEPs \mathcal{G}_1 and \mathcal{G}_2 .

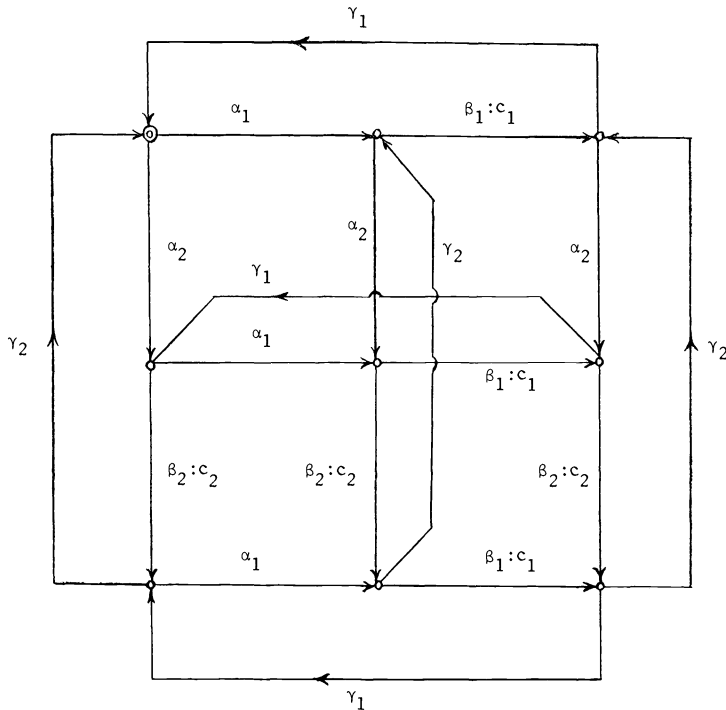


FIG. 11.2. Example 1: Shuffle \mathcal{G} of \mathcal{G}_1 and \mathcal{G}_2 .

(ii) *Fair usage*: The USE states of $\mathcal{G}_1, \mathcal{G}_2$ are occupied according to first-come-first-served discipline, namely the index sequence of events β_i must coincide with the index sequence of events α_j .

In practical terms this standard problem could, of course, be solved by a queue; but instead we shall approach it via the ideas of previous sections. However, we defer to a future article the theoretical issue of how conditions like (i) and (ii) may be formalized, simply taking it for granted that from them the “legal” behavior $L_g \subset L_m(\mathcal{G})$ can be explicitly determined. In fact the reader may convince himself that L_g is described by the generator displayed in Fig. 11.3.⁸

By inspection of Fig. 11.3 it is easy to see that L_g is both controllable and $L_m(\mathcal{G})$ -closed. That is,

$$L_g = \sup \{C_{\mathcal{G}}(L_g) \cap F_{\mathcal{G}}(L_g)\}.$$

By Theorem 6.1(ii) there exists a proper supervisor $\mathcal{S} = (S, \phi)$ such that $L_c(\mathcal{S}/\mathcal{G}) = L_g$. As demonstrated in the proof of Proposition 5.1, the state transition diagram for L_g (Fig. 11.3) can serve to define S ; it just remains to identify the state feedback map ϕ . For each state x of S , $\phi(x)$ is a map

$$\phi(x) : \{c_1, c_2\} \rightarrow \{0, 1\},$$

i.e., a binary evaluation of each of the controls c_1, c_2 . So, with reference to Fig. 11.3, it is enough to define

$$\phi(x)(c_1) = \begin{cases} 1 & \text{if an edge labeled } \beta_1 \text{ issues from } x, \\ 0 & \text{otherwise,} \end{cases}$$

and similarly for $\phi(x)(c_2)$. The resulting control patterns are tabulated in Fig. 11.4. The supervisor $\mathcal{S} = (S, \phi)$ then certainly determines

$$L(\mathcal{S}/\mathcal{G}) = \bar{L}_g, \quad L_m(\mathcal{S}/\mathcal{G}) = L_g.$$

We remark that in this example the alternative supervisor

$$\mathcal{S}^\circ = (S^\circ, \phi^\circ)$$

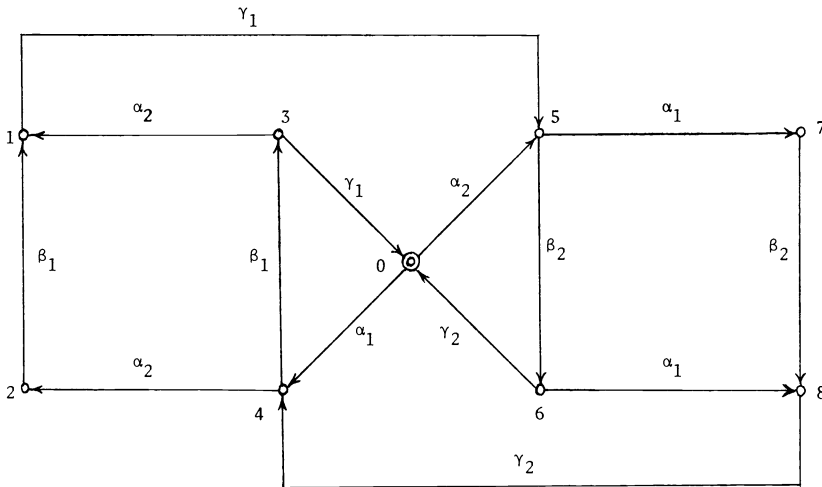


FIG. 11.3. Example 1: Recognizer for L_g .

⁸ Alternatively the generator of Fig. 11.3 could be taken as providing the definition of L_g .

State	x_0	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8
ϕ	00	00	10	00	10	01	00	01	00
ϕ°	00	10	10	10	10	01	01	01	01
π	x'_0	x'_1	x'_1	x'_2	x'_2	x'_3	x'_3	x'_4	x'_4

FIG. 11.4. Example 1: Control data for \mathcal{S} , \mathcal{S}° and \mathcal{S}' .

defined by setting $S^\circ = S$, and with ϕ° as tabulated in Fig. 11.4, determines exactly the same language controlled in \mathcal{G} as \mathcal{S} does, namely

$$L(\mathcal{S}^\circ/\mathcal{G}) = L(\mathcal{S}/\mathcal{G}) = \bar{L}_g.$$

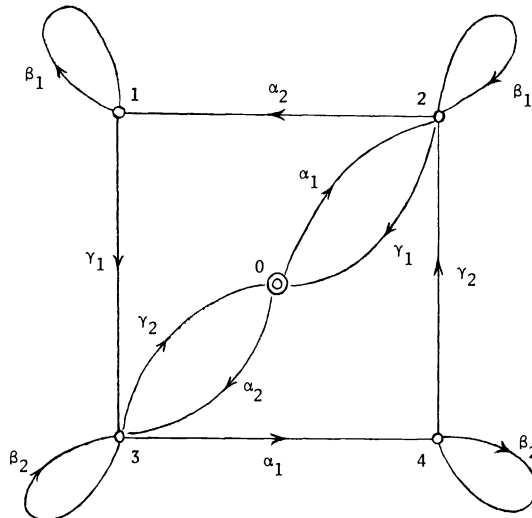
To verify this statement note that, for instance, states x_1, x_3 of S are entered only on the occurrence of the event β_1 ; but since β_1 can be immediately followed in \mathcal{G}_1 only by γ_1 , the enablement of β_1 by ϕ° in x_1 and x_3 can have no effect on the language controlled in \mathcal{G} .

It may be left to the reader to verify that \mathcal{S}° is complete with respect to \mathcal{G} . From \mathcal{S}° we construct a new supervisor $\mathcal{S}' = (S', \phi')$ and a projection $\pi: \mathcal{S}^\circ \rightarrow \mathcal{S}'$ as tabulated in Fig. 11.4; the result is displayed in Fig. 11.5. By Proposition 8.1

$$(L_m, L_c, L)(\mathcal{S}'/\mathcal{G}) = (L_m, L_c, L)(\mathcal{S}^\circ/\mathcal{G}),$$

namely control and marking action are preserved. The simplified supervisor \mathcal{S}' has just 5 states and is equivalent, in fact, to a queue (of maximum length 2) that stores events α in order of occurrence and is popped by the corresponding events γ .

12. Example 2. In a manufacturing system we consider two machines M_1, M_2 connected in tandem and separated by a buffer B (Fig. 12.1). Each machine M_i is modeled as a CDEP over the alphabet $\{\alpha_i, \beta_i, \lambda_i, \mu_i\}$ and having binary-valued controls $\{u_i, v_i\}$ (Fig. 12.2). The machine states are IDLE (I), WORKING (W) and DOWN (D).

FIG. 11.5. Example 1: Quotient supervisor \mathcal{S}' .

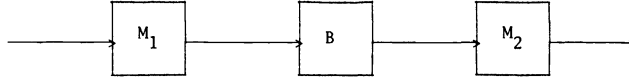


FIG. 12.1. Example 2: Machines coupled by a buffer.

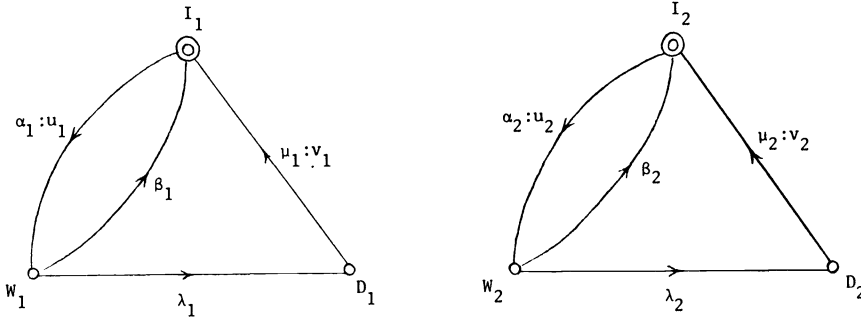


FIG. 12.2. Example 2: State diagrams of machines.

The control u enables/disables the transition from I to W ($u = 1$ allows M to “accept a workpiece”); while v enables/disables the transition from D to I ($v = 1$ means, when M is in state D , that M is “under repair”). The buffer B has one slot, i.e., is EMPTY (E) or FULL (F); it is not a CDEP but simply an automaton driven by M_1 and M_2 (Fig. 12.3). The system operates as follows. Machine M_1 takes a workpiece (event α_1), and either successfully completes processing and passes the workpiece to the buffer (event β_1); or breaks down and discards the workpiece (event λ_1), but in that case may later be repaired (event μ_1). Machine M_2 operates in the same way, but takes its workpiece from the buffer B , provided one is there.

The problem is to manipulate the controls in order to satisfy the four requirements stated informally below.

- (i) M_1 executes α_1 only if B is in E .
- (ii) M_2 executes α_2 only if B is in F (thereby driving B to E).
- (iii) M_1 cannot execute α_1 while M_2 is in D_2 .
- (iv) If M_1 is in D_1 and M_2 is in D_2 then $v_1 = 0$.

Condition (iv) means that if both machines are down then M_2 must be repaired before M_1 .

As in Example 1, we shall not formalize these requirements or present the details of how the legal language L_g is derived from them, but merely display the result. The language L_g that incorporates requirements (i)–(iv) with the system constraints is generated as shown in Fig. 12.4. The corresponding recognizer defines a supervisor \mathcal{S}° such that $L(\mathcal{S}^\circ/\mathcal{G}) = \bar{L}_g$; the control patterns are tabulated in Fig. 12.6. It can be verified that \mathcal{S}° admits the quotient \mathcal{S} displayed in Fig. 12.5; the required projection is also tabulated in Fig. 12.6. The quotient represents a reduction from 12 states to 6.

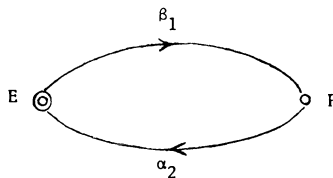


FIG. 12.3. Example 2: State diagram of buffer.

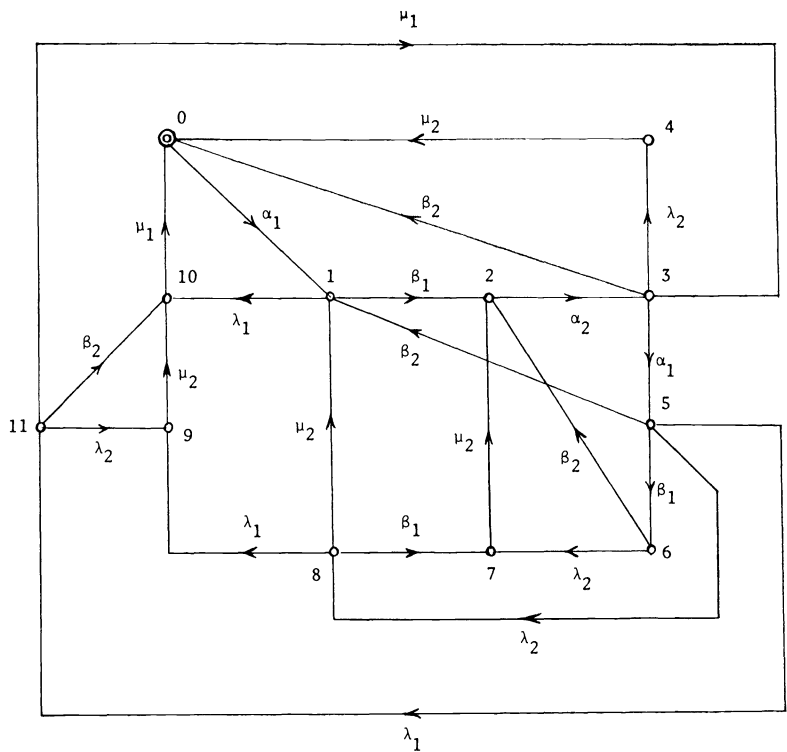


FIG. 12.4. Example 2: Recognizer for L_g .

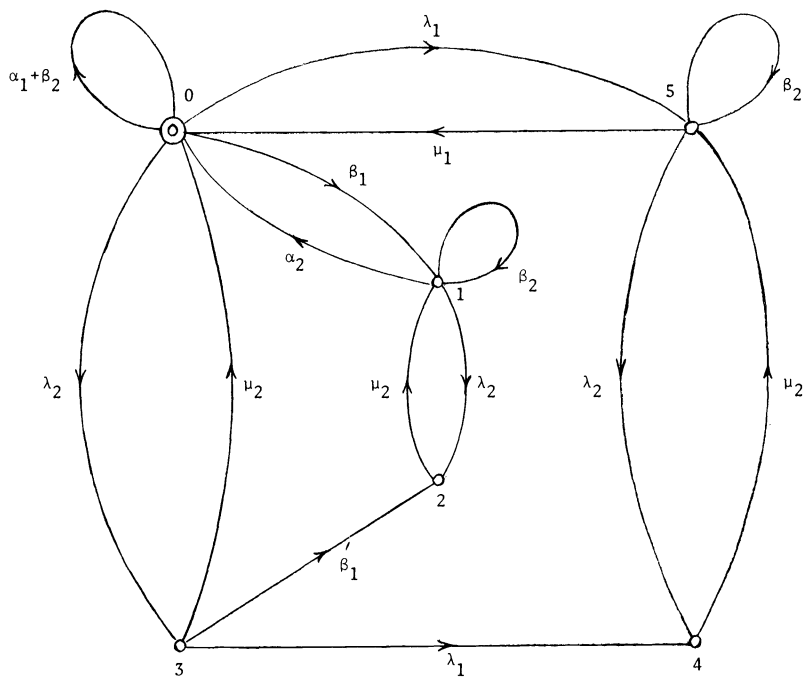


FIG. 12.5. Example 2: Quotient supervisor \mathcal{P} .

x^0	x	u_1	v_1	u_2	v_2
0	0	1	-	0	-
1	0	1*	-	0	-
2	1	0	-	1	-
3	0	1	-	0	-
4	3	0	-	0	1
5	0	1*	-	0	-
6	1	0	-	1*	-
7	2	0	-	-	1
8	3	0	-	0	1
9	4	0	0	0	1
10	5	-	1	0	-
11	5	-	1	0	-

FIG. 12.6. Example 2: Control data for \mathcal{S} and \mathcal{S}^0 . Assignments (*) are determined by consistency for the quotient; entries (-) may be assigned arbitrarily, consistent with the quotient.

As will be shown in a future article, it can actually be obtained directly from two modular “subsupervisors,” of which one is modeled on the buffer, and the other incorporates the logic of breakdown and repair.

13. Conclusion. In this article we have introduced a broad class of controlled discrete event processes together with some general concepts and results relating to their control or “supervision.” Our main conclusion, the Quotient Structure Theorem, is similar in spirit to the Internal Model Principle of regulator theory; it may be roughly paraphrased by saying that “supervisors must be modeled on the task to be accomplished.”

In future articles we shall discuss constructive methods for computing the supremal controllable (or closed controllable) sublanguage of a given language, as well as concrete methods for system specification and supervisor synthesis.

REFERENCES

- R. AVEYARD [1974], *A boolean model for a class of discrete event systems*, IEEE Trans. Syst. Man and Cyb., SMC-4, pp. 249-258.
- J. BEAUQUIER AND M. NIVAT [1980], *Application of formal language theory to problems of security and synchronization*, in Formal Language Theory—Perspective and Open Problems, R. V. Book, ed., Academic Press, New York, pp. 407-454.
- S. EILENBERG [1974], *Automata, Languages, and Machines*, Vol. A, Academic Press, New York.
- G. S. FISHMAN [1978], *Principles of Discrete Event Simulation*, John Wiley, New York.
- B. T. HAILPERN AND S. S. OWICKI [1983], *Modular verification of computer communication protocols*, IEEE Trans. Comm., COM-31, pp. 56-68.
- M. A. HARRISON [1965], *Introduction to Switching and Automata Theory*, McGraw-Hill, New York.
- C. A. R. HOARE [1983], *Notes on communicating sequential processes*, Tech. Monograph PRG-33, Programming Research Group, Oxford Univ. Computing Laboratory, Oxford.
- J. E. HOPCROFT AND J. D. ULLMAN [1979], *Introduction to Automata Theory, Languages, and Computation*. Addison-Wesley, Reading, MA.
- G. MILNE AND R. MILNER [1979], *Concurrent processes and their syntax*, J. Assoc. Comput. Mach., 26, pp. 302-321.
- H. NIJMEIJER [1983], *Nonlinear Multivariable Control: A Differential Geometric Approach*, thesis, Rijksuniversiteit Groningen, the Netherlands.
- D. PARK [1981], *Concurrency and automata on infinite sequences*, in Theoretical Computer Science, Lecture Notes in Computer Science 104, Springer-Verlag, New York, pp. 167-183.
- J. L. PETERSON [1981], *Petri Net Theory and the Modeling of Systems*, Prentice-Hall, Englewood Cliffs, NJ.

- A. PNUELI [1979], *The temporal semantics of concurrent programs*, in Semantics of Concurrent Computation, Lecture Notes in Computer Science 70, Springer-Verlag, New York, pp. 1-20.
- P. J. RAMADGE [1983], *Control and Supervision of Discrete Event Processes*, Ph.D. thesis, Dept. Electrical Engineering, Univ. Toronto, Toronto, Ontario.
- P. J. RAMADGE AND W. M. WONHAM [1982a], *Supervisory control of discrete event processes*, in Feedback Control of Linear and Nonlinear Systems, Lecture Notes in Control and Information Sciences 39, Springer-Verlag, New York, pp. 202-214.
- , [1982b], *Supervision of discrete event processes*, Proc. 21st IEEE Conference on Decision and Control, December, pp. 1228-1229.
- , [1984], *Supervisory control of a class of discrete event processes*, Proc. Sixth International Conference Analysis and Optimization of Systems, Nice, June 1984, in Analysis and Optimization of Systems, A. Bensoussan and J. L. Lions, eds., Lecture Notes in Computer and Information Science 63, Springer-Verlag, New York, 1984, Part 2, pp. 477-498.
- A. C. SHAW [1978], *Software descriptions with flow expressions*, IEEE Trans. Software Engrg., SE-4 (3), pp. 242-254.
- M. W. SHIELDS [1979], *COSY train journeys*, Rpt. ASM/67, Computing Laboratory, Univ. Newcastle-upon-Tyne.
- M. STEENSTRUP, M. A. ARBIB AND E. G. MANES [1981], *Port automata and the algebra of concurrent processes*, Computer and Information Science Tech. Rpt. 81-25, Univ. Massachusetts, Amherst, MA.
- G. SZÁSZ [1963], *Introduction to Lattice Theory*, Academic Press, New York.
- W. M. WONHAM [1979], *Linear Multivariable Control: A Geometric Approach*, sec. ed., Springer-Verlag, New York.
- B. P. ZEIGLER [1984], *Multifaceted Modelling and Discrete Event Simulation*, Academic Press, New York.

A SIMULTANEOUS ITERATIVE METHOD FOR COMPUTING PROJECTIONS ON POLYHEDRA*

ALFREDO N. IUSEM[†] AND ALVARO R. DE PIERRO[‡]

Abstract. A simultaneous version of Hildreth's iterative algorithm for norm minimization over linear inequalities is presented. Proofs are given showing that the algorithm converges from any starting point to its projection on the linear constraints set in the feasible case and to the nearest least squares solution in the general case.

Key words. projection methods, least squares, quadratic programming

AMS(MOS) subject classifications. 52A05, 65F10, 90C25

1. Introduction. Linearly constrained quadratic optimization problems derived from computing the least element of polyhedra appear in various fields of application and it is not rare to encounter large-scale or even huge-scale problems. Usually the matrix describing the constraints will be sparse, but all too often no special structure pattern is detectable in it. In such cases row-action methods are frequently used. A row action method is an iterative procedure which requires in each step only the current point and one row of the matrix, and performs no transformation on the matrix elements, see Censor [3]. Image reconstruction from projections is an important application of these types of methods, see, e.g., Gordon and Herman [8], Herman and Lent [9].

Hildreth's quadratic programming procedure [11] is a row-action method whose capabilities for solving large-scale problems were numerically demonstrated by Herman and Lent [10].

In order to describe the geometry behind Hildreth's algorithm and our modified version, we start by commenting briefly on the old algorithms by Kaczmarz [12] and Cimmino [4].

These methods were proposed for finding a feasible solution for a system of linear equations, but they can be easily modified for systems of linear inequalities. In such a case, they can be described as follows: Kaczmarz's approach, as presented by Agmon [1] for the inequality case, starts with an arbitrary point and generates the iterates by using the inequalities in a cyclic way and taking as next iterate the orthogonal projection of the current iterate onto the halfspace defined by the current inequality (projecting a point onto a halfspace means projecting it onto the associated hyperplane if the point lies outside the half space and leaving it unmodified otherwise). In Cimmino's method, on the other hand, the current iterate is projected onto all the half spaces and the new iterate is a convex combination of such projections. When the feasible set is nonempty, both algorithms converge from any starting point. Cimmino's method is convergent also in the infeasible case (see [6] and [15]).

Hildreth's method can be seen as a modification of Kaczmarz-Agmon's procedure for the case when the required solution is not just a feasible point but a norm minimizer one. In this case, when the current iterate is outside the halfspace corresponding to the current inequality, the next iterate is its orthogonal projection onto the associated

* Received by the editors January 10, 1984, and in revised form December 6, 1985.

[†] Instituto de Matemática Pura e Aplicada, Estrada Dona Castorina, 110, Rio de Janeiro, Brasil.

[‡] Instituto de Matemática, Universidade Federal do Rio de Janeiro CEP 68530, Rio de Janeiro, Brasil.

The work of this author was partially supported by CNPq, under grant 301699/81.

hyperplane, as in Kaczmarz's method. But when the current point satisfies the constraint the next iterate is its projection onto a hyperplane parallel to the hyperplane defined by the inequality, lying between it and the point itself. The location of such a parallel hyperplane is determined, at each step, by a sequence of dual variables. This sequence guarantees that at each step the current iterate is the orthogonal projection of the starting point onto a perturbed feasible set, defined by translates of the original hyperplanes (i.e. the perturbation acts only upon the right-hand side). When the feasible set is nonempty this sequence of perturbed right-hand sides converges to the original one, and the intermediate minimization property is preserved in the limit.

A proof of convergence can be found in Lent and Censor [13], who give a compact version of Hildreth's method with relaxation and almost-cyclic (instead of cyclic) utilization of the constraints.

Another extension of Hildreth's method was proposed by Bregman [2], who considers the case of minimizing nonquadratic functions. By substituting the orthogonal projections (both onto the original and the perturbed hyperplanes) by the so-called "Bregman projections," he defines an algorithm which converges for a wide class of functions. In [7] it is shown that all strictly convex functions which tend to infinity faster than linearly (i.e. such that $\lim_{\|x\| \rightarrow \infty} (f(x)/\|x\|) = \infty$) belong to that class.

Another view of Hildreth's method is obtained by looking at the sequence of dual variables. In the dual space, Hildreth's method is equivalent to the Successive Over-Relaxation (SOR) algorithm applied to a linear complementarity problem, where the matrix AA' substitutes for the original matrix A and the right-hand side is the original one. See [14] for a description of SOR methods applied to this kind of problem.

In this paper we propose a Cimmino-like relaxed version of Hildreth's method. Each iterate is orthogonally projected onto all the hyperplanes (or its parallel translates, when the constraint is satisfied) and the new iterate is a relaxed convex combination of such projections. Such implementation has two features which make it different from Censor and Lent's version. First, it is appropriate for parallel processing computers, since all the projections of the current iterate can be calculated simultaneously (that is why we call our method "simultaneous"). Second, the convergence of Cimmino's method in the infeasible case is preserved. In our case, the primal sequence converges to a norm minimizing least squares solution, i.e. to the orthogonal projection of the starting point onto the set of points which minimize the sum of the squares of the distances to the halfspaces associated with the inequalities. When the system is feasible, this is just the orthogonal projection of the starting point onto the feasible region.

Convergence in the infeasible case to a well characterized point is an interesting property, since in the applications the feasible set may be empty due to errors associated with the data gathering process, for instance.

At the same time, consideration of the infeasible case produces a rather involved proof, since no easy argument based on Fejér convergence seems available. Also, due to the primal-dual character of the method, we were not able to find a Lyapunov function for the algorithm.

In § 2 we present the algorithm and show some immediate properties, like the linear complementarity between primal and dual variables and the intermediate minimization property of the iterates.

In § 3 we show that the difference between consecutive iterates tends to zero, and that the sequence of the location parameters for the perturbed hyperplanes is bounded. From there, it follows that the sequence of perturbed right-hand sides and finally the sequence of primal iterates are bounded, i.e. they all have convergent subsequences.

In § 4 we prove that all such sequences are indeed convergent; in particular, the sequence of primal iterates converges to the projection of the starting point onto the limit of the perturbed sets, i.e. the feasible set for the system whose right-hand side is the limit of the perturbed right-hand sides. In § 5 such set is identified as the set of weighted least squares solutions if the system is infeasible, or the original feasible set otherwise.

2. The algorithm. Consider the system of inequalities

$$(1) \quad Ax \leq b$$

where A is an $m \times n$ matrix, x and b are n and m vectors respectively.

We introduce first some notation which will be used throughout the paper. If a_i are the rows of A ($a_i \neq 0$ for $1 \leq i \leq m$), let

$$H_i = \{x: \langle a_i, x \rangle = b_i\},$$

$$C_i = \{x: \langle a_i, x \rangle \leq b_i\},$$

$$C = \{x: Ax \leq b\} = \bigcap_{i=1}^m C_i.$$

Let P_i and Q_i be the orthogonal projections onto C_i and H_i , respectively,

$$(2) \quad P_i x = x + \min \left\{ 0, \frac{b_i - \langle a_i, x \rangle}{\|a_i\|^2} \right\} a_i,$$

$$(3) \quad Q_i x = x + \frac{b_i - \langle a_i, x \rangle}{\|a_i\|^2} a_i,$$

where $\langle \cdot, \cdot \rangle$ and $\|\cdot\|$ denote the Euclidean inner product and norm, respectively.

Now the algorithm is defined. Let α be a real number in the open interval $(0, 2)$ and $\lambda_1, \dots, \lambda_m$ be real positive numbers such that $\sum_{i=1}^m \lambda_i = 1$. Take $z^0 = 0$ and $x^0 \in \mathbb{R}^n$ arbitrary. Define

$$(4) \quad x^{k+1} = x^k + \alpha \left(\sum_{i=1}^m \lambda_i c_i^k a_i \right),$$

$$(5) \quad z_i^{k+1} = z_i^k - \lambda_i c_i^k \quad (1 \leq i \leq m),$$

$$(6) \quad c_i^k = \min \left\{ \frac{z_i^k}{\lambda_i}, \frac{b_i - \langle a_i, x^k \rangle}{\|a_i\|^2} \right\}.$$

Next we establish some facts which describe the geometry of the algorithm defined by (4)-(6).

PROPOSITION 1 (nonnegativity of the dual variables). *For any k , $z_i^k \geq 0$ ($1 \leq i \leq m$).*

Proof. True for $k=0$ by definition. From (6) $c_i^k \leq z_i^k / \lambda_i$; then

$$z_i^{k+1} = z_i^k - \lambda_i c_i^k \geq z_i^k - z_i^k = 0. \quad \square$$

Define

$$w_i^k = x^k + c_i^k a_i \quad (1 \leq i \leq m),$$

$$d_i^k = \frac{b_i - \langle a_i, x^k \rangle}{\|a_i\|^2} \quad (1 \leq i \leq m).$$

The w_i^k 's are the "modified projections" whose relaxed convex combination gives the next iterate, i.e.,

$$x^{k+1} = (1 - \alpha)x^k + \alpha \left(\sum_{i=1}^m \lambda_i w_i^k \right).$$

The next proposition shows that w_i^k is the orthogonal projection of x^k onto H_i if $x^k \notin C_i$ or a point in the segment between x^k and its projection onto H_i if $x^k \in C_i$.

PROPOSITION 2. *For all i , $1 \leq i \leq m$ and for all $k \geq 0$, the following statements hold:*

(i) *If $x^k \notin C_i$, then $w_i^k = P_i x^k = Q_i x^k$ and $d_i^k = c_i^k < 0$.*

(ii) *If $x^k \in C_i$, then $0 \leq c_i^k \leq d_i^k$.*

(iii) *$w_i^k \in C_i$.*

(iv) *If $c_i^k \neq d_i^k$, then $c_i^{k+1} \leq 0$.*

Proof. (i) If $x^k \notin C_i$, then $d_i^k < 0$. Since $z_i^k \geq 0$ (Proposition 1) $c_i^k = d_i^k$. From the definitions of w_i^k , P_i and Q_i we get $w_i^k = P_i x^k = Q_i x^k$.

(ii) If $x^k \in C_i$, then $d_i^k \geq 0$ and by Proposition 1, $0 \leq c_i^k \leq d_i^k$.

(iii) Obviously if $c_i^k = d_i^k$. Otherwise $c_i^k = z_i^k / \lambda_i \geq 0$ and $c_i^k < d_i^k$. So

$$\langle a_i, w_i^k \rangle = \langle a_i, x^k \rangle + c_i^k \|a_i\|^2 \leq \langle a_i, x^k \rangle + b_i - \langle a_i, x^k \rangle = b_i.$$

(iv) If $c_i^k \neq d_i^k$, then $c_i^k = z_i^k / \lambda_i \Rightarrow z_i^{k+1} = 0 \Rightarrow c_i^{k+1} \leq 0$. \square

The next lemma is the starting point for our convergence proof, and justifies our calling w_i^k a "modified projection" of x^k : it holds trivially when w_i^k is the projection of x^k onto C_i (case (i) of the proof).

LEMMA 1. $\|x^{k+1} - w_i^{k+1}\| \leq \|x^{k+1} - w_i^k\|$.

Proof. We consider all possible cases for x^{k+1} , w_i^{k+1} , w_i^k .

(i) If $x^{k+1} \notin C_i$, then $w_i^{k+1} = P_i x^{k+1}$ (Proposition 2(ii)). The inequality holds because $P_i x^{k+1}$ is the closest point to x^{k+1} in C_i and $w_i^k \in C_i$ (Proposition 2(iii)).

(ii) If $x^{k+1} \in C_i$ and $w_i^k = Q_i x^k$, then

$$(7) \quad \|x^{k+1} - w_i^k\| = \|x^{k+1} - Q_i x^k\| \geq \|Q_i x^{k+1} - x^{k+1}\|$$

using the closest point property for the projection Q_i . On the other hand $x^{k+1} \in C_i$, so

$$(8) \quad c_i^{k+1} \geq 0 \Rightarrow \|x^{k+1} - w_i^{k+1}\| = c_i^{k+1} \|a_i\| \leq d_i^{k+1} \|a_i\| = \|x^{k+1} - Q_i x^{k+1}\|$$

(Proposition 2(ii)). The result follows from (7) and (8).

(iii) The only remaining case is $x^{k+1} \in C_i$, $w_i^k \neq Q_i x^k$; in this case $w_i^k \neq Q_i x^k \Rightarrow c_i^k \neq d_i^k \Rightarrow c_i^{k+1} = 0$ (Proposition 2(ii) and (iv)) $\Rightarrow x^{k+1} = w_i^{k+1}$. \square

Let

$$\sigma_k = \sum_{i=1}^m \lambda_i (c_i^k)^2 \|a_i\|^2.$$

LEMMA 2.

$$\sigma_{k+1} \leq \sigma_k - \left(\frac{2}{\alpha} - 1 \right) \|x^{k+1} - x^k\|^2.$$

Proof. From Lemma 1,

$$\|x^{k+1} - w_i^{k+1}\|^2 \leq \|x^{k+1} - w_i^k\|^2 = \|x^{k+1} - x^k\|^2 + \|x^k - w_i^k\|^2 - 2\langle x^{k+1} - x^k, w_i^k - x^k \rangle.$$

Multiplying by λ_i and summing on i , (remember that $\sum_{i=1}^m \lambda_i = 1$)

$$\begin{aligned} \sigma_{k+1} &\leq \sigma_k + \|x^{k+1} - x^k\|^2 - 2 \left\langle x^{k+1} - x^k, \sum_{i=1}^m \lambda_i w_i^k - x^k \right\rangle \\ &= \sigma_k + \|x^{k+1} - x^k\|^2 - \frac{2}{\alpha} \|x^{k+1} - x^k\|^2 = \sigma_k - \left(\frac{2}{\alpha} - 1 \right) \|x^{k+1} - x^k\|^2. \quad \square \end{aligned}$$

Now we look at the dual variables z_i^k . The name “dual variables” is justified in the following two propositions.

We define recursively

$$(9) \quad q_i^0 = 0 \quad (1 \leq i \leq m),$$

$$(10) \quad q_i^{k+1} = -c_i^k \|a_i\|^2 + b_i - \langle a_i, x^k \rangle \quad (1 \leq i \leq m).$$

PROPOSITION 3 (Linear complementarity). *For any k , $q_i^k z_i^k = 0$ ($1 \leq i \leq m$).*

Proof. By definition $q_i^{k+1} = -c_i^k \|a_i\|^2 + b_i - \langle a_i, x^k \rangle$ and $z_i^{k+1} = z_i^k - \lambda_i c_i^k$. If $z_i^{k+1} > 0$, $c_i^k \neq z_i^k / \lambda_i \Rightarrow c_i^k = d_i^k \Rightarrow q_i^{k+1} = 0$. On the other hand if $q_i^{k+1} > 0$, $c_i^k \neq d_i^k \Rightarrow c_i^k = z_i^k / \lambda_i \Rightarrow z_i^{k+1} = 0$. \square

PROPOSITION 4. $x^k = x^0 - \alpha A^t z^k$, where “ t ” denotes matrix transposition.

Proof. By induction

$$x^1 = x^0 + \alpha \sum_{i=1}^m \lambda_i c_i^0 a_i = x^0 - \alpha A^t (z^1 - z^0) = x^0 - \alpha A^t z^1.$$

Assuming $x^k = x^0 - \alpha A^t z^k$, then

$$x^{k+1} = x^k + \alpha \sum_{i=1}^m \lambda_i c_i^k a_i = x^k - \alpha A^t (z^{k+1} - z^k),$$

from which the result follows. \square

These two propositions allow us to prove that the algorithm preserves the intermediate optimality property of Hildreth’s algorithm presented in [13].

Define the vector b^k (perturbed right-hand side) as

$$b^k = q^k + Ax^k$$

where q^k are the vectors with components defined by (9) and (10). Construct a sequence of perturbed constraint sets S^k by

$$S^k = \{x: Ax \leq b^k\}.$$

PROPOSITION 5.

$$(i) \quad q_i^k \geq 0 \quad (1 \leq i \leq m).$$

$$(ii) \quad x^k \in S^k.$$

Proof. (i) follows from (6), (9), (10) and (ii) is immediate from part (i) and the definition of S^k . \square

THEOREM 1. x^k is the solution to

$$\min \frac{1}{2} \|x - x^0\|^2 \quad \text{s.t. } x \in S^k.$$

Proof. Since the minimand is convex, the Kuhn–Tucker conditions are sufficient for optimality. They are

$$(11) \quad x - x^0 + A^t \mu = 0,$$

$$(12) \quad \mu_i (\langle a_i, x \rangle - b_i^k) = 0,$$

$$(13) \quad \mu \geq 0,$$

$$(14) \quad Ax \leq b^k.$$

Take $x = x^k$, $\mu = \alpha z^k$. Equation (14) is satisfied by Proposition 5(ii); (11) follows from Proposition 4, (12) from Proposition 3, observing that $Ax^k - b^k = -q^k$, and (13) from Proposition 1. \square

3. Convergence of a subsequence. In order to establish the convergence of the algorithm, we will prove that the sequences $\{c^k\}$, $\{b^k\}$ and $\{x^k\}$ are convergent, with

the limit of the last one being the projection of x^0 onto the feasible set of a perturbed system whose right-hand side is the limit of the sequence $\{b^k\}$. As a first step, we will conclude from Lemma 2 that these sequences are bounded, so they have convergent subsequences.

LEMMA 3. (i) *The sequence $\{c^k\} \subset \mathbb{R}^m$ has a convergent subsequence $c^{k_j} \xrightarrow{j \rightarrow \infty} v$.*

$$(ii) \quad x^{k+1} - x^k \xrightarrow{k \rightarrow \infty} 0.$$

$$(iii) \quad \sum_{i=1}^m \lambda_i c_i^k a_i \xrightarrow{k \rightarrow \infty} 0.$$

$$(iv) \quad \sum_{i=1}^m \lambda_i v_i a_i = 0.$$

Proof. (i) Since $0 < \alpha < 2$, Lemma 2 implies that σ_k is a positive decreasing sequence, hence convergent. It follows that $\{c^k\}$ is bounded. So it has a convergent subsequence.

(ii) $\|x^{k+1} - x^k\|^2 \leq (\alpha/(2-\alpha))(\sigma_k - \sigma_{k+1})$ from Lemma 2. Since $\{\sigma_k\}$ is convergent and $\alpha < 2$, the right-hand side tends to zero.

(iii) From (4)

$$\left\| \sum_{i=1}^m \lambda_i c_i^k a_i \right\|^2 = \frac{1}{\alpha^2} \|x^{k+1} - x^k\|^2 \leq \frac{1}{\alpha(2-\alpha)} (\sigma_k - \sigma_{k+1}) \rightarrow 0.$$

(iv) From (iii), since v is the limit of a subsequence of $\{c^k\}$. \square

From now on, $\{c^{k_j}\}$ will be a convergent subsequence guaranteed by Lemma 3 and v its limit. Let y^* be the vector with components defined by

$$y_i^* = -v_i \|a_i\|^2 \quad (1 \leq i \leq m),$$

and S the set defined by

$$S = \{x: Ax \leq b + y^*\}.$$

PROPOSITION 6.

$$(i) \quad b^{k_j} \xrightarrow{j \rightarrow \infty} b + y^*.$$

$$(ii) \quad S \neq \emptyset.$$

Proof. (i) $b_i^{k_j+1} = -c_i^{k_j} \|a_i\|^2 + b_i - \langle a_i, x^{k_j+1} - x^{k_j} \rangle \xrightarrow{j \rightarrow \infty} y_i^* + b_i$ by Lemma 3(i) and (ii).

(ii) In Linear Programming jargon, since $S^{k_j} \neq \emptyset$ for all j , we have a feasible basis B^{k_j} . The number of bases is finite, so there is one which is feasible for an infinite number of j 's, and hence for the limit of the b^{k_j} 's, which is $b + y^*$. \square

Now that the convergence of the subsequences $\{c^{k_j}\}$ and $\{b^{k_j}\}$ is established, we look at the subsequence $\{x^{k_j}\}$. We will show that it converges to the projection of the starting point x^0 onto S . Observe that S^k and S are convex and so the projections of any point x onto them are well defined (as the closest point to x in the respective sets). Let x^* be the projection of x^0 onto S , \hat{x}^k the projection of x^k onto S and \tilde{x}^k the projection of x^* onto S^k . We will use the following result due to Daniel [5].

THEOREM 2. Let $T = \{u: Au \leq p\}$ and $T' = \{u: Au \leq p'\}$. If T and T' are nonempty, there exists $\gamma \in \mathbb{R}$ depending only on A such that for any $u \in T$ there is a $u' \in T'$ which satisfies

$$\|u - u'\| \leq \gamma \|(p - p')^+\|$$

where the upper plus notation means, for any vector v , $(v^+)_i \equiv \max(0, v_i)$.

Proof. See [5]. \square

LEMMA 4.

- (i) $\|\hat{x}^{k_j} - x^{k_j}\| \xrightarrow{j \rightarrow \infty} 0.$
- (ii) $\|x^{k_j} - x^0\| \xrightarrow{j \rightarrow \infty} \|x^* - x^0\|.$
- (iii) $\|\hat{x}^{k_j} - x^0\| \xrightarrow{j \rightarrow \infty} \|x^* - x^0\|.$

Proof. (i) Use Theorem 2 with $T' = S$, $T = S^{k_j}$, $u = x^{k_j}$. Since \hat{x}^{k_j} is the projection of x^{k_j} onto S , and $x^{k_j} \in S^{k_j}$, if u' is the vector resulting from Theorem 1, $\|\hat{x}^{k_j} - x^{k_j}\| \leq \|x^{k_j} - u'\| \leq \gamma \|(b^{k_j} - b - y^*)^+\| \xrightarrow{j \rightarrow \infty} 0$ by virtue of Proposition 6(i).

(ii) Since $\hat{x}^{k_j} \in S$ and x^* is the projection of x^0 onto S

$$\|x^* - x^0\| \leq \|\hat{x}^{k_j} - x^0\| \leq \|\hat{x}^{k_j} - x^{k_j}\| + \|x^{k_j} - x^0\|.$$

Then for any $\varepsilon > 0$ there exists j_1 such that for $j \geq j_1$

$$(15) \quad \|x^* - x^0\| \leq \|x^{k_j} - x^0\| + \varepsilon \quad \text{because of part (i).}$$

By Theorem 1, $\|x^{k_j} - x^0\| \leq \|x - x^0\| \quad \forall x \in S^{k_j}$. Then

$$\|x^{k_j} - x^0\| \leq \|\tilde{x}^{k_j} - x^0\| \leq \|\tilde{x}^{k_j} - x^*\| + \|x^* - x^0\| \quad \text{since } \tilde{x}^{k_j} \in S^{k_j}.$$

Use Theorem 2 with $T = S$, $T' = S^{k_j}$, $u = x^*$. If u' is the vector in T' guaranteed by the theorem, since \tilde{x}^{k_j} is the projection of x^* onto S^{k_j} , we have $\|\tilde{x}^{k_j} - x^*\| \leq \|u' - x^*\| \leq \gamma \|(b - b^{k_j} + y^*)^+\| \xrightarrow{j \rightarrow \infty} 0$ by Proposition 6(i). So

$$(16) \quad \|x^{k_j} - x^0\| \leq \|x^* - x^0\| + \varepsilon \quad \text{for } j \geq j_2.$$

From (15) and (16) we have that for $j \geq \max\{j_1, j_2\}$

$$\| \|x^{k_j} - x^0\| - \|x^* - x^0\| \| < \varepsilon.$$

(iii) Immediate from (i) and (ii). \square

THEOREM 3. $x^{k_j} \xrightarrow{j \rightarrow \infty} x^*$.

Proof. $\|x^{k_j} - x^*\| \leq \|x^{k_j} - \hat{x}^{k_j}\| + \|\hat{x}^{k_j} - x^*\|.$

The first term of the right-hand side tends to zero by Lemma 4(i). Since $\|\hat{x}^{k_j} - x^0\| \xrightarrow{j \rightarrow \infty} \|x^* - x^0\|$, x^* is the unique closest point to x^0 in S and $\hat{x}^{k_j} \in S$, we conclude that $\hat{x}^{k_j} \xrightarrow{j \rightarrow \infty} x^* \Rightarrow \|\hat{x}^{k_j} - x^*\| \xrightarrow{j \rightarrow \infty} 0$. So $x^{k_j} \xrightarrow{j \rightarrow \infty} x^*$. \square

4. Convergence of the sequence. We will prove now that x^* , the projection of x^0 onto S , is the limit of the whole sequence. We will need four technical lemmas. The complication arises from the fact that x^* may be either inside or outside each of the half spaces and the behaviour of the sequence is different in each case, so that

consideration of each alternative is required. First we show that the difference between consecutive elements of the sequence $\{c^k\}$ tends to 0.

LEMMA 5. $c_i^{k+1} - c_i^k \xrightarrow[k \rightarrow \infty]{} 0$ for $1 \leq i \leq m$.

Proof. From Lemma 3(ii),

$$(17) \quad x^k - x^{k+1} \xrightarrow[k \rightarrow \infty]{} 0.$$

By continuity of the inner product

$$(18) \quad d_i^k - d_i^{k+1} \xrightarrow[k \rightarrow \infty]{} 0 \quad \text{for } 1 \leq j \leq m.$$

Consider the norm defined, for any $x \in \mathbb{R}^n$, as

$$\|x\|_*^2 = \sum_{i=1}^m \lambda_i \|a_i\|^2 x_i^2.$$

From Lemma 2 $\sigma_k = \|c^k\|_*^2$ converges. Since all norms are equivalent in finite dimension $\sum_{i=1}^m |c_i^k|$ converges, so

$$(19) \quad \sum_{i=1}^m (|c_i^k| - |c_i^{k+1}|) \xrightarrow[k \rightarrow \infty]{} 0.$$

Take k_0 big enough so that $\|x^k - x^{k+1}\|$, $|d_i^k - d_i^{k+1}|$ ($i = 1, \dots, m$) and $\|c_i^k| - |c_i^{k+1}||$ ($i = 1, \dots, m$) are less than a given ε for $k > k_0$. Now we consider three cases according to the location of x^k and x^{k+1} with respect to the hyperplane H_i . Let

$$I_1 = \{i: x^k \in C_i \text{ and } x^{k+1} \in C_i\},$$

$$I_2 = \{i: x^k \notin C_i \text{ and } x^{k+1} \notin C_i\},$$

$$I_3 = \{1, \dots, m\} - I_1 - I_2.$$

(i) For $i \in I_2$, $c_i^{k+1} = d_i^{k+1} < 0$, $c_i^k = d_i^k < 0$ from Proposition 2(i). So

$$(20) \quad |c_i^k - c_i^{k+1}| = |d_i^k - d_i^{k+1}| < \varepsilon.$$

Also

$$(21) \quad -\varepsilon < |c_i^k| - |c_i^{k+1}| < \varepsilon.$$

(ii) For $i \in I_3$, since x^k and x^{k+1} are on different sides of H_b we use Proposition 2(i) and (ii) to get

$$(22) \quad |c_i^k| \leq |d_i^k| \leq \frac{\|x^k - Q_i x^k\|}{\|a_i\|} \leq \frac{\|x^k - x^{k+1}\|}{\|a_i\|} \leq \frac{\varepsilon}{\|a_i\|},$$

$$(23) \quad |c_i^{k+1}| \leq |d_i^{k+1}| \leq \frac{\|x^{k+1} - Q_i x^{k+1}\|}{\|a_i\|} \leq \frac{\|x^k - x^{k+1}\|}{\|a_i\|} \leq \frac{\varepsilon}{\|a_i\|}.$$

(iii) From (22) and (23) both $c_i^k - c_i^{k+1}$ and $|c_i^k| - |c_i^{k+1}|$ are arbitrarily small for big enough k and $i \in I_3$. So, using Proposition 2(ii), (19) and (21):

$$(24) \quad \sum_{i \in I_1} (c_i^k - c_i^{k+1}) = \sum_{i \in I_1} (|c_i^k| - |c_i^{k+1}|) \xrightarrow[k \rightarrow \infty]{} 0.$$

But for $i \in I_1$, either $c_i^k \neq d_i^k \Rightarrow c_i^k - c_i^{k+1} \geq 0$ (from Proposition 2(ii) and (iv)) or

$$(25) \quad c_i^k = d_i^k \Rightarrow c_i^k - c_i^{k+1} \geq d_i^k - d_i^{k+1}.$$

Since the right-hand side of (25) tends to 0, it follows from (24) that each term $c_i^k - c_i^{k+1}$ is arbitrarily small for big enough k . \square

We use Lemma 5 to prove some relations between $x^* = \lim_j x^{k_j}$ and $v = \lim_j c^{k_j}$. Let $d^* = (b_i - \langle a_i, x^* \rangle) / \|a_i\|^2$ and $D_i = \{x \in \mathbb{R}^n : \langle a_i, x \rangle < b_i\}$.

LEMMA 6.

- (i) If $x^* \notin C_i$, then $v_i = d_i^* < 0$.
- (ii) If $x^* \in H_i$, then $v_i = 0$.
- (iii) If $x^* \in D_i$, then either $v_i = 0$ or $v_i = d_i^* > 0$.

Proof. (i) Since $x^* \notin C_i$, $d_i^* < 0$ and $x^{k_j} \notin C_i$ for big enough j . So $c_i^{k_j} = d_i^{k_j} < 0$.

Taking limits as $j \rightarrow \infty$ on both sides the result follows.

(ii) Since $x^* \in H_i$, $|d_i^{k_j}| < \varepsilon$ for big enough j and any ε . But by definition $|c_i^{k_j}| \leq |d_i^{k_j}| \Rightarrow$

$$0 = \lim_{j \rightarrow \infty} c_i^{k_j} = v_i.$$

(iii) Since $x^* \in D_i$, $x^{k_j} \in C_i$ for big enough $j \Rightarrow c_i^{k_j} \geq 0 \Rightarrow v_i \geq 0$. Assume $v_i > 0$. We claim that

$$(26) \quad c_i^{k_j} = d_i^{k_j}.$$

Otherwise, by Proposition 2(iv), $c_i^{k_j} \leq 0$; but from Lemma 5 and the definition of v_i , $c_i^{k_j+1}$ is arbitrarily close to $v_i > 0$ for big enough j . Take limits on both sides of (26). \square

Let now $I = \{i : x^* \in D_i, v_i = 0\}$, $J = \{i : x^* \in D_i, v_i \neq 0\}$ (later on it will be shown that $J = \emptyset$).

LEMMA 7. *There exists s such that for $j > s$*

- (a) $c_i^{k_j} = d_i^{k_j}$ for $i \in J$,
- (b) $c_i^{k_j} = 0$ for $i \in I$.

Proof. (a) It follows from (26).

(b) From Lemma 3(i) and Theorem 4, x^{k_j-1} gets arbitrarily close to x^* for big enough j , so $d_i^{k_j-1}$ is arbitrarily close to d_i^* , since $x^* \in D_i$. On the other hand, $c_i^{k_j-1}$ is arbitrarily close to $v_i = 0$ by Lemma 5 and the definition of v_i . It follows that $c_i^{k_j-1} \neq d_i^{k_j-1} \Rightarrow c_i^{k_j} \leq 0$ (Proposition 2(iv)). But $x^* \in D_i \Rightarrow x^{k_j} \in D_i \Rightarrow c_i^{k_j} \geq 0$. So $c_i^{k_j} = 0$ for big enough j . \square

The following lemma is the final piece of the convergence proof. It requires the following elementary proposition, which results from the fact that projections are contractions.

PROPOSITION 7. (i) For any $x, y \in \mathbb{R}^n$, $0 \leq \beta < 2$, $1 \leq i \leq m$,

$$\|(1-\beta)(x-y) + \beta(Q_i x - Q_i y)\| \leq \|x-y\|.$$

(ii) For any $x \in \mathbb{R}^n$, $y \in H_i$, $0 < \beta < 2$, $1 \leq i \leq m$,

$$\|(1-\beta)x + \beta Q_i(x-y)\| \leq \|x-y\|.$$

Proof. (i) It follows easily from (3), expanding the square of the norm.

(ii) From (i), since for $y \in H_i$, $Q_i y = y$. \square

Define now $\varepsilon = \frac{1}{4} \min_i \{\|v_i\| \|a_i\| : v_i \neq 0\}$. Use Lemmas 3(ii) and 5 to find r such that $\|x^{k+1} - x^k\| < \varepsilon$, $|c_i^{k+1} - c_i^k| < \varepsilon / \|a_i\|$ for $k > r$.

LEMMA 8. *Take $k > r$. If x^k generated by the algorithm (4)–(6) satisfies:*

- (a) $\|x^k - x^*\| < \varepsilon$,
- (b) $c_i^k = 0$ for $i \in I$,
- (c) $c_i^k = d_i^k$ for $i \in J$,

then

- (i) $\|x^{k+1} - x^*\| \leq \|x^k - x^*\|$,
- (ii) x^{k+1} satisfies (a), (b) and (c).

Proof. (i) $\|x^{k+1} + x^*\| = \|(x^k - x^*) + \alpha[\sum_{i=1}^m \lambda_i(c_i^k - v_i)a_i]\| \leq \sum_{i=1}^m \lambda_i \|(x^k - x^*) + \alpha[(c_i^k - v_i)a_i]\|$.

Since $\sum_{i=1}^m \lambda_i = 1$, the result is established, provided we show that

$$(27) \quad \|(x^k - x^*) + \alpha[(c_i^k - v_i)a_i]\| \leq \|x^k - x^*\| \quad (1 \leq i \leq m).$$

Consider the following cases:

(I) $x^* \notin C_i$. By Lemma 6(i), $v_i = d_i^* \Rightarrow \|x^* - Q_i x^*\| = |v_i| \|a_i\| > 3\varepsilon$. Since $\|x^k - x^*\| < \varepsilon$, $x^k \notin C_i$. So

$$(28) \quad c_i^k = d_i^k \Rightarrow x^k + c_i^k a_i = Q_i x^k.$$

Also by Lemma 6(i),

$$(29) \quad x^* + v_i a_i = x^* + d_i^* a_i = Q_i x^*.$$

Substitute (28) and (29) into the left-hand side of (27), apply Proposition 7(i) with $\beta = \alpha$ and get the required inequality.

(II) $x^* \in H_i$. By Lemma 6(ii) $v_i = 0$ and $x^* + v_i a_i = Q_i x^*$. If $x^k + c_i^k a_i = Q_i x^k$, apply Proposition 7(i) with $\beta = \alpha$, as before. Otherwise, $x^k \in D_i$ and $0 \leq c_i^k < d_i^k \Rightarrow 0 \leq \alpha(c_i^k/d_i^k) < 2$. Take $\beta = \alpha(c_i^k/d_i^k)$. The left-hand side of (27) becomes

$$\|(1-\beta)x^k + \beta(x^k + d_i^k a_i) - x^*\| = \|(1-\beta)x^k + \beta Q_i x^k - x^*\|.$$

Apply Proposition 7(ii) and get (27).

(III) $x^* \in D_i$, $v_i = 0$. By hypothesis (b), $c_i^k = 0$, and both sides of (27) are equal.

(IV) $x^* \in D$, $v_i \neq 0$. By Lemma 6(iii)

$$(30) \quad v_i = d_i^* \Rightarrow x^* + v_i a_i = Q_i x^*,$$

By hypothesis (c),

$$(31) \quad c_i^k = d_i^k \Rightarrow x^k + c_i^k a_i = Q_i x^k.$$

Substitute (30) and (31) into the left-hand side of (27) and apply Proposition 7(i) with $\beta = \alpha$.

(ii) x^{k+1} satisfies (a) because of part (i). For (b), since $\|x^k - x^*\| < \varepsilon$, $x^k \in D_i \Rightarrow d_i^k > 0$. Since $c_i^k = 0$, by Proposition 2(iv) $c_i^{k+1} \leq 0$. Since x^{k+1} satisfies (a), $x^{k+1} \in D_i \Rightarrow c_i^{k+1} \geq 0$; so $c_i^{k+1} = 0$. For (c), if $c_i^{k+1} \neq d_i^{k+1}$ then by Proposition 2(iv), $c_i^{k+2} \leq 0$. Since $\|x^{k+2} - x^*\| \leq \|x^{k+2} - x^{k+1}\| + \|x^{k+1} - x^k\| + \|x^k - x^*\| \leq 3\varepsilon$, we get $x^{k+2} \in D_i \Rightarrow c_i^{k+2} \geq 0$. So $c_i^{k+2} = 0$. Then

$$(32) \quad |c_i^k| = |c_i^k - c_i^{k+2}| \leq |c_i^k - c_i^{k+1}| + |c_i^{k+1} - c_i^{k+2}| \leq \frac{2\varepsilon}{\|a_i\|}.$$

On the other hand,

$$\begin{aligned} c_i^k \|a_i\| &= \|x^k - Q_i x^k\| \geq \|x^* - Q_i x^k\| - \|x^k - x^*\| \\ &\geq \|x^* - Q_i x^*\| - \|x^k - x^*\| > 4\varepsilon - \varepsilon = 3\varepsilon, \end{aligned}$$

in contradiction with (32). Conclude that $c_i^{k+1} = d_i^{k+1}$. \square

We present now our main convergence result.

THEOREM 4. $x^* = \lim_{k \rightarrow \infty} x^k$.

Proof. Let r and ε be as defined just before Lemma 8. Take j big enough so that $j > s$ as defined in Lemma 7, $k_j > r$ and $\|x^{k_j} - x^*\| < \varepsilon$. Then x^{k_j} satisfies the hypotheses

of Lemma 8, which can be applied recursively for any $k > k_j$. So $\|x^k - x^*\| \leq \|x^{k_j} - x^*\|$ for any $k > k_j$. Since $x^{k_j} \xrightarrow{j \rightarrow \infty} x^*$ the theorem holds. \square

5. Characterization of the set S . Let $F = \{x \in \mathbb{R}^n : \sum_{i=1}^m \lambda_i P_i x = x\}$ and define $f: \mathbb{R}^n \rightarrow \mathbb{R}$ as $f(x) = \sum_{i=1}^m \lambda_i \|x - P_i x\|^2$. In [6] it was shown that F is just the set of minimizers of f and that $F = C$ when $C \neq \emptyset$. The results of §§ 3 and 4 indicate that the point x^* is the projection of x^0 onto the set $S = \{x: Ax \leq b + y^*\}$. We prove here that S is in fact F . So x^* is the closest solution of (1) to x^0 , if (1) is feasible, or the weighted (with the λ_i 's) least squares solution closest to x^0 , if (1) is infeasible.

LEMMA 9. $J = \emptyset$.

Proof. Take $i \in J$. Since $x^k \xrightarrow{k \rightarrow \infty} x^*$, there exists K such that $x^k \in C_i$ for $k > K$. So $c_i^k \geq 0$ for $k \geq K$. Now

$$0 \leq z_i^k = z_i^K - \lambda_i \sum_{j=K}^{k-1} c_i^j \Rightarrow \sum_{j=K}^{k-1} c_i^j \leq \frac{z_i^K}{\lambda_i}.$$

Let k go to infinity. The series converges. So $c_i^k \rightarrow 0$ and $v_i = \lim_{j \rightarrow \infty} c_i^j = 0$. \square

COROLLARY 1. $x^* + v_i a_i = P_i x^*$.

Proof. The second alternative in Lemma 6(iii) is impossible, so $v_i = \min \{0, d_i^*\}$. \square

COROLLARY 2. $x^* \in F$.

Proof. By definition of F , $x^* \in F$ if and only if $x^* = \sum_{i=1}^m \lambda_i P_i x^*$. From Corollary 1, $\sum_{i=1}^m \lambda_i P_i x^* = x^* + \sum_{i=1}^m \lambda_i v_i a_i = x^*$ by Lemma 3(iv). \square

LEMMA 10. (x^*, y^*) is a solution of

$$\begin{aligned} & \min_{x, y} \sum \lambda_i \frac{y_i^2}{\|a_i\|^2}, \\ & \text{s.t.} \begin{cases} Ax \leq b + y, \\ y \geq 0. \end{cases} \end{aligned}$$

Proof. Because of the convexity of the minimand, the Kuhn-Tucker conditions are sufficient for optimality. They are

$$(33) \quad 2 \frac{\lambda_i y_i}{\|a_i\|^2} - \mu - \rho = 0,$$

$$(34) \quad A' \mu = 0,$$

$$(35) \quad \mu_i (\langle a_i, x \rangle - b_i - y_i) = 0, \quad i = 1, \dots, m,$$

$$(36) \quad \rho_i y_i = 0, \quad i = 1, \dots, m,$$

$$(37) \quad Ax - b - y \leq 0,$$

$$(38) \quad \rho \geq 0,$$

$$(39) \quad \mu \geq 0.$$

Take $\rho = 0$, $\mu_i = -(2\lambda_i y_i^* / \|a_i\|^2)$ ($i = 1, \dots, m$), so (33), (36) and (38) are satisfied. In view of Lemmas 6 and 9, $y_i^* = -v_i \|a_i\|^2 \geq 0$. So (39) is satisfied. From Lemma 3(iv), $A' \mu = 0$ and (34) holds. Since $x^* \in S$, (37) holds. Only (35) remains to be checked. If $\mu_i \neq 0$, use Lemmas 6 and 9 to conclude that

$$x^* \notin C_i \Rightarrow v_i = d_i^* = \frac{b_i - \langle a_i, x^* \rangle}{\|a_i\|^2} \Rightarrow \langle a_i, x^* \rangle - b_i - y_i^* = 0. \quad \square$$

Observe that the minimand is strictly convex in y , so y^* is the unique solution for the “ y ” part of the problem.

LEMMA 11. $F = \{x | Ax \leq b + y^*\}$.

Proof. As noted before, F is the set of solutions of the problem $\min_x \sum_{i=1}^m \lambda_i \|P_i x - x\|^2$ (see [6]). Such a problem is equal to

$$\min_x \sum_{i=1}^m \lambda_i \left[\min \left\{ 0, \frac{b_i - \langle a_i, x \rangle}{\|a_i\|^2} \right\} \right]^2 = \min_x \sum_{i=1}^m \frac{\lambda_i}{\|a_i\|^2} [\max \{0, \langle a_i, x \rangle - b_i\}]^2,$$

which is equivalent to

$$\min_y \sum_{i=1}^m \frac{\lambda_i y_i^2}{\|a_i\|^2} \quad \text{s.t. } y_i = \max \{0, \langle a_i, x \rangle - b_i\},$$

or

$$(40) \quad \min_y \sum_{i=1}^m \frac{\lambda_i y_i^2}{\|a_i\|^2} \quad \text{s.t. } \begin{cases} Ax \leq b + y, \\ y \geq 0. \end{cases}$$

So $F = \{x | Ax \leq b + y^0\}$ where y^0 solves the “ y ” part of (40). By Lemma 10, such a y^0 has to be y^* . \square

THEOREM 5. *The sequence $\{x^k\}$ converges to the projection of x^0 onto F (i.e., to the closest point to x^0 in F).*

Proof. Immediate from Lemma 11, remembering that x^* is, by definition, the projection of x^0 onto S . \square

6. Conclusion. Hildreth published his algorithm more than twenty years ago [11] but it was only recently applied for solving huge-scale, sparse problems [10]. In 1980, Lent and Censor [13] gave an extended convergence proof of Hildreth’s method introducing an almost cyclic control and showing that it has an important optimality property. Along the lines of Lent and Censor’s paper, we have modified Hildreth’s algorithm preserving its main characteristics like the optimality property given by our Theorem 5 and obtaining a more general convergence theorem that includes a possibly empty constraint set.

REFERENCES

- [1] S. AGMON, *The relaxation method for linear inequalities*, Canad. J. Math., 6 (1954), pp. 382–392.
- [2] L. M. BREGMAN, *The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming*, U.S.S.R. Comput. Math. and Math. Phys., 7 (1967), pp. 200–217.
- [3] Y. CENSOR, *Row-action methods for huge and sparse systems and their applications*, SIAM Rev., 23 (1981), pp. 444–466.
- [4] G. CIMMINO, *Calcolo approssimato per le soluzioni dei sistemi di equazioni lineari*, La Ricerca Scientifica, vol. XVI, Ser III, Anno IX (1938), pp. 326–333.
- [5] J. W. DANIEL, *On perturbation of systems of linear inequalities*, SIAM J. Numer. Anal., 10 (1973), pp. 229–307.
- [6] A. R. DE PIERRO AND A. N. IUSEM, *A simultaneous projections method for linear inequalities*, Linear Algebra Appl., 64 (1985), pp. 243–253.
- [7] ———, *A relaxed version of Bregman’s method for convex programming*, J. Optim. Theory and Appl., to appear.
- [8] R. GORDON AND G. T. HERMAN, *Three-dimensional reconstruction from projections: A review of algorithms*, Internat. Rev. Cytology, 38 (1974), pp. 111–151.
- [9] G. T. HERMAN AND A. LENT, *Iterative reconstruction algorithms*, Computers in Biology and Medicine, 6 (1976), pp. 273–294.

- [10] G. T. HERMAN AND A. LENT, *A family of iterative quadratic optimization algorithms for pairs of inequalities with application in diagnostic radiology*, Math. Programming Stud., 9 (1978), pp. 15–29.
- [11] C. HILDRETH, *A quadratic programming procedure*, Naval Res. Logist. Quart., 4 (1957), pp. 79–85; Erratum Ibid., p. 361.
- [12] S. KACZMARZ, *Angenaherte Auflösung von Systemen linearer Gleichungen*, Bull. Acad. Polon. Sci. Lett., A 35 (1937), pp. 355–357.
- [13] A. LENT AND Y. CENSOR, *Extensions of Hildreth's row-action method for quadratic programming*, SIAM J. Control, 18 (1980), pp. 444–454.
- [14] O. L. MANGASARIAN, *Solution of symmetric linear complementarity problems by iterative methods*, J. Optim. Theory Appl., 22 (1977), pp. 465–485.
- [15] M. Z. NASHED, *Continuous and semicontinuous analogues of iterative methods of Cimmino and Kaczmarz with applications to the inverse Radon transform*, in Mathematical Aspects of Computerized Tomography, G. T. Herman and F. Natterer, eds. Lectures Notes in Medical Informatics, Springer-Verlag, Berlin, 8, 1981, pp. 160–178.

SYMMETRIES IN OPTIMAL CONTROL*

A. J. VAN DER SCHAFT†

Abstract. It is argued that the existence of symmetries may simplify, as in classical mechanics, the solution of optimal control problems. A procedure for obtaining symmetries for the optimal Hamiltonian resulting from the Maximum Principle is given; this avoids the actual calculation of the optimal Hamiltonian. This procedure is based upon the notion of symmetry for the Hamiltonian system with inputs and outputs associated with an optimal control problem.

Key words. optimal control, Hamiltonian system, symmetry, reduction

AMS(MOS) subject classifications. 49A10, 49B10, 58F05, 93C10

1. The Maximum Principle and Hamiltonian systems. Let us consider a smooth nonlinear control system

$$(1) \quad \dot{x} = f(x, u), \quad x \in X, \quad u \in U$$

where f is a smooth mapping. (Smooth always means C^∞ or C^k with k "big enough.") For simplicity of exposition we will take $X = \mathbb{R}^n$ and $U = \mathbb{R}^m$, although X and U may be arbitrary smooth manifolds. (We can even take the input space U to be *state dependent*. Then (x, u) are fiber respecting coordinates for a fiber bundle B over X , instead of coordinates for the product space $X \times U$, see [2], [6], [8].)

Let now $L: X \times U \rightarrow \mathbb{R}$ and $K: X \rightarrow \mathbb{R}$ be smooth functions. We consider the (unrestricted and smooth) Bolza problem of minimizing (with respect to $u(\cdot)$) the cost functional

$$(2) \quad J(x_0, u(\cdot)) = K(x(T)) + \int_0^T L(x(t), u(t)) dt$$

under the constraints

$$(3) \quad \dot{x}(t) = f(x(t), u(t)), \quad x(0) = x_0 \in X.$$

This is called the (finite time) *optimal control problem*. Of course we have to worry about the class of functions $U(x_0)$ from $[0, T]$ to U (which may depend on the initial condition), over which the cost functional is minimized. Since we only want to deal with some *structural* properties of the above optimal control problem, we make the following simplifying assumptions (see also [4]):

1) $U(x_0)$ consists of measurable functions such that $\dot{x} = f(x, u)$ has a well-defined solution for all $t \in [0, T]$ and $x(0) = x_0$.

2) For each $x_0 \in X$ there exists a $u^*(\cdot) \in U(x_0)$ such that

$$(4) \quad J(x_0, u^*(\cdot)) = \min_{u(\cdot) \in U(x_0)} J(x_0, u(\cdot)),$$

($u^*(\cdot): [0, T] \rightarrow U$ is called the *optimal control*).

In order to solve the optimal control problem, the Maximum Principle tells us to introduce the Hamiltonian function $H: X \times \mathbb{R}^n \times U \rightarrow \mathbb{R}$ given by

$$(5) \quad H(x, p, u) := p^T f(x, u) - L(x, u)$$

* Received by the editors October 1, 1984; accepted for publication (in revised form) December 17, 1985.

† Department of Applied Mathematics, Twente University of Technology, 7500 AE Enschede, the Netherlands.

with $p \in \mathbb{R}^n$ the *co-state*, and to consider the following set of differential equations

$$(6a) \quad \dot{x}_i(t) = \frac{\partial H}{\partial p_i}(x(t), p(t), u(t)) = f_i(x(t), u(t)), \quad i = 1, \dots, n,$$

$$(6b) \quad \dot{p}_i(t) = -\frac{\partial H}{\partial x_i}(x(t), p(t), u(t)),$$

with the (mixed) boundary conditions

$$(7) \quad \begin{aligned} x(0) &= x_0, \\ p_i(T) &= -\frac{\partial K}{\partial x_i}(x(T)), \quad i = 1, \dots, n, \end{aligned}$$

where $x(T)$ is the solution at time T of (6a) for $x(0) = x_0$. A necessary condition for a control function $u^* \in U(x_0)$ to be optimal, i.e., satisfying (4), is that for every $t \in [0, T]$

$$(8) \quad H(x^*(t), p^*(t), u^*(t)) = \max_{u \in U} H(x^*(t), p^*(t), u)$$

where $(x^*(\cdot), p^*(\cdot))$ is the solution of (6) with $u(\cdot) = u^*(\cdot)$ and boundary conditions (7). So the Maximum Principle leads us to the following static optimization problem: Find for every $(x, p) \in X \times \mathbb{R}^n$ a $u^* \in U$ such that

$$(9) \quad H(x, p, u^*) = \max_{u \in U} H(x, p, u).$$

Since we assumed U to be \mathbb{R}^m (or a manifold), (9) implies the first order conditions

$$(10) \quad \frac{\partial H}{\partial u_j}(x, p, u^*) = 0, \quad j = 1, \dots, m.$$

Hence the Maximum Principle leads in a natural way to the *system*

$$(11) \quad \begin{aligned} \dot{x}_i &= \frac{\partial H}{\partial p_i}(x, p, u), & i = 1, \dots, n, \\ \dot{p}_i &= -\frac{\partial H}{\partial x_i}(x, p, u), \\ y_j &= \frac{\partial H}{\partial u_j}(x, p, u), & j = 1, \dots, m, \end{aligned}$$

and a necessary condition for $u^*(\cdot)$ to be optimal is that the *outputs* y_j of this system, resulting from $u^*(\cdot)$ and boundary conditions (7), are constant zero.

Now equations (11) form a *Hamiltonian system* as introduced in [2] and developed in [5], [6], [7], [8]. In fact the state space of this Hamiltonian system is $X \times \mathbb{R}^n = \mathbb{R}^{2n}$, with the natural symplectic form $\sum_{i=1}^n dp_i \wedge dx_i$, the input space is $U = \mathbb{R}^m$, and the output space is $Y = \mathbb{R}^m$, with coordinates (y_1, \dots, y_m) . The product space $U \times Y$ has the natural symplectic form $\sum_{j=1}^m dy_j \wedge du_j$. From a geometric point of view equations (11) describe a $(2n + m)$ -dimensional submanifold L of $T(X \times \mathbb{R}^n) \times (U \times Y)$ (the coordinates (x, p, u) parametrize the possible state space evolutions (\dot{x}_i, \dot{p}_i) and outputs y_j). This submanifold has a special structure related to the given symplectic structures on $X \times \mathbb{R}^n$ and $U \times Y$.

Recall the definition of a *Lagrangian submanifold* [1], [8]. A submanifold L of a manifold N with symplectic form ω is Lagrangian if ω restricted to L is zero and $\dim L = \frac{1}{2} \dim N$.

DEFINITION 1 [5], [6], [7], [8]. A Hamiltonian system with state space $\mathbb{R}^n \times \mathbb{R}^n$, input space \mathbb{R}^m and output space \mathbb{R}^m is given by a submanifold $L \subset T(\mathbb{R}^n \times \mathbb{R}^n) \times (\mathbb{R}^m \times \mathbb{R}^m)$ such that

(i) L can be parametrized by the state space variables (x, p) and the input variables u .

(ii) L is a Lagrangian submanifold of $T(\mathbb{R}^n \times \mathbb{R}^n) \times (\mathbb{R}^m \times \mathbb{R}^m)$ with its natural symplectic form $\sum_{i=1}^n (dp_i \wedge dx_i + dp_i \wedge d\dot{x}_i) - \sum_{j=1}^m dy_j \wedge du_j$.

Since L is Lagrangian and satisfies condition (i) there exists a *generating function* $H(x, p, u)$ for L such that L is given by equations (11) [7], [8]. In the optimal control case this generating function $H(x, p, u)$ has the extra property of being affine in the p -variables. We call (11) the Hamiltonian system *associated* with the optimal control problem.

Remark. If X and U are arbitrary manifolds, we have to generalize the definition of the associated Hamiltonian system in the following way. Instead of $\mathbb{R}^n \times \mathbb{R}^n$ we take as state space T^*X , and the space $\mathbb{R}^m \times \mathbb{R}^m$ of inputs and outputs becomes T^*U , where both cotangent bundles are endowed with their natural symplectic forms.

Now we investigate the consequences of imposing the necessary conditions (10) on the associated Hamiltonian system (11). Since the symplectic form $\sum_{i=1}^n (dp_i \wedge dx_i + dp_i \wedge d\dot{x}_i) - \sum_{j=1}^m dy_j \wedge du_j$ is zero restricted to the submanifold L associated with (11), the form $\sum_{i=1}^n (dp_i \wedge dx_i + dp_i \wedge d\dot{x}_i)$ is zero restricted to the subset $L \cap \{y_j = \partial H / \partial u_j = 0, j = 1, \dots, m\}$, and therefore also restricted to the projection V of $L \cap \{y_j = 0, j = 1, \dots, m\}$ onto $T\mathbb{R}^{2n}$. Now if V is a nice $2n$ -dimensional submanifold of $T\mathbb{R}^{2n}$ it follows that V is a *Lagrangian* submanifold of $T\mathbb{R}^{2n}$, $\sum_{i=1}^n (dp_i \wedge dx_i + dp_i \wedge d\dot{x}_i)$. Moreover, if V can be parametrized by the state space variables (x, p) , this implies that V is actually the graph of a Hamiltonian vectorfield on \mathbb{R}^{2n} [1], [8]. The simplest case is where the matrix $(\partial^2 H / \partial u_i \partial u_j)$ has rank m in every solution (x, p, u^*) of (10) (the so-called *nonsingular* case). Then the equations $\partial H / \partial u_j(x, p, u^*) = 0, j = 1, \dots, m$, have locally a unique solution $u^*(x, p)$, and V is locally given as

$$(12) \quad V = \left\{ \left(x, p, \frac{\partial H}{\partial p}(x, p, u^*(x, p)), -\frac{\partial H}{\partial x}(x, p, u^*(x, p)) \right) \mid x \in \mathbb{R}^n, p \in \mathbb{R}^n \right\}.$$

Hence V is locally the graph of the Hamiltonian vector field of the (locally defined) *optimal Hamiltonian* $H^0(x, p) := H(x, p, u^*(x, p))$.

Remark. If $(\partial^2 H / \partial u_i \partial u_j)$ is singular but the rank of the map $\partial H / \partial u(\cdot, \cdot, \cdot)$ from $\mathbb{R}^n \times \mathbb{R}^n \times \mathbb{R}^m$ to \mathbb{R}^m is equal to m , then the set $(\partial H / \partial u)^{-1}(0)$ is a $2n$ -dimensional submanifold of $\mathbb{R}^n \times \mathbb{R}^n \times \mathbb{R}^m$. Under certain regularity conditions (see [8]) on the map $(x, p, u) \rightarrow (x, p, \partial H / \partial p(x, p, u), -\partial H / \partial x(x, p, u))$ from $\mathbb{R}^n \times \mathbb{R}^n \times \mathbb{R}^n$ to $T(\mathbb{R}^n \times \mathbb{R}^n)$ it follows that V is an (immersed) Lagrangian submanifold of $T(\mathbb{R}^n \times \mathbb{R}^n)$. However, since in general V need not be parametrized by the state space variables (x, p) , we generally only obtain a set of *implicit* Hamiltonian differential equations on $\mathbb{R}^n \times \mathbb{R}^n$ [8].

For clarity of exposition we will make the following additional assumptions which will hold throughout the next section:

1) The matrix $(\partial^2 H / \partial u_i \partial u_j)(x, p, u^*)$ is nonsingular in every solution (x, p, u^*) of (10). The solution u^* of (10) is unique and is a smooth mapping $u^*(x, p)$ from $\mathbb{R}^n \times \mathbb{R}^n$ to \mathbb{R}^m .

2) This solution $u^*(x, p)$ is *optimal*, resulting in the optimal Hamiltonian $H^0(x, p) := H(x, p, u^*(x, p))$.

2. Symmetries. Under the simplifying assumptions made before the optimal control problem reduces to the solution of a set of Hamiltonian equations

$$(13) \quad \begin{aligned} \dot{x}_i &= \frac{\partial H^0}{\partial p_i}(x, p), & x(0) &= x_0, \\ \dot{p}_i &= -\frac{\partial H^0}{\partial x_i}(x, p), & p_i(T) &= -\frac{\partial K}{\partial x_i}(x(T)), \end{aligned} \quad i = 1, \dots, n,$$

with $H^0(x, p) = H(x, p, u^*(x, p))$, where $u^*(x, p)$ is the unique solution of

$$(14) \quad \frac{\partial H}{\partial u_j}(x, p, u^*(x, p)) = 0, \quad j = 1, \dots, m.$$

Now solving (13) and (14) is typically a formidable task, and it is worthwhile to look for circumstances which make the solution easier.

If the equations (14) are explicitly solved for $u^*(x, p)$ and if we therefore have an *explicit* expression for $H^0(x, p)$, it is a classical method (in mechanics) to look for *symmetries* of H^0 in order to simplify the solution of (13). (This point was also raised in [3].) Let us introduce some notation. If $F: \mathbb{R}^n \times \mathbb{R}^n \rightarrow \mathbb{R}$ is a smooth function, then we denote the corresponding Hamiltonian vectorfield

$$(15) \quad \dot{x}_i = \frac{\partial F}{\partial p_i}(x, p), \quad \dot{p}_i = -\frac{\partial F}{\partial x_i}(x, p), \quad i = 1, \dots, n$$

by X_F . Moreover if G is another smooth function on $\mathbb{R}^n \times \mathbb{R}^n$, then the *Poisson bracket* $\{F, G\}$ of F and G is defined as

$$(16) \quad \{F, G\} = \sum_{i=1}^n \left(\frac{\partial F}{\partial p_i} \frac{\partial G}{\partial x_i} - \frac{\partial F}{\partial x_i} \frac{\partial G}{\partial p_i} \right).$$

It is easy to see that $\{F, G\} = -\{G, F\}$ and that

$$(17) \quad \{F, G\} = X_F(G).$$

Now the most general definition of an (infinitesimal) symmetry for H^0 is of a Hamiltonian vectorfield X_F satisfying

$$(18) \quad X_F(H^0) = \{F, H^0\} = 0.$$

This implies that $0 = \{F, H^0\} = -\{H^0, F\} = -X_{H^0}(F)$, and hence that F is a *first integral* or *conserved quantity* for (13). Therefore if X_F is a symmetry of H^0 , then F is a first integral for X_{H^0} . Conversely if F is a first integral for X_{H^0} , i.e., $X_{H^0}(F) = 0$, it follows that X_F is a symmetry for H^0 . The existence of such a conserved quantity F for (13) may be used for *reducing* the $2n$ -dimensional set of equations (13) to a $(2n-2)$ -dimensional set. Indeed, suppose that dF nowhere vanishes (this can be relaxed). Then there exists a constant c such that the solution of (13) remains within the submanifold $F^{-1}(c)$. Moreover we may factor out $F^{-1}(c)$ by the integral curves of X_F to obtain a $(2n-2)$ -dimensional manifold. It follows from $X_F(H^0) = 0$ that the equations (13) project to Hamiltonian equations on this reduced manifold. If there are more symmetries available, or a *group* of symmetries, this reduction procedure can be generalized [1]. In general the existence of symmetries for H^0 reduces the solution of (13) to the solution of a *lower-dimensional* set of Hamiltonian equations. (Notice however that our situation is somewhat more complicated than in mechanics, since we do not know the initial conditions of (13), but a mixed set of initial and terminal conditions, see also [4].)

In conclusion, if we have an explicit expression for H^0 the knowledge of symmetries simplifies the solution $x^*(\cdot), p^*(\cdot)$ of (13). Henceforth it also simplifies the construction of the optimal control in open loop form $u^*(t) = u^*(x^*(t), p^*(t))$, or in feedback form $u^*(x^*(t), t) = u^*(x^*(t), p^*(t))$ from the solution $u^*(x, p)$ of (14).

Remark. A symmetry X_F for H^0 may be also profitably used for solving the Hamilton-Jacobi-Bellman equation

$$(19) \quad \begin{aligned} \frac{\partial S}{\partial t}(x, t) &= -\max_u \left(-L(x, u) + \frac{\partial S}{\partial x}(x, t)f(x, u) \right) \\ &= -H^0\left(x, \frac{\partial S}{\partial x}(x, t)\right), \quad S(x, T) = -K(x). \end{aligned}$$

(We have adopted the sign convention from mechanics— S is minus the Bellman value function.) Now (19) defines a flow on the set of Lagrangian submanifolds of $\mathbb{R}^n \times \mathbb{R}^n$. In fact for every t the Lagrangian submanifold is given as $\{(x, p = (\partial S / \partial x)(x, t))\}$. If X_F is a symmetry for H^0 , it follows that the action of X_F commutes with this flow. Explicitly, if the integral flow $X_F(\tau)$ of X_F maps for a *small* and *fixed* τ the Lagrangian submanifold $\{(x, p = (\partial S / \partial x)(x, T))\}$ onto another Lagrangian submanifold $\{(x, p = (\partial R / \partial x)(x))\}$, then $X_F(\tau)$ maps for every $t \in [0, T]$ the Lagrangian submanifold $\{(x, p = (\partial S / \partial x)(x, t))\}$ onto Lagrangian submanifolds $\{(x, p = (\partial R / \partial x)(x, t))\}$ where $R(x, t)$ is the solution of

$$(20) \quad \frac{\partial R}{\partial t}(x, t) = -H^0\left(x, \frac{\partial R}{\partial x}(x, t)\right), \quad R(x, T) = R(x).$$

Therefore instead of solving (19) we may also solve (20) for the maybe easier terminal condition $R(x, T) = R(x)$. In the linear quadratic case (i.e. $F(x, u) = Ax + Bu$, $L(x, u) = \frac{1}{2}x^T Qx + \frac{1}{2}u^T Ru$, and (19) becoming a Riccati equation) this was noted in [10].

Of course in many cases an explicit expression for $u^*(x, p)$ and $H^0(x, p)$ is hard to obtain. However in the author's thesis [8] it was indicated that even *without* explicitly calculating $H^0(x, p)$ we can a priori deduce symmetries for $H^0(x, p)$ by looking for symmetries of the associated Hamiltonian system (11). The same idea was used by Grizzle and Marcus [4] from a different point of view. Let $\dot{x} = f(x, u)$, with $x \in M$ (n -dimensional) and $u \in U$, be a control system. Suppose there exists a vectorfield $G(x)$ on M such that $[G(x), f(x, u)] = 0$ for all $u \in U$ (G is called a symmetry of the control system (cf. [8], [4]). Furthermore suppose that $G(L(x, u)) = 0$ for all $u \in U$ and that $G(K(x)) = 0$. (G is called a symmetry of the optimal control problem, [4].)

Then if G is nowhere zero, M can be locally factored out by the integral curves of G to obtain an $(n-1)$ -dimensional manifold N . It is then shown [4] that the optimal control problem (2) reduces to an optimal control problem defined on this lower dimensional manifold N , and that the optimal control in feedback form can be defined on N . Furthermore this can be generalized from single vectorfields G to a Lie algebra of symmetry vectorfields generated by a symmetry Lie group.

In the sequel it will be shown that this kind of symmetry for the optimal control problem considered in [4] corresponds to a special, although important, type of symmetry for the associated Hamiltonian system and the optimal Hamiltonian $H^0(x, p)$. Furthermore if the end cost function K is *not* invariant under G ($G(K(x)) \neq 0$) it is noted in [4] that the above procedure cannot be followed without modifications, but recourse has to be taken to the same Hamiltonian approach as will be used in this paper. On the other hand the class of symmetries considered in [4] is *enlarged* in [4] by allowing for *feedback* transformations. A feedback $u = \alpha(x, v)$ transforms the

optimal control problem (2) into

$$\min_v K(x(T)) + \int_0^T L(x(t), \alpha(x(t), v(t))) dt$$

under the constraints $\dot{x}(t) = f(x(t), \alpha(x(t), v(t)))$, $x(0) = x_0$. Now a vectorfield G may be a symmetry of this *transformed* optimal control problem without being a symmetry of the original optimal control problem. Such a symmetry G also results in a symmetry of the optimal Hamiltonian but may *not* be obtainable by our approach (although in most cases it will, see the examples). This leads to the question of determining *what* class of symmetries for H^0 can be obtained by our approach and how this class is affected by feedback. This problem is addressed (but not fully solved) in the last part of the paper.

We will now show how we can deduce symmetries for (13) *without* explicitly constructing $H^0(x, p)$ by looking for symmetries of the associated Hamiltonian system (11). Recall the notion of a *prolongation* of a vectorfield or a function. Let S be a vectorfield on M with integral flow S_t (i.e., $(d/dt)S_t(x) = S(S_t(x))$). Then $(S_t)_*: TM \rightarrow TM$ is the integral flow of a vectorfield on TM which we denote by \dot{S} . Let further $F: M \rightarrow \mathbb{R}$, then $\dot{F}: TM \rightarrow \mathbb{R}$ is defined by $\dot{F}(v) = dF(v)$, $v \in TM$.

DEFINITION 2 [5], [7], [8]. Let (11) be a Hamiltonian system given by a Lagrangian submanifold $L \subset T(\mathbb{R}^n \times \mathbb{R}^n) \times (\mathbb{R}^m \times \mathbb{R}^m)$. An (infinitesimal) *symmetry* is a pair of vectorfields (S, S^e) , S a Hamiltonian vectorfield on $\mathbb{R}^n \times \mathbb{R}^n$ and S^e a Hamiltonian vectorfield on $\mathbb{R}^m \times \mathbb{R}^m$, such that the vectorfield (\dot{S}, S^e) on $T(\mathbb{R}^n \times \mathbb{R}^n) \times (\mathbb{R}^m \times \mathbb{R}^m)$ is *tangent* to L , i.e., $(\dot{S}, S^e)(z) \in T_z L$ for all $z \in L$.

A *conservation law* is a pair of functions (F, F^e) , with $F: \mathbb{R}^n \times \mathbb{R}^n \rightarrow \mathbb{R}$ and $F^e: \mathbb{R}^m \times \mathbb{R}^m \rightarrow \mathbb{R}$, such that the function $\dot{F} - F^e: T(\mathbb{R}^n \times \mathbb{R}^n) \times (\mathbb{R}^m \times \mathbb{R}^m) \rightarrow \mathbb{R}$ restricted to L is zero.

Remark. The above definitions are really extensions of the usual definitions of symmetry and conserved quantity for Hamiltonian differential equations, as can be seen as follows. If we forget about inputs and outputs, so if $L \subset T(\mathbb{R}^n \times \mathbb{R}^n)$ is just the graph of a Hamiltonian vectorfield X_H , then \dot{S} being tangent to L means the following. In coordinates \dot{S} is given as $S(x)(\partial/\partial x) + (\partial S/\partial x)(x)\dot{x}(\partial/\partial \dot{x})$ (we forget about indices). Consider now a point $z = (x, X_H(x))$ on L . Elements of $T_z L$ are of the form $(\partial/\partial x) + (\partial X_H(x)/\partial x)(\partial/\partial \dot{x})$. Hence \dot{S} is tangent to L in z if $S(x)(\partial/\partial x) + (\partial S/\partial x)(x)X_H(x)(\partial/\partial \dot{x})$ is a multiple of $(\partial/\partial x) + (\partial X_H(x)/\partial x)(\partial/\partial \dot{x})$. This only happens if $(\partial S/\partial x)(x)X_H(x) = (\partial X_H(x))/(\partial x)S(x)$, or equivalently if the Lie bracket $[S, X_H]$ equals zero. Furthermore, $\dot{F} - F^e = 0$ restricted to L just means that $dF/dt = F^e(y, u)$, with $y_j = \partial H/\partial u_j$, where d/dt is differentiation along the system (13).

As is the case for Hamiltonian *vectorfields*, symmetries and conservation laws for Hamiltonian *systems* are in one-to-one correspondence [5], [8]. In fact if (S, S^e) is a symmetry, then there exists a conservation law (F, F^e) such that $S = X_F$, $S^e = X_{F^e}$, and conversely if (F, F^e) is a conservation law, then (X_F, X_{F^e}) is a symmetry.

We notice that (F, F^e) being a conservation law for (11) can be also succinctly expressed by the equality (see (17))

$$(21) \quad \{H(x, p, u), F(x, p)\} = F^e\left(\frac{\partial H}{\partial u}(x, p, u), u\right) \quad \forall x, p, u$$

where $\{, \}$ means Poisson bracket on $\mathbb{R}^n \times \mathbb{R}^n$.

We now show how symmetries (or conservation laws) for the Hamiltonian system (11) yield symmetries (or conserved quantities) for the optimal Hamiltonian.

THEOREM 3 [8]. Let $(S = X_F, S^e = X_{F^e})$ be a symmetry for (11). Then S is a symmetry for H^0 if $F^e(0, u) = 0$, for all $u \in U$.

Proof. Since (X_F, X_{F^e}) is a symmetry, (21) holds. Therefore

$$\begin{aligned} \{H(x, p, u^*(x, p)), F(x, p)\} &= F^e\left(\frac{\partial H}{\partial u}(x, p, u^*(x, p)), u^*(x, p)\right) \\ &\quad + \sum_{j=1}^m \frac{\partial H}{\partial u_j}(x, p, u^*(x, p))\{u_j^*(x, p), F(x, p)\}. \end{aligned}$$

Since $(\partial H / \partial u_j)(x, p, u^*(x, p)) = 0$, we obtain $\{H^0(x, p), F(x, p)\} = F^e(0, u^*(x, p))$, and hence if $F^e(0, u) = 0$, for all u , $\{H^0(x, p), F(x, p)\} = -S(H^0) = 0$. \square

In conclusion, one can obtain symmetries of H^0 by looking for pairs (F, F^e) satisfying (21) and $F^e(0, u) = 0$. Furthermore, these symmetries may also be useful in finding the solution $u^*(x, p)$ of (14):

THEOREM 4. Let (X_F, X_{F^e}) be a symmetry for (11) with $F^e(0, u) = 0 \ \forall u$. Let $u^*(x, p) = (u_1^*(x, p), \dots, u_m^*(x, p))$ be the solution of (14). Then for $j = 1, \dots, m$

$$(22) \quad \{F(x, p), u_j^*(x, p)\} = \frac{\partial F^e}{\partial y_j}(0, u^*(x, p)).$$

Proof. Differentiate the equalities $(\partial H / \partial u_j)(x, p, u^*(x, p)) = 0$, $j = 1, \dots, m$, with respect to x_k and p_k , $k = 1, \dots, n$:

$$(23) \quad \frac{\partial^2 H}{\partial x_k \partial u_j}(x, p, u^*) + \sum_{i=1}^m \frac{\partial^2 H}{\partial u_i \partial u_j}(x, p, u^*) \frac{\partial u_i^*}{\partial x_k}(x, p) = 0,$$

$$(24) \quad \frac{\partial^2 H}{\partial p_k \partial u_j}(x, p, u^*) + \sum_{i=1}^m \frac{\partial^2 H}{\partial u_i \partial u_j}(x, p, u^*) \frac{\partial u_i^*}{\partial p_k}(x, p) = 0.$$

Furthermore, differentiate

$$\{H(x, p, u), F(x, p)\} = \sum_{k=1}^n \left(\frac{\partial H}{\partial p_k} \frac{\partial F}{\partial x_k} - \frac{\partial H}{\partial x_k} \frac{\partial F}{\partial p_k} \right) = F^e\left(\frac{\partial H}{\partial u}, u\right)$$

with respect to u_j , $j = 1, \dots, m$:

$$(25) \quad \sum_{k=1}^n \left(\frac{\partial^2 H}{\partial u_j \partial p_k} \frac{\partial F}{\partial x_k} - \frac{\partial^2 H}{\partial u_j \partial x_k} \frac{\partial F}{\partial p_k} \right) = \frac{\partial}{\partial u_j} \left(F^e\left(\frac{\partial H}{\partial u}, u\right) \right).$$

Evaluate (25) in the points $(x, p, u^*(x, p))$, and substitute (23) and (24) into (25):

$$(26) \quad - \sum_{k=1}^n \sum_{i=1}^m \frac{\partial^2 H}{\partial u_i \partial u_j} \frac{\partial u_i^*}{\partial p_k} \frac{\partial F}{\partial x_k} + \sum_{k=1}^n \sum_{i=1}^m \frac{\partial^2 H}{\partial u_i \partial u_j} \frac{\partial u_i^*}{\partial x_k} \frac{\partial F}{\partial p_k} = \frac{\partial}{\partial u_j} \left(F^e\left(\frac{\partial H}{\partial u}, u\right) \right)$$

with everything evaluated in $(x, p, u^*(x, p))$. The left-hand side of (26) is equal to

$$(27) \quad \sum_{i=1}^m \frac{\partial^2 H}{\partial u_i \partial u_j}(x, p, u^*) \{F, u_i^*\}(x, p),$$

while the right-hand side of (26) equals

$$\begin{aligned} &\frac{\partial}{\partial u_j} \left(F^e\left(\frac{\partial H}{\partial u}, u\right) \right) \Big|_{u=u^*(x, p)} \\ (28) \quad &= \sum_{i=1}^n \frac{\partial F^e}{\partial y_i} \left(\frac{\partial H}{\partial u}, u \right) \Big|_{u=u^*} \frac{\partial^2 H}{\partial u_j \partial u_i}(x, p, u^*) + \frac{\partial F^e}{\partial u_j} \left(\frac{\partial H}{\partial u}, u \right) \Big|_{u=u^*} \\ &= \sum_{i=1}^n \frac{\partial^2 H}{\partial u_j \partial u_i}(x, p, u^*) \frac{\partial F^e}{\partial y_i}(0, u^*) \end{aligned}$$

since $F^e(0, u) = 0$ implies $\partial F^e / \partial u_j(0, u) = 0$, $j = 1, \dots, m$.

Since $(\partial^2 H / \partial u_i \partial u_j)(x, p, u^*)$ is nonsingular, we obtain (22). \square

Remark. $(\partial F^e / \partial y_j)(0, u^*)$ may also be written as $\{F^e(y, u), u_j\}_{\mathbb{R}^{2m}}(0, u^*)$, with $\{, \}_{\mathbb{R}^{2m}}$ the Poisson bracket on \mathbb{R}^{2m} (given by $\{G(y, u), K(y, u)\}_{\mathbb{R}^{2m}} = \sum_{j=1}^m ((\partial G / \partial y_j)(\partial K / \partial u_j) - (\partial G / \partial u_j)(\partial K / \partial y_j))$). Hence (22) may be rewritten as $\{F, u_j^*\}_{\mathbb{R}^{2n}} = \{F^e, u_j\}_{\mathbb{R}^{2m}}(0, u^*)$.

Therefore, if (F, F^e) is a conservation law with $F^e(0, u) = 0$ the solution of (14) has to belong to the mappings $u^*(x, p)$ whose components satisfy the partial differential equations

$$(29) \quad \{F, u_j^*\} = G_j(u^*), \quad j = 1, \dots, m$$

with $G_j(u^*) = (\partial F^e / \partial y_j)(0, u^*)$.

An important special case of Theorem 3 are the conservation laws (F, F^e) with F^e *identically* zero. The fact that such conservation laws may exist is due to the possible nonminimality of the Hamiltonian system. Indeed, if $F^e = 0$, we obtain

$$(30) \quad \{H(x, p, u), F\} = 0 \quad \forall x, p, u.$$

Hence all integral curves of the Hamiltonian system (11) starting from a fixed initial condition u_0 remain within a submanifold $F^{-1}(c)$, with c a constant, and therefore the system is not “controllable.” Moreover, it follows from (30) that

$$(31) \quad X_F(y_j) = \left\{ F, \frac{\partial H}{\partial u_j}(x, p, u) \right\} = \frac{\partial}{\partial u_j} \{F, H(x, p, u)\} = 0$$

and that

$$(32) \quad [X_{H(x, p, u)}, X_F] = 0 \quad \forall u.$$

(Recall the identity $[X_F, X_G] = X_{\{F, G\}}$ for arbitrary functions F, G on \mathbb{R}^{2n} , [1].) Hence the system is not “observable” and we may factor out the state space by the integral curves of X_F . (It follows from (32) that the vectorfields $X_{H(x, p, u)}$ leave these integral curves invariant.) For a more detailed treatment of these issues we refer to [6], [8]. It follows from Theorem 4 that if $F^e = 0$ then the optimal $u^*(x, p)$ satisfies

$$(33) \quad \{F, u_j^*\} = 0, \quad j = 1, \dots, m.$$

Therefore if we reduce the $2n$ -dimensional state space to a $(2n - 2)$ -dimensional space as sketched above, the optimal $u(x, p)$ also projects to a mapping on this reduced space. The symmetries considered in [4] form a subclass of this special type. Indeed, let the vectorfield G satisfy $[G(x), f(x, u)] = 0$, for all $u \in U$, and $G(L(x, u)) = 0$. Then

$$\begin{aligned} \{H(x, p, u), p^T G(x)\} &= \{p^T f(x, u) - L(x, u), p^T G(x)\} \\ &= p^T \frac{\partial G}{\partial x}(x) f(x, u) - p^T \frac{\partial F}{\partial x}(x, u) G(x) - p^T \frac{\partial L}{\partial x}(x, u) G(x) \\ &= p^T [f(x, u), G(x)] - p^T G(L(x, u)) = 0. \end{aligned}$$

Hence $(p^T G(x), 0)$ is a conservation law for the associated Hamiltonian system, and $p^T G(x)$ is a conserved quantity for the optimal Hamiltonian H^0 . In the physics literature a symmetry with conserved quantity of the form $p^T G(x)$ is called a *geometrical* symmetry (because the symmetry is induced by a vectorfield on the x -space), in contrast to a symmetry with a general conserved quantity $F(x, p)$, which is called a *dynamical* symmetry.

As in the case of Hamiltonian vector fields [1], the treatment of a single symmetry may be extended to *groups* of symmetries. In our context the basic observation in order to do so is the following.

THEOREM 5. Let (F_i, F_i^e) , $i = 1, 2$, be two conservation laws for (11), with $F_i^e(0, u) = 0$, for all u . Then $(\{F_1, F_2\}_{\mathbb{R}^{2n}}, \{F_1^e, F_2^e\}_{\mathbb{R}^{2m}})$ is again a conservation law with $\{F_1^e, F_2^e\}_{\mathbb{R}^{2m}}(0, u) = 0$, for all u . (As before $\{, \}_{\mathbb{R}^{2n}}$ and $\{, \}_{\mathbb{R}^{2m}}$ denote Poisson brackets on \mathbb{R}^{2n} , resp. \mathbb{R}^{2m} .)

Proof. This can be proved by geometric considerations [8], but also by the following explicit calculation. We have $\{H(x, p, u), F_i(x, p)\} = F_i^e((\partial H / \partial u), u)$, $i = 1, 2$. Hence, by Jacobi's identity for the Poisson bracket,

$$\begin{aligned} \{H(x, p, u), \{F_1, F_2\}\} &= \{\{H(x, p, u), F_1\}, F_2\} - \{\{H(x, p, u), F_2\}, F_1\} \\ &= \left\{F_1^e\left(\frac{\partial H}{\partial u}, u\right), F_2\right\} - \left\{F_2^e\left(\frac{\partial H}{\partial u}, u\right), F_1\right\}. \end{aligned}$$

Now

$$\left\{F_1^e\left(\frac{\partial H}{\partial u}, u\right), F_2\right\} = \sum_{j=1}^m \frac{\partial F_1^e}{\partial y_j} \left(\frac{\partial H}{\partial u}, u\right) \left\{\frac{\partial H}{\partial u_j}, F_2\right\},$$

and

$$\begin{aligned} \left\{\frac{\partial H}{\partial u_j}, F_2\right\} &= \frac{\partial}{\partial u_j} \{H, F_2\} = \frac{\partial}{\partial u_j} \left(F_2^e\left(\frac{\partial H}{\partial u}, u\right)\right) \\ &= \frac{\partial F_2^e}{\partial u_j} + \sum_{k=1}^m \frac{\partial F_2^e}{\partial y_k} \frac{\partial^2 H}{\partial u_j \partial u_k}. \end{aligned}$$

Hence

$$\begin{aligned} \{H(x, p, u), \{F_1, F_2\}_{\mathbb{R}^{2n}}\} &= \sum_{j=1}^m \frac{\partial F_1^e}{\partial y_j} \frac{\partial F_2^e}{\partial u_j} - \frac{\partial F_1^e}{\partial u_j} \frac{\partial F_2^e}{\partial y_j} \\ &\quad + \sum_{j=1}^m \sum_{k=1}^m \frac{\partial F_1^e}{\partial y_j} \frac{\partial F_2^e}{\partial y_k} \frac{\partial^2 H}{\partial u_j \partial u_k} - \sum_{j=1}^m \sum_{k=1}^m \frac{\partial F_2^e}{\partial y_j} \frac{\partial F_1^e}{\partial y_k} \frac{\partial^2 H}{\partial u_j \partial u_k} \\ &= \{F_1^e, F_2^e\}_{\mathbb{R}^{2m}}. \end{aligned}$$

Furthermore, $F_i^e(0, u) = 0$ implies $(\partial F_i^e(0, u)) / \partial u_j = 0$, $j = 1, \dots, m$, and hence

$$\{F_1^e, F_2^e\}(0, u) = \sum_j \left(\frac{\partial F_1^e(0, u)}{\partial y_j} \frac{\partial F_2^e(0, u)}{\partial u_j} - \frac{\partial F_1^e(0, u)}{\partial u_j} \frac{\partial F_2^e(0, u)}{\partial y_j} \right) = 0. \quad \square$$

Therefore the mapping $F \mapsto F^e$, given by $\{H(x, p, u), F\} = F^e$, is an algebra morphism from functions on \mathbb{R}^{2n} to \mathbb{R}^{2m} (with respect to the respective Poisson brackets). It also follows from Theorems 4 and 5 that in the case of two conservation laws (F_i, F_i^e) the optimal $u^*(x, p)$ also has to satisfy

$$(34) \quad \{F_1, F_2\}_{\mathbb{R}^{2n}}, u_i^*_{\mathbb{R}^{2n}} = \{F_1^e, F_2^e\}_{\mathbb{R}^{2m}}, u_i^*_{\mathbb{R}^{2m}}(0, u^*).$$

We will now give some illustrative examples of the theory developed above.

Example 1. First we treat the example dealt with in Grizzle and Marcus [4] in our framework. Consider a particle of unit mass in a planar inverse-square-law gravitational field, which has thrusters in the “ x - y ” directions. The equations of motion in rectangular coordinates are given as

$$(\dot{q}_1, \dot{q}_2, \dot{v}_1, \dot{v}_2) = (v_1, v_2, -q_1(q_1^2 + q_2^2)^{-3/2} + u_1, -q_2(q_1^2 + q_2^2)^{-3/2} + u_2) = f(x, u)$$

and are defined on $M = (\mathbb{R}^2 \setminus \{0\}) \times \mathbb{R}^2$ and $U = \mathbb{R}^2$.

Let us take $L(x, u) = \frac{1}{2}(u_1^2 + u_2^2)$. An evident candidate for a symmetry vectorfield on $\mathbb{R}^2 \setminus \{0\}$ is $q_1(\partial/\partial q_2) - q_2(\partial/\partial q_1)$ (infinitesimal rotation). This vectorfield is prolonged to the vectorfield $G = q_1(\partial/\partial q_2) - q_2(\partial/\partial q_1) + v_1(\partial/\partial v_2) - v_2(\partial/\partial v_1)$ on M . Denote $x_1 = q_1$, $x_2 = q_2$, $x_3 = v_1$, $x_4 = v_2$; then the corresponding Hamiltonian function is $p^T G(x) = -p_1 x_2 + p_2 x_1 - p_3 x_4 + p_4 x_3$. Calculation yields

$$\begin{aligned} \{H(x, p, u), p^T G(x)\} &= \{p^T f(x, u) - L(x, u), p^T G(x)\} \\ &= \{p_1 x_3 + p_2 x_4 + p_3(-x_1(x_1^2 + x_2^2)^{-3/2} + u_1) \\ &\quad + p_4(-x_2(x_1^2 + x_2^2)^{-3/2} + u_2) \\ &\quad - \frac{1}{2}u_1^2 - \frac{1}{2}u_2^2, -p_1 x_2 + p_2 x_1 - p_3 x_4 + p_4 x_3\} \\ &= (\text{since the gravitational field is rotation invariant}) \\ &\quad \{p_3 u_1 + p_4 u_2, -p_1 x_2 + p_2 x_1 - p_3 x_4 + p_4 x_3\} \\ &= u_1 p_4 - u_2 p_3. \end{aligned}$$

Furthermore,

$$y_1 = \frac{\partial H}{\partial u_1} = p_3 - u_1, \quad y_2 = \frac{\partial H}{\partial u_2} = p_4 - u_2.$$

Hence,

$$(35) \quad \{H(x, p, u), p^T G(x)\} = u_1(y_2 + u_2) - u_2(y_1 + u_1) = u_1 y_2 - u_2 y_1.$$

Therefore $(F, F^e) = (-p_1 x_2 + p_2 x_1 - p_3 x_4 + p_4 x_3, u_1 y_2 - u_2 y_1)$ is a conservation law for the associated Hamiltonian system satisfying $F^e(0, u) = 0$, for all u . Hence by Theorems 3 and 4

$$\begin{aligned} &\{H(x, p, u^*(x, p)), -p_1 x_2 + p_2 x_1 - p_3 x_4 + p_4 x_3\} = 0, \\ (36) \quad &\{-p_1 x_2 + p_2 x_1 - p_3 x_4 + p_4 x_3, u_1^*(x, p)\} = \frac{\partial F^e}{\partial y_1}(0, u^*) = -u_2^*(x, p), \\ &\{-p_1 x_2 + p_2 x_1 - p_3 x_4 + p_4 x_3, u_2^*(x, p)\} = \frac{\partial F^e}{\partial y_2}(0, u^*) = u_1^*(x, p). \end{aligned}$$

On the other hand, in the Grizzle-Marcus approach one first applies *feedback*

$$(37) \quad \begin{pmatrix} u_1 \\ u_2 \end{pmatrix} = (q_1^2 + q_2^2)^{-1/2} \begin{pmatrix} q_1 & q_2 \\ -q_2 & q_1 \end{pmatrix} \begin{pmatrix} w_1 \\ w_2 \end{pmatrix} = \alpha(x, w)$$

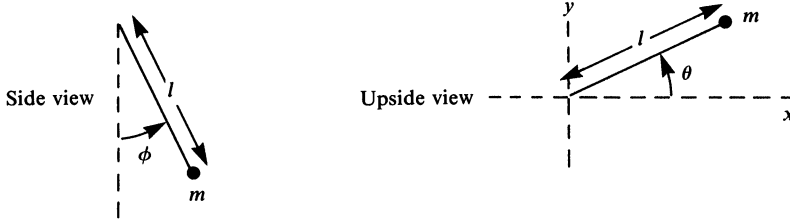
(defined on $\mathbb{R}^2 \setminus \{0\}$), transforming $f(x, u)$ into $\tilde{f}(x, w)$, with $w = (w_1, w_2)$ the new inputs. The modified Hamiltonian $\tilde{H}(x, p, w) = p^T \tilde{f}(x, w) - \tilde{L}(x, w)$, with $\tilde{L}(x, w) = L(x, \alpha(x, w)) = \frac{1}{2}w_1^2 + \frac{1}{2}w_2^2$ then satisfies

$$(38) \quad \{\tilde{H}(x, p, w), p^T G(x)\} = 0$$

while

$$(39) \quad \{p^T G(x), w_i^*(x, p)\} = 0, \quad i = 1, 2.$$

Example 2. Consider a mathematical pendulum in space (\mathbb{R}^3) with mass $m = 1$ and length $l = 1$.



Suppose there is a horizontal field by which one can exert a force u_1 in the x -direction and a force u_2 in the y -direction. In spherical coordinates the dynamical equations are

$$(40) \quad \begin{aligned} \ddot{\theta} &= -u_1 \sin \theta + u_2 \cos \theta, \\ \ddot{\phi} &= -g \sin \phi + u_1 \cos \theta \cos \phi + u_2 \sin \theta \cos \phi \end{aligned}$$

with $(\phi, \theta) \in S^2$ (the unit sphere). Therefore the state space is $M = TS^2$ with local coordinates $x_1 = \phi$, $x_2 = \theta$, $x_3 = \dot{\phi}$, $x_4 = \dot{\theta}$. Once more we take $L(x, u) = \frac{1}{2}(u_1^2 + u_2^2)$. The symmetry vectorfield on M is given in local coordinates by $G(x) = (\partial/\partial x_2)$, with corresponding Hamiltonian $p^T G(x) = p_2$. Then

$$\begin{aligned} \{p^T f(x, u) - L(x, u), p^T G(x)\} \\ &= \{p_1 x_3 + p_2 x_4 + p_3(-g \sin x_1 + u_1 \cos x_2 \cos x_1 + u_2 \sin x_2 \cos x_1) \\ &\quad + p_4(-u_1 \sin x_2 + u_2 \cos x_2) - \frac{1}{2}(u_1^2 + u_2^2), p_2\} \\ &= u_1 p_3 \sin x_2 \cos x_1 - u_2 p_3 \cos x_2 \cos x_1 + u_1 p_4 \cos x_2 + u_2 p_4 \sin x_2. \end{aligned}$$

Furthermore

$$(41) \quad \begin{aligned} y_1 &= \frac{\partial H}{\partial u_1} = p_3 \cos x_2 \cos x_1 - p_4 \sin x_2 - u_1, \\ y_2 &= \frac{\partial H}{\partial u_2} = p_3 \sin x_2 \cos x_1 + p_4 \cos x_2 - u_2. \end{aligned}$$

Hence $\{H(x, p, u), p^T G(x)\} = u_1(y_2 + u_2) - u_2(y_1 + u_1) = u_1 y_2 - u_2 y_1$.

So $(p_2, u_1 y_2 - u_2 y_1)$ is a conservation law satisfying the conditions of Theorems 3 and 4. Hence

$$(42) \quad \{H^0, p_2\} = 0,$$

$$(43) \quad \{p_2, u_1^*\} = -u_2^*, \quad \{p_2, u_2^*\} = u_1^*.$$

Of course, this can be easily checked. Setting $y_1 = y_2 = 0$ in (41), one obtains

$$(44) \quad u_1^* = p_3 \cos x_2 \cos x_1 - p_4 \sin x_2, \quad u_2^* = p_3 \sin x_2 \cos x_1 + p_4 \cos x_2$$

in accordance with (43). Furthermore one calculates

$$(45) \quad H^0(x, p) = H(x, p, u^*(x, p)) = p_1 x_3 + p_2 x_4 - p_3 g \sin x_1 + p_3^2 \cos^2 x_1 + p_4^2$$

and hence (42) is satisfied.

We note that this example, although very close to Example 1, *cannot* be treated by the methods of [4]. This is because there does not exist a smooth feedback $u = \alpha(x, w)$ such that $\{\tilde{H}(x, p, w), p_2\} = 0$; the feedback (37) in rectangular coordinates is *not* defined for $\phi = \theta = 0$. (In Example 1 the origin was excluded from the state space!)

In the above case u_1^* and u_2^* can be immediately computed, thanks to the simple form of $L(x, u)$. However suppose $L(x, u) = \frac{1}{4}(u_1^2 + u_2^2)^2$. Then still $\{H(x, p, u), p_2\} = u_1 p_3 \sin x_2 \cos x_1 - u_2 p_3 \cos x_2 \cos x_1 + u_1 p_4 \cos x_2 + u_2 p_4 \sin x_2$ while

$$y_1 = p_3 \cos x_2 \cos x_1 - p_4 \sin x_2 - (u_1^2 + u_2^2)u_1,$$

$$y_2 = p_3 \sin x_2 \cos x_1 + p_4 \cos x_2 - (u_1^2 + u_2^2)u_2.$$

Hence $\{H(x, p, u), p_2\} = u_1(y_2 + u_2(u_1^2 + u_2^2)) - u_2(y_1 + u_1(u_1^2 + u_2^2)) = u_1 y_2 - u_2 y_1$. Consequently $(p_2, u_1 y_2 - u_2 y_1)$ is still a conservation law. Therefore although u_1^* and u_2^* are not so easy to obtain, one knows a priori that (42) and (43) are satisfied. From (43) one obtains

$$(46) \quad \begin{aligned} \frac{\partial^2 u_1^*}{\partial x_2^2} &= \{p_2, \{p_2, u_1^*\}\} = -\{p_2, u_2^*\} = -u_1^*, \\ \frac{\partial^2 u_2^*}{\partial x_2^2} &= \{p_2, \{p_2, u_2^*\}\} = \{p_2, u_1^*\} = -u_2^*. \end{aligned}$$

Consequently as a function of x_2 one knows that u_1^* and u_2^* are of the form $a \sin x_2 + b \cos x_2$, $a, b \in \mathbb{R}$. More generally for any $L(x, u)$ of the form $L(x, u) = h(x_1) \cdot k(\frac{1}{2}(u_1^2 + u_2^2))$, with h and k arbitrary smooth functions, one has

$$\begin{aligned} \{p^T f(x, u) - L(x, u), p_2\} &= u_1 \left(y_2 + h(x_1) \frac{dk}{dz} \left(\frac{1}{2} (u_1^2 + u_2^2) \right) \right) \cdot 2u_2 \\ &\quad - u_2 \left(y_1 + h(x_1) \frac{dk}{dz} \left(\frac{1}{2} (u_1^2 + u_2^2) \right) \right) \cdot 2u_1 \\ &= u_1 y_2 - u_2 y_1. \end{aligned}$$

So again $(p_2, u_1 y_2 - u_2 y_1)$ is a conservation law.

Example 3. We shall show that in the linear-quadratic case there *cannot* exist quadratic conservation laws (F, F^e) with $F^e = 0$, if the system is controllable. Hence the methods of [4] are in this case not applicable. Consider a linear system $\dot{x} = Ax + Bu$ with $L(x, u) = \frac{1}{2}x^T Qx + \frac{1}{2}u^T Ru + u^T Sx$. A linear geometrical symmetry $\dot{x} = Gx$ with G a square matrix corresponds to a quadratic Hamiltonian $p^T Gx$. Calculating,

$$\begin{aligned} \{p^T (Ax + Bu) - \frac{1}{2}x^T Qx - \frac{1}{2}u^T Ru - u^T Sx, p^T Gx\} \\ = -p^T (AG - GA)x + p^T GBu + x^T QGx + u^T SGx. \end{aligned}$$

Now suppose $(p^T Gx, 0)$ is a conservation law. Then

$$(47) \quad AG = GA, \quad GB = SG = 0, \quad QG \text{ skew-symmetric.}$$

The first two equations yield $G(A^k B) = A^k GB = 0$, $k = 0, 1, \dots$. Hence if (A, B) is controllable necessarily $G = 0$! Feedback $u = Fx + Hv$, $\det H \neq 0$, cannot change this situation since $H(x, p, u)$ remains of the same form and (A, B) is controllable if and only if $(A + BF, BH)$ is controllable.

However, there *may* exist linear symmetries with $F^e(y, u) = \frac{1}{2}y^T My + y^T Nu$ (and so $F^e(0, u) = 0$). Consider for example the system $\dot{x} = u$ on \mathbb{R}^n with $L(x, u) = \frac{1}{2}x^T Qx + \frac{1}{2}u^T Ru$, where $Q = Q^T$ and $R = R^T > 0$. The Hamiltonian is $H(x, p, u) = p^T u - \frac{1}{2}x^T Qx - \frac{1}{2}u^T Ru$ and the optimal Hamiltonian is obtained by setting $\partial H / \partial u_j = 0$, which yields $u^* = R^{-1}p$, and hence $H^0(x, p) = \frac{1}{2}p^T R^{-1}p - \frac{1}{2}x^T Qx$.

Let us look at symmetries for $H^0(x, p)$ of the form $p^T Fx$, with F an $n \times n$ -matrix. $\{H^0(x, p), p^T Fx\} = p^T R^{-1} F^T p + x^T Q Fx$, and hence F has to be such that $R^{-1} F^T$ and QF are skew-symmetric. Now $\{H(x, p, u), p^T Fx\} = \{p^T u - \frac{1}{2} x^T Qx - \frac{1}{2} u^T Ru, p^T Fx\} = u^T F^T p - x^T Q Fx = u^T F^T p$. Let us take $F^e(y, u) = y^T Fu$. Since $y = \partial H / \partial u = p - Ru$ we obtain $F^e(y, u) = -u^T R F u + p^T F u$. Because $R^{-1} F^T$, or equivalently, RF has to be skew-symmetric, $F^e(y, u) = p^T F u$, and hence $\{H(x, p, u), p^T Fx\} = y^T F u$.

Remark. This last example shows the close connection of our theory with the original Noether theorem on symmetries of Lagrangian functions. This is further investigated in [9].

One of the most pressing questions is now the following. By looking at conservation laws (F, F^e) , with $F^e(0, u) = 0$, for the associated Hamiltonian system, can we obtain *all* the symmetries for the optimal Hamiltonian, and if not, which subclass of symmetries do we obtain?

The first part of this question is answered as follows. Let the dimension of the codistribution, generated by taking Poisson brackets of the functions $H(x, p, u)$ for each u , be $k \leq 2n$ (for simplicity we assume constant dimensions). Then there are exactly $2n - k$ independent functions K_i such that $\{H(x, p, u), K_i\} = 0$. Furthermore we can arbitrarily choose m independent functions F_i^e on \mathbb{R}^{2n} satisfying $F_i^e(0, u) = 0$ for all u . Hence by Theorem 5 there exist *at most* $\min(m, k)$ independent functions F_i on \mathbb{R}^{2n} , also independent from the functions K_i , such that there exist functions F_i^e on \mathbb{R}^{2m} in such a way that (F_i, F_i^e) are conservation laws for the Hamiltonian system. Hence, in general we do *not* obtain all the symmetries of the optimal Hamiltonian $H^0(x, p)$.

The second part of the question, which subclass of symmetries do we obtain, is much harder. Let X_F be a symmetry for the optimal Hamiltonian, i.e., $X_F(H^0) = \{F, H^0\} = 0$. Then it follows that

$$(48) \quad \{H(x, p, u), F(x, p)\} = F'(x, p, u)$$

with the function F' satisfying

$$(49) \quad F'(x, p, u^*(x, p)) = 0.$$

Now X_F corresponds to a conservation law for the Hamiltonian system if and only if $F'(x, p, u)$ can be written as a function of $y = \partial H / \partial u$ and u , i.e., if there exists a function $F^e : \mathbb{R}^{2m} \rightarrow \mathbb{R}$ such that $F'(x, p, u) = F^e((\partial H / \partial u)(x, p, u), u)$.

PROPOSITION 6. *Let $X_F(H^0) = 0$. Then there exists an $F^e : \mathbb{R}^{2m} \rightarrow \mathbb{R}$ such that (F, F^e) is a conservation law for the associated Hamiltonian system if and only if for every $G : \mathbb{R}^{2n} \rightarrow \mathbb{R}$ such that $(\partial / \partial u_j)\{H(x, p, u), G(x, p)\} = 0$, $j = 1, \dots, m$, it follows that $\{H(x, p, u), F(x, p)\}, G(x, p)\} = 0$.*

Proof. Let $\partial / \partial u_j \{H, G\} = 0$; then equivalently $X_G(y_j) = X_G(\partial H / \partial u_j) = \{\partial H / \partial u_j, G\} = 0$. Also let $\{H(x, p, u), F(x, p)\} = F'(x, p, u)$. Then $\{H(x, p, u), F(x, p)\}, G(x, p)\} = -X_G(F') = 0$, for every such G , implies that $F'(x, p, u)$ only depends on y and u , and hence is of the form $F^e(\partial H / \partial u, u)$. Since $F'(x, p, u^*(x, p)) = 0$, it follows that $F^e(0, u) = 0$. \square

By Theorem 4 it is also a necessary condition for a symmetry X_F of H^0 to be obtainable from a conservation law (F, F^e) that F satisfies equations of the form $\{F, u_i^*\} = G_i(u^*)$. This brings us to another interesting point. If we apply *feedback* $u = \alpha(x, v)$, $v \in \mathbb{R}^m$, with the matrix $\partial \alpha / \partial v$ nonsingular, to the system $\dot{x} = f(x, u)$ and the running cost $L(x, u)$, we obtain

$$(50) \quad \dot{x} = \tilde{f}(x, v) := f(x, \alpha(x, v)), \quad \tilde{L}(x, v) := L(x, \alpha(x, v))$$

resulting in a new Hamiltonian

$$(51) \quad \tilde{H}(x, p, v) = p^T \tilde{f}(x, v) - \tilde{L}(x, v).$$

Now it is clear that

$$(52) \quad \begin{aligned} \max_v \tilde{H}(x, p, v) &= \max_v p^T f(x, \alpha(x, v)) - L(x, \alpha(x, v)) \\ &= \max_u p^T f(x, u) - L(x, u) = \max_u H(x, p, u) = H^0(x, p). \end{aligned}$$

Hence the optimal Hamiltonian $H^0(x, p)$ does not change under feedback, and consequently the symmetries for H^0 remain the same. However the Hamiltonian systems associated respectively to $H(x, p, u)$ and $\tilde{H}(x, p, v)$ are really different. (It is in general not true that by applying feedback $u = \beta(x, p, v)$ to the Hamiltonian system resulting from $H(x, p, u)$ one can obtain the Hamiltonian system corresponding to $\tilde{H}(x, p, v)$.) Consequently the set of conservation laws (F, F^e) for both Hamiltonian systems are in general different. Hence it may happen that for a symmetry X_F of H^0 there exists an F^e such that (F, F^e) is a conservation law for $H(x, p, u)$, while there does *not* exist an \tilde{F}^e such that (F, \tilde{F}^e) is a conservation law for $\tilde{H}(x, p, v)$. Moreover if $u^*(x, p)$ and $v^*(x, p)$ are the optimal controls resulting from maximizing H and \tilde{H} , then there may exist functions G_i such that $\{F, u_i^*\} = G_i(u^*)$, but no functions \tilde{G}_i such that $\{F, v_i^*\} = \tilde{G}_i(v^*)$. The following question is therefore worthwhile to investigate.

Question. Let $H(x, p, u) = p^T f(x, u) - L(x, u)$ be the Hamiltonian of an optimal control problem yielding the optimal Hamiltonian $H^0(x, p) = \max_{u \in \mathbb{R}^m} H(x, p, u)$. Let X_F be a symmetry for H^0 , i.e., $X_F(H^0) = 0$. Does there exist a feedback $u = \alpha(x, v)$, $v \in \mathbb{R}^m$, and a smooth function \tilde{F}^e on \mathbb{R}^{2m} such that (F, \tilde{F}^e) is a conservation law for the Hamiltonian system corresponding to $\tilde{H}(x, p, v) = p^T \tilde{f}(x, v) - \tilde{L}(x, v)$?

If the above question can be answered affirmatively, then in a sense all the symmetries for the optimal Hamiltonian can be recovered from symmetries of an associated Hamiltonian system.

Remark. The above question is also related to the problem of bringing $H(x, p, u)$ into some kind of normal form by feedback transformations $u = \alpha(x, v)$ and state space transformations. If we allow for the *larger* class of transformations $u = \alpha(x, p, v)$ and take the usual assumption that $(\partial^2 H / \partial u_i \partial u_j)$ is nonsingular (say for simplicity negative definite), then the Morse Lemma yields for $H(x, p, u)$ the normal form $\bar{H}(x, p) - \sum_{j=1}^m u_j^2$. Hence we end up with the Hamiltonian system

$$(53) \quad \begin{aligned} \dot{x}_i &= \frac{\partial \bar{H}}{\partial p_i}(x, p), \\ \dot{p}_i &= -\frac{\partial \bar{H}}{\partial x_i}(x, p), \end{aligned} \quad y_j = u_j.$$

In this degenerate case $H^0(x, p) = \bar{H}(x, p)$, and $\{\bar{H}(x, p) - \sum_{j=1}^m u_j^2, F(x, p)\}$ equals a function $F^e((\partial H / \partial u), u)$ if and only if $\{\bar{H}, F\} = 0$. It would be interesting to extend the Morse Lemma to the smaller class of transformations $u = \alpha(x, v)$.

Acknowledgments. I would like to thank Jessy Grizzle and Steve Marcus for some conversations, which stimulated me in extending the original observations made in [8] to the present paper. Also I would like to thank Henk Nijmeijer for his help in the proof of Theorem 4.

REFERENCES

- [1] R. A. ABRAHAM AND J. E. MARSDEN, *Foundations of Mechanics*, 2nd ed., Benjamin/Cumming, Reading, MA, 1981.
- [2] R. W. BROCKETT, *Control theory and analytical mechanics*, in Geometric Control Theory, C. Martin and R. Herman, eds., Lie Groups: History, Frontiers and Applications, Vol. VII, Math-Sci Press, Brookline, MA, 1977, pp. 1-46.
- [3] ———, *Book review*, Foundations of Mechanics, R. A. Abraham and J. E. Marsden, IEEE Trans. Automat. Control, AC-26 (1981), pp. 977-978.
- [4] J. W. GRIZZLE AND S. I. MARCUS, *Optimal control of systems possessing symmetries*, IEEE Trans. Automat. Control, AC-29 (1984), pp. 1037-1040.
- [5] A. J. VAN DER SCHAFT, *Symmetries and conservation laws for Hamiltonian systems with inputs and outputs: A generalization of Noether's theorem*, Systems Control Lett., 1 (1981), pp. 108-115.
- [6] ———, *Observability and controllability for smooth nonlinear systems*, this Journal, 20 (1982), pp. 338-354.
- [7] ———, *Hamiltonian dynamics with external forces and observations*, Math. Systems Theory, 15 (1982), pp. 145-168.
- [8] ———, *System theoretic descriptions of physical systems*, Doctoral dissertation, Groningen, 1983, also CWI Tracts No. 3, CWI Amsterdam, 1984.
- [9] ———, *Optimal control and Hamiltonian input-output systems*, Proc. Conf. on Algebraic and Geometric Methods in Non-linear Control Theory, Paris, 1985, to appear.
- [10] M. A. SHAYMAN, *A symmetry group for the matrix Riccati equation*, Systems Control Lett., 2 (1982), pp. 17-24.

PATHWISE NONLINEAR FILTERING FOR NONDEGENERATE DIFFUSIONS WITH NOISE CORRELATION*

M. H. A. DAVIS† AND M. P. SPATHOPOULOS†

Abstract. We consider the filtering problem of calculating $E[f(x_t)|y_s, s \leq t]$ where x_t is a nondegenerate diffusion process on a finite dimensional manifold with given generator and $y_t = \int_0^t h(x_s) ds + w_t^0$ is a real valued observation process. w_t^0 is a scalar Brownian motion whose correlation with x_t is specified by a given vector field Z . It is shown that this filtering problem is robust, i.e., the estimates for x_t can be calculated separately for each observation sample path y_t .

Key words. nonlinear filtering, noise correlation, conditional densities, Markov process, extended generator, stochastic differential equation, manifold, horizontal lifting

AMS(MOS) subject classifications. 93E11, 60H20, 60J60

Introduction. This paper concerns the nonlinear filtering problem of calculating recursively estimates $E[f(x_t)|y_s, 0 \leq s \leq t]$, where x_t is a Markov process on a manifold M and y_t is a real-valued observation process, with possible correlation between the respective "noises." The main set-up of this paper is the following:

The signal process x_t is an A -diffusion process on the manifold M where A is a given nondegenerate second-order differential operator. The scalar observation process y_t is given by the following equation:

$$(0.1) \quad dy_t = h(x_t) dt + dw_t^0.$$

Here w^0 is a standard Brownian motion (BM). Finally, the possible correlation between the signal and the observation process noise is specified by a given vector field Z on M (see (2.1) below).

In Part I a pathwise solution is obtained when the signal is given by the stochastic differential equation¹

$$(0.2) \quad df(x_t) = L_0 f(x_t) dt + L_i f(x_t) \circ dw_t^i, \quad x_0 = x$$

where $L_0, L_i, i = 1, \dots, d$ are C^∞ vector fields, and $w_t^i, i = 1, \dots, d$ are independent scalar BM. It is supposed that there are functions $a_i(\cdot), i = 1, 2, \dots, d$ such that

$$\langle w^i, w^0 \rangle_t = \int_0^t a_i(x_s) ds.$$

It is known [9] that the unconditional distribution $u(t, x) = E[f(x(t, x, w))]$ of the "signal" flow $x = x(t, x, w)$ of diffeomorphisms constructed by (0.2) is obtained by solving the heat equation

$$\frac{\partial u(t, x)}{\partial t} = Au(t, x), \quad \lim_{\substack{t \downarrow 0 \\ p \rightarrow x}} u(t, p) = f(x)$$

where A is the extended generator of the diffusion process x_t given by $A = L_0 + \frac{1}{2} \sum_{i=1}^d L_i^2$.

* Received by the editors November 12, 1984; accepted for publication (in revised form) December 20, 1985.

† Department of Electrical Engineering, Imperial College, London SW7 2BT, England.

¹ In (0.2) and throughout the paper, the Einstein summation convention is used. Stochastic integrals written $\sigma_t \circ dw_t^i$ are Stratonovich integrals whereas those written $\sigma_t dw_t^i$ are Ito integrals.

The main result of the “pathwise” theory of nonlinear filtering is that it is possible to compute the conditional distribution of x_t given $\{y_s, 0 \leq s \leq t\}$, in terms of the solution of another heat equation whose coefficients depend on the observation sample path $\{y_s, 0 \leq s \leq t\}$, i.e. a heat equation of the form

$$\frac{\partial U(t, x)}{\partial t} = A_t^\gamma U(t, x), \quad \lim_{\substack{t \downarrow 0 \\ p \rightarrow x}} U(t, p) = f(x),$$

where A_t^γ is a sample-path dependent generator to be computed. In Part I we compute this generator for the above correlated noise case. These results are a slight generalization of those in [2]; they are included here for completeness.

However, this formulation is quite restrictive. Given a nondegenerate generator A on a manifold M of dimension d it is not always possible to find d independent BM such that equation (0.2) generates an A -diffusion process on $M(d)$. This only can be done if $M(d)$ is parallelizable (for example, S^k is parallelizable only for $k = 1, 3, 7$). In general, $n > d$ BM will be required for this, where we only know an upper bound of n which is $2d + 1$ (Whitney’s imbedding theorem). So, there is no systematic way to calculate the d vector fields in equation (0.2).

Therefore if instead of the equation (0.2) an A -diffusion process is given on a manifold $M(d)$, a different set-up will be required.

In Part II we tackle the nonlinear filtering problem when the signal is an A -diffusion process on the manifold $M(d)$, the observation is given by the same equation (0.1) and the noise correlation is represented by a vector field Z .

Given a generator A , an A diffusion can be obtained on M using the “lifting” technique. Motivated by this technique, we find that lifting the signal and the observation equation on the orthonormal frame bundle $O(M)$ of M we can use the same approach as in Part I to derive a pathwise solution on $O(M)$. Then the projection of this solution on M will be the required pathwise solution of our problem under the new formulation. In Part II we find the pathwise solution on $O(M)$ and we calculate the generator \tilde{A}_t^γ on $O(M)$. Then we calculate the projection, A_t^γ , of this generator \tilde{A}_t^γ on M which corresponds to a pathwise solution of the problem on M . Thus the frame bundle, horizontal lifting etc. disappear in the formulation of the final result (Theorem 2.4).

Our main result, Theorem 2.4, involves apparently rather restrictive conditions: *nondegenerate diffusions* and *one-dimensional observations*. As regards nondegeneracy, we remark that theory in Part I covers a wide class of degenerate signal processes, so long as they are given in differential equation form (1.2). When the signal is specified only by its generator, we require nondegeneracy in order to apply the horizontal lifting technique. There is in fact little general theory available for handling degenerate diffusions specified in this form (cf. Stroock and Varadhan [15]).

The theory extends to vector-valued observations when there is no noise correlations ($Z = 0$) but generally not otherwise. As pointed out in [1] and [2] the pathwise method fails if the noise vector fields Z_i corresponding to vector observations $\{y_t^i, i = 1, \dots, m\}$ do not commute.

The reason for this is that although a decomposition of the type (2.11) is still possible, ξ_t is now given by the SDE

$$df(\xi_t(x)) = Z_i f(\xi_t(x)) \circ dy_t^i, \quad \xi_0(x) = x.$$

Stochastic flow theory tells us that $\xi_t(\cdot)$ is almost surely a diffeomorphism, but no continuous dependence on $\{y_t^1, \dots, y_t^m\}$ can be expected; see [10].

It is argued in [1], [2] that pathwise filtering theory is the only theory with any relevance to practical applications, since such applications involve sample paths which are a null set in any "almost sure" theory. The purpose of this paper is to see how far the pathwise theory will go, and the result is that it can be extended to certain cases where the signal is not originally specified in terms of SDE's.

PART I

1.1. The signal and observation equations. For notational consistency with Part II below, the signal process is here denoted r_t , and its generator \tilde{A} . Thus the filtering problem is given by the observation equation

$$(1.1) \quad dy_t = h(r_t) dt + dw_t^0$$

where y_t is the real valued observed process and the signal process r_t is the solution of the stochastic differential equation

$$(1.2) \quad df(r_t) = L_0 f(r_t) dt + L_j f(r_t) \circ dw_t^j, \quad j = 1, \dots, d, \quad f \in C^\infty(N).$$

The process r_t evolves on a σ -compact, connected C^∞ manifold N of dimension n , and L_0, L_1, \dots, L_d are C^∞ vector fields on N . r_0 has a given distribution ν , and w^1, \dots, w^d are independent scalar Brownian motions. We assume r_t has infinite lifetime. To allow possible correlation between the signal noise (w^1, \dots, w^d) and the observation noise w^0 , we suppose there are C^∞ functions a_1, \dots, a_d such that

$$(1.3) \quad \langle w^i, w^0 \rangle_t = \int_0^t a_i(r_s) ds.$$

The corresponding Ito form of (1.2) is

$$df(r_t) = \tilde{A}f(r_t) dt + d\tilde{M}_t^f$$

where $\tilde{A} = L_0 + \frac{1}{2} \sum_{i=1}^d L_i^2$ is the extended generator of r_t and

$$\tilde{M}_t^f := f(r_t) - f(0) - \int_0^t \tilde{A}f(r_s) ds = \int_0^t L_i f(r_s) dw_s^i.$$

Clearly

$$\langle M_t^f, w_t^0 \rangle = \int_0^t \sum_i^d a_i(r_s) L_i f(r_s) ds = \int_0^t \tilde{Z}f(r_s) ds$$

where \tilde{Z} is the vector field

$$(1.4) \quad \tilde{Z}(r) = \sum_{j=1}^d a_j(r) L_j(r).$$

PROPOSITION 1.1. *If (1.3) holds, then $\sum_{i=1}^d a_i^2(r) \leq 1$ for all $r \in N$ except a set of potential zero for r_t .*

Proof. If M_1, M_2 are continuous local martingales and $H_i \in L_2(\langle M_i \rangle)$ $i = 1, 2$, then inequality VII 54.3 from [5] states that

$$\left| \int_0^t H_1 H_2 d\langle M_1, M_2 \rangle \right| \leq \left(\int_0^t H_2^2 d\langle M_2, M_2 \rangle \right)^{1/2} \left(\int_0^t H_1^2 d\langle M_1, M_1 \rangle \right)^{1/2}.$$

Taking

$$H_1 = 1, \quad H_2 = 1, \quad M_1 = \int_0^t a_i(r_s) dw_s^i, \quad M_2 = w_t^0,$$

we have

$$\left| \sum_i \int_0^t a_i^2(r_s) ds \right| \leq t^{1/2} \left(\sum_i \int_0^t a_i^2(r_s) ds \right)^{1/2}$$

or

$$\sum_i \int_0^t a_i^2(r_s) ds \leq t$$

i.e.

$$\int_0^t \left(\sum_i a_i^2(r_s) - 1 \right) ds \leq 0.$$

The conclusion follows, since this holds for all t and starting points r_0 . We shall make the following assumption.

ASSUMPTION 1.1.

$$(1.5) \quad \sum_i a_i^2(r) < 1 \quad \text{for all } r \in N.$$

This essentially means that w_t^0 always contains some “fresh” noise. As pointed out in [12] the above assumption is a nondegeneracy hypothesis, which is crucial for the Zakai equation to have a nice solution. If it is not satisfied, the conditional law of x_t given $\mathcal{Y}_t = \sigma(y, s \leq t)$ may not have a density w.r.t. Lebesgue measure. Let

$$u_t^y(f) := E[f(r_t) | \mathcal{Y}_t].$$

It is convenient to calculate an unnormalized form σ_t of u_t^y so that u_t^y is then given by

$$u_t^y(f) = \sigma_t(f) / \sigma_t(1)$$

where “1” denotes the function $1(r) = 1$. $\sigma_t(f)$ satisfies the Zakai equation

$$(1.6) \quad d\sigma_t(f) = \sigma_t(\tilde{A}f) dt + \sigma_t(\tilde{D}f) dy_t$$

where

$$(1.7) \quad \tilde{D} := \tilde{Z} + h.$$

To get the appropriate form of the Kallianpur–Striebel formula for this problem, we introduce a measure P_0 via the Girsanov transformation

$$\frac{dP_0}{dP} = \exp \left(- \int_0^T h(r_s) dw_s^0 - \frac{1}{2} \int_0^T h^2(r_s) ds \right)$$

and for $i = 1, \dots, d$ define

$$(1.8) \quad dv^i := dw^i + a_i(r_t)h(r_t) dt.$$

Then, under P_0 ,

- (i) y_t and v_t^i , $i = 1, \dots, d$ are BM's;
- (ii) v_t^i, v_t^j are independent for $i \neq j$;
- (iii) $\langle v^i, y \rangle_t = \int_0^t a^i(r_s) ds$.

Now project the v^i onto y , i.e., define

$$(1.9) \quad \tilde{b}_t^i := v_t^i - \int_0^t a_i(r_s) dy_s.$$

Then

$$\langle \tilde{b}^i, y \rangle_t = 0$$

and

$$\langle \tilde{b} \rangle_t = \int_0^t (I - a(r_s) a'(r_s)) ds$$

where

$$[\langle \tilde{b} \rangle_t]_{i,j} = \langle \tilde{b}^i, \tilde{b}^j \rangle_t,$$

$a' = (a'_1, \dots, a'_d)$ and $I = d \times d$ identity matrix.

Since $\sum_i a_i^2(r) < 1$ the matrix $I - a(r)a'(r)$ is positive definite and can be factored into a product of positive definite matrices $\tilde{\Delta}(r)\tilde{\Delta}'(r)$. Defining

$$(1.10) \quad b_t := \int_0^t \tilde{\Delta}^{-1}(r_s) d\tilde{b}_s,$$

we find that $\langle b \rangle_t = It$ and $\langle b^i, y \rangle_t = 0$ i.e., b^1, \dots, b^d, y are (under measure P_0) independent standard BM's. We now express the signal equation (1.2) as a stochastic differential equation driven by b_t and y_t . From (1.2) and (1.8)

$$(1.2') \quad df(r_t) = \hat{Y}_0 f(r_0) dt + L_t f(r_t) \circ dv_t^i$$

where

$$\hat{Y}_0 := L_0 - h\tilde{Z}.$$

Since (1.9) involves Ito integrals, we express (1.2') in Ito form as

$$df(r_t) = \left(\hat{Y}_0 f + \frac{1}{2} \sum L_i^2 f \right) dt + L_t f dv_t^i.$$

When we use (1.9) and (1.10), this becomes

$$(1.2'') \quad df(r_t) = \left(\hat{Y}_0 f + \frac{1}{2} \sum L_i^2 f \right) dt + Y_j f db^j + \tilde{Z} f dy$$

where \tilde{Z} is given by (1.4), and in vector notation the Y_j are given by

$$Y := \tilde{\Delta}' L.$$

Transforming back to Stratonovich integrals in (1.2''), we finally obtain

$$(1.11) \quad df(r_t) = Y_0 f(r_t) dt + \tilde{Z} f(r_t) \circ dy_t + Y_j f(r_t) \circ db_t^j$$

where

$$Y_0 = L_0 - h\tilde{Z} - \frac{1}{2} \left[\sum_i \tilde{\delta}_i^j L_j \tilde{\delta}_i^k + a^j L_j a^k \right] L_k, \quad \tilde{\Delta} = (\tilde{\delta}_k^n)$$

$$\left(= \hat{Y}_0 + \frac{1}{2} \sum (L_i^2 - Y_i^2) - \frac{1}{2} \tilde{Z}^2 \right).$$

Equation (1.11) is the key formula for the filtering problem, as it expresses r_t in the form of an equation driven by the observation process y_t and the other "inputs" b^1, \dots, b^d which are independent of y . The next task is to decompose (1.11) in such a way that the dependence of r_t and y is explicitly brought out. This is done by the Doss-Sussmann technique as in [2], [10].

1.2. The KS formula and the associated multiplicative functional. Let $\tilde{\zeta}_t(r) = \tilde{\zeta}(t, r)$ denote the flow of the vector field \tilde{Z} , i.e., the unique solution of the equation

$$\frac{d}{dt}f(\tilde{\zeta}_t(r)) = \tilde{Z}f(\tilde{\zeta}_t(r)), \quad f \in C^\infty(N), \quad \tilde{\zeta}_0(r) = r.$$

This is a diffeomorphism for all $t \geq 0$. Define

$$\tilde{\xi}_t(r) := \tilde{\zeta}_{y_t}(r).$$

As is easily checked, $\tilde{\xi}_t(r)$ is the solution of

$$d\tilde{\xi}_t(r) = \tilde{Z}(\tilde{\xi}_t(r)) \circ dy_t$$

and obviously $\tilde{\xi}_t(r)$ is a.s. a diffeomorphism for all $t \geq 0$. Now consider the equation

$$(1.12) \quad df(\eta_t) = \tilde{\xi}_t^{-1} Y_0 f(\eta_t) dt + \tilde{\xi}_t^{-1} Y_j f(\eta_t) \circ db_t^j$$

where $\tilde{\xi}_{t*}: T_p(N) \rightarrow T_q(N)$, ($q = \tilde{\xi}_t(p)$) is the differential map and $\tilde{\xi}_t^*(f) := f \circ \tilde{\xi}_t$.

The equation has a unique solution $\eta_t = \eta_t(r)$. Applying the extended Ito formula [10], we get

$$(1.13) \quad r_t(r) = \tilde{\xi}_t \circ \eta_t(r) = \tilde{\zeta}(y_t, \eta_t).$$

The representation (1.12), (1.13) describes the behaviour of r_t conditioned on y under P_0 . Recall that the map $\tilde{\xi}_t^{-1}$ is parametrized by y and that y, b are independent. Thus, conditioned on y, η_t is a diffusion process whose differential generator is

$$(1.14) \quad \tilde{A}_t^* = \tilde{\xi}_t^{-1} Y_0 + \frac{1}{2} \sum_{i=1}^d (\tilde{\xi}_t^{-1} Y_i)^2$$

and for each $t > 0$, r_t is diffeomorphically related to η_t by (1.13).

THEOREM 1.1. *The conditional distribution $u_t^y(f)$ of r_t given \mathcal{Y}_t is given by*

$$(1.15) \quad u_t^y(f) = \frac{\sigma_t(f)}{\sigma_t(1)}$$

where

$$(1.16) \quad \sigma_t(f) = \langle \tilde{T}_{0,t}^y(\tilde{B}_{y_t} f), \nu \rangle.$$

In (1.16), $\tilde{T}_{0,t}^y$ is a semigroup whose extended generator is

$$(1.17) \quad \begin{aligned} \tilde{A}_t^y &= \exp(\tilde{H}_{y_t}) \tilde{A}_t^* \exp(-\tilde{H}_{y_t}) - \frac{1}{2} \tilde{\xi}_t^*(\tilde{D}h), \\ \tilde{B}_t &:= \exp\left(\int_0^t \tilde{\xi}_u^* h(r) du\right) \tilde{\xi}_t^* f(r) \end{aligned}$$

and

$$\tilde{H}_t := \int_0^t \tilde{\xi}_s^* h(r) ds.$$

Proof. We know from a standard formula of conditional expectations that $\sigma_t(f)$ is given in terms of the measure P_0 by

$$\sigma_t(f) := E_0 \left[f(r_t) \exp \left(\int_0^t h(r_s) dy_s - \frac{1}{2} \int_0^t h^2(r_s) ds \right) \middle| \mathcal{Y}_t \right].$$

It is immediate from (1.11) that

$$(1.18) \quad d\langle h(r), y \rangle_t = \tilde{Z}h(r_t) dt$$

and hence the Stratonovich version of this is

$$(1.19) \quad \sigma_t(f) = E_0 \left[f(r_t) \exp \left(\int_0^t h(r_s) \circ dy_s - \frac{1}{2} \int_0^t \tilde{D}h(r_s) ds \right) \middle| \mathcal{Y}_t \right]$$

where \tilde{D} is given by (1.7).

Now since $r_t = \tilde{\xi}_t \circ \eta_t(r)$ and η_t is a functional of the independent processes y_t and $b_t^i = (b_t^1, \dots, b_t^d)$, we can express (1.19) in the form

$$(1.20) \quad \sigma_t(f) = E^b \left(\tilde{\xi}_t^* f(\eta_t) \exp \left(\int_0^t \tilde{\xi}_s^* h(\eta_s) \circ dy_s - \frac{1}{2} \int_0^t \tilde{\xi}_s^* \tilde{D}(\eta_s) ds \right) \right)$$

where E^b means integration over the sample space measure for b_t (=Wiener measure on $C([0, T]; \mathbb{R}^n)$). This is the KS formula for the correlated noise problem. In order to get it in robust form, we need to calculate the stochastic integral in (1.20) as an explicit functional of y .

Introduce the function

$$(1.21) \quad \tilde{H}_y(r) = \tilde{H}(y, r) := \int_0^y \tilde{\xi}_s^* h(r) ds,$$

and calculate $\tilde{H}(y_t, r_t)$ using the Ito formula and (1.13). This gives

$$(1.22) \quad \tilde{H}(y_t, r_t) = \int_0^t h(r_s) \circ dy_s + \int_0^t (\tilde{\xi}_s^{-1} Y_0) \tilde{H}_{y_s}(\eta_s) ds + \int_0^t (\tilde{\xi}_s^{-1} Y_j) \tilde{H}_{y_s}(\eta_s) \circ db_s^j.$$

The stochastic integral with respect to b^j in (1.22) can be reexpressed in Ito form in the standard way using (1.12). Do this and introduce the notation

$$(1.23) \quad Y_j^* := \tilde{\xi}_s^{-1} Y_j, \\ \tilde{B}_{y_t} f(r) := \exp \left(\int_0^{y_t} \tilde{\xi}_u^* h(r) du \right) \tilde{\xi}_{y_t}^* f(\dot{r}).$$

Then using (1.23) in (1.20) gives

$$(1.24) \quad \sigma_t(f) = E[\tilde{B}_{y_t} f(\eta_t) \tilde{a}_t^0(y)]$$

where

$$(1.25) \quad \tilde{a}_t^0(y) := \exp \left[- \int_0^t Y_j^* \tilde{H}_{y_u}(\eta_u) db_u^j - \frac{1}{2} \int_0^t \sum (Y_j^*)^2 \tilde{H}_{y_u}(\eta_u) du \right. \\ \left. - \int_0^t Y_0^* \tilde{H}_{y_u}(\eta_u) du - \frac{1}{2} \int_0^t \xi_u^* \tilde{D}h(\eta_u) du \right].$$

For each $y \in C$, $\tilde{a}_t^0(y)$ is a multiplicative functional (m.f.) of η_t and hence the formula

$$\tilde{T}_{s,t}^y f(\eta) = E_{s,\eta}[f(\eta_t) \tilde{a}_t^s(y)]$$

defines a semigroup of operators on $B(N)$ in terms of which (1.24) becomes

$$\sigma_t(f) = \langle \tilde{T}_{0,t}^y \tilde{B}_{y_t} f, \nu \rangle.$$

We can thus compute $\sigma_t(f)$ by solving the forward equation corresponding to $\tilde{T}_{0,t}^y$ involving $(\tilde{A}_t^y)^*$ where \tilde{A}_t^y is the extended generator of $\tilde{T}_{0,t}^y$ (as defined in (31) of [3]).

We now calculate \tilde{A}_t^y by factoring $\tilde{a}_t^0(y)$ as follows:

$$\begin{aligned} \tilde{a}_t^0(y) = & \exp \left[- \int_0^t Y_j^* \tilde{H}_{y_u}(\eta_u) db_u^j - \frac{1}{2} \int_0^t \sum (Y_j^* \tilde{H}_{y_u}(\eta_u))^2 du \right] \\ & \times \exp \left[\frac{1}{2} \int_0^t \sum (Y_j^* \tilde{H}_{y_u}(\eta_u))^2 du - \frac{1}{2} \int_0^t \sum (Y_j^*)^2 \tilde{H}_{y_u}(\eta_u) du \right. \\ & \left. - \int_0^t Y_0^* \tilde{H}_{y_u}(\eta_u) du - \frac{1}{2} \int_0^t \tilde{\xi}_u^* \tilde{D}h(\eta_u) du \right]. \end{aligned}$$

Since the first term corresponds to a Girsanov type m.f. and the second to a Feynman-Kac type m.f. it follows [2] that

$$\tilde{A}_t^y f = [\tilde{A}_t^* f - \sum Y_j^* \tilde{H}_{y_t} Y_j^* f] + (\frac{1}{2} \sum (Y_j^* \tilde{H}_{y_t})^2 - \tilde{A}_t^* \tilde{H}_{y_t} - \frac{1}{2} \tilde{\xi}_t^* \tilde{D}h) f$$

where \tilde{A}_t^* is the generator of η_t , i.e.,

$$\tilde{A}_t^* = Y_0^* + \frac{1}{2} \sum_{j=1}^d (Y_j^*)^2.$$

In [2] it has been proved that

$$e^g \tilde{A} e^{-g} = \tilde{A} - \sum (L_j g) L_j - \tilde{A} g + \frac{1}{2} \sum (L_j g)^2.$$

When we use the above expression, the generator \tilde{A}_t^y becomes

$$\tilde{A}_t^y = e^{\tilde{H}_{y_t}} \tilde{A}_t^* e^{-\tilde{H}_{y_t}} - \frac{1}{2} \tilde{\xi}_t^* (\tilde{D}h)$$

which is equation (1.17) and the proof is completed.

PART II

2.1. Problem formulation. We consider the filtering problem where now the “signal” process x_t is a diffusion process on the manifold $M(d)$ with a given nondegenerate extended generator A . By this we mean that for all $f \in C^\infty(M)$

$$M_t^f := f(x_t) - f(x_0) - \int_0^t A f(x_s) ds$$

is a local martingale. In local coordinates A is given as

$$A f(x) = \frac{1}{2} a^{ij}(x) \frac{\partial^2 f}{\partial x_i \partial x_j}(x) + b^i(x) \frac{\partial f}{\partial x_i}(x), \quad f \in C^\infty(M)$$

and $a^{ij}(x) q_i q_j > 0$ for all x and $q = (q_i) \in \mathbb{R}^d - \{0\}$.

We assume that x_t has infinite lifetime. This is true if $M(d)$ is compact; conditions under which it holds for noncompact M are discussed by Elworthy [8].

The observation process is given by the following equation:

$$dy_t = h(x_t) dt + dw_t^0.$$

Since the joint process (x_t, w_t^0) is a Hunt process, we know from results in [3] that the joint variation $\langle M^f, w^0 \rangle_t$ must take the form

$$\langle M^f, w^0 \rangle_t = \int_0^t Z f(x_s, w_s^0) ds$$

for some function Zf , and it is shown in [3] that the map $f \rightarrow Zf(x, w^0)$ is a derivation, i.e., satisfies the Leibnitz rule, except perhaps on (x, w) sets of potential zero.

For later development it is necessary to assume that Zf does not depend on w^0 . Then Z is a derivation on M , i.e., $(f \rightarrow Zf(x)) \in T_x(M)$. If we further assume that $Zf(x)$ varies smoothly with x , then Z is a vector field. Thus a natural formulation of the idea of noise correlation is to suppose that a C^∞ vector field Z is given such that

$$(2.1) \quad \langle M^f, w^0 \rangle_t = \int_0^t Zf(x_s) ds.$$

We write Z in local coordinates as

$$(2.2) \quad Z = \bar{a}^i(x) \frac{\partial}{\partial x_i}.$$

Finally, we assume that Z is complete, i.e.

$$\zeta_t(x) \text{ is defined for all } t > 0$$

where $\zeta_t(x)$ is the flow of Z .

2.2. Construction of the process x_t, w_t^0 by horizontal lifting. We use the horizontal lifting technique described in § 5.4 of Ikeda and Watanabe [9] to which the reader is referred for complete details. We summarize briefly here, beginning with some preliminary definitions from differential geometry.

Let M be a C^∞ -manifold and let $\mathcal{X}(M)$ denote the set of vector fields on M . By an *affine connection* ∇ we mean a rule which associates to every $X \in \mathcal{X}(M)$ a linear mapping $\nabla_X : \mathcal{X}(M) \rightarrow \mathcal{X}(M)$ having the following properties:

- (i) $\nabla_X Y$ is bilinear in X and Y ,
- (ii) $\nabla_{fX+gY} = f\nabla_X + g\nabla_Y$,
- (iii) $\nabla_X(fY) = f\nabla_X Y + (Xf)Y$.

The operator ∇_X is called covariant differentiation with respect to X . The components of the connection ∇ are defined as the functions $\{\Gamma_{jk}^i(x)\}$ such that in local coordinates

$$\nabla_{\partial_i} \partial_j = \Gamma_{ij}^k(x) \frac{\partial}{\partial x_k} \quad \left(\partial_i = \frac{\partial}{\partial x_i} \right).$$

An affine connection is called symmetric if $\Gamma_{ij}^k = \Gamma_{ji}^k$. In local coordinates $\nabla_X Y$ may be expressed as

$$\nabla_X Y = \left[X^i(x) \frac{\partial}{\partial x_i} Y^k(x) + \Gamma_{ij}^k(x) X^i(x) Y^j(x) \right] \frac{\partial}{\partial x_k}$$

where

$$X = X^i(x) \frac{\partial}{\partial x_i} \quad \text{and} \quad Y = Y^i(x) \frac{\partial}{\partial x_i}.$$

For I an interval of the real line, let $c : I \ni t \rightarrow c(t) \in M$ be a smooth curve in M , and let $X(t) \in T_{c(t)}(M)$ for $t \in I$. $X(t)$ is said to be parallel along c (with respect to ∇) if

$$\frac{d}{dt} X^i(t) + \Gamma_{kj}^i(c(t)) X^j(t) \frac{dc^k(t)}{dt} = 0, \quad t \in I.$$

For $t_0, t_1 \in I$, $t_0 \leq t_1$, $X(t_1)$ is uniquely determined from $X(t_0)$ and we say that $X(t_1)$ is obtained from $X(t_0)$ by *parallel displacement* along the curve $c(t)$. A C^∞ -manifold is called *Riemannian* if a tensor field $g = (g_{ij})$ of type $(0, 2)$ is given on M such that

- (i) g is symmetric, i.e., $g_{ij}(x) = g_{ji}(x)$,

(ii) g is positive definite, i.e., $g_{ij}(x)q^i q^j > 0$ for all x and $q \in R^d$, $q \neq 0$.
 g is called the Riemannian metric. It defines an inner product on each tangent space $T_x(M)$ by

$$\langle X, Y \rangle = g_{ij}(x) X^i Y^j.$$

An affine connection $\nabla = \{\Gamma_{ij}^k\}$ is said to be compatible with the Riemannian metric g if the inner product is preserved during a parallel displacement of tangent vectors. There exists a unique symmetric affine connection compatible with g , and this is the so-called *Riemannian connection*.

Given the second order operator A on M , we can define a Riemannian metric g on $M(d)$ by

$$g_{ij} = (a^{ij})^{-1} \quad (\text{i.e. } g^{ij} = a^{ij}).$$

Let $G(x) := (a^{ij}(x))$. Then $M(d)$ is a Riemannian manifold.

The components of the corresponding Riemannian connection, called Christoffel symbols, are given by

$$(2.3) \quad \Gamma_{ij}^k = \frac{1}{2} \left[\frac{\partial}{\partial x_i} g_{mj} + \frac{\partial}{\partial x_j} g_{im} - \frac{\partial}{\partial x_m} g_{ij} \right] g^{km}.$$

Now let us introduce the bundle of orthonormal frames on M . By an orthonormal frame $e = [e_1, e_2, \dots, e_d]'$ at x we mean an orthonormal base of $T_x(M)$. Then $O(M)$ is defined as the collection of all orthonormal frames at all points $x \in M$:

$$O(M) := \{r = (x, e) : x \in M, e \text{ is a frame at } x\}$$

and in local coordinates

$$r = (x^i, e_j^i) \in R^d \times R^{d(d+1)/2}$$

where

$$e_j = e_j^i \frac{\partial}{\partial x_i}, \quad j = 1, \dots, d,$$

and since e_i 's are orthonormal in $T_x(M)$

$$g_{kl} e_i^k e_j^l = \delta_j^i \quad ^2$$

or equivalently

$$\sum_{m=1}^d e_m^i e_m^j = g^{ij}(x) = a^{ij}(x).$$

An element β of the orthogonal group $O(d)$ acts on $O(M)$ by

$$(2.4) \quad T_\beta(x, e) = (x, e\beta)$$

where $e\beta = [(e\beta)_1, (e\beta)_2, \dots, (e\beta)_d]'$ is an orthonormal frame at x defined by

$$(e\beta)_j = \beta_j^i e_i, \quad j = 1, \dots, d.$$

Therefore $O(M)$ is a principal fibre bundle over M with the structural group $O(d)$.

The Riemannian connection of M determines a splitting of the tangent bundle to $O(M)$ into vertical and horizontal components. For each $r \in O(M)$

$$H_r = \left\{ X = a^i \left(\frac{\partial}{\partial x_i} \right)_x - \Gamma_{kl}^i(x) e_j^l a^k \frac{\partial}{\partial e_j^i}; \quad (a^i) \in R^d \right\},$$

² By δ_q^{μ} we mean the Kronecker delta.

the *horizontal subspace*, is a linear subspace of $T_r(O(M))$ which is independent of the choice of local coordinates (x^i, e_j^i) .

A vector field X on $O(M)$ is called horizontal if $X(r) \in H_r$ for each $r \in O(M)$. Let π denote the projection map $\pi: O(M) \rightarrow M$ and let $d\pi: TO(M) \rightarrow TM$ denote its differential map. Then given $X \in \mathcal{X}(M)$, there exists a unique $\tilde{X} \in \mathcal{X}(O(M))$ such that \tilde{X}_r is the horizontal lift of $\tilde{X}_{\pi(r)}$ for all $r \in O(M)$. \tilde{X} is called the horizontal lift of X and in local coordinates is given by

$$\tilde{X} = X^i(x) \frac{\partial}{\partial x_i} - \Gamma_{ij}^q(x) X^j(x) e_p^i \frac{\partial}{\partial e_p^q} \quad \text{if } X = X^i(x) \frac{\partial}{\partial x_i}.$$

Similarly, given a smooth curve $c(t)$ on M , the curve $\tilde{c}(t)$ on $O(M)$ is called a horizontal lift of c if (a) $(d\tilde{c}/dt)(t)$ is horizontal and (b) $\pi(\tilde{c}(t)) = c(t)$. Then the curve $\tilde{c}(t)$ is given by $\tilde{c}(t) = (c(t), e(t))$ where the e 's are parallel translated along $c(t)$. Finally we can define the system of the canonical horizontal vector fields $\{L_1, \dots, L_d\}$ such that L_j is the horizontal lift of $e_j \in T_x(M)$ for all $r = (x, e)$. L_j may be expressed as

$$(2.5) \quad L_j = e_j^i \frac{\partial}{\partial x_i} - \Gamma_{kt}^q e_j^k e_p^t \frac{\partial}{\partial e_p^q}.$$

Let us now return to the filtering problem as formulated in § 2.1.

ASSUMPTION 2.1. $\bar{a}(x)\bar{a}'(x) < G(x)$.

Remark. In R^n , $g_{ij} = \delta_{ij}$, $\Gamma_{ij}^q = 0$ and the above inequality corresponds to $aa' < I$, which was the assumption introduced in Part I.

The process (x_t, w_t^0) can be constructed by the "horizontal lifting" technique. The following result is proved in [9].

THEOREM 2.1. In Part I, take $N = O(M)$, L_i , $i = 1, \dots, d$ the canonical horizontal vector fields and let L_0 be the horizontal lift of the vector field

$$\bar{b} = \bar{b}^i(x) \frac{\partial}{\partial x_i}, \quad \bar{b}^i = b^i + \frac{1}{2} g^{ij} \Gamma_{ij}^i.$$

If we define $r(t)$ by (1.2) then $x_t = \pi(r_t)$ is a diffusion process on M with extended generator A .

Remarks. The diffusion $r_t = r(t, r, w)$ on $O(M)$ is the flow of diffeomorphisms defined by

$$(2.5') \quad dr(t) = L_i(r(t)) \circ dw^i + L_0(r(t)) dt, \quad r(0) = r.$$

Let $\tilde{A} = L_0 + \frac{1}{2} \sum_{i=1}^d L_i^2$ be the generator of the diffusion process $r(t)$. One has to show that this corresponds to the horizontal lift of the generator A , i.e., $d\pi(\tilde{A}) = A$. To show this, we take $f(r_t) = f(x_t, e_t) \equiv f(x_t)$ and prove that $\tilde{A}f(x_t) \equiv Af(x_t)$. In [9] it is proved that

$$\frac{1}{2} d\pi \left[\sum_{i=1}^d L_i^2 f \right] = \frac{1}{2} \sum_{i=1}^d L_i(L_i f(x_t)) = \frac{1}{2} \left[a^{ij} \frac{\partial f}{\partial x_i} \frac{\partial f}{\partial x_j} - g^{ij} \Gamma_{ij}^i \frac{\partial f}{\partial x_i} \right]$$

so finally we have

$$d\pi(\tilde{A}) = d\pi(L_0) + \frac{1}{2} \left[\sum_{i=1}^d L_i^2 \right] = \bar{b} + \frac{1}{2} \left[a^{ij} \frac{\partial}{\partial x_i} \frac{\partial}{\partial x_j} - g^{ij} \Gamma_{ij}^i \frac{\partial}{\partial x_i} \right] = A.$$

The Markov property of x_t is a consequence of the fact that $x(t, T_\beta r, w) = x(t, r, w\beta)$ and the remark that $w\beta$ is another d -dimensional Wiener process. Therefore, the probability law of $x(\cdot, T_\beta r, w)$ is independent of $\beta \in O(d)$, i.e., depends only on $x = \pi(r)$.

We have constructed x_t in terms of d -dimensional BM(w^1, \dots, w^d). We now have to construct a BM w_t^0 such that (2.1) holds.

LEMMA 2.1. *Augment w^1, \dots, w^d with a further independent BM w^{d+1} . Then we can choose a BM w^0 s.t. $d\langle w^i, w^0 \rangle_t = \gamma^i(r_t) dt$ where $\gamma^i(r)$ $i = 1, \dots, d$ are any real-valued measurable functions satisfying*

$$\sum_{i=1}^d (\gamma^i(r))^2 \leq 1 \quad \text{for all } r.$$

Proof. Define

$$(2.6) \quad w_t^0 = \sum_{i=1}^d \int_0^t \gamma^i(r_s) dw_s^i + \int_0^t \theta_s dw_s^{d+1}$$

where

$$\theta_s = \left[1 - \sum_{i=1}^d (\gamma^i(r_s))^2 \right]^{1/2}.$$

THEOREM 2.2. *Let w^0 be a BM defined by (2.6) where $\gamma^i(r)$ is given by*

$$(2.7) \quad \gamma^i(r) = (e^{-1})_j^i \bar{a}^j(x)$$

and (e^{-1}) denotes the inverse of the matrix $e = (e_j^i)$. Then $\langle M^f, w^0 \rangle_t = \int_0^t Zf(x_s) ds$ where Z is the given vector field $Z(x) = \bar{a}^i(x)(\partial/\partial x_i)$ $i = 1, \dots, d$.

Proof. Since $df(r_t) = L_0 f(r_t) dt + L_i f(r_t) \circ dw_t^i$, it follows that

$$d\langle M^f, w^0 \rangle_t = L_i f(r_t) d\langle w^i, w^0 \rangle_t.$$

Now if $f: M \rightarrow R$ then

$$L_j f(r_t) = e_j^i \frac{\partial f(x_t)}{\partial x_i}$$

which implies that

$$d\langle M^f, w^0 \rangle_t = e_j^i \frac{\partial f(x_t)}{\partial x_i} \gamma^j(r) dt = e_j^i \frac{\partial f(x)}{\partial x_i} (e^{-1})_k^j \bar{a}^k(x) dt$$

and since

$$(e)(e^{-1}) = I \Rightarrow e_j^i (e^{-1})_k^j = \bar{\delta}_k^i$$

we conclude that

$$d\langle M^f, w^0 \rangle_t = \bar{a}^k(x) \frac{\partial f(x)}{\partial x_k} dt$$

or

$$\langle M^f, w^0 \rangle_t = \int_0^t Zf(x, ds).$$

Remark. The function $\gamma^i(r)$ is an intrinsically defined function, i.e. independent of local coordinates, as can easily be proved. (A similar function appears at the end of § 5.2 of [9]). The inequality $\sum_{i=1}^d (\gamma^i(r))^2 < 1$ is satisfied according to Corollary 2.2 below.

LEMMA 2.2. *Given the vector field $Z = \bar{a}^i(x)(\partial/\partial x_i)$ on $M(d)$, then the horizontal lift \tilde{Z} with respect to the Riemannian connection can be expressed as*

$$\tilde{Z}(r) = \gamma^i(r)L_i$$

with γ^i given by (2.7).

Proof. Using (2.5) and (2.7) in the above expression, we take

$$\begin{aligned}\tilde{Z}(r) &= \bar{a}^i(e^{-1})^j_i e_j^k \frac{\partial}{\partial x_k} - \bar{a}^i(e^{-1})^j_i \Gamma_{kl}^q e_j^k e_p^l \frac{\partial}{\partial e_p^q} \\ &= \bar{a}^i \frac{\partial}{\partial x_i} - \Gamma_{kl}^q \bar{a}^k e_p^l \frac{\partial}{\partial e_p^q}\end{aligned}$$

which is by definition the horizontal lift of Z .

DEFINITION 2.1. For $f \in C^\infty(O(M))$ we denote by \tilde{M}^f the martingale

$$\tilde{M}_t^f = f(r_t) - f(r_0) - \int_0^t \tilde{A}f(r_s) ds.$$

Then if $f: M \rightarrow \mathbb{R}$, $M_t^f \equiv \tilde{M}_t^f$.

LEMMA 2.3. *Let $\bar{a}(x) = (\bar{a}_1(x), \dots, \bar{a}_d(x))'$ and $G(x) = (a^{ij}(x))$. Then under our problem formulation structure the following inequality holds for all $x \in M$ except a set of potential zero for x_i , independently of Assumption 2.1 above.*

$$G(x) - \bar{a}(x)\bar{a}'(x) \geq 0.$$

Proof. Let $x_t = \pi(r_t)$ where r_t denotes the solution of (2.5') with arbitrary r_0 . Since $\langle M_t^f, w_t^0 \rangle = \int_0^t Zf(x_s) ds$ inequality VII. 54.3 of [5] becomes

$$\left| \int_0^t H_1 H_2 Zf(x_s) ds \right| \leq \left[\int_0^t H_1^2 ds \right]^{1/2} \left[\int_0^t H_2^2 d\langle M_t^f, M_t^f \rangle \right]^{1/2}$$

and taking $H_1 = Zf(x_s)$, $H_2 = 1$ we have

$$\left| \int_0^t (Zf(x_s))^2 ds \right| \leq \left[\int_0^t (Zf(x_s))^2 ds \right]^{1/2} [\langle M_t^f, M_t^f \rangle]^{1/2},$$

which implies that

$$(2.8) \quad \int_0^t (Zf(x_s))^2 ds \leq \langle M^f, M^f \rangle_t \quad \forall t.$$

From [3] it is known that

$$(2.9) \quad \langle M^f, M^f \rangle_t = \int_0^t \Delta_A^{ff}(x_s) ds$$

where

$$\Delta_A^{fg} = A(fg) - fAg - gAf.$$

Therefore using (2.8) and (2.9), we have

$$\int_0^t |Zf(x_s)|^2 ds \leq \int_0^t \Delta_A^{ff}(x_s) ds$$

and since in local coordinates

$$\Delta_A^f(x) = a^{ij}(x) \frac{\partial f(x)}{\partial x_i} \frac{\partial f(x)}{\partial x_j} = \nabla f'(x) G(x) \nabla f(x),$$

$$Zf(x) = \bar{a}^i(x) \frac{\partial f(x)}{\partial x_i} = \bar{a}'(x) \nabla f(x),$$

we conclude that

$$\int_0^t \nabla f'(x_s) \bar{a}(x_s) \bar{a}'(x_s) \nabla f(x_s) ds \leq \int_0^t \nabla f'(x_s) G(x_s) \nabla f(x_s) ds \quad \forall t$$

or

$$\int_0^t (G(x_s) - \bar{a}(x_s) \bar{a}'(x_s)) ds \geq 0 \quad \text{for all } t.$$

The result follows.

COROLLARY 2.1. $I - \gamma\gamma' \geq 0$ up to sets of potential zero.

Proof. From Lemma 2.3

$$G - \bar{a}\bar{a}' \geq 0$$

and since $ee' = G$, rearranging, we obtain

$$I - (e^{-1}) \bar{a}\bar{a}'(e^{-1})' \geq 0$$

or equivalently

$$I - \gamma\gamma' \geq 0.$$

Remark. Since $\langle \tilde{M}^f, w^0 \rangle_t = \int_0^t \tilde{Z}f(r_s) ds$, working as in Lemma 2.3 on $O(M)$ we can also conclude that $I - \gamma\gamma' \geq 0$ up to sets of potential zero.

COROLLARY 2.2. If $G - \bar{a}\bar{a}' > 0$ then $I - \gamma\gamma' > 0$.

Proof. Follows directly from Corollary 2.1.

2.3. The generator A_t^y . With the above construction done Part I can be applied, i.e. we have a pathwise filter for r_t given \mathcal{Y}_t .

Our objective is to show that there is a pathwise filter for x_t on M involving only “downstairs” objects. First consider the following lemma.

LEMMA 2.4. Given a curve $x(t)$ on M , consider two horizontal lifts $\tilde{c}_1(t)$ and $\tilde{c}_2(t)$ of the curve $x(t)$, with $\tilde{c}_2(0) = T_\beta \tilde{c}_1(0)$ for some $\beta \in O(d)$ (see (2.4)). Then for all $t \geq 0$

$$(2.10) \quad \tilde{c}_2(t) = T_\beta \tilde{c}_1(t).$$

Proof. The parallel translating of e 's along the curve $x(t)$, in local coordinates, is given by

$$\frac{de_a^i(t)}{dt} = -\Gamma_{kj}^i(x_t) e_a^j(t) \frac{dx^k}{dt}(t),$$

$$x^i(0) = x^i,$$

$$e_a^i(0) = e_a^i.$$

Obviously the rotation of the initial frame e_a^i does not affect the above equation, and therefore both sides of (2.10) satisfy the same differential equation.

In Part I take $N = O(M)$, L_i , $i = 1, \dots, d$ the canonical horizontal vector fields, L_0 as given in Theorem 2.1, $h(r) \equiv h(x)$ $a(r) = \gamma(r)$, $\tilde{Z}(r)$ the horizontal lift of the vector field Z and $r = (x, e) \in O(M)$ where $\pi(r) = x \in M$.

Then the signal equation becomes

$$df(r_t) = Y_0 f(r_t) dt + \tilde{Z}f(r_t) \circ dy_t + Y_j f(r_t) \circ dw_t^j, \quad j = 1, \dots, d$$

where

$$Y_0 = L_0 - h\tilde{Z} - \frac{1}{2} \left[\sum_i \tilde{\delta}_i^j L_j \tilde{\delta}_i^k + \gamma^j L_j \gamma^k \right] L_k, \quad Y = \tilde{\Delta}' L, \quad \tilde{\Delta} \tilde{\Delta}' = I - \gamma \gamma'.$$

The matrix $I - \gamma \gamma'$ is positive definite and can be factorized into the product $\tilde{\Delta} \tilde{\Delta}'$ since according to Corollary 2.2 $I - \gamma \gamma' > 0$. Let $\Delta \Delta' := G - \bar{a} \bar{a}'$. Clearly $\tilde{\Delta} \tilde{\Delta}' = (e^{-1}) \Delta \Delta' (e^{-1})'$.

By decomposing the signal equation we have

$$(2.11) \quad r_t = \tilde{\xi}_t \circ \eta_t(r) = \tilde{\zeta}(y_t, \eta_t)$$

where

$$d\tilde{\xi}_t = \tilde{Z}(\tilde{\xi}_t) \circ dy_t$$

and $\eta_t(r)$ is the diffusion process

$$d\eta_t = \tilde{\xi}_t^{-1} Y_0 f(\eta_t) dt + \tilde{\xi}_t^{-1} Y_j f(\eta_t) \circ dw_t^j, \quad \eta_0 = r$$

whose differential generator is

$$\tilde{A}_t^* = \tilde{\xi}_t^{-1} Y_0 + \frac{1}{2} \sum_{i=1}^d (\tilde{\xi}_t^{-1} Y_i)^2.$$

The process $\eta_t(r)$ has a unique solution up to some explosion time τ . Recall that $\tilde{\zeta}_t(r)$ is defined for all t since the vector field Z is assumed to be complete. Then if $r_t(r)$ is defined for all t , $\eta_t(r)$ is also defined for all t since from (2.11)

$$\eta_t(r) = \tilde{\zeta}(-y_t, r_t(r)).$$

Thus we conclude that, if the signal equation has no explosion and the vector field Z is complete, then $\tau = \infty$.

Since $\tilde{\zeta}_t(r) = \tilde{\zeta}(t, r)$ is the flow of the horizontal vector field \tilde{Z} on $O(M)$, it follows that $\tilde{\zeta}_t$ is the horizontal lift of the curve ζ_t on M (i.e., the flow of the vector field Z on M). Thus $\tilde{\xi}_t = \tilde{\zeta}_{y_t}$ is the horizontal lift of the curve ξ_t on M .

It is quite straightforward to conclude that in local coordinates

$$\tilde{\xi}_t^{-1}(x, e) = (\tilde{\xi}_t^{-1}(x, e)^k, \tilde{\xi}_t^{-1}(x, e)^r_s) = (\xi_t^{-1}(x)^k, e_s^r(t))$$

where the $e_s^r(t)$ are parallel translated along the curve $\xi_t^{-1}(x)$ and therefore

$$\frac{\partial \tilde{\xi}_t^{-1}(x, e)^k}{\partial e_p^q} = 0.$$

Since $\tilde{\xi}_t$ is a diffeomorphism on $O(M)$, $\tilde{\xi}_t^{-1} L_j \in \mathcal{H}(O(M))$, it is given in local coordinates by

$$(2.12) \quad \begin{aligned} L_j^* f(l) &:= (\tilde{\xi}_t^{-1} L_j(x)) f(l) \\ &= e_j^i \frac{\partial \tilde{\xi}_t^{-1}(x)^k}{\partial x_i} \frac{\partial f}{\partial l_k} \\ &\quad + \left[e_j^i \frac{\partial \tilde{\xi}_t^{-1}(x, e)^r_s}{\partial x_i} - \Gamma_{k\mu}^q(x) e_j^k e_p^\mu \frac{\partial \tilde{\xi}_t^{-1}(x, e)^r_s}{\partial e_p^q} \right] \frac{\partial f}{\partial e_s^r(l)} \\ &\quad \forall f \in C^\infty(\pi^{-1}(U_l(M))), \quad l = \xi_t^{-1}(x), \quad e_s^r(l) = \tilde{\xi}_t^{-1}(x, e)^r_s. \end{aligned}$$

THEOREM 2.3. $x_t := \pi(\eta_t)$ is a diffusion process on M with generator

$$(2.13) \quad A_t^* = \frac{1}{2} [a^{k\mu}(x) - \bar{a}^k(x) \bar{a}^\mu(x)] \frac{\partial^*}{\partial x_k} \frac{\partial^*}{\partial x_\mu} + \left[\frac{1}{2} \sum_m \frac{\partial \delta_m^p(x)}{\partial x_j} \delta_m^j(x) - \frac{1}{2} \frac{\partial \bar{a}^p(x)}{\partial x_k} \bar{a}^k(x) + b^p(x) - h(x) \bar{a}^p(x) \right] \frac{\partial^*}{\partial x_p}$$

where by $\partial^*/\partial x_k$ we mean

$$\xi_t^{-1} \frac{\partial}{\partial x_k} = \frac{\partial \xi_t^{-1}(x)^a}{\partial x_k} \frac{\partial}{\partial l_a} = \left[\frac{\partial \xi_t(x)^a}{\partial x_k} \right]^{-1} \frac{\partial}{\partial l_a},$$

$$x = \xi_t(l).$$

Proof. $\eta_t(\eta) = (\eta(t, \eta, w))$ is the solution of the stochastic differential equation

$$(2.14) \quad d\eta(t) = (\tilde{\xi}_t^{-1} Y_j)(\eta(t)) \circ dw^j(t) + (\tilde{\xi}_t^{-1} Y_0)(\eta(t)) dt, \quad \eta(0) = \eta,$$

i.e., $\eta(t, \eta, w)$ is the flow of diffeomorphisms on $O(M)$ corresponding to the vector fields $\tilde{\xi}_t^{-1} Y_j$, $j = 1, \dots, d$ and the drift vector field $\tilde{\xi}_t^{-1} Y_0$. Let $B = d\pi(\tilde{\xi}_t^{-1} Y_0)$.

It has been proved [14] that in local coordinates (2.14) is equivalent to

$$dx^i(t) = \tilde{\delta}_k^m(\eta_t) e_m^j(t) \frac{\partial \xi_t^{-1}(x)^i}{\partial x_j} \circ dw^k(t) + B^i(x_t) dt,$$

$$de_s^r(t) = -\Gamma_{\mu\mu}^q(x_t) \tilde{\delta}_i^m(\eta_t) e_m^k(t) e_p^\mu(t) \frac{\partial \tilde{\xi}_t^{-1}(x, e)_s^r}{\partial e_p^q} \circ dw^i(t) + \tilde{\delta}_i^m(\eta_t) e_m^j(t) \frac{\partial \tilde{\xi}_t^{-1}(x, e)_s^r}{\partial x_j} \circ dw^i(t) - \Gamma_{m\mu}^r(x_t) B^m(x_t) e_s^\mu(t) dt$$

$$i = 1, \dots, d, \quad r, s = 1, \dots, d,$$

$$x^i(0) = x^i,$$

$$e_s^r(0) = e_s^r.$$

$$B^i(x_t) = \left[b^k(x_t) + \frac{1}{2} g^{\mu j}(x_t) \Gamma_{\mu j}^k(x_t) - h(x_t) \bar{a}^k(x_t) - \frac{1}{2} \sum_m \frac{\partial \delta_m^k(x_t)}{\partial x_j} \delta_m^j(x_t) - \frac{1}{2} \Gamma_{\mu j}^k(x_t) \sum_m \delta_m^\mu(x_t) \delta_m^j(x_t) - \frac{1}{2} \bar{a}^j(x_t) \frac{\partial \bar{a}^k(x_t)}{\partial x_j} - \frac{1}{2} \Gamma_{\mu j}^k(x_t) \bar{a}^\mu(x_t) \bar{a}^j(x_t) \right] \frac{\partial \xi_t^{-1}(x)^i}{\partial x_k},$$

$$\Delta \Delta' = G - \bar{a} \bar{a}',$$

$$\Delta = (\delta_k^n)$$

and $n(t) = (x^i(t), e_s^r(t))$. That the solution $\eta(t)$ lies on $O(M)$ if $\eta(0) \in O(M)$ is clear since $\tilde{\xi}_t^{-1} Y_j$ is a vector field on $O(M)$.

A stochastic curve $x(t) = (x^i(t))$ on M is defined by $x(t) = \pi(\eta(t))$. The curve $x(t)$ in M depends on the choice of the initial frame e at x .

It follows that

$$(2.15) \quad x(t, T_\beta \eta, w) = x(t, \eta, w\beta), \quad t \in [0, \infty), \quad \beta \in O(d)$$

where $w\beta = (w(t)\beta)$ is another d -dimensional Wiener process. Indeed according to Lemma 2.4, since $\tilde{\xi}_t^{-1}(x, e\beta) = (\tilde{\xi}_t^{-1}(x, e))(\beta)$ it can be shown [14] that

$$\left[\frac{\partial \tilde{\xi}_t^{-1}(x, e\beta)_s^r}{\partial e_p^a} \right] = (\beta)' \left[\frac{\partial \tilde{\xi}_t^{-1}(x, e)_s^r}{\partial e_p^a} \right] (\beta)$$

and thus this matrix remains orthogonally equivalent³ after the rotation of the initial frame. The above matrices are of dimension $d \times d$ for r, s fixed. Similarly

$$\left| \frac{\partial \tilde{\xi}_t^{-1}(x, e\beta)_s^r}{\partial x_j} \right| = \left| \frac{\partial \tilde{\xi}_t^{-1}(x, e)_s^r}{\partial x_j} \right| (\beta)$$

where now the above matrix is of dimension $d \times 1$ for r, s fixed. The functions $\tilde{\delta}_k^m(t)$ are independent of any rotation of the initial frame, since they represent intrinsic functions given by the equation $\tilde{\Delta}\tilde{\Delta}' = I - \gamma\gamma'$ where γ is an intrinsic function. The drift term $B^i(x_t)$ is independent of the frame e . Therefore (2.15) follows at once.

Hence the probability law of $x(\cdot, T_\beta\eta, w)$ is independent of $\beta \in O(d)$ and depends only on $x = \pi(\eta)$. We denote it by P_x . It is now easy to deduce the strong Markov property of the system $[P_x]$ from that of $\eta(\cdot, \eta, w)$.

Thus, as in Theorem 2.1, there will be a diffusion process $x_t = \pi(\eta_t)$ on M having as extended generator $A_t^* = d\pi(\tilde{A}_t^*)$.

Taking $f: M \rightarrow R$, we can find the generator $A_t^* = d\pi(\tilde{A}_t^*)$ as in Theorem 2.1. After extensive calculations it has been shown [14] that A_t^* is given by (2.13) and this completes the proof.

With reference to KS formula (1.20) it is evident that the conditional distribution σ_t is entirely determined by the law of the "downstairs" process $x_t = \pi(\eta_t)$. This observation leads to the main result.

THEOREM 2.4. *The conditional distribution $u_t^y(f)$ of the A -diffusion process x_t given \mathcal{Y}_t is given by*

$$u_t^y(f) = \frac{\sigma_t(f)}{\sigma_t(1)}$$

where

$$\sigma_t(f) = \langle T_{0,t}^y(B_{y_t}f), \nu \rangle;$$

$T_{0,t}^y$ is a semigroup whose extended generator is

$$A_t^y = e^{H_{y_t}} A_t^* e^{-H_{y_t}} - \frac{1}{2} \xi_t^*(Dh), \quad D = Z + h,$$

$$B_t := \exp \left(\int_0^t \xi_u^* h(x) du \right) \xi_t^* f(x),$$

$$H_t := \int_0^t \xi_s^* h(x) ds.$$

Proof. According to Theorem 1.1, Part I, the associated generator \tilde{A}_t^y on $O(M)$ will be given by

$$\tilde{A}_t^y = e^{\tilde{H}_{y_t}} \tilde{A}_t^* e^{-\tilde{H}_{y_t}} - \frac{1}{2} \tilde{\xi}_t^*(\tilde{D}h)$$

where

$$\tilde{H}_t(r) = \int_0^t \tilde{\xi}_s^* h(r) ds \quad \text{and} \quad \tilde{B}_t = \exp \left(\int_0^t \tilde{\xi}_u^* h(r) du \right) \tilde{\xi}_t^* f(r).$$

³ $O(d)$ -equivalent in Ikeda and Watanabe's terminology [9].

But since $h(r) \equiv h(x)$, we conclude that

$$\tilde{H}_{y_t}(r) \equiv H_{y_t}(x).$$

Taking $f: M \rightarrow R$ the generator $A_t^y = d\pi(\tilde{A}_t^y)$ will be given by

$$A_t^y = e^{H_{y_t}} d\pi(\tilde{A}_t^*) e^{-H_{y_t} - \frac{1}{2}\xi_t^*(Dh)}$$

and

$$B_t = \exp\left(\int_0^t \xi_u^* h(x) du\right) \xi_t^* f(x),$$

which completes the proof.

2.4. Concluding remarks. The problem of pathwise nonlinear filtering for non-degenerate diffusions on manifolds with noise correlation has been studied. Two different models for the diffusion processes have been considered. The first concerned diffusions constructed as solutions of stochastic differential equations on manifolds. The second concerned diffusions given by their extended generator. In both cases scalar observation processes with possible correlation with the signal processes have been considered.

In Theorem 2.4 the smoothness assumptions can be weakened considerably. Thus in Part I it is only necessary to consider C^1 vector fields L_i and C^1 function a_i . Similarly in Part II the coefficients $a^j(x)$ of the generator A must be C^1 so that the connections Γ_{ij}^k can be defined, and the "noise" vector field Z must be C^1 .

Finally, some remarks on manifold-valued *observations*. These were first studied by Duncan [6]. There the stochastic differential equations in the tangent bundle are formulated and solved by using the notation of parallelism of vectors along a curve. Recently Ng and Caines [11] have considered the same problem by the use of the strong solutions of the SDE for the state and observation processes in the orthonormal frame bundles to give a direct derivation of the Zakai equation. In [13] Pontier and Szpirglas show that the filtration of the manifold-valued observation coincides with that of another process taking values in Euclidean space R^n , thus reducing the problem to the classical Zakai equation. A similar approach could possibly be used to extend our results to, say, observations on the circle but of course observations on higher-dimensional manifolds will meet the geometric obstructions mentioned in the introduction. A pathwise theory of filtering manifold valued observations (with no noise correlation) is developed in [16]. It is worth mentioning that in none of the above references is any form of noise correlation considered.

Acknowledgment. We are grateful to a referee who discovered an error in an earlier version of this paper.

REFERENCES

- [1] J. M. C. CLARK, *The design of robust approximations to the stochastic differential equations of nonlinear filtering*, in Communication Systems and Random Process Theory, J. K. Skwirzynski, ed., Sijthoff and Noordhoff, Alphen aan den Rijn, 1978.
- [2] M. H. A. DAVIS, *Pathwise non-linear filtering*, in Stochastic Systems, M. Hazewinkel and J. C. Willems, eds, Reidel, Dordrecht, 1981.
- [3] ———, *Factorization of a multiplicative functional of non-linear filtering theory*, Systems Control Lett., 1 (1981), pp. 49–53.
- [4] ———, *On a multiplicative functional transformation arising in nonlinear filtering theory*, Z. Wahrsch. Verw. Gebiete, 54 (1980), pp. 125–139.

- [5] C. DELLACHERIE AND P. MEYER, *Probabilités et potentiel*, Chapitres V et VIII: Théorie des martingales, Hermann, 1980.
- [6] T. E. DUNCAN, *Some filtering results in Riemannian manifolds*, Inform. and Control, 35 (1977), pp. 182–195.
- [7] R. J. ELLIOTT AND M. KOHLMANN, *Robust filtering for correlated multidimensional observations*, Math. Z., 178 (1981), pp. 559–578.
- [8] K. D. ELWORTHY, *Stochastic Differential Equations on Manifolds*, Cambridge University Press, Cambridge, 1982.
- [9] N. IKEDA AND S. WATANABE, *Stochastic Differential Equations and Diffusion Processes*, North-Holland, Amsterdam, 1981.
- [10] H. KUNITA, *On the decomposition of solutions of stochastic differential equations*, London Mathematical Society Symposium on Stochastic Integrals, Durham 1980.
- [11] S. K. NG AND P. E. CAINES, *Nonlinear filtering in Riemannian manifolds*, IMA J. Math. Control Inform., 2 (1985), pp. 25–36.
- [12] E. PARDOUX, *Non-linear filtering, prediction and smoothing*, in Stochastic Systems, M. Hazewinkel and J. C. Willems, eds., Reidel, Dordrecht, 1981.
- [13] M. PONTIER AND J. SZPIRGLAS, *Filtrage non-linéaire avec observation sur une variété*, Stochastics, 15 (1985), pp. 121–148.
- [14] M. P. SPATHOPOULOS, Ph.D. thesis, Imperial College of Science and Technology, London, 1986.
- [15] D. STROOCK AND R. S. VARADHAN, *Multidimensional Diffusion Processes*, Springer-Verlag, New York, 1979.
- [16] M. H. A. DAVIS AND M. P. SPATHOPOULOS, *Pathwise nonlinear filtering with observations on a manifold*, Proc. 25th IEEE Conference on Decision and Control, Athens, 1986.

IDENTIFIABILITY UNDER APPROXIMATION FOR AN ELLIPTIC BOUNDARY VALUE PROBLEM*

KARL KUNISCH† AND L. W. WHITE‡

Abstract. Necessary and sufficient conditions for identifiability of the diffusion coefficient in Galerkin approximations to a two point boundary value problem are derived for various choices of Galerkin subspaces. The results are further used to investigate output least squares identifiability and output least squares stability of the diffusion coefficient.

Key words. parameter identifiability, stability of output least squares problem, boundary value problem, Galerkin approximation

AMS(MOS) subject classification. 34A55

1. Introduction. In this note we consider the following boundary value problems:

$$(1.1) \quad \begin{aligned} -(au_x)_x + cu &= f \quad \text{in } (0, 1), \\ u_x(0) &= u_x(1) = 0. \end{aligned}$$

Let $I = (0, 1)$ and $f \in L^2(I)$. Recall that if $a \in H^1(I)$ with $a(x) \geq \alpha > 0$ and $c \in L^2(I)$ with $c(x) \geq c > 0$ a.e., then there exists a unique solution $u = u(a)$ of (1.1) in $H^2(I)$. We are concerned with the identification of the coefficient a , given information of the solution $u(a)$ and in particular we will study the injectivity of the mapping $a^M \rightarrow u^N(a^M)$ where a^M is some approximation to a and u^N an approximation to u . At first we describe the problem in a more general context.

Let $\mathcal{C}: H^2(I) \rightarrow Z$ be a continuous linear operator from the solution space to the observation space Z describing the type of available information of the state u . To determine the coefficient corresponding to an observation $z \in Z$ of the system that is modeled by (1.1), the following output least squares formulation is used frequently:

$$(1.2) \quad \underset{Q_{ad}}{\text{minimize}} \quad |\mathcal{C}u(a) - z|_Z^2.$$

Here the set Q_{ad} of admissible parameters is chosen such that the existence of a solution of (1.2) is guaranteed.

For example, if n observations $\{z_k\}_{k=1}^n$ taken at the points $\{x_k\}_{k=1}^n$ are available, we may take $Z = \mathbb{R}^n$ with $\mathcal{C}: H^2(I) \rightarrow \mathbb{R}^n$ defined by $\mathcal{C}u = \{u(x_k)\}_{k=1}^n$. In this case (1.2) becomes

$$(1.3) \quad \underset{Q_{ad}}{\text{minimize}} \quad \sum_{k=1}^n |u(x_k; a) - z_k|^2.$$

Alternatively one might have distributed observations $z \in L^2(I)$, or using the data $\{z_k\}_k$ at $\{x_k\}_{k=1}^n$ one might want to obtain a function $z \in L^2(I)$ either by interpolation or

* Received by the editors May 6, 1985; accepted for publication (in revised form) December 6, 1985.

† Division of Applied Mathematics, Brown University, Providence, Rhode Island 02912, and Institut für Mathematik, Technische Universität Graz, A-8010 Graz, Austria. This research was supported in part by the Fonds zur Förderung der Wissenschaftlichen Forschung, Austria, S3206, and by the Air Force Office of Scientific Research under contract AF-AFOSR 84-0398 (5-28393).

‡ Department of Mathematics, The University of Oklahoma, Norman, Oklahoma 73019. This research was supported in part by the Air Force Office of Scientific Research under contract AF-AFOSR 84-0271.

least squares regression; for \mathcal{C} we would then take $\mathcal{C}u = u$ and the optimization problem becomes

$$(1.4) \quad \underset{Q_{\text{ad}}}{\text{minimize}} \quad |u(a) - z|_{L^2}^2.$$

In either case an appropriate choice for Q_{ad} would be

$$Q_{\text{ad}} = \{a \in H^1: a(x) \geq \alpha > 0, |a|_{H^1} \leq \gamma\},$$

with $\gamma > \alpha$, [6].

Defining the attainable set $\mathcal{V} = \{\mathcal{C}u(a): a \in Q_{\text{ad}}\}$, one may view the optimization problem (1.2) as having two parts:

- (i) given $z \in Z$, find z_{proj} , the projection of z on \mathcal{V} ,
- (ii) given z_{proj} , find $\bar{a} \in Q_{\text{ad}}$ such that $\mathcal{C}u(\bar{a}) = z_{\text{proj}}$.

Assuming the existence of z_{proj} , the uniqueness of z_{proj} depends on the geometry of \mathcal{V} . In (ii) there exists an \bar{a} such that $\mathcal{C}u(\bar{a}) = z_{\text{proj}}$ by definition of Q_{ad} . The question of uniqueness of such an \bar{a} arises and it is guaranteed if $\Phi: a \rightarrow \mathcal{C}u(a)$ is injective at \bar{a} . Injectivity of Φ at \bar{a} is called *identifiability* of a at \bar{a} . The above mentioned uniqueness problems are rather involved in general, see, e.g., [2], [7, Appendix], and [11] for a hyperbolic equation.

When solving (1.2) on a computer it is necessary to replace (1.1)–(1.2) by a finite dimensional problem. This is done by approximating both the solutions of (1.1) and the set Q_{ad} by functions from finite dimensional function spaces. A finite dimensional version of the minimization problem (1.2) is then solved to obtain an estimate for the unknown coefficient a (compare e.g., [1], [5]). Again the existence and uniqueness questions analogous to the two steps (i) and (ii) above can be considered.

The main purpose of this investigation is the study of the uniqueness for the finite dimensional analogue of (ii). If for a chosen approximation of a by a^M the mapping $a^M \rightarrow u^N(a^M)$ is injective at \bar{a}^M , then a is called *identifiable under approximation* at \bar{a}^M . The related question for parabolic equations in dimension one has been treated in [4].

Our results below indicate that the injectivity of $a^M \rightarrow u^N(a^M)$ depends upon certain rank conditions that imply compatibility conditions upon the spaces used to approximate the coefficient a and the solution $u(a)$. It will be seen that a may be identifiable under approximation without the known sufficient conditions for identifiability of a in (1.1) being satisfied [9]. The results here, although depending on the choice of Neumann boundary conditions, can easily be adapted to different boundary conditions.

In § 2 we formulate the discrete problems and give general conditions for identifiability under approximation. In § 3 we examine several concrete examples and obtain necessary and sufficient conditions for identifiability under approximation for these cases. Identifiability will be guaranteed if there is a sufficient amount of movement in the coefficients of the basis element expansion for the approximate solution u^N , where u^N depends on the parameter a^M in question. On the other hand, if the parameter a^M is assumed to be known over those parts of the domain where u^N is stationary, then it can still be identifiable over the remaining parts of the domain $(0, 1)$.

Section 4 is devoted to the problem of continuous dependence of the solution of the discretized version of (1.3) or (1.4) on the observation z and Q_{ad} . Sufficient conditions for output least squares identifiability (OLSI) [2] and output least squares stability (OLS-stability) [3] are given.

Finally, in § 5 we report the findings of a numerical experiment that supports the practical relevance of our results.

2. Basic results. To approximate (1.2) by the standard finite element method [10], let $\{B_i\}_{i=0}^N$ and $\{\phi_j\}_{j=1}^M$ be sets of linearly independent functions defined on I with $B_i \in H^1(I)$ and ϕ_j piecewise continuous. Let $A^M = \text{span}\{\phi_j : j = 1, \dots, M\}$ and $H^N = \text{span}\{B_i : i = 0, \dots, N\}$. Setting

$$\tilde{u}^N = \sum_{i=0}^N \mu_i B_i \quad \text{and} \quad a^M = \sum_{j=1}^M a_j^M \phi_j,$$

we have upon integration by parts of (1.1) with u replaced by u^N :

$$\sum_{i=0}^N \mu_i \langle a B_{i,x}, B_{k,x} \rangle + \sum_{i=0}^N \mu_i \langle c B_i, B_k \rangle = \langle f, B_k \rangle \quad \text{for } k = 0, \dots, N,$$

where $\langle \cdot, \cdot \rangle$ denotes the inner product in L^2 . Replacing a by a^M it follows that

$$\sum_{i=0}^N u_i \sum_{j=1}^M a_j^M \langle \phi_j B_{i,x}, B_{k,x} \rangle + \sum_{i=0}^N \mu_i \langle c B_i, B_k \rangle = \langle f, B_k \rangle \quad \text{for } k = 0, \dots, N.$$

Rearranging the summations in this last expression, we arrive at

$$(2.1) \quad \sum_{j=1}^M a_j^M \sum_{i=0}^N \langle \phi_j B_{i,x}, B_{k,x} \rangle \mu_i + \sum_{i=0}^N \langle c B_i, B_k \rangle \mu_i = \langle f, B_k \rangle \quad \text{for } k = 0, \dots, N.$$

We now make the following definitions: H_j and K are $(N+1) \times (N+1)$ matrices with the (i, k) th elements given by

$$(H_j)_{i,k} = \langle \phi_j B_{i,x}, B_{k,x} \rangle \quad \text{and} \quad (K)_{i,k} = \langle c B_i, B_k \rangle,$$

for $i, k = 0, \dots, N; j = 1, \dots, M$. Similarly $\tilde{f} \in \mathbb{R}^{N+1}$, $\tilde{\mu} \in \mathbb{R}^{N+1}$ and $\tilde{a}^M \in \mathbb{R}^M$ are given by

$$(\tilde{f})_k = \langle f, B_k \rangle, \quad (\tilde{\mu})_i = \mu_i, \quad (\tilde{a}^M)_j = a_j^M.$$

With this notation (2.1) becomes

$$(2.2) \quad \sum_{j=1}^M a_j^M H_j \tilde{\mu} + K \tilde{\mu} = \tilde{f},$$

where we used the symmetry of H_j and K . Thus we obtain a mapping $\tilde{a}^M \rightarrow \tilde{\mu}(\tilde{a}^M)$ from \mathbb{R}^M into \mathbb{R}^{N+1} or, equivalently, $a^M \rightarrow u^N(a^M)$, $a^M = \sum_{j=1}^M a_j^M \phi_j$, from A^M to H^N , that is well defined as long as $\sum_{j=1}^M a_j^M H_j + K$ is invertible. For example, if $c \geq 0$ and $\{B_{i,x}\}_{i=0}^N$ are linearly independent, then $\sum_{j=1}^M a_j^M H_j + K$ is invertible for all

$$\tilde{a}^M \in \mathcal{A} = \left\{ \tilde{a}^M \in \mathbb{R}^M : \sum_{j=1}^M (a^M)_j \phi_j > 0 \quad \text{on } I \right\}.$$

Similarly, if $c \geq c > 0$ as assumed throughout, then again $\sum a_j^M H_j + K$ is invertible for all $\tilde{a}^M \in \mathcal{A}$.

We now define identifiability of $\tilde{a}^M = \text{col}(a_1^M, \dots, a_k^M)$ in (2.2).

DEFINITION 2.1. The parameter $\tilde{a}^M \in \mathcal{A}$ in (2.2) is called identifiable if $\tilde{b} \in \mathcal{A}$ and $\tilde{\mu}(\tilde{a}^M) = \tilde{\mu}(\tilde{b}^M)$ implies $\tilde{a}^M = \tilde{b}^M$.

For a specific choice of approximation of a in (1.1) by a^M , we say that a is identifiable under approximation at a^M if a^M is identifiable.

THEOREM 2.1. *Let $\{\phi_j\}_{j=1}^M$ and $\{B_i\}_{i=0}^N$ be linearly independent, $c(x) \geq c > 0$ and $\tilde{a}^M \in \mathcal{A}$. Then $\tilde{a}^M \in A$ is identifiable if and only if the vectors $\{H_j \tilde{\mu}(\tilde{a}^M)\}_{j=1}^M$ are linearly independent.*

Proof. Using (2.2), linear independence of $H_j \tilde{\mu}(\tilde{a}^M)$ clearly implies identifiability of \tilde{a}^M . Conversely assume that there exists a nontrivial vector $(\alpha_1, \dots, \alpha_M) \in \mathbb{R}^M$ with

$$(2.3) \quad \sum_{j=1}^M \alpha_j H_j \tilde{\mu}(\tilde{a}^M) = 0.$$

Then $\sum_{j=1}^M (a_j^M - \varepsilon \alpha_j) \phi_j > 0$ for some sufficiently small $\varepsilon > 0$ and \tilde{a}_1^M given by $(\tilde{a}_1^M)_j = a_j^M - \varepsilon \alpha_j$ satisfies $\tilde{a}_1^M \in \mathcal{A}$. Multiplying (2.3) by ε and subtracting it from (2.2), we find that $\tilde{\mu}(\tilde{a}^M) = \tilde{\mu}(\tilde{a}_1^M)$. This ends the proof.

Since $H_j: \mathbb{R}^{N+1} \rightarrow \mathbb{R}^{N+1}$, for $j = 1, \dots, M$, we have the following:

COROLLARY 2.1. *If $M > N + 1$, then \tilde{a}^M in (2.2) is not identifiable.*

COROLLARY 2.2. *If $\tilde{\mu}(\tilde{a}^M) \in \text{Ker}(H_j)$ for some $j = 1, \dots, M$, then \tilde{a}^M is not identifiable.*

Proof. If $\tilde{\mu}(\tilde{a}^M) \in \text{Ker}(H_j)$ then the set $\{H_j \tilde{\mu}(\tilde{a}^M)\}_{j=1}^M$ is linearly dependent and the result follows from Theorem 2.1.

Remark 2.1. Corollary 2.2 should be compared with the condition $|u_x| \geq k_1 > 0$ which is known to be a sufficient condition for identifiability of a in the infinite dimensional problem (1.1) [9].

DEFINITION 2.2. The coordinates $\{\tilde{a}_{j_k}^M\}_{k=1}^{\hat{M}}$, $\hat{M} \leq M$, of the parameter vector $\tilde{a}^M \in \mathcal{A}$ are called identifiable if $\tilde{b}^M \in \mathcal{A}$, $\tilde{\mu}(\tilde{a}^M) = \tilde{\mu}(\tilde{b}^M)$ and $\tilde{a}_{j_k}^M = \tilde{b}_{j_k}^M$ for all $j \neq j_k$, $k = 1, \dots, \hat{M}$ imply $\tilde{a}^M = \tilde{b}^M$.

PROPOSITION 2.1. *The coordinates $\{\tilde{a}_{j_k}^M\}_{k=1}^{\hat{M}}$ of $\tilde{a}^M \in \mathcal{A}$ are identifiable if and only if the vectors $\{H_{j_k} \tilde{\mu}(\tilde{a}^M)\}_{k=1}^{\hat{M}}$ are linearly independent.*

The proof is obvious from that of Theorem 2.1.

3. Several examples. In this section we consider several concrete examples and determine their identifiability properties. We point out that here we use N to denote the number of subintervals of I and N and M of the previous section are a function of this N .

Case 1. Let I be partitioned into N subintervals of length $1/N$. For $i = 0, \dots, N$ define the linear spline basis functions

$$(3.1) \quad B_i(x) = \begin{cases} Nx - i + 1, & \frac{i-1}{N} \leq x \leq \frac{i}{N}, \\ -Nx + i + 1, & \frac{i}{N} \leq x \leq \frac{i+1}{N}, \\ 0, & \text{otherwise,} \end{cases}$$

and for $j = 1, \dots, N$ the 0th order splines

$$(3.2) \quad \phi_j(x) = \begin{cases} 1, & \frac{j-1}{N} \leq x \leq \frac{j}{N}, \\ 0, & \text{otherwise.} \end{cases}$$

Thus N and M of the previous section are both N here.

We approximate the solution u of (1.1) by linear splines and the coefficient a by constant splines. It is straightforward to compute the $(N+1) \times (N+1)$ matrices H_j , $j=1, \dots, N$:

$$H_j = N \begin{pmatrix} & & & j-1 & j & & & \\ & & & \vdots & \vdots & & & \\ & & & & & & & \\ 0 & & & & & & & 0 \\ & \ddots & & & & & \ddots & \\ & & 0 & & 0 & & & \\ \vdots & & & 1 & -1 & & \cdot & \cdots j-1 \\ \vdots & & & -1 & 1 & & \cdot & \cdots j \\ & & 0 & & 0 & & & \\ & \ddots & & & & \ddots & & \\ 0 & & & & & & & 0 \end{pmatrix}.$$

Let $\tilde{\mu} = \tilde{\mu}(\tilde{a}^M) = \text{col}(\mu_0, \dots, \mu_N)$ with $\tilde{a}^M \in \mathcal{A}$. Then

$$H_j \tilde{\mu} = N \begin{pmatrix} 0 \\ \vdots \\ 0 \\ \mu_{j-1} - \mu_j & \cdots j-1 \\ -\mu_{j-1} + \mu_j & \cdots j \\ 0 \\ \vdots \\ 0 \end{pmatrix}.$$

Now set $\beta_j = \mu_j - \mu_{j-1}$ for $j=1, \dots, N$. To study the linear independence of the vectors $\{H_j \tilde{\mu}\}_{j=1}^N$, note that

$$(H_1 \tilde{\mu}, \dots, H_N \tilde{\mu}) = NB$$

where B is the $(N+1) \times N$ matrix

$$B = \begin{pmatrix} -\beta_1 & 0 & \cdots & 0 \\ \beta_1 & -\beta_2 & & \\ 0 & \beta_2 & & \\ \vdots & & \ddots & \\ 0 & & & -\beta_{N-1} \\ & & & \beta_{N-1} & -\beta_N \\ & & & & \beta_N \end{pmatrix}.$$

LEMMA 3.1. *The vectors $\{H_j \tilde{\mu}\}_{j=1}^N$ are linearly independent if and only if $\beta_i \neq 0$ for all $i=1, \dots, N$.*

Proof. It is easily shown that B is row equivalent [8] to

$$\tilde{B} = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & & \\ \vdots & & \ddots & \\ 0 & & & 0 & 1 \\ 0 & & & & 0 \end{pmatrix},$$

provided $\beta_i \neq 0$ for all i and $\{H_j \tilde{\mu}\}_{j=1}^N$ are linearly independent in this case. Conversely, if $\beta_i = 0$ for some i , then the column rank of B is less than N and linear dependence of $\{H_j \tilde{\mu}\}_{j=1}^N$ follows.

THEOREM 3.1. *In Case 1, \tilde{a}^M is identifiable, if and only if $\mu_i(\tilde{a}^M) \neq \mu_{i-1}(\tilde{a}^M)$ for all $i = 1, \dots, N$.*

PROPOSITION 3.1. *The coordinates $\{\tilde{a}_{j_k}^M\}_{k=1}^{\hat{M}}$ of \tilde{a}^M are identifiable if and only if $\mu_{j_k}(\tilde{a}^M) \neq \mu_{j_k-1}(\tilde{a}^M)$ for all $k = 1, \dots, \hat{M}$.*

The interpretation of this result is that the parameter \tilde{a}^M can only be identified at coordinates where the corresponding solution is nonstationary.

Case 2. Let N be even and let I be partitioned into subintervals of length $1/N$. The functions B_i , $i = 0, \dots, N$, are taken as in (3.1). Here, however, we define the functions ϕ_j for $j = 1, \dots, N/2$ by

$$\phi_j(x) = \begin{cases} 1, & \frac{2(j-1)}{N} \leq x \leq \frac{2j}{N}, \\ 0, & \text{otherwise.} \end{cases}$$

Thus M of § 2 is $N/2$ now. We find in this case that the $(N+1) \times (N+1)$ matrices H_j , for $j = 1, \dots, N/2$ are given by

$$H_j = N \begin{pmatrix} 0 & & & & 0 \\ & 1 & -1 & & 1 \\ & -1 & 2 & -1 & \\ & & 0 & -1 & 1 \\ 0 & & & & 0 \end{pmatrix},$$

where the first entry of the nontrivial submatrix is in the $2j-2, 2j-2$ position of H_j . With $\tilde{\mu} = \tilde{\mu}(\tilde{a}^M) = \text{col}(\mu_0, \dots, \mu_N)$ as before, we have

$$(H_1 \tilde{\mu}, \dots, H_{N/2} \tilde{\mu}) = NB$$

where the $(N+1) \times N/2$ -matrix B is given by

$$B = \begin{pmatrix} -\beta_1 & 0 & & 0 \\ \beta_1 - \beta_2 & 0 & & \\ \beta_2 & -\beta_3 & & \\ 0 & \beta_3 - \beta_4 & \cdots & \\ & \beta_4 & & \\ & 0 & & 0 \\ & & & \beta_{N-1} \\ & & & \beta_{N-1} - \beta_N \\ & 0 & & \beta_N \end{pmatrix},$$

and $\beta_i = \mu_i - \mu_{i-1}$ for $i = 1, \dots, N$.

The collection of vectors $\{H_j \tilde{\mu}\}_{j=1}^{N/2}$ is linearly independent if and only if $\text{rank}(B) = N/2$.

LEMMA 3.2. *In Case 2 the vectors $\{H_j \tilde{\mu}\}_{j=1}^{N/2}$ are linearly independent if and only if $\beta_{2i-1} \neq 0$ or $\beta_{2i} \neq 0$ for $i = 1, \dots, N/2$.*

Proof. The rank of the matrix B is equal to the rank of \tilde{B} where

$$\tilde{B} = \begin{pmatrix} \beta_1 & 0 & & 0 \\ \beta_2 & 0 & & 0 \\ 0 & \beta_3 & & 0 \\ 0 & \beta_4 & \cdots & 0 \\ & 0 & & \beta_{N-1} \\ 0 & 0 & & \beta_N \end{pmatrix}.$$

From this and the fact that the dimension of the column space of a matrix is equal to the rank of that matrix, the result follows.

THEOREM 3.2. *In Case 2, \tilde{a}^M is identifiable if and only if $\mu_{2i-1} \neq \mu_{2i-2}$ or $\mu_{2i} \neq \mu_{2i-1}$ for $i = 1, \dots, (N/2)$.*

Case 3. Let I be partitioned into N subintervals of length $1/N$. Again we take the functions B_i , $i = 0, \dots, N$ to be those defined in (3.1). Further we set $\phi_i = B_i$, $i = 0, \dots, N$. Thus M of § 2 is $N+1$ here and both ϕ_j and B_j are linear splines defined on the same mesh. In this case the structure of the $(N+1) \times (N+1)$ matrices $\{H_j\}_{j=0}^N$ is slightly more complicated than in Cases 1 and 2. These matrices are now given as follows:

$$H_0 = \frac{N}{2} \begin{pmatrix} 1 & -1 & & 0 \\ -1 & 1 & & 0 \\ & & & \\ & 0 & & 0 \end{pmatrix},$$

with the first entry in the $(0, 0)$ -element. For $j = 1, \dots, N-1$ we have

$$H_j = \frac{N}{2} \begin{pmatrix} 0 & & 0 & & 0 \\ & 1 & -1 & 0 & \\ 0 & -1 & 2 & -1 & 0 \\ & 0 & -1 & 1 & \\ 0 & & 0 & & 0 \end{pmatrix},$$

where the first entry of the nontrivial submatrix appears in the $(j-1, j-1)$ position of H_j . Finally

$$H_N = \frac{N}{2} \begin{pmatrix} 0 & & 0 \\ & 1 & -1 \\ 0 & -1 & 1 \end{pmatrix},$$

where the first nonzero entry occurs in the $N-1, N-1$ element.

To investigate the linear independence of $\{H_j \tilde{\mu}\}_{j=0}^N$, note that

$$(H_0 \tilde{\mu}, \dots, H_N \tilde{\mu}) = \frac{N}{2} B,$$

where the $(N+1) \times (N+1)$ matrix B is given by

$$B = \begin{pmatrix} -\beta_1 & -\beta_1 & & 0 & 0 \\ \beta_1 & \beta_1 - \beta_2 & & 0 & 0 \\ 0 & \beta_2 & \dots & 0 & 0 \\ \vdots & \vdots & & -\beta_{N-1} & 0 \\ 0 & 0 & & \beta_{N-1} - \beta_N & -\beta_N \\ & & & \beta_N & \beta_N \end{pmatrix},$$

and $\beta_i = \mu_i - \mu_{i-1}$ for $i = 1, \dots, N$. Performing row operations on B , we find that B is equivalent to

$$\tilde{B} = \begin{pmatrix} \beta_1 & \beta_1 & 0 & 0 & 0 \\ 0 & \beta_2 & \beta_2 & 0 & 0 \\ \vdots & 0 & \beta_3 & \vdots & \vdots \\ 0 & 0 & 0 & \beta_{N-1} & 0 \\ & & & \beta_N & \beta_N \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}.$$

We then see that B has rank less or equal to N and therefore we have

THEOREM 3.3. *In Case 3, $\tilde{a}^M \in \mathcal{A}$ is not identifiable.*

If one decreases the number of ϕ_j 's in Case 3, then it is reasonable to expect that sufficient and necessary conditions for the identifiability of \tilde{a}^M in the spirit of Cases 1 and 2 can be obtained. We verify this next for a particular choice of N and M . Moreover, in § 5 we present a numerical experiment which tends to support this contention.

Case 4. Let I be partitioned into $2N$ subintervals of length $1/2N$. We choose the functions B_i for $i = 0, \dots, 2N$ as

$$B_i(x) = \begin{cases} 2Nx - i + 1, & \frac{i-1}{2N} \leq x \leq \frac{i}{2N}, \\ -2Nx + i + 1, & \frac{i}{2N} \leq x \leq \frac{i+1}{2N}, \\ 0, & \text{otherwise.} \end{cases}$$

For the functions ϕ_j we take

$$\phi_j(x) = \begin{cases} Nx - j + 1, & \frac{j-1}{N} \leq x \leq \frac{j}{N}, \\ -Nx + j + 1, & \frac{j}{N} \leq x \leq \frac{j+1}{N}, \\ 0, & \text{otherwise,} \end{cases}$$

for $j=0, \dots, N$; thus M of § 2 is $N+1$. The $(2N+1) \times (2N+1)$ -matrices $\{H_j\}_{j=0}^N$ are given as follows:

$$H_0 = \frac{N}{2} \left(\begin{array}{ccc|ccc} 3 & -3 & 0 & & & \\ -3 & 4 & -1 & & & 0 \\ 0 & -1 & 1 & & & \\ \hline & & & & & \\ & 0 & & & & 0 \end{array} \right),$$

with 3 in the $(0,0)$ element,

$$H_j = \frac{N}{2} \left(\begin{array}{ccc|ccc|ccc} & & & & 0 & & & & \\ \hline & & & & & & & & \\ & 1 & -1 & 0 & 0 & 0 & & & \\ & -1 & 4 & -3 & 0 & 0 & & & \\ 0 & 0 & -3 & 6 & -3 & 0 & 0 & & \\ & 0 & 0 & -3 & 4 & -1 & & & \\ & 0 & 0 & 0 & -1 & 1 & & & \\ \hline & & & & 0 & & & & \end{array} \right),$$

where the first entry of the nontrivial submatrix occurs in the $2(i-1), 2(j-1)$ -element and

$$H_N = \frac{N}{2} \left(\begin{array}{ccc|ccc} & 0 & & & 0 & & \\ \hline & & & & & & \\ & & & & 1 & -1 & 0 \\ & 0 & & & -1 & 4 & -3 \\ & & & & 0 & -3 & 3 \end{array} \right).$$

Let $\tilde{\mu} = \tilde{\mu}(\tilde{a}^M) = \text{col}(\mu_0, \dots, \mu_{2N})$. To investigate the linear independence of $\{H_j \tilde{\mu}\}_{j=0}^N$ we put $\beta_i = \mu_i - \mu_{i-1}$ for $i=1, \dots, 2N$ and observe that

$$(H_0 \tilde{\mu}, \dots, H_N \tilde{\mu}) = \frac{N}{2} B$$

where the $(2N+1) \times (N+1)$ matrix B is given by

$$B = \left(\begin{array}{ccc|ccc|ccc|ccc} -3\beta_1 & -\beta_1 & 0 & & & & & & & & \\ 3\beta_1 - \beta_2 & \beta_1 - 3\beta_2 & 0 & & & & & & & & \\ \beta_2 & 3\beta_2 - 3\beta_3 & -\beta_3 & & & & & & & & \\ 0 & 3\beta_3 - \beta_4 & \beta_3 - 3\beta_4 & & & & & & & & \\ & \beta_4 & 3\beta_4 - 3\beta_5 & & & & & & & & \\ & 0 & 3\beta_5 - \beta_6 & \dots & & & & & & & \\ & & \beta_6 & & & & & & & & \\ & & 0 & & & & & & & & \\ & & & & & & & & & & \\ & & & & & & 0 & & & & \\ & & & & & & -\beta_{2N-3} & & & & \\ & & & & & & \beta_{2N-3} - 3\beta_{2N-2} & & & 0 & \\ & & & & & & 3\beta_{2N-2} - 3\beta_{2N-1} & & & -\beta_{2N-1} & \\ & & & & & & 3\beta_{2N-1} - \beta_{2N} & & & \beta_{2N-1} - 3\beta_{2N} & \\ & & & & & & \beta_{2N} & & & 3\beta_{2N} & \end{array} \right)$$

Performing row operations on B , we obtain a matrix \tilde{B} which has the same rank as B and is given by

$$\tilde{B} = \begin{pmatrix} 3\beta_1 & \beta_1 & 0 & & 0 & 0 \\ \beta_2 & 3\beta_2 & 0 & & & \\ 0 & 3\beta_3 & \beta_3 & & & \\ & \beta_4 & 3\beta_4 & & & \\ & 0 & 3\beta_5 & \dots & & \\ & & \beta_6 & & 0 & \\ & & 0 & & \beta_{2N-3} & 0 \\ & & & & 3\beta_{2N-2} & 0 \\ & & & & 3\beta_{2N-1} & \beta_{2N-1} \\ & 0 & 0 & 0 & \beta_{2N} & 3\beta_{2N} \end{pmatrix},$$

which is of dimension $2N \times (N+1)$. The vectors $H_j \tilde{\mu}$ are linearly independent if $\text{rank}(\tilde{B}) = N+1$.

THEOREM 3.4. *In Case 4, if $\mu_{2i-1} \neq \mu_{2i-2}$ or $\mu_{2i} \neq \mu_{2i-1}$ for all $i = 1, \dots, N$ and $(\mu_{2i-1} - \mu_{2i-2})(\mu_{2i} - \mu_{2i-1}) \neq 0$ for some $i = 1, \dots, N$, where $\tilde{\mu} = \tilde{\mu}(\tilde{a}^M)$, then $\tilde{a}^M \in \mathcal{A}$ is identifiable.*

Proof. Let $\tilde{\alpha} = \text{col}(\tilde{\alpha}_0, \dots, \tilde{\alpha}_N) \in \mathbb{R}^{N+1}$ and $\tilde{B}\tilde{\alpha} = 0$. Choose i_1 such that $\beta_{2i_1-1} \cdot \beta_{2i_1} \neq 0$. Then $\alpha_{i_1-1} = \alpha_{i_1} = 0$. Further $\alpha_i = 0$ for all other i , since $\beta_{2i-1} \neq 0$ or $\beta_{2i} \neq 0$. This implies linear independence of the columns of \tilde{B} and thus of B . Theorem 3.1 then implies the result.

THEOREM 3.5. *In Case 4, if $\mu_1 - \mu_0 = \mu_2 - \mu_1 = 0$ or $\mu_{N-1} - \mu_{N-2} = \mu_N - \mu_{N-1} = 0$ or $\mu_{2i-1} - \mu_{2i-2} = \mu_{2i} - \mu_{2i-1} = \mu_{2i+1} - \mu_{2i} = \mu_{2i+2} - \mu_{2i+1} = 0$ for some $i = 2, \dots, N-2$, where $\tilde{\mu} = \tilde{\mu}(\tilde{a})$ then $\tilde{a}^M \in \mathcal{A}$ is not identifiable.*

Proof. Under the assumptions of the theorem the column-rank of \tilde{B} is not maximal and this implies the result.

Remark 3.1. Four consecutive zeros in the β_i 's do not necessarily imply nonidentifiability, provided the zeros start with an even index and are not at the "beginning" or "end" of the sequence $\{\beta_i\}$; in particular $\beta_{2i} = \beta_{2i+1} = \beta_{2i+2} = \beta_{2i+3} = 0$ with $2 \leq i < N-2$ does not imply nonidentifiability. For example let $N = 5$, $\beta_1 = \beta_2 = \beta_3 = \beta_8 = \beta_9 = \beta_{10} = 1$ and $\beta_4 = \beta_5 = \beta_6 = \beta_7 = 0$. Then $\{H_j \tilde{\mu}\}$, $j = 0, \dots, N$, with $\beta_i = \mu_i - \mu_{i-1}$, are linearly independent. Choosing μ_0 and \tilde{a}^M , we can thus calculate μ_i , $i = 1, \dots, 2N$ and \tilde{f} such that \tilde{a}^M is identifiable and $\beta_{2i} = \dots = \beta_{2i+3} = 0$.

Remark 3.2. The conditions of Theorems 3.1–3.4 are conditions on the variation of adjacent $\mu_i(\tilde{a}^M)$ -values. If this variation is sufficient, the identifiability of \tilde{a}^M is guaranteed. The results indicate that the larger the difference between the dimension of the state space approximation and the dimension of the parameter space approximation is, the more likely it is that identifiability of the approximated coefficient holds. In [9] identifiability of a in (1.1) is studied under various conditions on the sign of u_x and u_{xx} . The most general condition implying identifiability of a is $\inf_I \max(|u_x|, u_{xx}) > 0$. Clearly one can construct examples where this condition is not met but identifiability under approximation, e.g., according to one of the Cases 1–4, of \tilde{a}^M holds.

4. Two stability concepts. In this section we discuss the application of two concepts of stability to the finite dimensional output least squares problem

$$(P_M^N) \quad \text{minimize } |u^N(a^M) - z|_{L^2}^2 \text{ over } C,$$

where C is a convex and closed subset of

$$Q_{\text{ad}}^M = \left\{ a^M = \sum_{j=1}^M a_j^M \phi_j; a_j^M \in \mathbb{R}, a^M(x) \geq \alpha > 0, |a^M|_{H^1} \leq \gamma \right\},$$

and $u^N(a^M) = \sum_{i=0}^N \mu_i B_i$, with $\text{col}(\mu_0, \dots, \mu_N) = \tilde{\mu}$ satisfying (2.2). The existence of a minimum a_*^M of (P_M^N) can easily be argued. The reason for introducing the set $C \subset Q_{\text{ad}}^M$ here is that for the first stability concept, uniqueness of solutions of (P_M^N) is required and this cannot be guaranteed over all of Q_{ad}^M . Here we consider H^1 -smooth approximating coefficients, some remarks on L^∞ approximations are given further below. We investigate the continuous dependence of a_*^M on z and also on Q_{ad}^M when dealing with the second stability concept. For $C \subset Q_{\text{ad}}^M$ let $\mathcal{V}(C) = \{u^N(a^M); a^M \in C\}$ denote the attainable set.

DEFINITION 4.1 [2]. The parameter a^M in (2.2) is called output least squares identifiable (OLSI) by (P_M^N) over $C \subset Q_{\text{ad}}^M$, if there exists a neighborhood $\tilde{\mathcal{V}}$ of $\mathcal{V}(C)$ such that for every $z \in \tilde{\mathcal{V}}$ the problem (P_M^N) has a unique solution a_*^M depending continuously on z .

Let $\mathcal{A}_{\text{inj}} \subset Q_{\text{ad}}^M$ be such that $\{H_j \tilde{\mu}(\tilde{a}^M)\}$ is linearly independent for every $a^M = \sum_{j=1}^M (\tilde{a}^M)_j \phi_j \in \mathcal{A}_{\text{inj}}$. As examples for such sets we can take neighborhoods in Q_{ad} of points of identifiability in the sense of § 2.

THEOREM 4.1. *Let \mathcal{A}_{inj} be as just described. Then a^M in (2.2) is OLSI by (P_M^N) over every closed convex subset C of \mathcal{A}_{inj} , provided that $\text{diam } C$ is sufficiently small and z is sufficiently close to the $\mathcal{V}(C)$. Moreover, the unique solution Q_*^M of (P_M^N) depends Lipschitz-continuously on z .*

Proof. By [2, Thm. 4] it suffices to show that $a^M \rightarrow u^N(a^M)$ is twice continuously Fréchet differentiable with $a^M \rightarrow u_{a^M}^N(a^M)$ injective on C . This is equivalent to the existence of continuous first and second order derivatives of $\tilde{a}^M \rightarrow \tilde{\mu}(\tilde{a}^M)$ with \tilde{a}^M such that $\sum_{j=1}^M (\tilde{a}^M)_j \phi_j = a^M \in \mathcal{A}_{\text{inj}}$ and injectivity of $\tilde{\mu}_{\tilde{a}^M}(\tilde{a}^M)$ for every $a^M \in C$. Let $\tilde{\mu}_{\tilde{a}^M}(\tilde{a}^M; \tilde{h}) = \tilde{\eta}$ and $\tilde{\mu}_{\tilde{a}^M \tilde{a}^M}(\tilde{a}; \tilde{h}, \tilde{k}) = \tilde{\xi}$ be the first, resp. second, derivative in directions \tilde{h} and (\tilde{h}, \tilde{k}) . Then

$$(4.1) \quad \tilde{\eta} = -L^{-1}(\Sigma h_j H_j \tilde{\mu}(\tilde{a}^M))$$

and

$$(4.2) \quad \tilde{\xi} = -L^{-1}(\Sigma k_j H_j \tilde{\eta} + \Sigma h_j H_j \tilde{\mu}_{\tilde{a}^M}(\tilde{a}^M; \tilde{k})),$$

where $L = \Sigma \tilde{a}_j^M H_j + K$ and $\tilde{h} = \text{col}(h_1, \dots, h_M)$, $\tilde{k} = \text{col}(k_1, \dots, k_M)$, and the continuity assumptions follow. The injectivity of $\tilde{\mu}_{\tilde{a}^M}(\tilde{a}^M)$ is guaranteed by linear independence of $\{H_j \tilde{\mu}(\tilde{a}^M)\}$; this ends the proof.

To describe the second notion of stability, we consider the case $C = Q_{\text{ad}}^M$:

$$(P_M^N)_w \quad \text{minimize } |u^N(a^M) - z|^2 \text{ over } Q_{\text{ad}}^M.$$

We study continuous dependence of local solutions of $(P_M^N)_w$ on $w = (z, \alpha, \gamma) \in W$, where $W = H^0 \times \mathbb{R} \times \mathbb{R}$. Here W is endowed with the Hilbert-space product norm. We always assume $0 < \alpha < \gamma$, so that Q_{ad}^M is not empty and solutions of (2.2) and $(P_M^N)_w$ exist.

DEFINITION 4.2 [2]. The parameter a^M is called output least squares (OLS)-stable in Q_{ad}^M at the local solution a_0^M of $(P_M^N)_{w^0}$, $w^0 \in W$, if there exists a neighborhood $V(w^0)$ of w^0 in W , a neighborhood $V(a_0^M)$ of a_0^M in H^1 and a constant κ , such that for all $w = (z, \alpha, \gamma) \in V(w^0)$ there exists a local solution a_w^M of $(P_M^N)_w$ with $a_w^M \in V(a_0^M)$ and for all local solutions $a_w^M \in V(a_0^M)$ of $(P_M^N)_w$ we have

$$|a_w^M - a_{w^0}^M|_{H^1} \leq \kappa |w - w^0|_W^{1/2}.$$

Remark 4.1. In comparing OLSI to OLS-stability we observe the following differences: OLSI requires uniqueness of the solutions of the minimization problem, whereas for OLS-stability, uniqueness is not required, with continuity being checked at each local solution. If OLSI holds, then the solutions depend on the observations in a Lipschitz continuous way, whereas OLS-stability only guarantees Hölder continuous dependence. Further, OLSI requires continuous dependence of the solutions on the observation only, whereas OLS-stability involves continuous dependence on the observations as well as on the admissible set Q_{ad} .

Output least squares stability is proved by techniques that guarantee stability of solutions of abstract optimization problems with respect to perturbations in the problem data, see [3] and the references given there. Let $A^M = \text{span}\{\phi_j\}_{j=1}^M$ and let $F(a^M)$ be the Lagrange functional associated with (P_M^N) :

$$F(a^M) = |u^N(a^M) - z|^2 - \lambda^* g(a^M),$$

where $\lambda^* \in C^* \times \mathbb{R}$ and

$g: A^M \times W \rightarrow C^* \times \mathbb{R}$ is given by

$$g(a^M, w) = (\alpha - a^M, |a^M|_{H^1}^2 - \gamma^2), \quad w = (z, \alpha, \gamma).$$

Note that $a^M \in Q_{\text{ad}}^M(w)$ if and only if $g(a^M, w) \in \hat{K} = C_- \times \mathbb{R}_-$, with C_- and \mathbb{R}_- the natural negative cones in $C(I)$ and \mathbb{R} . We shall frequently drop the index w and write $g(a^M)$ and Q_{ad}^M for $g(a^M, w)$ and $Q_{\text{ad}}^M(w)$.

THEOREM 4.2. *Let $A^M = \text{span}\{\phi_j\}_{j=1}^M$ be such that it contains the constant functions, let $(z^0, \alpha^0, \gamma^0) = w^0 \in W$ with $0 < \alpha^0 < \gamma^0$ and let $a_0^M = \sum_{j=1}^M (\tilde{a}_0^M)_j \phi_j$ be a local solution of $(P_M^N)_{w^0}$. If $\{H_j \tilde{\mu}(\tilde{a}_0^M)\}_{j=1}^M$ are linearly independent vectors in \mathbb{R}^{N+1} and $|u^N(a_0^M) - z|$ is sufficiently small, then a^M is OLS-stable in $Q_{\text{ad}}^M(w^0)$ at the local solution a_0^M of $(P_M^N)_{w^0}$.*

For the proof of this theorem the following lemma on the regularity of the constraint set Q_{ad}^M will be required; its proof is quite similar to that of Lemma 4.2 in [3] but will be included for the sake of completeness.

LEMMA 4.1. *Let A^M contain the constant functions. Then every $a^M \in Q_{\text{ad}}^M$ is a regular point, i.e., $0 \in \text{int}\{g(a^M) + \mathcal{R}(g_{a^M}(a^M)) - \hat{K}\} \subset C \times \mathbb{R}$, where \mathcal{R} denotes the range of the mapping $g_{a^M}(a^M)$.*

Proof of Lemma 4.1. We need to show that

$$(4.3) \quad \begin{aligned} 0 &\in \text{int}\{g(a^M) + g_{a^M}(a^M)A^M - C_- \times \mathbb{R}_-\} \\ &= \text{int}\{\alpha - a^M - h^M + C_+, |a^M|_{H^1}^2 - \gamma^2 + 2\langle a^M, h^M \rangle_{H^1} + \mathbb{R}_+ : h^M \in A^M\}, \end{aligned}$$

where we used that $g_{a^M}(a^M)h^M = (-h^M, 2\langle a^M, h^M \rangle_{H^1})$. Let $(\phi, r) \in C \times \mathbb{R}$ with $|(\phi, r)|_{C \times \mathbb{R}} < \delta$ and $\delta > 0$ to be chosen sufficiently small. Note that $\phi - \min \phi \in C_+$ and $\min \phi \in A^M$. In view of the first component in (4.3) we decompose ϕ as

$$\phi = \alpha - a^M - (\alpha - a^M - \min \phi) + \phi - \min \phi$$

and therefore $\phi \in \alpha - a^M - A^M + C_+$. As for the second component in (4.3) observe that

$$\begin{aligned} |a^M|_{H^1}^2 - \gamma^2 + 2\langle a^M, \alpha - a^M - \min \phi \rangle_{H^1} &= -|a^M|_{H^1}^2 + 2\langle a^M, \alpha - \min \phi \rangle_{H^1} - \gamma^2 \\ &\leq \alpha^2 - \gamma^2 + 2\delta |a^M|_{H^1}. \end{aligned}$$

Thus, for δ sufficiently small one can choose $\tilde{r} \in \mathbb{R}_+$ such that

$$r = |a^M|_{H^1}^2 - \gamma^2 + 2\langle a^M, \alpha - a^M - \min \phi \rangle_{H^1} + \tilde{r}$$

and, since (ϕ, r) was arbitrary, a^M is shown to be a regular point.

Proof of Theorem 4.2. We apply results on the stability of abstract optimization problems as summarized in [3, § 3]. Due to the fact that $a^M \rightarrow |u^N(a^M) - z^0|^2$ and $a^M \rightarrow g(a^M, w^0)$ are twice continuously differentiable at a_0^M and since the point $a_0^M \in Q_{ad}^M$ is a regular point, it suffices to establish a lower bound on the second derivative of F at \tilde{a}_0^M . Let $\eta = u_{a^M}^N(a_0^M; h^M)$ and $\xi = u_{a^M, a^M}^N(a_0^M; h^M, h^M)$ for $h^M \in A^M$. Then

$$\begin{aligned} F_{a^M, a^M}(a_0^M; h^M, h^M) &= \langle u^N(a_0^M) - z, \xi \rangle_{H^0} + |\eta|_{H^0}^2 - 2\lambda |h^M|_{H^1}^2 \\ &\geq -|u^N(a_0^M) - z|_{H^0} |\xi|_{H^0} + |\eta|_{H^0}^2 - 2\lambda |h^M|_{H^1}^2, \end{aligned}$$

where $\lambda \leq 0$ is the Lagrange multiplier associated with the norm constraint. In view of (4.1), (4.2), the finite dimensionality of A^M , and the linear independence of $\{H_j \tilde{\mu}(\tilde{a}_0^M)\}_{j=1}^M$ it follows that there exist constants c_1 and c_2 such that

$$F_{a^M, a^M}(a_0^M; h^M, h^M) \geq -c_1 |u^N(a_0^M) - z|_{H^0} |h^M|_{H^1}^2 + c_2 |h^M|_{H^1}^2,$$

so that for $|u^N(a_0^M) - z|$ sufficiently small there exists a constant c_3 with

$$F_{a^M, a^M}(a_0^M; h^M, h^M) \geq c_3 |h^M|_{H^1}^2,$$

from which the result follows [3, Thms. 3.2, 3.3].

Remark 4.2. If $Q_{ad}^M \subset L^\infty$ only and $|a^M|_{H^1} \leq \gamma$ is replaced by $|a^M|_{L^\infty} \leq \gamma$ in the definition of Q_{ad}^M , then again one can show existence of solutions of (P_N^M) and Theorem 4.1 holds with obvious modifications. The results leading to Theorem 4.2 need yet to be generalized to handle the nondifferentiable L^∞ -norm constraint.

5. Numerical results. In this section we present some results of a numerical experiment to estimate the coefficient a in (1.1) given observations z of u . To solve (1.2) with $\mathcal{C} = I$, we consider (2.2) which defines a mapping $\tilde{a}^M \rightarrow \tilde{\mu}(\tilde{a}^M)$ for $\tilde{a}^M \in \mathcal{A}$, and the finite dimensional minimization problems

$$(5.1) \quad \text{minimize} \int_0^1 \left(\sum_{i=0}^N \mu_i(\tilde{a}^M) B_i - z \right)^2 dx.$$

For our experiments we imposed no constraints on $\tilde{a}^M \in \mathbb{R}^M$, although $\tilde{\mu}(\tilde{a}^M)$ is not well defined for some \tilde{a}^M . The basis functions ϕ_j and B_i were chosen as linear spline functions with equidistant grid on $(0, 1)$. As data z we took the values of a solution of (1.1) by choosing the coefficient a and the observation $z(x) = u(x) = x^2(1-x)^2$, and calculating $f = u - (au_x)_x$ from it. Using this f , we then compute $\tilde{\mu}$ from (2.2) as we solve (5.1). For the minimization the Newton-Raphson algorithm was used.

In our calculations (see Figs. 1–10) $N = 10$ represents the number of subintervals used in the linear spline approximation for the solution of (1.1). Thus the dimension of the approximation space for the solution is 11. Further NBI is the number of subintervals of I that determine the linear spline approximation of a ; the dimension of the approximation space for a is NBI + 1. A necessary condition for identifiability of \tilde{a}^M is thus NBI + 1 \leq 11, see Corollary 2.1. We show calculations for NBI = 4, 5, 6, 8–11, for the choice of $a(x) = 1 + x$. In the first five cases good results are obtained. Note that NBI = 5 and $N = 10$ is a special case of Theorem 3.4. In the case NBI = 10, \tilde{a}^M is not identifiable by Theorem 3.3. Numerically this is reflected by the appearance of oscillations as NBI approaches 10 from below, see the graphs for NBI = 9, 10, 11. The start-up value for the minimization routine was chosen as $\tilde{a}_0^M \equiv 2$. We point out that a different scaling of the axes in the various paths was utilized. We also show the graphs for $u^N(a^M)$, when $N = 10$ and NBI = 9, 10, 11. The graphs for the approximating solutions for NBI = 4, 5, 6, 8 are indistinguishable from NBI = 9.

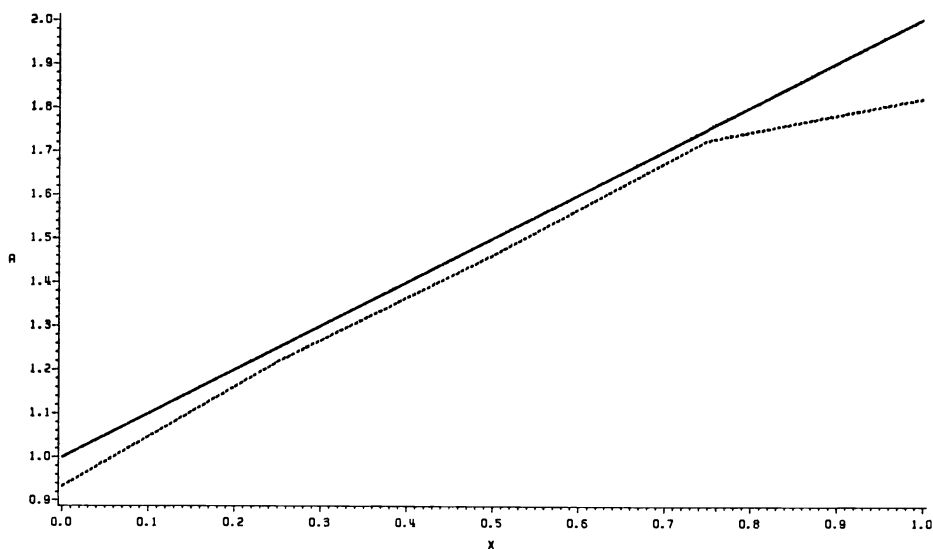


FIG. 1

..... A ESTIMATE
 ——— A TRUE VALUE

1D FLOW - NEUMAN B.C.S.

$A = 1 + X$, $U = X * X * (1.0 - X) ** 2$

$N = 10$, $NBI = 4$

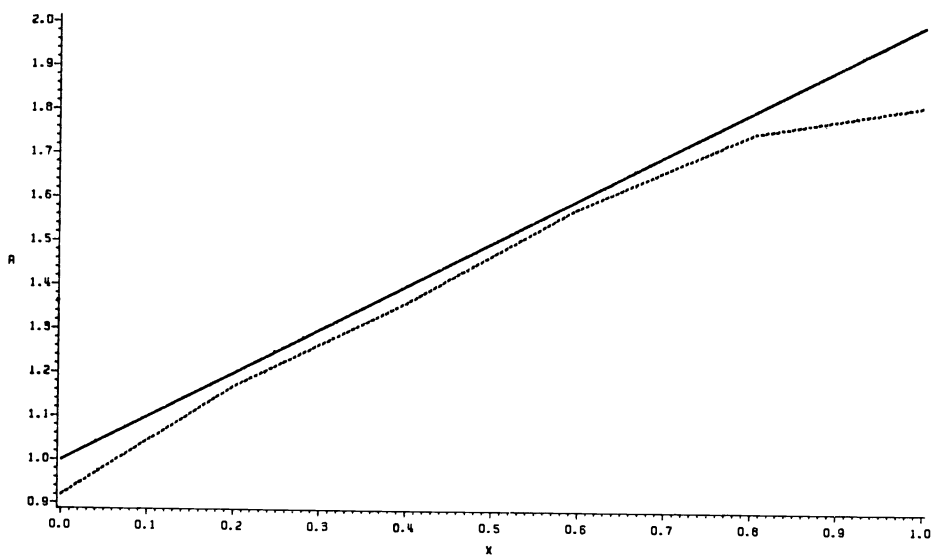


FIG. 2

..... A ESTIMATE
 ——— A TRUE VALUE

1D FLOW - NEUMAN B.C.S.

$A = 1 + X$, $U = X * X * (1.0 - X) ** 2$

$N = 10$, $NBI = 5$

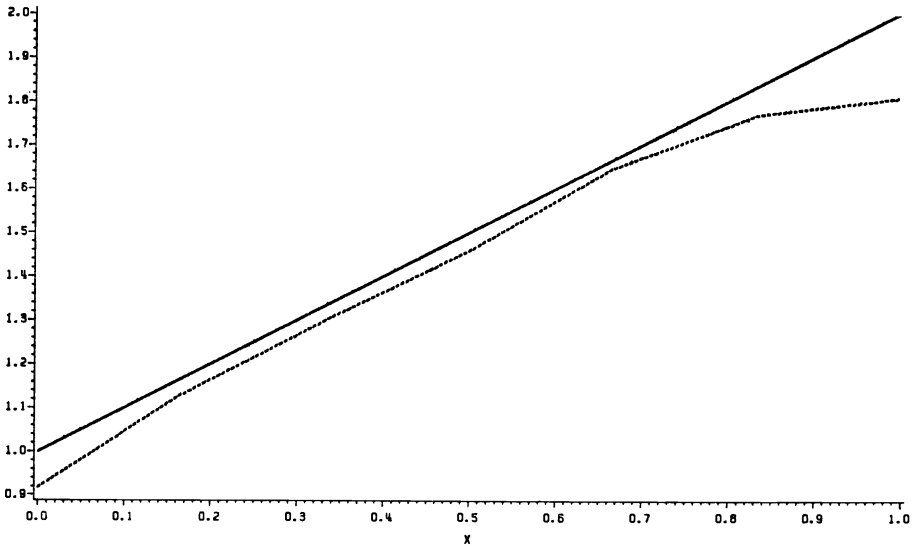


FIG. 3

..... A ESTIMATE
 _____ A TRUE VALUE

1D FLOW - NEUMAN B.C.S.

$A = 1 + X$, $U = X * X * (1.0 - X) ** 2$

$N = 10$, $NBI = 6$

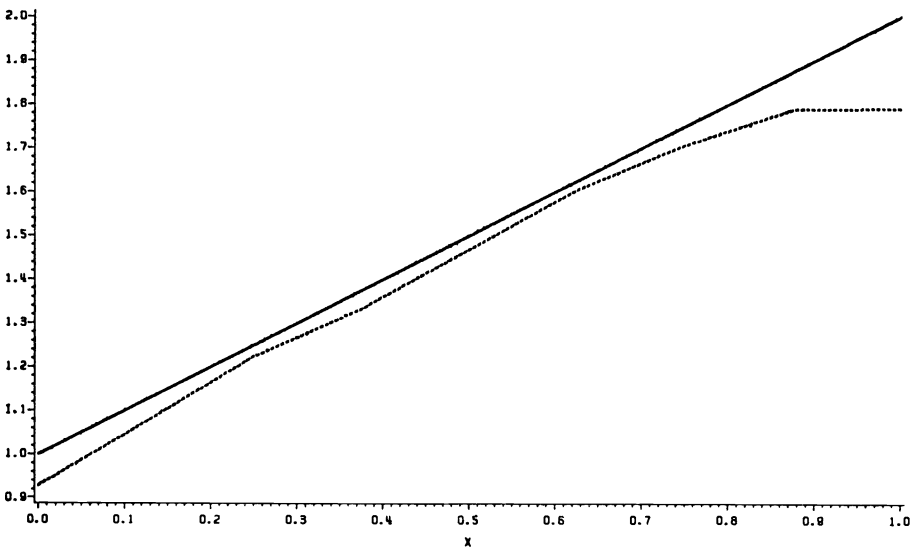


FIG. 4

..... A ESTIMATE
 _____ A TRUE VALUE

1D FLOW - NEUMAN B.C.S.

$A = 1 + X$, $U = X * X * (1.0 - X) ** 2$

$N = 10$, $NBI = 8$

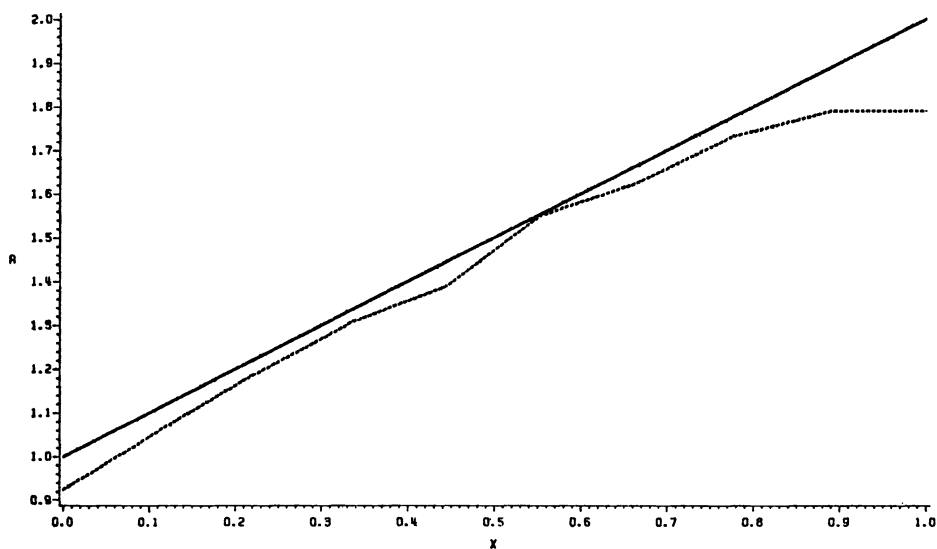


FIG. 5

..... A ESTIMATE

———— A TRUE VALUE

1D FLOW - NEUMAN B.C.S.

$A = 1 + X$, $U = X * X * (1.0 - X) ** 2$

$N = 10$, $NBI = 9$

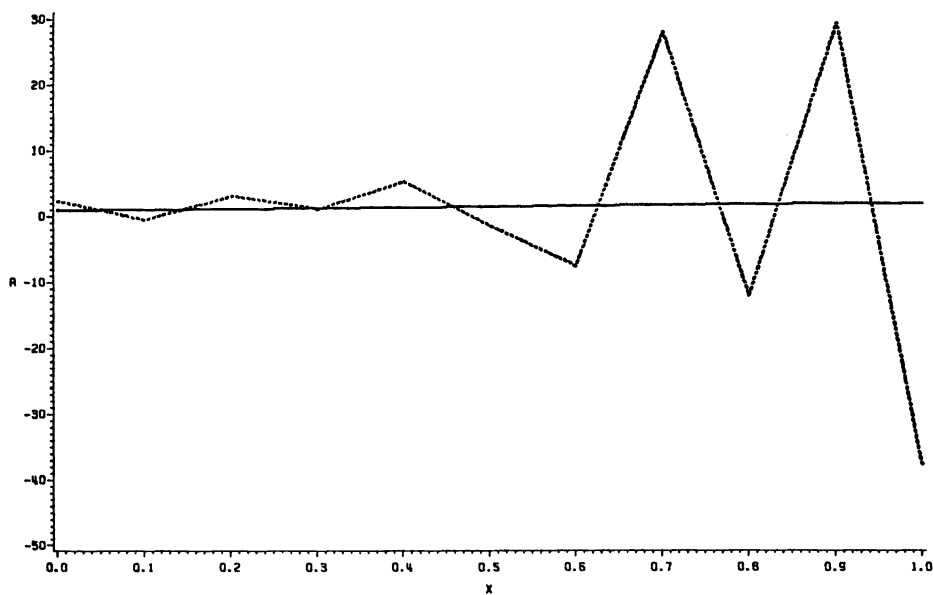


FIG. 6

..... A ESTIMATE

———— A TRUE VALUE

1D FLOW - NEUMAN B.C.S.

$A = 1 + X$, $U = X * X * (1.0 - X) ** 2$

$N = 10$, $NBI = 10$

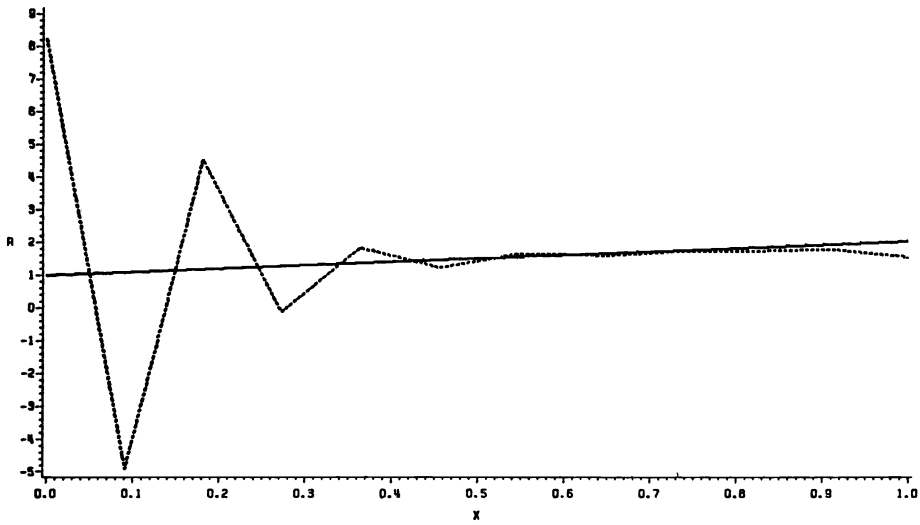


FIG. 7

..... A ESTIMATE

———— A TRUE VALUE

1D FLOW - NEUMAN B.C.S.

$A = 1 + X$, $U = X * X * (1.0 - X) ** 2$

$N = 10$, $NBI = 11$

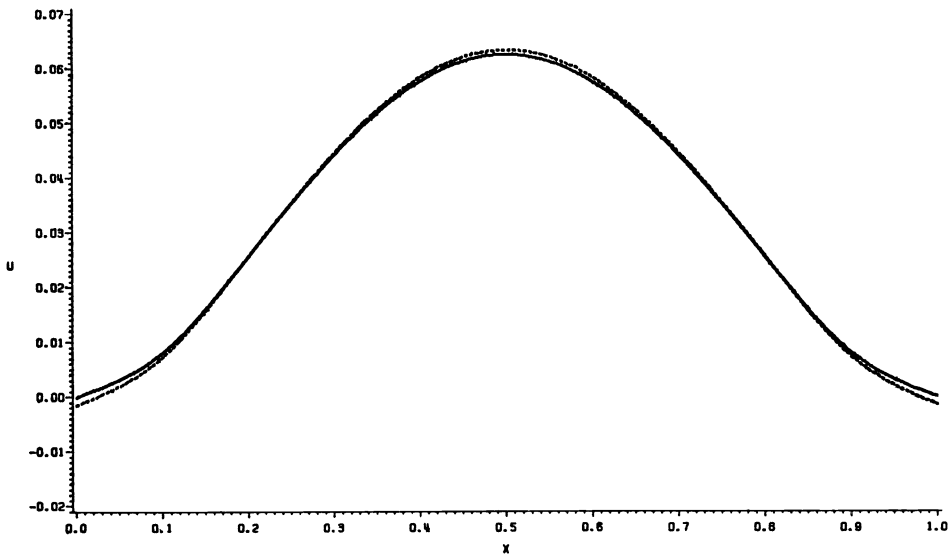


FIG. 8

..... U ESTIMATE

———— U TRUE VALUE

1D FLOW - NEUMAN B.C.S.

$A = 1 + X$, $U = X * X * (1.0 - X) ** 2$

$N = 10$, $NBI = 9$

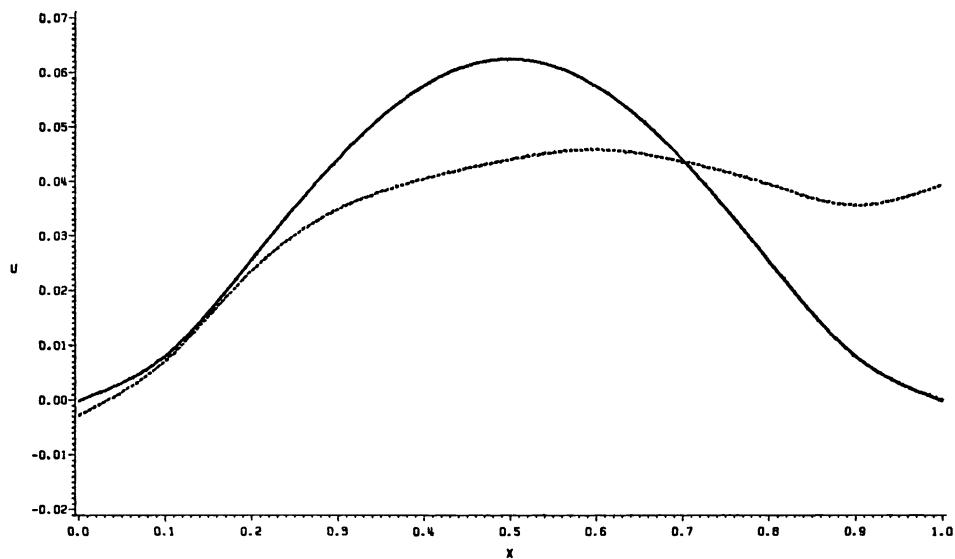


FIG. 9

..... U ESTIMATE

———— U TRUE VALUE

1D FLOW - NEUMAN B.C.S.

$A = 1 + X$, $U = X * X * (1.0 - X) ** 2$

$N = 10$, $NBI = 10$

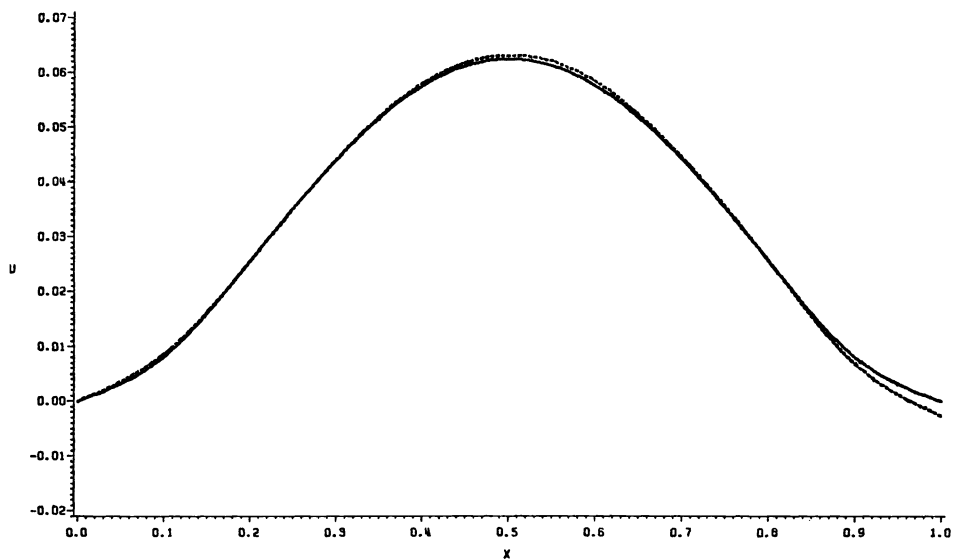


FIG. 10

..... U ESTIMATE

———— U TRUE VALUE

1D FLOW - NEUMAN B.C.S.

$A = 1 + X$, $U = X * X * (1.0 - X) ** 2$

$N = 10$, $NBI = 11$

REFERENCES

- [1] H. T. BANKS, P. L. DANIEL AND E. S. ARMSTRONG, *Parameter estimation for a static model of Maypole-hoop column antenna surfaces*, Proc. of the IEEE Internat. Large Scale Systems Symposium, Va, Beach, VA, 1982, pp. 253-255.
- [2] G. CHAVENT, *Local stability of the output least square parameter estimation technique*, INRIA Research Matematica Aplicada e Computacional, 2 (1983), pp. 3-22.
- [3] F. COLONIUS AND K. KUNISCH, *Stability for parameter estimation in two point boundary value problems*, Journal für die Reine und Angewandte Mathematik, to appear.
- [4] K. KUNISCH AND L. W. WHITE, *Parameter identifiability under approximation*, Quart. Appl. Math., 10 (1986), pp. 121-146.
- [5] ———, *Parameter estimation for elliptic equations in multi-dimensional domains with point and flux observations*, Nonlinear Analysis, Theory, Methods and Applications, 10 (1986), pp. 121-146.
- [6] ———, *Regularity properties in parameter estimation of diffusion coefficients in one dimensional elliptic boundary value problems*, Applicable Analysis, 21 (1986), pp. 71-87.
- [7] B. J. LEVITAN, *Distribution of Zeros of Entire Functions*, Amer. Math. Soc. Translations 5, Providence, RI, 1964.
- [8] B. NOBLE, *Applied Linear Algebra*, Prentice-Hall, Englewood Cliffs, NJ, 1969.
- [9] G. R. RICHTER, *An inverse problem for the steady state diffusion equation*, SIAM J. Appl. Math., 41 (1981), pp. 210-221.
- [10] M. SCHULTZ, *Spline Analysis*, Prentice-Hall, Englewood Cliffs, NJ, 1973.
- [11] L. W. WHITE, *Identification of a Friction Coefficient in a First Order Linear Hyperbolic Equation*, Proc. of the 1983 IEEE Control and Decision Conference, pp. 56-59.

NEARLY OPTIMAL STATE FEEDBACK CONTROLS FOR STOCHASTIC SYSTEMS WITH WIDEBAND NOISE DISTURBANCES*

HAROLD J. KUSHNER† AND W. Runggaldier‡

Abstract. Much of optimal stochastic control theory is concerned with diffusion models. Such models are often only idealizations (or limits in an appropriate sense) of the actual physical process, which might be driven by a wide band-width (not white) process or be a discrete parameter system with correlated driving noises. Optimal or nearly optimal controls, derived for the diffusion models, would not normally be useful, or even of much interest, if they were not also “nearly optimal” for the physical system that the diffusion approximates. It turns out that, under quite broad conditions, the “nearly optimal” controls for the diffusions do have this desired robustness property and are “nearly optimal” for the physical (say wideband noise driven) process, even when compared to controls that can depend on all the (past) driving noise. We treat the problem over a finite time interval, as well as the average cost per unit time problem. Extensions to discrete parameter systems, and to systems stopped on first exit from a bounded domain, are also discussed. Weak convergence methods provide the appropriate analytical tools.

Key words. optimal stochastic control, wideband noise disturbance, approximately optimal control, weak convergence for diffusions

AMS(MOS) subject classifications. 93E20, 93E25, 60F05, 60J60

1. Introduction. The paper is concerned with “approximately optimal” controls for a wide variety of systems driven by wide band-width noise, and their discrete parameter counterparts. Consider a system of the type

$$(1.1) \quad \dot{x}^\varepsilon = F_\varepsilon(x^\varepsilon, \xi^\varepsilon, u^\varepsilon), \quad x \in R^r, \text{ Euclidean } r\text{-space},$$

where $\xi^\varepsilon(\cdot)$ is a wide band-width noise process (the band-width $\rightarrow \infty$ as $\varepsilon \rightarrow 0$), and the cost is

$$(1.2) \quad R^\varepsilon(u^\varepsilon) = E \int_0^{T_1} k(x^\varepsilon(s), u^\varepsilon(s)) ds$$

for some $T_1 < \infty$. When we wish to emphasize the control, we write the solution to (1.1) as $x^\varepsilon(u^\varepsilon, \cdot)$.

For the moment (and loosely speaking) suppose that (1.1) is “close” to a controlled diffusion process, modeled by (1.3), in the sense that if $u^\varepsilon(\cdot)$ is a sequence of “nice” controls for (1.1), then there is a control $u(\cdot)$, and a corresponding controlled diffusion $x(u, \cdot)$ defined by (1.3), such that as $\varepsilon \rightarrow 0$, $x^\varepsilon(u^\varepsilon, \cdot) \Rightarrow x(u, \cdot)$, where \Rightarrow denotes weak convergence (see the next section). Let $\bar{u}(\cdot)$ denote an optimal control for the limit diffusion (1.3), and $\bar{u}^\delta(\cdot)$ a “smooth” δ -optimal control, where $\delta > 0$.

$$(1.3) \quad dx = \bar{b}(x, u) dt + \sigma(x) dw.$$

Now apply $\bar{u}^\delta(\cdot)$ to (1.1). Under fairly broad conditions, it is shown that

$$(1.4) \quad \inf_{u \in RC^\varepsilon} R^\varepsilon(u) \geq R^\varepsilon(\bar{u}^\delta) - \delta$$

* Received by the editors August 28, 1985; accepted for publication (in revised form) January 9, 1986.

† Lefschetz Center for Dynamical Systems, Division of Applied Mathematics, Brown University, Providence, Rhode Island 02912. The work of this author was supported in part by the U.S. Army Research Office under grant DAAG-29-84-K-0082; by the Air Force Office of Scientific Research under grant AF-AFOSR 81-0116-C; and by the Office of Naval Research under grant N00014-83-K-0542.

‡ University of Padova, Padova, Italy. The work of this author was supported by the U.S. Army Research Office under grant DAAG-29-K-0082 and by the Air Force Office of Scientific Research under grant AF-AFOSR 81-0116-C.

for small $\varepsilon > 0$, where RC^ε are the admissible (relaxed) controls for (1.1) (see § 3). Since $\bar{u}^\delta(\cdot)$ is only a function of x and t , it would be considerably simpler than an optimal control for (1.1).

The methods also work well for the discrete parameter case

$$(1.5) \quad x_{n+1}^\varepsilon = x_n^\varepsilon + \varepsilon F_\varepsilon(x_n^\varepsilon, \xi_n^\varepsilon, u_n).$$

The $\{\xi_n^\varepsilon\}$ and $\xi^\varepsilon(\cdot)$ can be state dependent, and there are straightforward extensions to the discounted cost problem, to the problem where the process is stopped on first exit from a set, to the impulsive control problem, and to the average cost per unit time case.

The basic technique is that of weak convergence theory [1], [2], [3], which will be seen to provide a very natural and relatively simple basis for results of the type presented here. The relevant background results are listed in § 2. In § 3 the problem on a finite interval $[0, T_1]$ for a form of (1.1) is set up, and the assumptions stated. For convenience in dealing with the weak convergence, as well as to minimize detail and the number of hypotheses, we work with relaxed controls. The relevant estimates and approximations (the "chattering" lemma, etc.) are also stated in § 3. In § 4, the results for the finite interval are proved. Section 5 concerns the discrete parameter case. The average cost per unit time problem is in § 6, and extensions are discussed in § 8.

A related problem is discussed by Blankenship and Papanicolaou in [4] and Bensoussan and Blankenship in [5]. They deal with the particular nondegenerate system

$$(1.6) \quad \begin{aligned} dx^\varepsilon &= f(x^\varepsilon, y^\varepsilon, u) dt + \sqrt{2} dw, \\ \varepsilon dy^\varepsilon &= g(x^\varepsilon, y^\varepsilon, u) dt + \sqrt{2\varepsilon} dB, \end{aligned}$$

where $w(\cdot)$ and $B(\cdot)$ are mutually independent standard Wiener processes. The technique in [4], [5] concerns an asymptotic expansion of the Bellman equation associated with the optimal control of (1.6). These expansions are hard to carry out, and rely heavily on various nondegeneracy properties associated with (1.6). In a "linear-quadratic" problem, they show that applying the optimal control for the limit problem to the prelimit problem gives a cost increase of $O(\varepsilon)$. There is negligible overlap in methodology with the ideas here. We can treat (1.6) if $g(\cdot)$ does not depend on $u(\cdot)$.

The results in [4], [5] seem to require an analytical approach, rather than our purely probabilistic approach. The methods used here seem quite simple in comparison, and cover a broader collection of problems. Expansions of the value functions do not seem to be obtainable by our methods. On the other hand, we can show, for many typical problem formulations, that the optimal or δ -optimal control for the limit system is a good (nearly optimal) control for the system which is driven by wide band-width noise. Such robustness is an important part of the statement of the control problem. In fact, the optimal or nearly optimal controls for diffusion models would not usually be of interest, were they also not good controls for the actual physical system which is "idealized" by the diffusion model. The general ideas carry over to more general spaces (e.g., to measure valued processes).

2. Weak convergence. Let $C^r[0, \infty)$ denote the space of R^r -valued continuous functions with the sup norm topology on bounded intervals, and let $D^r[0, \infty)$ denote the space of R^r -valued functions which are right continuous and have left-hand limits. Endow $D^r[0, \infty)$ with the Skorohod topology [2]. Our processes (except for the discrete parameter case) have values in C^r , but it is easier to prove tightness in D^r , and then to show that all limits are continuous.

Let \bar{F}_t^ε denote the minimal σ -algebra over which $\{x^\varepsilon(s), \xi^\varepsilon(s), s \leq t\}$ is measurable, and let E_t^ε denote expectation conditioned on \bar{F}_t^ε . Let $f(\cdot)$ be progressively measurable with respect to $\{\bar{F}_t^\varepsilon\}$. We say that $f(\cdot)$ is in $D(\hat{A}^\varepsilon)$, the domain of the operator \hat{A}^ε and $\hat{A}^\varepsilon f = g$ if for each $T < \infty$

$$\begin{aligned} \sup_{t \leq T} E|g(t)| &< \infty, \quad E|g(t+\delta) - g(t)| \rightarrow 0 \quad \text{as } \delta \downarrow 0, \quad \text{each } t, \\ \sup_{\substack{T \geq t \\ \delta > 0}} E \left| \frac{E_t^\varepsilon f(t+\delta) - f(t)}{\delta} - g(t) \right| &< \infty, \\ \lim_{\delta \downarrow 0} E \left| \frac{E_t^\varepsilon f(t+\delta) - f(t)}{\delta} - g(t) \right| &\rightarrow 0 \quad \text{each } t. \end{aligned}$$

If $f(\cdot) \in D(\hat{A}^\varepsilon)$ then ([3], [6])

$$(2.1) \quad f(t) - \int_0^t \hat{A}^\varepsilon f(s) ds \quad \text{is a martingale}$$

and

$$(2.2) \quad E_t^\varepsilon f(t+s) - f(t) = \int_t^{t+s} E_t^\varepsilon \hat{A}^\varepsilon f(u) du.$$

The following condition for tightness in $D'[0, \infty)$ ([3, Thm. 3.4]) is a sufficient condition for a criterion of Aldous and Kurtz [2]. Let \hat{C}_0 denote the continuous real valued functions on R^r with compact support, and \hat{C}_0^k the subset of functions all of whose mixed partial derivatives of order up to k are continuous.

THEOREM 0. Let $x^\varepsilon(\cdot)$ have paths in $D'[0, \infty)$ and let

$$(2.3) \quad \lim_{K \rightarrow \infty} \lim_{\varepsilon} P\{\sup_{t \leq T} |x^\varepsilon(t)| \geq K\} = 0, \quad \text{each } T < \infty.$$

For each $f(\cdot) \in \hat{C}_0^\infty$ and $T < \infty$ let there be a sequence $f^\varepsilon(\cdot) \in D(\hat{A}^\varepsilon)$ such that either (i) or (ii) below hold. Then $\{x^\varepsilon(\cdot)\}$ is tight in $D'[0, \infty)$.

(i) For each $T < \infty$, $\{\hat{A}^\varepsilon f^\varepsilon(t), \varepsilon > 0, t \leq T\}$ is uniformly integrable and for each $\alpha > 0$

$$(2.4) \quad \lim_{\varepsilon} P\{\sup_{t \leq T} |f^\varepsilon(t) - f(x^\varepsilon(t))| \geq \alpha\} = 0.$$

(ii) Equation (2.4) holds and for each $T < \infty$ there is a random variable $B_T^\varepsilon(f)$ such that

$$(2.5) \quad \sup_{t \leq T} |\hat{A}^\varepsilon f^\varepsilon(t)| \leq B_T^\varepsilon(f), \quad \lim_{K \rightarrow \infty} \lim_{\varepsilon} P\{B_T^\varepsilon(f) \geq K\} = 0.$$

Consider a discrete parameter case

$$x_{n+1}^\varepsilon = x_n^\varepsilon + F_\varepsilon(x_n^\varepsilon, \xi_n^\varepsilon).$$

Let \bar{F}_n^ε denote the minimal σ -algebra over which $\{x_i^\varepsilon, \xi_{i-1}^\varepsilon, i \leq n\}$ is measurable, with E_n^ε denoting the associated conditional expectation. We say that $f(\cdot) \in D(\hat{A}^\varepsilon)$ if it is constant on each $[n\varepsilon, (n+1)\varepsilon)$ interval, $f(n\varepsilon)$ is \bar{F}_n^ε -measurable, and $\sup_n E|f(n\varepsilon)| < \infty$. Then we define

$$\hat{A}^\varepsilon f(n\varepsilon) = [E_n^\varepsilon f((n+1)\varepsilon) - f(n\varepsilon)]/\varepsilon,$$

and the discrete parameter analogues of Theorem 0 and (2.1), (2.2) hold. In particular for $f \in D(\hat{A}^\varepsilon)$,

$$E_n^\varepsilon f((n+m)\varepsilon) - f(n\varepsilon) = \varepsilon \sum_{i=n}^{n+m-1} E_n^\varepsilon \hat{A}^\varepsilon f(i\varepsilon).$$

Let $M(\infty)$ denote the collection of measures $\{m(\cdot)\}$ on the Borel subsets of $U \times [0, \infty)$, where U is compact and $m([0, t] \times U) = t$, for all $t \geq 0$. We will be working with weak convergence of a sequence of $M(\infty)$ -valued random variables. Write $(m, f) = \int f(s, \alpha) m(ds \times d\alpha)$. We say that $m_n \rightarrow 0$ if $(m, f) \rightarrow 0$ for each continuous $f(\cdot)$ with compact support. When we say that $m_n(\cdot) \Rightarrow m(\cdot)$ for a sequence of random measures, we always mean weak convergence in $M(\infty)$.

3. Assumptions and relaxed controls. We adopt a particular noise model, which is a standard way of modeling wide band-width noise. The model can readily be generalized, since only a few properties of the processes are used. The model is convenient also because the relevant weak convergence results can be easily referred to. A control $u(\cdot)$ for (1.1) is said to be *admissible* if it takes values in U , a compact set, and it is progressively measurable with respect to the σ -algebras $\sigma\{\xi^\varepsilon(s), s \leq t\} \equiv F_t^\varepsilon$.

A random measure $m(\cdot)$ with values in $M(\infty)$ is said to be an *admissible relaxed control* if $\iint_0^t f(s, \alpha) m(ds \times d\alpha) \equiv (f, m)_t$ is progressively measurable with respect to $\{F_t^\varepsilon\}$ for each bounded continuous $f(\cdot)$. If $m(\cdot)$ is admissible, then there is a measure valued "derivative" function of (ω, t) with value $m_t(\cdot)$ at time t such that for smooth $f(\cdot)$

$$\int f(s, \alpha) m(ds \times d\alpha) = \int dt \int f(s, \alpha) m_s(d\alpha),$$

and $m_t(\cdot)$ is (weakly) progressively measurable in the sense that $\int_0^t ds \int f(s, \alpha) m_s(d\alpha)$ is progressively measurable. We sometimes write the derivative function as m_\cdot . Let AC^ε and RC^ε denote the class of admissible and admissible relaxed controls, respectively, for (1.1).

Assumption A1. $\xi^\varepsilon(t) = \xi(t/\varepsilon^2)$, where $\xi(\cdot)$ is a stationary zero mean process which is either (a) strongly mixing,¹ right continuous and bounded, with the mixing rate function $\phi(\cdot)$ satisfying $\int_0^\infty \phi^{1/2}(s) ds < \infty$ or (b) stationary Gauss-Markov with an integrable correlation function (which thus must go to zero exponentially).

Assumption A2. $F_\varepsilon(x, \xi, u) = b(x, u) + \tilde{b}(x, \xi) + g(x, \xi)/\varepsilon$, where $E\tilde{b}(x, \xi) = Eg(x, \xi) = 0$ under A1(a), and $\tilde{b}(x, \xi) = g(x)\xi$, $\tilde{b}(x, \xi) = b(x)\xi$ under A1(b), $k(\cdot, \cdot)$ is bounded and continuous, and $b(\cdot, \cdot)$, $\tilde{b}(\cdot, \cdot)$, $g(\cdot, \cdot)$ are continuous. The derivative $g_x(\cdot, \xi)$ is continuous (in x, ξ). Also $b(\cdot, \alpha)$ satisfies a linear growth condition and a Lipschitz condition in x , uniformly in $\alpha \in U$. Under A1(a), $\tilde{b}(\cdot, \xi)$, $g(\cdot, \xi)$, $g_x(\cdot, \xi)$ satisfy the same uniform Lipschitz and growth condition, and under A1(b), $\tilde{b}(\cdot)$, $g(\cdot)$ and $g_x(\cdot)$ do.

Define

$$\begin{aligned} \{a_{ij}(x)\} &= \int_{-\infty}^{\infty} Eg(x, \xi(t))g'(x, \xi(0)) dt = a(x), \\ \bar{b}_i(x, u) &= b_i(x, u) + \int_0^\infty E \sum_j g_{i,x_j}(x, \xi(t))g_j(x, \xi(0)) dt, \quad i \leq r. \end{aligned}$$

Assumption A3. Suppose that $\{a_{ij}(\cdot)\}$ has a Lipschitz continuous square root $\sigma(\cdot)$.

For the problem on $[0, T_1]$, the boundedness condition on $k(\cdot, \cdot)$ can be replaced by a polynomial growth condition. For the average cost per unit time problem, the stability methods and assumptions of § 7 can be used for the same purpose.

¹ That is, for $A \in \sigma(\xi(v), v \leq s)$, $B \in \sigma(\xi(v), v \geq s+t)$, $\sup_{A,B} |P(B|A) - P(B)| \leq \phi(s)$.

The weak convergence and existence (of an optimal control) arguments are easier if one works with relaxed controls. It is convenient to work with relaxed controls on $[0, \infty)$. If the control problem is of interest on $[0, T_1]$ only, then define $u(\cdot)$ or $m(\cdot)$ in any admissible way on $[T_1, \infty)$.

Admissible controls for (1.3) or (3.1) below. An *admissible control* for (1.3) is any U -valued function $u(\cdot)$ which is nonanticipative with respect to $w(\cdot)$. An *admissible relaxed control* for (1.3) or (3.1) below is any $M(\infty)$ valued random variable $m(\cdot)$ such that for any collection $\{f_\beta(\cdot)\}$ of bounded continuous functions $f_\beta(\cdot)$, and each $t > 0$, $\{\int_0^t f_\beta(s, \alpha) m(ds \times d\alpha)\}$ is independent of $\{w(t+s) - w(t), s > 0\}$. If $m(\cdot)$ is an admissible relaxed control then there is a $(\omega, t$ -dependent) measure $m_t(\cdot)$ on the Borel sets of U such that

$$\int_0^t \int f(s, \alpha) m(ds \times d\alpha) = \int_0^t ds \int f(s, \alpha) m_s(d\alpha), \quad t < \infty$$

for each bounded and continuous $f(\cdot)$ and almost all ω . When working with (1.3) or (3.1), we assume that $\bar{b}(\cdot)$ and $\sigma(\cdot)$ have the continuity, growth and Lipschitz conditions ascribed to $b(\cdot)$ and $\sigma(\cdot)$ in A1-A3. Let AC and RC denote the class of admissible and admissible relaxed controls, respectively. The classes of admissible controls are defined as they are because we wish to avoid working explicitly with feedback controls. Our processes are uniquely defined for any admissible control.

THEOREM 1. *Let $m(\cdot)$ be an admissible relaxed control (with respect to a Wiener process $w(\cdot)$). Then there exists a nonanticipative solution to*

$$(3.1) \quad dx = dt \int \bar{b}(x, \alpha) m_t(d\alpha) + \sigma(x) dw, \quad x(0) = x$$

and

$$(3.2) \quad E \sup_{t \leq T} |x(t)|^2 \leq K[1 + |x|^2],$$

where K depends only on T and on the growth rates and Lipschitz constants on $\bar{b}(\cdot)$ and $\sigma(\cdot)$.

Define $\{x_n^\Delta\}$ by $x_0^\Delta = x_1^\Delta = x$ and for $n \geq 1$,

$$(3.3) \quad x_{n+1}^\Delta = x_n^\Delta + \int_{n\Delta-\Delta}^{n\Delta} ds \int \bar{b}(x_n^\Delta, \alpha) m_s(d\alpha) + \sigma(x_n^\Delta)[w(n\Delta + \Delta) - w(n\Delta)].$$

Define $x^\Delta(\cdot)$ to be the piecewise constant interpolation (interval Δ) of $\{x_n^\Delta\}$. Then there is a $K_\Delta \rightarrow 0$ as $\Delta \rightarrow 0$ (and depending only on T and on the Lipschitz and growth constants) such that

$$(3.4) \quad E \sup_{t \leq T} |x^\Delta(t) - x(t)|^2 \leq K_\Delta(1 + |x|^2)$$

(K_Δ does not depend on $m(\cdot)$).

Let $m^n(\cdot) \Rightarrow \bar{m}(\cdot)$, where the $m^n(\cdot)$ are admissible with respect to some Wiener process, and let $x^n(\cdot)$ satisfy (3.1) with $m(\cdot) = m^n$. Then $(x^n(\cdot), m^n(\cdot)) \Rightarrow (x(\cdot), \bar{m}(\cdot))$ where $x(\cdot), \bar{m}(\cdot)$ satisfy (3.1) for some Wiener process $w(\cdot)$ and $m(\cdot)$ is admissible with respect to $w(\cdot)$.

Proof. The existence and uniqueness proof for the relaxed control case follows the same (standard) lines as when an admissible control $u(\omega, t)$ is used, and is discussed by Fleming [7] and Fleming and Nisio [8]. The proofs of the estimates (3.2), (3.4) also follow the classical lines. To get the weak convergence in the last paragraph, it

is sufficient to work with the discrete parameter case (3.3), in view of the uniformity (in $m(\cdot)$) of K and K_Δ . But the result is obvious for the discrete parameter case, owing to the continuity of $\bar{b}(\cdot, \cdot)$ and the Lipschitz conditions and linear growth conditions. QED

For (3.1), define

$$(3.5) \quad R(m) = E \int_0^{T_1} \int k(x(s), \alpha) m_s(d\alpha) ds,$$

where $x(\cdot)$ corresponds to $m(\cdot)$ via (3.1). We *sometimes* write the solution to (1.3) or (3.1) as $x(u, \cdot)$ or $x(m, \cdot)$.

THEOREM 2. *In the class of admissible relaxed controls for (3.1), there is an optimal control.*

Proof. The theorem follows from Theorem 1. Simply choose a weakly convergent subsequence $m^\delta(\cdot)$, $\delta \rightarrow 0$, such that $R(m^\delta) \rightarrow \inf_{m \in RC} R(m) \equiv \bar{R}$. Denote the limit of $\{x(m^\delta, \cdot), m^\delta(\cdot)\}$ by $(x(\bar{m}, \cdot), \bar{m}(\cdot))$. Then by Theorem 1, $\bar{m}(\cdot)$ is admissible for some Wiener process $w(\cdot)$ and $(x(\bar{m}, \cdot), \bar{m}(\cdot), w(\cdot))$ solves (3.1). By the weak convergence,

$$E \int_0^{T_1} \int k(x^\delta(s), \alpha) m^\delta(ds \times d\alpha) \rightarrow E \int_0^{T_1} \int k(x(s), \alpha) \bar{m}(ds \times d\alpha) = \bar{R} = R(\bar{m}).$$

QED

Since we wish to show (in the following sections) that any smooth and nearly optimal feedback control for (1.3) is a nearly optimal control of (1.1) for small $\varepsilon > 0$, it is important to know that there is a smooth nearly optimal control for (1.3). This is shown in the next two theorems.

The chattering lemma.

THEOREM 3. *For each $\delta > 0$, there is a piecewise constant admissible control $u^\delta(\cdot)$ for (1.3) such that*

$$R(u^\delta) \leq \inf_{m \in RC} R(m) + \delta.$$

Remark. A proof is in [7], [8]. We only give a rough outline of the construction. Let $\bar{m}(\cdot)$ be an optimal admissible relaxed control. Let $u_1^\rho, \dots, u_k^\rho$ be a ρ -grid in U . Define A_1^ρ by $A_1^\rho = \{\alpha \in U: |\alpha - u_1^\rho| \leq \rho\}$. For $k \geq n > 1$, define

$$A_n^\rho = \{\alpha \in U: |\alpha - u_n^\rho| \leq \rho\} - \bigcup_i^{n-1} A_i^\rho.$$

For $\Delta > 0$ and $i \geq 0$, define

$$\tau_{in}^{\Delta\rho} = \int_{i\Delta}^{i\Delta+\Delta} \bar{m}_s(A_s^\rho) ds,$$

the total integrated time that the optimal relaxed control "takes values" in the set A_n^ρ in the time interval $[i\Delta, i\Delta + \Delta)$. Define the piecewise constant admissible control $\tilde{u}^\delta(\cdot)$ by $\tilde{u}^\delta(t) = u_0^\rho$ for $t \leq \Delta$, where u_0^ρ is any value in U ; in general, set $\tilde{u}^\delta(t) = u_n^\rho$ on

$$\left[(i+1)\Delta + \sum_{l=1}^{n-1} \tau_{il}^{\Delta\rho}, (i+1)\Delta + \sum_{l=1}^n \tau_{il}^{\Delta\rho} \right], \quad i \geq 0, \quad n \leq k.$$

Then, for small ρ and Δ , $\tilde{u}^\delta(\cdot)$ satisfies our needs, even though the intervals of constancy are random.

We can also get a control whose intervals of constancy are nonrandom. Let $\Delta_1 > 0$ be such that $\Delta/\Delta_1 \equiv \bar{k}$ is a large integer, and write $k_{ii}^{\Delta\rho} = [\tau_{ii}^{\Delta\rho}/\Delta_1]$. Then define $\hat{u}^\delta(\cdot)$ as $\tilde{u}^\delta(\cdot)$ was defined but with $k_{ii}^{\Delta\rho}\Delta_1$ replacing $\tau_{ii}^{\Delta\rho}$, and on the nonassigned set, simply set $\hat{u}^\delta(t) = u_0^\rho$, where u_0^ρ is any value in U . For small Δ_1 , Δ and ρ , and large \bar{k} , $\hat{u}^\delta(\cdot)$ also satisfies our needs.

THEOREM 4. *For each $\delta > 0$, there is a piecewise constant (in t) and locally Lipschitz continuous in x (uniformly in t) control $\bar{u}^\delta(\cdot)$ such that $\bar{u}^\delta(t) = \bar{u}^\delta(x(i\Delta), i\Delta)$ for $t \in [i\Delta, i\Delta + \Delta]$*

$$R(\bar{u}^\delta) \leq \inf_{m \in RC} R(m) + \delta.$$

Proof. Fix $\delta > 0$. By the previous theorem, we can find a $\Delta > 0$ and an admissible control $u^\delta(\cdot)$, constant on each interval $[i\Delta, i\Delta + \Delta)$, and such that

$$R(u^\delta) \leq \inf_{m \in RC} R(m) + \delta/4.$$

By examining the imbedded Markov chain $\{x(i\Delta), i\Delta \leq T_1\}$, we see that there is an admissible control $\hat{u}^\delta(t)$ which is piecewise constant and has the form $\hat{u}^\delta(t) = \hat{u}^\delta(x(i\Delta), i\Delta)$ for $t \in [i\Delta, i\Delta + \Delta)$ for some function $\hat{u}^\delta(x, t)$, and is such that

$$R(\hat{u}^\delta) \leq R(u^\delta) + \delta/4.$$

In fact we can suppose that the $\hat{u}^\delta(t)$ take only a finite number of values u_1, \dots, u_k , where k might depend on δ but not otherwise on Δ . Let $x(\cdot)$ denote the process corresponding to the control $\hat{u}^\delta(\cdot)$. Define $B_l^i = \{x: \hat{u}^\delta(x, i\Delta) = u_l\}$. There are open sets \tilde{B}_l^i with smooth boundaries (say, unions of a finite number of spheres) and whose closures are disjoint and such that $(\partial B$ denotes the boundary of the set $B)$

$$(3.6) \quad \begin{aligned} P\{x(i\Delta) \in \partial \tilde{B}_l^i\} &= 0 \quad \text{for all } i, l, \quad i\Delta \leq T_1, \\ \sum_{i=0}^{T_1/\Delta-1} P\left\{x(i\Delta) \in \bigcup_l (B_l^i \Delta \tilde{B}_l^i)\right\} &\leq \frac{\delta}{4T_1} \left[1 + \sup_{x, \alpha} |k(x, \alpha)|\right]. \end{aligned}$$

For each i , define $\tilde{u}^\delta(x, i\Delta)$ to equal u_l on \tilde{B}_l^i , and use any locally Lipschitz continuous interpolation for $x/\in \cup_l \tilde{B}_l^i$. Thus the costs with use of $\hat{u}^\delta(\cdot)$ (on one hand) and use (on the other hand) of $\tilde{u}^\delta(x, i\Delta)$ for $t \in [i\Delta, i\Delta + \Delta)$ and each i differ by at most $\delta/2$. In fact the latter control and $\hat{u}^\delta(\cdot)$ differ on a set whose probability is less than the right side of (3.6). QED

4. Weak convergence of and approximation of the optimal controls for $x^\varepsilon(\cdot)$. In this section we work with the control problem on $[0, T_1]$ and prove (Theorem 5) that the weak limit of any (weakly convergent) sequence of admissible relaxed control for (4.1) is an admissible relaxed control for (3.1) and that the corresponding costs converge. Then, in Theorem 6, we show that any smooth “nearly optimal” feedback control for (3.1) also is “nearly optimal” for (4.1) for small ε .

Let $\delta_\varepsilon \rightarrow 0$, and let $\hat{m}^\varepsilon(\cdot)$ be a δ_ε -optimal admissible relaxed control for the process defined by

$$(4.1) \quad \dot{x}^\varepsilon = \int b(x^\varepsilon, \alpha) m_t(d\alpha) + \tilde{b}(x^\varepsilon, \xi^\varepsilon) + g(x^\varepsilon, \xi^\varepsilon)/\varepsilon,$$

with cost function (3.5). For convenience, we define all $m(\cdot)$ on $[0, \infty)$. In the analysis below it is convenient (but not necessary) to have $\int b(x^\varepsilon(t), \alpha) m_t(d\alpha)$ right continuous (in order to be able to readily evaluate \hat{A}^ε). Owing to the Lipschitz condition and to

the continuity and growth conditions, for each ε we can suppose (w.l.o.g.) that $\hat{m}_t^\varepsilon(\cdot)$, is, in fact, constant on intervals $[i\Delta_\varepsilon, i\Delta_\varepsilon + \Delta_\varepsilon)$ for small enough Δ_ε .

Define L^m , the infinitesimal operator of $x(m, \cdot)$ defined by (3.1), by

$$L^m f(x) = f'_x(x) \int \bar{b}(x, \alpha) m_t(d\alpha) + \frac{1}{2} \sum_{ij} f_{x_i x_j}(x) a_{ij}(x).$$

THEOREM 5. Assume A1–A3. Then $\{x^\varepsilon(\hat{m}^\varepsilon, \cdot), \hat{m}^\varepsilon(\cdot)\}$ is tight in $D^r[0, \infty) \times M(\infty)$. Let $(x^\varepsilon(\hat{m}^\varepsilon, \cdot), \hat{m}^\varepsilon(\cdot)) \Rightarrow (x(\hat{m}, \cdot), \hat{m}(\cdot))$. There is a $w(\cdot)$ such that $\hat{m}(\cdot)$ is admissible with respect to $w(\cdot)$ and

$$(4.2) \quad dx = dt \int \bar{b}(x, \alpha) \hat{m}_t(d\alpha) + \sigma(x) dw.$$

Also

$$\begin{aligned} R^\varepsilon(\hat{m}^\varepsilon) &= E \int_0^{T_1} \int k(x^\varepsilon(s), \alpha) \hat{m}^\varepsilon(ds \times d\alpha) \\ &\rightarrow E \int_0^{T_1} \int k(x(s), \alpha) \hat{m}(ds \times d\alpha) = R(\hat{m}). \end{aligned}$$

Proof. We first work with a truncated system, since tightness is easier to prove if the $x^\varepsilon(\cdot)$ paths are all bounded (see e.g. [3, Chap. 3.3 or 4.6.4] or [9]). Let $q_N(\cdot)$ be a twice continuously differentiable function satisfying $q_N(x) = 1$ for $|x| \leq N$, $q_N(x) = 0$ for $|x| \geq N+1$ and $q_N(x) \in [0, 1]$ for all x . Define $b_N(x, \alpha) = b(x, \alpha)q_N(x)$, $g_N(x, \xi) = g(x, \xi)q_N(x)$, etc., and let $x^{\varepsilon, N}(\cdot)$ denote the \sim solution to (4.1) corresponding to the use of b_N , \tilde{b}_N , g_N , and $\hat{m}^\varepsilon(\cdot)$.

Part 1. Tightness of $\{x^{\varepsilon, N}(\cdot)\}$. Since $U \times [0, t_1]$ is compact for each $t_1 < \infty$, $\{\hat{m}^\varepsilon(\cdot)\}$ is tight in $M(\infty)$. To prove the tightness of $\{x^{\varepsilon, N}(\cdot)\}$, we use the first order perturbed test function method of [3, Chap. 3] (see also [9]). Let $f(\cdot) \in \hat{C}_0^2$. Then (write x for $x^{\varepsilon, N}(t)$ for convenience)

$$\hat{A}^\varepsilon f(x) = f'_x(x) \left[\int b_N(x, \alpha) \hat{m}_t^\varepsilon(d\alpha) + \tilde{b}_N(x, \xi^\varepsilon(t)) + g_N(x, \xi^\varepsilon(t)) / \varepsilon \right].$$

For arbitrary $T < \infty$ and for $t \leq T$, define $f_1^\varepsilon(t) = f_1^\varepsilon(x^{\varepsilon, N}(t), t)$, where

$$\begin{aligned} f_1^\varepsilon(x, t) &= \int_t^T f'_x(x) E_t^\varepsilon g_N(x, \xi^\varepsilon(s)) ds / \varepsilon \\ &= \varepsilon \int_{t/\varepsilon^2}^{T/\varepsilon^2} f'_x(x) E_t^\varepsilon g_N(x, \xi(s)) ds. \end{aligned}$$

Under A1(a), $f_1^\varepsilon(t) = O(\varepsilon)$. Under A1(b), $f_1^\varepsilon(t) = O(\varepsilon)|\xi^\varepsilon(t)|$. In either case

$$\sup_{t \leq T} |f_1^\varepsilon(t)| \xrightarrow{P} 0 \quad \text{as } \varepsilon \rightarrow 0.$$

We have

$$\begin{aligned} \hat{A}^\varepsilon f_1^\varepsilon(t) &= -f'_x(x^{\varepsilon, N}(t)) g_N(x^{\varepsilon, N}(t), \xi^\varepsilon(t)) / \varepsilon \\ &\quad + \frac{1}{\varepsilon} \int_t^T ds [f'_x(x^{\varepsilon, N}(t)) E_t^\varepsilon g_N(x^{\varepsilon, N}(t), \xi^\varepsilon(s))]'_x x^{\varepsilon, N}(t). \end{aligned}$$

Define $f^\varepsilon(t) = f(x^{\varepsilon,N}(t)) + f_1^\varepsilon(t)$. Then, writing x for $x^{\varepsilon,N}(t)$, using the above results and a scale change $s/\varepsilon^2 \rightarrow s$,

$$\begin{aligned}
 \hat{A}^\varepsilon f^\varepsilon(t) = & f'_x(x) \int b_N(x, \alpha) \hat{m}_t^\varepsilon(d\alpha) + f'_x(x) \tilde{b}_N(x, \xi^\varepsilon(t)) \\
 & + \int_{t/\varepsilon^2}^{T/\varepsilon^2} ds E_t^\varepsilon[f'_x(x) g_N(x, \xi(s))]'_x g_N(x, \xi^\varepsilon(t)) \\
 & + \varepsilon \int_{t/\varepsilon^2}^{T/\varepsilon^2} ds E_t^\varepsilon[f'_x(x) g_N(x, \xi(s))]'_x \\
 & \cdot \left[\int b_N(x, \alpha) \hat{m}_t^\varepsilon(d\alpha) + \tilde{b}_N(x, \xi^\varepsilon(t)) \right].
 \end{aligned}
 \tag{4.3}$$

Under A1(a), the second and third terms in (4.3) are $O(1)$. Under A1(b), they are $O(1)[1 + |\xi^\varepsilon(t)|^2]$. Under A1(a), the last term is $O(\varepsilon)$, and under A1(b) it is $O(\varepsilon)[1 + |\xi^\varepsilon(t)|^2]$. In either case the conditions of Theorem 0 hold. Hence $\{x^{\varepsilon,N}(\cdot)\}$ is tight in $D'[0, \infty)$.

Part 2. The martingale problem satisfied by the limit. Let ε index a weakly convergent subsequence with limit denoted by $x^N(\cdot)$, $\hat{m}(\cdot)$; i.e., $\{x^{\varepsilon,N}(\cdot), \hat{m}^\varepsilon(\cdot)\} \Rightarrow (x^N(\cdot), \hat{m}(\cdot))$. There is an (ω, t) -measurable $\hat{m}_t(\cdot)$ such that $\hat{m}_t(U) = 1$ and

$$\int_0^t \int f(s, \alpha) \hat{m}_s(d\alpha) ds = \int_0^t \int f(s, \alpha) \hat{m}(ds \times d\alpha)$$

for each continuous $f(\cdot)$. This is a consequence of the fact that $\hat{m}\{A \times [0, t]\}$ is absolutely continuous for each Borel A , uniformly in ω , A , which implies that the (measurable) limit

$$\lim_{\Delta} [\hat{m}\{A \times [0, t]\} - \hat{m}\{A \times [0, t - \Delta]\}] / \Delta \equiv \hat{m}_t(A)$$

exists for a.a. (ω, t) for each Borel A .

Define L_N^m as L^m was defined, but with the use of \bar{b}_N and g_N instead of \bar{b} and \bar{g} . Let $f(\cdot) \in \hat{C}_0^2$ and define $M_f^N(\cdot)$ by

$$M_f^N(t) = f(x^N(t)) - f(x(0)) - \int_0^t L_N^m f(x^N(s)) ds.$$

We next show that $M_f^N(\cdot)$ is a martingale with respect to $B^N(t) \equiv \sigma\{x^N(s); \hat{m}(A \times [0, s]), \text{Borel } A, s \leq t\}$.

We know that $x^N(\cdot)$ has paths in $D'[0, \infty)$, but we have not yet proved that the paths are in $C'[0, \infty)$. There are at most a countable set of t -points such that $P\{x^N(\cdot)$ is discontinuous at $t\} > 0$. Denote this set by $\mathcal{T} = \{\tau_i\}$. In what follows, until continuity is established, the t_i , t , $t + s$ do not take values in \mathcal{T} . Let $h(\cdot)$ be bounded and continuous and let $t_i < t < t + s$. Let q_1 and q_2 be arbitrary integers and $k_j(\cdot)$ arbitrary bounded and continuous functions. By (2.1), (2.2), and a change of scale ($s/\varepsilon^2 \rightarrow s$) for one of the terms, we have

$$\begin{aligned}
 & Eh(x^{\varepsilon,N}(t_i), (k_j, \hat{m}^\varepsilon)_{t_i}, i \leq q_1, j \leq q_2) \\
 & \cdot \left\{ f(x^{\varepsilon,N}(t+s)) - f(x^{\varepsilon,N}(t)) + f_1^\varepsilon(t+s) - f_1^\varepsilon(t) \right.
 \end{aligned}$$

$$\begin{aligned}
(4.4) \quad & - \int_t^{t+s} \int f'_x(x^{\varepsilon,N}(\tau)) b_N(x^{\varepsilon,N}(\tau), \alpha) \hat{m}^\varepsilon(d\tau \times d\alpha) \\
& - E_t^\varepsilon \int_t^{t+s} f'_x(x^{\varepsilon,N}(\tau)) \tilde{b}_N(x^{\varepsilon,N}(\tau), \xi^{\varepsilon,N}(\tau)) d\tau \\
& - \int_t^{t+s} d\tau E_\tau^\varepsilon \int_{\tau/\varepsilon^2}^{T/\varepsilon^2} [f'_x(x^{\varepsilon,N}(\tau)) g_N(x^{\varepsilon,N}(\tau), \xi(v))]'_x g_N(x^{\varepsilon,N}(\tau), \xi^\varepsilon(\tau)) dv \\
& + \text{terms which go to 0 in mean as } \varepsilon \rightarrow 0 \Big\} = 0.
\end{aligned}$$

Owing to (2.1) and (2.2), (4.4) holds with or without the E_t^ε term on the right-hand side. Recall that $(f, m)_t \equiv \int_0^t \int f(s, \alpha) m(ds \times d\alpha)$.

Now take limits ($\varepsilon \rightarrow 0$) in (4.4) and use Skorohod imbedding ([10, Thm. 3.1.1]). The imbedding allows us to define the probability space so that the weak convergence becomes w.p.l. in the topology of the space $D'[0, \infty) \times M(\infty)$. We use the imbedding without changing the notation, where convenient. The f_1^ε terms in (4.4) disappear as $\varepsilon \rightarrow 0$. Also by the weak convergence and Skorohod imbedding,

$$\begin{aligned}
\int_t^{t+s} \int b_N(x^{\varepsilon,N}(\tau), \alpha) \hat{m}^\varepsilon(d\tau \times d\alpha) & \rightarrow \int_t^{t+s} \int b_N(x^N(\tau), \alpha) \hat{m}(d\tau \times d\alpha), \\
(k_j, \hat{m}^\varepsilon)_t & \rightarrow (k_j, \hat{m})_t,
\end{aligned}$$

w.p.l., uniformly on each finite interval. Next consider the second integral term in (4.4). We will show that

$$(4.5) \quad \lim_{\varepsilon} E \left| \int_t^{t+s} E_t^\varepsilon \tilde{b}_N(x^{\varepsilon,N}(\tau), \xi^\varepsilon(\tau)) d\tau \right| = 0.$$

Since $\{x^{\varepsilon,N}(\cdot)\}$ is tight in $D'[0, \infty)$ it is essentially a right equicontinuous set in the following sense. Given $\rho > 0$ and $T < \infty$, there is a compact set $\Omega_\rho \subset D'[0, T]$ such that

$$P\{x^{\varepsilon,N}(\cdot) \in \Omega_\rho\} \geq 1 - \rho.$$

For $y(\cdot) \in D'[0, T]$, define $w_y[a, b] = \sup\{|y(s) - y(t)| : s, t \in [a, b]\}$ and define

$$w'_y(\delta) = \inf_{\substack{\{t_i\} \\ q}} \max_{i \leq q} w_y[t_i, t_{i+1}],$$

where $0 = t_0 < \dots < t_q = T$ and $t_{i+1} - t_i \geq \delta$. Then [1, p. 116]

$$(4.6) \quad \lim_{\delta} \sup_{y(\cdot) \in \Omega_\rho} w'_y(\delta) = 0.$$

Because of this “equirightcontinuity” characterization, to get the limit (4.5) it is sufficient to evaluate

$$\begin{aligned}
& \lim_{\Delta \downarrow 0} \overline{\lim}_{\varepsilon} E \left| \int_t^{t+s} E_t^\varepsilon \tilde{b}_N(x^{\varepsilon,N}(\tau - \Delta), \xi^\varepsilon(\tau)) d\tau \right| \\
& \leq \lim_{\Delta \downarrow 0} \overline{\lim}_{\varepsilon} E \left| \int_t^{t+s} E_{\tau-\Delta}^\varepsilon \tilde{b}(x^{\varepsilon,N}(\tau - \Delta), \xi^\varepsilon(\tau)) d\tau \right| \equiv \lim_{\Delta \downarrow 0} \overline{\lim}_{\varepsilon} K_\Delta^\varepsilon.
\end{aligned}$$

There are constants C_N and C'_N depending only on N such that, under A1(a)

$$|E_{\tau-\Delta}^\varepsilon \tilde{b}(x^{\varepsilon,N}(\tau - \Delta), \xi^\varepsilon(\tau))| \leq C_N \phi(\Delta/\varepsilon^2),$$

and under A1(b)

$$|E_{\tau-\Delta}^{\varepsilon} \tilde{b}(x^{\varepsilon, N}(\tau-\Delta), \xi^{\varepsilon}(\tau))| \leq C'_N [\exp -\lambda \Delta / \varepsilon^2] |\xi^{\varepsilon}(\tau-\Delta)|,$$

where $\phi(\cdot)$ is the mixing rate A1(a) for $\xi(\cdot)$, and $\exp -\lambda t$ is a bound on the norm of the correlation matrix (under A1(b)). Thus, under A1(a), $\lim_{\varepsilon} K_{\Delta}^{\varepsilon} = 0$ for each $\Delta > 0$. Under A1(b) $K_{\Delta}^{\varepsilon} \leq O(\exp -\lambda \Delta / \varepsilon^2) \int_t^{t+s} |\xi^{\varepsilon}(\tau)| d\tau$. Thus (4.5) holds.

By a very similar technique we can show that, as $\varepsilon \rightarrow 0$, the double integral term in the brackets in (4.4) converges (in mean) to

$$(4.7) \quad \int_t^{t+s} d\tau \int_0^{\infty} E[f'_x(x^N(\tau))g_N(x^N(\tau), \xi(s))]'_x g_N(x^N(\tau), \xi(0)) ds.$$

The expectation in (4.7) is over the $\xi(\cdot)$ only. The $x^N(\tau)$ is considered to be a fixed parameter when taking the expectation. This last limit result is, in fact, a special case of [3, Thm. 5.11]. Thus

$$(4.8) \quad \begin{aligned} & Eh(x^N(t_i), (k_j, \hat{m})_{t_i}, i \leq q_1, j \leq q_2) \\ & \cdot \left[f(x^N(t+s)) - f(x^N(t)) - \int_t^{t+s} L_N^{\hat{m}} f(x^N(\tau)) d\tau \right] = 0. \end{aligned}$$

Since $q_1, q_2, h(\cdot)$ and the $k_j(\cdot)$, t_i, t, s are arbitrary (with $t_i, t, t+s \notin \mathcal{T} = \{\tau_i\}$), the assertion that the $M_f^N(\cdot)$ are $\{B^N(t)\}$ martingales is proved.

It follows from the fact that $x^N(\cdot)$ solves the martingale problem in $D'[0, \infty)$ associated with the local operator $L_N^{\hat{m}}$ that $x^N(\cdot)$ has continuous paths w.p.l.

Part 3. Representation of the limit. Define $\sigma_N(x) = \sigma(x)q_N(x)$. Since the $M_f^N(\cdot)$ are martingales with respect to $B^N(t)$, there is a standard Wiener process $w^N(\cdot)$ (augmenting the probability space if necessary, via the addition of an independent Wiener process if $a(\cdot)$ is degenerate) such that $w^N(t)$ is $B^N(t)$ adapted, $x^N(\cdot)$ is nonanticipative with respect to $w^N(\cdot)$ and

$$(4.9) \quad dx^N = dt \int \bar{b}_N(x^N, \alpha) \hat{m}_t(d\alpha) + \sigma_N(x^N) dw^N.$$

Also, since $w^N(\cdot)$ is $B^N(t)$ adapted, the $\hat{m}(A \times [0, t])$ and $\hat{m}_t(A)$ are nonanticipative with respect to $w^N(\cdot)$. Hence $\hat{m}(\cdot)$ is an admissible relaxed control for the problem with coefficients \bar{b}_N, σ_N .

We now let $N \rightarrow \infty$ and use a "piecing together" argument to get the representation (4.2).

The last assertion of the theorem follows from the weak convergence $(x^{\varepsilon}(\cdot), \hat{m}^{\varepsilon}(\cdot)) \Rightarrow (x(\cdot), \hat{m}(\cdot))$, and the continuity of the process $x(\cdot)$. QED

Remark. With a simpler proof (not requiring working with $\{\hat{m}^{\varepsilon}(\cdot)\}$) we have the following. Let $u(\cdot)$ be a (time-dependent) feedback control which is continuous in x , uniformly in t on each bounded (x, t) set, and for which the martingale problem associated with (1.3) has a unique solution. Then $x^{\varepsilon}(u, \cdot) \Rightarrow x(u, \cdot)$. Also $R^{\varepsilon}(u) \rightarrow R(u)$.

THEOREM 6. Assume A1–A3. Let $\delta > 0$. For the feedback control $\bar{u}^{\delta}(\cdot)$ of Theorem 4 on any Lipschitz continuous (uniformly in t) δ -optimal control $\bar{u}^{\delta}(\cdot)$ for $x(\cdot)$, we have

$$(4.10) \quad \lim_{\varepsilon} [R^{\varepsilon}(\bar{u}^{\delta}) - \inf_{m \in RC^{\varepsilon}} R^{\varepsilon}(m)] \leq \delta.$$

Proof. By the weak convergence argument of Theorem 5, $x^{\varepsilon}(\bar{u}^{\delta}, \cdot) \Rightarrow x(\bar{u}^{\delta}, \cdot)$ and $R^{\varepsilon}(\bar{u}^{\delta}) \rightarrow R(\bar{u}^{\delta})$. The theorem follows from this since

$$R^{\varepsilon} \in (\hat{m}^{\varepsilon}) \rightarrow R(\hat{m}) \geq \inf_{m \in RC} R(m) \geq R(\bar{u}^{\delta}) - \delta. \quad \text{QED}$$

5. The discrete parameter case. An advantage of the weak convergence point of view is that the discrete parameter case can be treated in almost the same way as the continuous parameter case.

Let the system be given by

$$(5.1) \quad x_{n+1}^{\varepsilon} = x_n^{\varepsilon} + \varepsilon \int b(x_n^{\varepsilon}, \alpha) m_n(d\alpha) + \varepsilon \tilde{b}(x_n^{\varepsilon}, \xi_n) + \sqrt{\varepsilon} g(x_n^{\varepsilon}, \xi_n),$$

where $\{\xi_n\}$ satisfies the discrete parameter form of A1(a) or A1(b) and the conditions on $g(\cdot)$, $b(\cdot)$, $\tilde{b}(\cdot)$ and $k(\cdot)$ in A2–A3 hold. Also, assume that the discrete parameter relaxed control $m_n(\cdot)$ depends on $\{\xi_{j-1}, x_j, j \leq n\}$ only. For any admissible relaxed control $m(\cdot)$ for (3.1), define the infinitesimal operator L^m by (which implicitly defines $\bar{b}(\cdot)$ and $\sigma(\cdot)$)

$$(5.2) \quad \begin{aligned} L^m f(x) &= f'_x(x) \int b(x, \alpha) m_t(d\alpha) + \frac{1}{2} \sum_{-\infty}^{\infty} E[f'_x(x) g(x, \xi_n)]'_x g(x, \xi_0) \\ &\equiv f'_x(x) \int \bar{b}(x, \alpha) m_t(d\alpha) + \frac{1}{2} \sum_{i,j} f_{x_i x_j}(x) a_{ij}(x). \end{aligned}$$

The discrete parameter case can easily be put into the framework of the last section. The optimal policy for the discrete parameter case would not usually be “relaxed,” but it is convenient to represent it as a relaxed control, since the limit controls might be relaxed. Define $x^{\varepsilon}(\cdot)$ by $x^{\varepsilon}(t) = x_n^{\varepsilon}$ on $[n\varepsilon, (n+1)\varepsilon)$, and define $m(\cdot)$ by

$$(5.3) \quad m(A \times [0, t]) = \varepsilon \sum_{n=0}^{[t/\varepsilon]-1} m_n(A) + \varepsilon(t - \varepsilon[t/\varepsilon])m_{[t/\varepsilon]}(A).$$

Let $\delta_{\varepsilon} \rightarrow 0$ and let $\hat{m}^{\varepsilon}(\cdot)$ be a δ_{ε} -optimal control for (5.1).

THEOREM 7. *Under the conditions of this section, Theorems 5 and 6 hold for the discrete parameter case.*

Remark. The proof is nearly identical to that of Theorems 5 and 6. One uses the discrete parameter versions (in [3]) of the theorems which were cited to that reference and the definition of $\hat{A}^{\varepsilon}f(n\varepsilon)$ and E_n^{ε} given in § 2.

6. Average cost per unit time. In this section, $(x^{\varepsilon}(\cdot), \xi^{\varepsilon}(\cdot))$ will be a Markov–Feller process with a stationary transition function when the control is of the feedback form $u(x, \xi)$, and $\xi(\cdot)$ is a Markov–Feller process. Let PM denote the class of U -valued functions of x for which (1.3) has a unique (weak sense) solution for each initial condition, and let PM^{ε} denote the class of U -valued continuous functions of (x, ξ) for which the corresponding $(x^{\varepsilon}(\cdot), \xi^{\varepsilon}(\cdot))$ is a Markov–Feller process (e.g., PM^{ε} includes all U -valued locally Lipschitz continuous functions). We work with (6.1), the same system dealt with in the previous section.

$$(6.1) \quad \dot{x}^{\varepsilon} = b(x^{\varepsilon}, u) + \tilde{b}(x^{\varepsilon}, \xi^{\varepsilon}) + g(x^{\varepsilon}, \xi^{\varepsilon})/\varepsilon.$$

Let SR denote the class of stationary admissible relaxed controls for (3.1) such that for each $m(\cdot) \in SR$, there is a process $x(m, \cdot)$ where the pair $(x(m, \cdot), m(\cdot))$ is stationary, and define SR^{ε} analogously for (6.1). When writing $\inf_{m \in SR} F(x(\cdot))$ for some function $F(\cdot)$, we infimize the functional values over these *stationary pairs* $(x(m, \cdot), m(\cdot))$.

The cost function (for a relaxed admissible control) is

$$\lim_{T \rightarrow \infty} \frac{1}{T} \int_0^T \int Ek(x^{\varepsilon}(t), \alpha) m_t(d\alpha) dt \equiv \gamma^{\varepsilon}(m)$$

and, for a feedback control,

$$\overline{\lim}_T \frac{1}{T} \int_0^T Ek(x^\varepsilon(t), u(x^\varepsilon(t), \xi^\varepsilon(t))) dt \equiv \gamma^\varepsilon(u).$$

We define the costs $\gamma(u)$ and $\gamma(m)$ for the controlled diffusion $x(\cdot)$ in the analogous way.

It is convenient to start our analysis with some additional assumptions. They will be discussed and sufficient conditions given for them in the next section.

Conditions C1-C4 hold in very many cases of interest. C1 and C3 are basically uniform (in the control) recurrence conditions. They certainly hold if the $x^\varepsilon(t)$ are confined to a compact set. But, more generally, if the system has a stability property for large $|x|$, then it can often be exploited to get C1 and C3. See § 7.3. Also, a nearly optimal stabilizing control for (1.3) is often a stabilizing control for (6.1).

Condition C1. There is $\varepsilon_0 > 0$ such that for each $\delta > 0$, there are continuous δ -optimal controls $u^{\varepsilon, \delta}(\cdot, \cdot) \in PM^\varepsilon$ such that $\{x^\varepsilon(u^{\varepsilon, \delta}, t), t < \infty, \varepsilon \leq \varepsilon_0\}$ is tight in R^r .

Condition C2. For each $\delta > 0$, there is a continuous δ -optimal control $\bar{u}^\delta(\cdot)$ in PM for (1.3) for which (1.3) has a unique invariant measure $\mu^\delta(\cdot)$, and such that $\bar{u}^\delta(\cdot) \in PM^\varepsilon$ for small ε .

Condition C3. For the $\bar{u}^\delta(\cdot)$ in C2, $\{x^\varepsilon(\bar{u}^\delta, t), t < \infty, \varepsilon > 0\}$ is tight in R^r .

Condition C4.

$$\inf_{u \in PM} \gamma(u) = \inf_{m \in SR} \gamma(m).$$

Theorem 8 says that if $\bar{u}^\delta(\cdot)$ is a δ -optimal control for the diffusion, then its use with the $x^\varepsilon(\cdot)$ gives a nearly (3δ) -optimal result for small ε .

THEOREM 8. Assume A1-A3 and C1-C4. Then for each $\delta > 0$, and small ε ,

$$(6.2) \quad \gamma^\varepsilon(\bar{u}^\delta) \leq \inf_{u \in PM^\varepsilon} \gamma^\varepsilon(u) + 3\delta.$$

Proof. Fix $\delta > 0$. $\bar{u}^\delta(\cdot)$ will be the function defined in C2, and $u^{\varepsilon, \delta}(\cdot)$ will be the function defined in C1. Let $P^{\varepsilon, \delta}(x, \xi, t, \cdot)$ denote the transition function for the Markov-Feller process $(x^\varepsilon(\cdot), \xi^\varepsilon(\cdot))$, under the control $u^{\varepsilon, \delta}(\cdot)$. Define the measures

$$P_T^{\varepsilon, \delta}(\cdot) = \frac{1}{T} E \int_0^T P^{\varepsilon, \delta}(x^\varepsilon(0), \xi^\varepsilon(0), t, \cdot) dt,$$

where the average E is over the possibly random initial condition $(x^\varepsilon(0), \xi^\varepsilon(0))$. Then

$$(6.3) \quad \gamma^\varepsilon(u^{\varepsilon, \delta}) = \overline{\lim}_T \int P_T^{\varepsilon, \delta}(dx \times d\xi) k(x, u^{\varepsilon, \delta}(x, \xi)).$$

Let $\xi^\varepsilon(t)$ take values in R^k , and let $M(0)$ denote the set of probability measures on R^{r+k} with the weak topology. By (C1), the set of $M(0)$ -valued measures $\{P_T^{\varepsilon, \delta}(\cdot), T < \infty\}$ is in a compact set in $M(0)$. It follows from Benes [11] that the limit of any weakly convergent (in the topology of $M(0)$) subsequence is an invariant measure for $(x^\varepsilon(\cdot), \xi^\varepsilon(\cdot))$, with the control $u^{\varepsilon, \delta}(\cdot)$ used.

Let $T_n \rightarrow \infty$ be a sequence such that it yields the $\overline{\lim}_T$ in (6.3) and also $P_{T_n}^{\varepsilon, \delta}(\cdot)$ converges weakly to an invariant measure $\mu^{\varepsilon, \delta}(\cdot)$ for $(x^\varepsilon(\cdot), \xi^\varepsilon(\cdot))$. Thus

$$\gamma^\varepsilon(u^{\varepsilon, \delta}) = \int k(x, u^{\varepsilon, \delta}(x, \xi)) \mu^{\varepsilon, \delta}(dx \times d\xi).$$

Let $(\hat{x}^\varepsilon(\cdot), \hat{\xi}^\varepsilon(\cdot))$ denote a stationary process corresponding to the invariant measure $\mu^{\varepsilon, \delta}(\cdot)$.

Write the control $u^{\varepsilon, \delta}(\cdot)$ for $(\hat{x}^\varepsilon(\cdot), \hat{\xi}^\varepsilon(\cdot))$ in the form of a relaxed control, which we call $m^{\varepsilon, \delta}(\cdot)$, with derivative $m_t^{\varepsilon, \delta}(\cdot)$. Let $m^{\varepsilon, \delta}$ denote the measure valued process which is the time derivative of $m^{\varepsilon, \delta}(\cdot \times [0, t])$; i.e., the process with value $m_t^{\varepsilon, \delta}(\cdot)$ at time t . Then the pair (state, relaxed control derivative) of processes $(\hat{x}^\varepsilon(\cdot), m^{\varepsilon, \delta})$ is stationary. Alternatively, for any sequence $\{t_i\}$ and set of increasing numbers $\{s_i\}$, the distributions of $\{\hat{x}^\varepsilon(t + t_i), m^{\varepsilon, \delta}(\cdot \times [s_j + t, s_{j+1} + t]), i, j\}$ do not depend on t .

By the stationarity, we can write

$$(6.4) \quad \gamma^\varepsilon(u^{\varepsilon, \delta}) = E \int_0^1 dt \int k(\hat{x}^\varepsilon(t), \alpha) m_t^{\varepsilon, \delta}(d\alpha).$$

By C1, the collection of invariant measures $\{\mu^{\varepsilon, \delta}(\cdot), \varepsilon > 0\}$ lies in a compact set in $M(0)$. Thus, by Theorem 5, $\{\hat{x}^\varepsilon(\cdot), m^{\varepsilon, \delta}(\cdot)\}$ is tight in $D^r[0, \infty) \times M(\infty)$. Let ε index a weakly convergent subsequence with limit $(\hat{x}^\delta(\cdot), m^\delta(\cdot))$. The limit is of the form (4.2), with the admissible $m^\delta(\cdot)$ replacing the $\hat{m}(\cdot)$ there. Let m^δ denote the measured-valued process which is the time derivative of $m^\delta(\cdot \times [0, t])$ (with value $m_t^\delta(\cdot)$ at time t). By the stationarity of $(x^\varepsilon(\cdot), m^{\varepsilon, \delta})$, the limit pair (state, relaxed control derivative) $(\hat{x}^\delta(\cdot), m^\delta)$ is also stationary, and by the weak convergence

$$(6.5) \quad \gamma^\varepsilon(u^{\varepsilon, \delta}) \rightarrow E \int_0^1 dt \int k(\hat{x}^\delta(t), \alpha) m_t^\delta(d\alpha).$$

Owing to the stationarity of $(\hat{x}^\delta(\cdot), m^\delta)$, the right side of (6.5) equals

$$(6.6) \quad \gamma(m^\delta) = \lim \frac{1}{T} E \int_0^T dt \int k(\hat{x}^\delta(t), \alpha) m_t^\delta(d\alpha).$$

We now apply $\bar{u}^\delta(\cdot)$ to $(x^\varepsilon(\cdot), \xi^\varepsilon(\cdot))$. Define $\tilde{P}_T^{\varepsilon, \delta}(\cdot)$ as $P_T^{\varepsilon, \delta}(\cdot)$ was defined, but with $(x^\varepsilon(\bar{u}^\delta(\cdot), \cdot), \xi^\varepsilon(\cdot))$ used. Choose $T_n \rightarrow \infty$ such that $\tilde{P}_{T_n}^{\varepsilon, \delta}(\cdot) \Rightarrow \tilde{\mu}^{\varepsilon, \delta}(\cdot)$, an invariant measure for $(x^\varepsilon(\bar{u}^\delta(\cdot), \cdot), \xi^\varepsilon(\cdot))$, and such that

$$\gamma^\varepsilon(\bar{u}^\delta) = \lim_n \int \tilde{P}_{T_n}^{\varepsilon, \delta}(dx \times d\xi) k(x, \bar{u}^\delta(x)).$$

Let $(\tilde{x}^\varepsilon(\cdot), \tilde{\xi}^\varepsilon(\cdot))$ denote the stationary process corresponding to the invariant measure $\tilde{\mu}^{\varepsilon, \delta}(\cdot)$ and control $\bar{u}^\delta(\cdot)$.

By C3, $\{\tilde{\mu}^{\varepsilon, \delta}(\cdot), \varepsilon > 0\}$ lies in a compact set in $M(0)$. Then, by Theorem 5, $\{\tilde{x}^\varepsilon(\cdot)\}$ is tight in $D^r[0, \infty)$. Let ε index a weakly convergent subsequence with limit $\tilde{x}^\delta(\cdot)$, and control $\bar{u}^\delta(\cdot)$. Then $\tilde{x}^\delta(\cdot)$ is stationary and is, in fact, the unique stationary process of the form (1.3) corresponding to the control $\bar{u}^\delta(\cdot)$. We have, by Theorem 5,

$$(6.7) \quad \gamma^\varepsilon(\bar{u}^\delta) = E \int_0^1 k(\tilde{x}^\varepsilon(t), \bar{u}^\delta(\tilde{x}^\varepsilon(t))) dt \rightarrow E \int_0^1 k(\tilde{x}^\delta(t), \bar{u}^\delta(\tilde{x}^\delta(t))) dt = \gamma(\bar{u}^\delta).$$

Also by the definition of $\bar{u}^\delta(\cdot)$ and C4,

$$(6.8) \quad \begin{aligned} \gamma(\bar{u}^\delta) &\leq \inf_{u \in PM} \gamma(u) + \delta, \\ \inf_{u \in PM} \gamma(u) &= \inf_{m \in SR} \gamma(m) \leq \gamma(m^\delta). \end{aligned}$$

The theorem follows from inequalities (6.8) and the convergence in (6.5), (6.7). QED

7. On Conditions C1–C4.

7.1. On Condition C4. Let there be an optimal (average cost per unit time) policy $\bar{u}(\cdot)$ in PM for (1.3) and such that the associated diffusion $\bar{x}(\cdot)$ has a unique invariant

measure which we denote by $\mu^{\bar{u}}(\cdot)$. Let the potential

$$C(x) = \int_0^\infty E_x[k(\bar{x}(s), \bar{u}(\bar{x}(s))) - \bar{\gamma}] ds$$

and constant $\bar{\gamma}$ satisfy the Bellman equation

$$(7.1) \quad \bar{\gamma} = \min_{u \in U} [L^u C(x) + k(x, u)].$$

See [12] for one set of conditions guaranteeing this. Let $m(\cdot) \in SR$, with the associated stationary process $x(m, \cdot) \equiv x^m(\cdot)$ and stationary measure $\mu^m(\cdot)$, where $x^m(\cdot)$ satisfies (4.2) for $\hat{m}(\cdot) = m(\cdot)$. Suppose that for any such $m(\cdot)$ with finite $\gamma(m)$,

$$(7.2) \quad \int |C(x)| \mu^m(dx) < \infty.$$

Then (7.1) implies that for any $T < \infty$

$$\bar{\gamma} T \leq EC(x^m(T)) - EC(x^m(0)) + E \int_0^T \int k(x^m(t), \alpha) m_t(d\alpha) dt.$$

Then, by the stationarity of $x^m(\cdot)$, $\bar{\gamma} \leq \gamma(m)$, and C4 holds. A sufficient condition for (7.2) will be given in § 7.3 below.

7.2. On Condition C2. We use results from [13], where the system $\dot{x} = \bar{b}(x, u)$ was assumed to have a stability property, uniformly in $u(\cdot) \in PM$. Write $\bar{b}(x, u) = B(x) + \hat{B}(x, u)$, where $B(\cdot)$ and $\hat{B}(\cdot)$ satisfy the conditions on $b(\cdot)$ in A2, and $\hat{B}(\cdot)$ and $\sigma(\cdot)$ are bounded, $k(\cdot, \cdot)$ is bounded and continuous, and $\{a_{ij}(x)\}$ is uniformly positive definite and satisfies A3. The model is such that the stabilizing effects of $B(\cdot)$ overpower the effects of $\hat{B}(x, u)$ for large $|x|$. This, together with the positive definiteness, will essentially guarantee C2. To quantify the stability property for large $|x|$, let there be a twice continuously differentiable function $V(\cdot)$ such that $0 \leq V(x) \rightarrow \infty$ as $|x| \rightarrow \infty$ and, for some compact set K and $\beta > 0$, $L^u V(x) \leq -\beta$, for $x \notin K$ and all $u(\cdot) \in PM$. (L^u is the differential generator of (1.3).) Let there be $c > 0$, $\alpha > 0$, $q(x) \geq 0$ such that $L^u V^2(x) \leq c - q(x)$, where $\inf_x q(x)/V(x) \geq \alpha$. Typically $V(\cdot)$ would be a Lyapunov function for the system $\dot{x} = B(x)$; e.g., if $B(x) = Ax$ where A is stable and for $Q > 0$, P can be defined by $A'P + PA = -Q$, and we use the Lyapunov function $x'Px = V(x)$. Note that our c and $V(x)$ are called c_2 and $W_1(x)$ in [13].

Under the above conditions, Theorems 3.1, 4.2, 4.3 and the proof of [13, Thm. 4.4] imply the following facts: To any $u(\cdot) \in PM$, there is a unique invariant measure $\mu^u(\cdot)$ for (1.3) and $\{\mu^u(\cdot), u(\cdot) \in PM\}$ is in a compact set in $M(0)$; let $u^\delta(\cdot)$ be a $\delta/2$ -optimal control in PM , smooth or not, and let

$$(7.3) \quad u^n(x) \rightarrow u^\delta(x) \text{ in } L_1(R^r), \quad u^n(\cdot) \in PM.$$

Then for each Borel set A , $\mu^{u^n}(A) \rightarrow \mu^{u^\delta}(A)$ and

$$(7.4) \quad \int k(x, u^n(x)) \mu^{u^n}(dx) \rightarrow \int k(x, u^\delta(x)) \mu^{u^\delta}(dx).$$

These facts imply that for any given $\delta/2$ -optimal $u^\delta(\cdot)$, there is a locally Lipschitz continuous $\bar{u}^\delta(\cdot)$ such that

$$\gamma(\bar{u}^\delta) - \gamma(u^\delta) \leq \delta/2.$$

Reference [13] uses a convexity condition (A3 there) on the set $\{\bar{b}(x, U), k(x, U)\}$ and on U . But, this convexity condition was used only to prove the existence of an optimal control. The smooth $\delta/2$ -optimal control always exists.

7.3. On the assumption (7.2). Again, we use results of [13]. Let $C(\cdot)$ satisfy (7.1) and assume the conditions of § 7.2. Then [13, proof of Lemma 5.1],

$$|C(x)| \leq K(1 + V(x))$$

for some $K < \infty$ (our $C(x)$ is called $V^u(x)$ in [13]). Adapting the proof of [13, Lemma 5.1] to our “relaxed” control case and using the c and α of § 7.2, we get for *any* $M < \infty$ and relaxed control $m(\cdot)$,

$$c \geq \lim_{t \rightarrow \infty} \int_0^t \alpha E \min [M, V(x^m(s))] dt.$$

By the stationarity, the integral equals $\alpha E \min [M, V(x^m(0))]$. Since M is arbitrary and c does not depend on $m(\cdot)$, (7.2) holds.

7.4. On Conditions C1, C3. Under a suitable stability condition on the limit system $x(\cdot)$, both C1 and C3 can be shown via a perturbed Lyapunov function method. In particular, we use some of the results of [3, Chap. 6.6] and [14]. We use the form $b(x, u) = B(x) + \hat{B}(x, u)$ and

$$(7.5) \quad \dot{x}^\varepsilon = B(x) + \hat{B}(x, u) + \tilde{b}(x, \xi^\varepsilon) + g(x, \xi^\varepsilon)/\varepsilon$$

and A2, A3, A1(a). Assume that $B(\cdot)$ and $\hat{B}(\cdot)$ satisfy the conditions on $b(\cdot)$ in A2. Analogous results can be obtained under A1(b), via the method in [3, Chap. 6.8]. We require the existence of a Lyapunov function $V(\cdot)$ satisfying certain inequalities. In applications, the assumptions are essentially equivalent to $B(\cdot)$ strongly dominating the effects of the other terms for large $|x|$.

We begin with an adaptation of a perturbed Lyapunov function method of [14], but with a simpler perturbation. Let $V(\cdot)$ be a twice continuously differentiable nonnegative function such that $V(x) \rightarrow \infty$ as $|x| \rightarrow \infty$ and D1–D4 hold. The K below are constants.

Condition D1. There are $\alpha > 0$, $c < \infty$, such that

$$V'_x(x)B(x) \leq -\alpha V(x) + c \quad \text{and} \quad |V'_x(x)\hat{B}(x, u)|/V(x) \rightarrow 0 \text{ as } |x| \rightarrow \infty;$$

Condition D2. $|V'_x(x)g(x, \xi)| + |V'_x(x)\tilde{b}(x, \xi)| \leq K(1 + V(x))$;

Condition D3. $|(V'_x(x)q(x))'_x p(x)| \leq K(1 + V(x))$, for the pairs

$$q(\cdot) = \tilde{b}(\cdot), p(\cdot) = B(\cdot), \hat{B}(\cdot), \tilde{b}(\cdot) \text{ and } g(\cdot) \quad \text{and}$$

$$q(\cdot) = g(\cdot), p(\cdot) = B(\cdot), \hat{B}(\cdot), \tilde{b}(\cdot);$$

Condition D4. $||V'_x(x)g(x, \xi)|'_x g(x, \xi)|/V(x) \rightarrow 0$ as $|x| \rightarrow \infty$.

Define $V_1^\varepsilon(t) = V_1^\varepsilon(x^\varepsilon(t), t)$, where

$$(7.6) \quad V_1^\varepsilon(x, t) = \int_t^\infty V'_x(x)E_t^\varepsilon \tilde{b}(x, \xi^\varepsilon(s)) ds + \frac{1}{\varepsilon} \int_t^\infty V'_x(x)E_t^\varepsilon g(x, \xi^\varepsilon(s)) ds.$$

By a change of scale $s/\varepsilon^2 \rightarrow s$ and A1(a), D2, we get that the first term is $O(\varepsilon^2)[1 + V(x)]$ and the second is $O(\varepsilon)[1 + V(x)]$. Define the perturbed Lyapunov function $V^\varepsilon(t) = V(x^\varepsilon(t)) + V_1^\varepsilon(t)$. Then (write for $x^\varepsilon(t)$ and \dot{x}^ε for $\dot{x}^\varepsilon(t)$, where convenient)

$$\hat{A}^\varepsilon V(x) = V'_x(x)[B(x) + \hat{B}(x, u) + \tilde{b}(x, \xi^\varepsilon(t)) + g(x, \xi^\varepsilon(t))/\varepsilon],$$

$$\hat{A}^\varepsilon V_1^\varepsilon(x, t) = -V'_x(x)\tilde{b}(x, \xi^\varepsilon(t)) - \frac{1}{\varepsilon} V'_x(x)g(x, \xi^\varepsilon(t))$$

$$+ \int_t^\infty ds [V'_x(x)E_t^\varepsilon \tilde{b}(x, \xi^\varepsilon(s))]'_x \dot{x}^\varepsilon$$

$$+\frac{1}{\varepsilon} \int_t^\infty ds [V'_x(x) E_t^\varepsilon g(x, \xi^\varepsilon(s))]'_x \dot{x}^\varepsilon.$$

By using the scale change $s/\varepsilon^2 \rightarrow s$, A1(a) and D1 to D4, we get that there is a function $h(x) \geq 0$ such that $h(x)/V(x) \rightarrow 0$ as $|x| \rightarrow \infty$ and such that

$$(7.7) \quad \hat{A}^\varepsilon V^\varepsilon(t) \leq -\alpha V(x^\varepsilon(t)) + h(x^\varepsilon(t)).$$

By the bound on $V_1^\varepsilon(x, t)$ below (7.6), we can write (for small $\varepsilon > 0$)

$$(7.8) \quad \hat{A}^\varepsilon V^\varepsilon(t) \leq -\frac{\alpha}{2} V^\varepsilon(x^\varepsilon(t)) + c_1,$$

for some $c_1 < \infty$. Inequality (7.8) yields, for some $c_2 < \infty$,

$$(7.9) \quad EV^\varepsilon(t) \leq e^{-\alpha t/2} EV^\varepsilon(0) + c_2.$$

Now use the bound on $V^\varepsilon(x, 0)$ obtained from the estimates below (7.6) to get that (for some $\varepsilon_0 > 0$)

$$\sup_{\varepsilon_0 \leq \varepsilon, t} EV(x^\varepsilon(t)) < \infty,$$

which yields C1 and C3.

By using the method and conditions in [3, Chap. 6.8], the conditions D1–D4 can be weakened. In particular, $V'_x(x)B(x) \leq -\alpha V(x) + c$ can be replaced by the condition that $V'_x(x)B(x) \leq -\alpha < 0$ for large $|x|$, and some $\alpha > 0$.

8. Extensions. Extensions of the results in §§ 4 to 6 to all the standard control problem formulations are quite possible. Here, we mention only a few possibilities.

8.1. Stopping times. Let G be a bounded open set with a piecewise differentiable boundary, and define

$$R^\varepsilon(m) = E \int_0^{\tau^\varepsilon(m)} ds \int k(x^\varepsilon(s), \alpha) m_s(d\alpha),$$

$$\tau^\varepsilon(m) = \inf \{t: x^\varepsilon(t) \notin G\},$$

where $x^\varepsilon(\cdot)$ is the solution to (4.1) which corresponds to m . Define $R(m)$, the cost for (3.1) in a similar way, with $\tau(m) = \inf \{t: x(t) \notin G\}$.

In extending Theorem 5 to this case, only two problems arise. First, is $\sup_\varepsilon E_x \tau^\varepsilon(m^\varepsilon) < \infty$ for the various sequences $\{m^\varepsilon(\cdot)\}$ which are used? Second, if $(x^\varepsilon(\cdot), m^\varepsilon(\cdot)) \Rightarrow (x(\cdot), m(\cdot))$, do the exit times also converge? The answers are affirmative under broad conditions, certainly if $\{a_{ij}(x)\}$ is uniformly positive definite in G . We discuss the questions in the simple case where $\xi^\varepsilon(\cdot)$ is Markov and bounded.

Suppose that there are $\delta > 0$ and $\rho > 0$ such that

$$(8.1) \quad \inf_{\substack{x \in G \\ m \in RC}} P_x \{x(m, t) \notin N_\delta(G), \text{ some } t \leq T\} \geq \rho,$$

where $N_\delta(G)$ is a δ -neighborhood of G and P_x denotes the probability given the initial condition x . Then it follows that there is a $\rho_1 > 0$ such that for any sequence of $m^\varepsilon(\cdot) \in RC^\varepsilon$

$$(8.2) \quad \liminf_{\varepsilon} \inf_{\xi, x \in G} P_{x, \xi} \{x^\varepsilon(m^\varepsilon, t) \notin G, \text{ some } t \leq 2T\} \geq \rho_1.$$

where $P_{x, \xi}$ denotes the probability given the initial conditions x, ξ .

Suppose that (8.2) is false. Then there are $\varepsilon \rightarrow 0$, and (bounded) initial conditions $x_\varepsilon \in G$ and ξ_ε , such that

$$(8.3) \quad \lim_{\varepsilon} P_{x_\varepsilon, \xi_\varepsilon} \{x^\varepsilon(m^\varepsilon, t) \in G, \text{ some } t \leq 2T\} = 0.$$

There is a subsequence (indexed by ε) and $m(\cdot) \in RC$ such that $\{x^\varepsilon(m^\varepsilon, \cdot), m^\varepsilon(\cdot)\} \Rightarrow \{x(m, \cdot), m(\cdot)\}$. Then (8.3) is contradicted by (8.1). It follows from (8.2) that there is an $\varepsilon_0 > 0$ such that

$$\sup_{\varepsilon_0 \geq \varepsilon \geq 0} E_{x, \xi} \tau^\varepsilon(m) < \infty, \quad x \in G, \xi.$$

In the nondegenerate case, if $\{x^\varepsilon(m^\varepsilon, \cdot), m^\varepsilon(\cdot)\} \Rightarrow (x(m, \cdot), m(\cdot))$, then the exit times also converge. This follows from the weak convergence and the fact that $x(m, \cdot)$ crosses the boundary of G indefinitely often in $[\tau(m), \tau(m) + \Delta]$, for any $\Delta > 0$.

8.2. State dependent noise. The results of §§ 4 to 6 can be extended to the case where the evolution of $\xi^\varepsilon(\cdot)$ depends on $x^\varepsilon(\cdot)$ or $\{\xi_n^\varepsilon\}$ depends on $\{x_n^\varepsilon\}$. The technique is a combination of the control "representation" results of this paper, and the weak convergence methods of the state dependent noise or singular perturbations sections of [3]. The main problems concern, as before, tightness and the representation of the limit as a particular control problem.

One particular case in [3] concerns Markov $(x_n^\varepsilon, \xi_{n-1}^\varepsilon)$, where if x_n^ε is fixed at x , the $\{\xi_n^\varepsilon\}$ is a Markov process with a unique invariant measure (see e.g., [3, Chap. 5.8.3]). Systems such as (1.6), or the wide band-noise driven forms can also be treated if the $g(\cdot)$ there does not depend on u .

REFERENCES

- [1] P. BILLINGSLEY, *Convergence of Probability Measures*, John Wiley, New York, 1968.
- [2] T. G. KURTZ, *Approximation of Population Processes*, Vol. 36 in CBMS-NSF Regional Conf. Series in Appl. Math., Society for Industrial and Applied Mathematics, Philadelphia, 1981.
- [3] H. J. KUSHNER, *Approximation and Weak Convergence Methods for Random Processes; with Applications to Stochastic Systems Theory*, MIT Press, Cambridge, MA, 1984.
- [4] G. L. BLANKENSHIP AND G. C. PAPANICOLAOU, *Stability and control of stochastic systems with wide-band noise disturbances II*, Preprint, Electrical Engineering Dept., Univ. of Maryland, College Park, MD, 1978.
- [5] A. BENSOUSSAN AND G. L. BLANKENSHIP, *Singular perturbations in stochastic control*, Univ. of Maryland, College Park, MD, Electrical Engineering Dept. Report, April 1985.
- [6] T. G. KURTZ, *Semigroups of conditioned shifts and approximations of Markov processes*, Ann. Probab., 4 (1975), pp. 618-642.
- [7] W. H. FLEMING, *Generalized solutions in optimal stochastic control*, Proc. URI Conf. on Control, 1982, p. 147-165.
- [8] W. H. FLEMING AND M. NISIO, *On stochastic relaxed controls for partially observable diffusions*, Nagoya Math. J., 93 (1984), pp. 71-108.
- [9] H. J. KUSHNER, *Jump-diffusion approximations for ordinary differential equations with wideband random right hand sides*, this Journal, 17 (1979), pp. 729-744.
- [10] A. V. SKOROHOD, *Limit theorems for stochastic processes*, Theory Probab. Appl., 1 (1956), pp. 262-290.
- [11] V. BENES, *Finite regular invariant measures for Feller processes*, J. Appl. Probab., 5 (1968).
- [12] M. COX AND I. KARATZAS, *Stationary control of Brownian motion in several dimensions*, Adv. Appl. Probab., 18 (1985), pp. 531-561.
- [13] H. J. KUSHNER, *Optimality conditions for the average cost per unit time problem with a diffusion model*, this Journal, 16 (1978), pp. 330-346.
- [14] G. BLANKENSHIP AND B. C. PAPANICOLAOU, *Stability and control of stochastic systems with wide band noise disturbances*, SIAM J. Appl. Math., 34 (1978), pp. 437-476.

THE OBSERVATION SPACE AND REALIZATIONS OF FINITE VOLTERRA SERIES*

P. E. CROUCH† AND P. C. COLLINGWOOD‡

Abstract. In this paper we extend the results on the theory of realizations of finite Volterra series by exploiting the structural properties of the observation space. In general, two distinct minimal realizations of a finite Volterra series may be constructed, one based on the properties of the observation space and the other based on the properties of the Lie algebra. Both these realizations display the canonical structure found earlier for such systems. The results given here also yield information on the observation algebra generated by functions in the observation space, just as previous work gave information on the Lie algebra. As an application, the structure found here is applied to the finite-dimensional filtering problem for these systems.

Key words. realization theory, Volterra series, observation space, canonical form, polynomial form, filtering, algebra

AMS(MOS) subject classifications. Primary 93B10, 93B15, 93B17, 93B20, 93B27; secondary 93B25

1. Introduction. In the paper by Crouch [2], some fundamental properties of systems with finite Volterra series were worked out. In this paper, we present some further structure for these systems. The main aspect under consideration is the observation space, introduced for nonlinear systems, to deal with problems concerning observability. We restrict our definitions in this paper to linear analytic systems of the form

$$(1) \quad \begin{aligned} \dot{x} &= f(x) + \sum_{i=1}^m u_i g_i(x), & x \in M, \\ y_i &= h_i(x), & 1 \leq i \leq p, \quad x(0) = x_0, \quad f(x_0) = 0 \end{aligned}$$

where M is a real analytic manifold, f, g_1, \dots, g_m are analytic vector fields on M , and h_1, \dots, h_p are analytic functions on M ; however, most concepts are generalizable to smooth systems. The observation space denoted \mathcal{H} is the smallest vector space of functions on M which contains h_1, \dots, h_p and is closed under the Lie derivative by the vector fields f, g_1, \dots, g_m . In general, \mathcal{H} is infinite-dimensional. We denote the Lie derivative of a function h by a vector field X by $L_X h$ and define $L_X^k(h) = L_X(L_X^{k-1}(h))$, $L_X^1(h) = L_X(h)$. We let \mathcal{A} denote the algebra over \mathbb{R} generated by \mathcal{H} under the operations of pointwise multiplication, addition and scalar multiplication. \mathcal{A} is called the observation algebra. Let \mathcal{L} denote the Lie algebra generated by the vector fields f, g_1, \dots, g_m , and let $[X, Y]$ denote the Lie bracket of vector fields X and Y .

For many problems in nonlinear systems theory the object of primary importance has been the Lie algebra \mathcal{L} . However, as stated above for questions concerning observability, the vector space \mathcal{H} seems to be the correct geometric/algebraic object to consider, especially for analytic systems—see Hermann and Krener [6]. Further use of \mathcal{H} has been made in realization theory (see Hijab [7] and Fliess and Kupka [8]). In estimation theory, some use has been made of the algebra \mathcal{A} , Hijab [7]. Further

* Received by the editors December 8, 1983; accepted for publication (in revised form) January 15, 1985.

† Department of Electrical and Computer Engineering, Arizona State University, Tempe, Arizona 85287. The work of this author was partially supported by Science and Engineering Research Council grant GR/B/9116.7.

‡ Department of Applied Mathematics, University of Sheffield, Sheffield S10 2TN, England. The work of this author was partially supported by Science and Engineering Research Council grant GR/B/3116.3.

use has been made of certain subspaces of \mathcal{H} and distributions orthogonal to them, Gauthier and Bornard [13], Aeyels [14] and Nijmeijer [15], to obtain canonical representations of systems with specific observability properties.

The main aim of this paper is to exploit \mathcal{H} and \mathcal{A} to obtain further structural properties of systems with finite Volterra series. In § 2 we begin by making some comments on the definitions used when referring to properties of system (1). Then, because of their connection with later work, we review two results. The first is a little-recognized result of Fliess [16] concerning the local structure of the algebra \mathcal{A} . The second result gives necessary and sufficient conditions for a system (1) to have an input-output map represented by a finite Volterra series. This result sharpens that obtained by Hijab [7] and Fliess and Kupka [8] for bilinear systems.

In § 3 we recall the canonical form for realizations of finite Volterra series obtained in Crouch [2] using the machinery of graded vector spaces. In order to distinguish this canonical form from others introduced in the works cited above, and motivated by the polynomial nature of these realizations, we refer to this canonical form as the graded polynomial form, or g.p.f. We then define the graded controllable polynomial form, or g.c.p.f., and show that the minimal realizations of finite Volterra series obtained in Crouch [2] are indeed in g.c.p.f. A system in g.c.p.f. is defined with the aid of a finite set of integers, such that any two minimal realizations of a finite Volterra series, both of which are in g.c.p.f., have the same set of integers. These integer invariants coincide with those introduced in Crouch [2].

In § 4 the graded observable polynomial form, or g.o.p.f., is defined. It is shown, by utilizing the structure of \mathcal{H} , that minimal realizations of finite Volterra series can be constructed which are in g.o.p.f. This canonical form differs substantially from the one appearing in Nijmeijer [15]. The g.o.p.f. is defined with the aid of another set of integer invariants which are not related to the observability indices introduced in Nijmeijer [15]. Even for minimal realizations of the same finite Volterra series, the two sets of integer invariants defined here need not coincide, but they do coincide for systems with finite Volterra series consisting of a single term.

It is easily verified from the structure of the g.o.p.f. that the observation algebra \mathcal{A} is a polynomial algebra. However, we show that this result is true for any minimal system in g.p.f. and in particular for a minimal system in g.c.p.f.

In § 5 we consider the estimation problem for a system with finite Volterra series driven by white noise, and use the results obtained in the previous sections to obtain a better description of the estimation algebra in this case.

Some of the constructions, using the observation space as detailed here, were obtained simultaneously by Kupka [12], but it is made clear there that the main aim is to give alternative proofs of some results in Crouch [2]. It is important to understand that we regard many of the results obtained here as complementary to the results in Crouch [2], and we therefore compare and contrast the two constructions where possible.

Since this paper was written, further work on the two canonical forms has been developed (Crouch and Collingwood [19] and Collingwood [20]). The structure of the observation algebra also helped motivate the new and innovative work by Bartosiewicz [21] and [22].

2. Preliminary properties of the observation space. In dealing with system (1) we need to use concepts of controllability, observability and minimality, which have been worked out for nonlinear systems (see, for example, Sussmann and Jurdjevic [9], and Hermann and Krener [6]). A minimal system is defined as one which is orbit minimal

and observable. For analytic systems, orbit minimality is equivalent to the controllability rank condition

$$\mathcal{L}(x) = \{X(x); X \in \mathcal{L}\} = T_x M, \quad x \in M,$$

where $T_x M$ is the tangent space to M at x . This is in turn equivalent to accessibility of the system; that is, the reachable set has nonempty interior in M from every initial state. Let \mathcal{S} be the ideal of \mathcal{L} generated by the vector fields

$$\begin{aligned} ad^k f(g_i), \quad k \geq 0, \quad 1 \leq i \leq m, \\ ad^k f(g) = [f, ad^{k-1} f(g)], \quad ad^1 f(g) = [f, g]. \end{aligned}$$

Since we assume $f(x_0) = 0$, we have $\mathcal{S}(x) = \mathcal{L}(x)$ for all $x \in M$. Thus the controllability rank condition is equivalent to $\mathcal{S}(x) = T_x M$ for all $x \in M$. This in turn implies strong accessibility of the system, that is, that the reachable set at time $T > 0$ has nonempty interior in M from every initial state. Thus, in our situation, we are dealing with strictly autonomous systems.

For the sake of simplicity, we shall refer a system which satisfies the controllability rank condition as controllable rather than by the more traditional terms accessible or weakly controllable. Similarly, we shall refer to a system which satisfies the observability rank condition

$$d\mathcal{H}(x) = \{dh(x); h \in \mathcal{H}\} = T_x^* M, \quad x \in M$$

and $T_x^* M$ is the cotangent space to M at x , as observable, rather than by the usual term weakly observable. In general, a system may satisfy the observability rank condition and fail to be observable in the usual sense. With the definitions above, a controllable and observable system need not be minimal in general. So when minimality is required this will be explicitly stated. As a matter of interest, it is proved in Crouch [2], that a controllable and observable system as defined here, which has a finite Volterra series, is indeed minimal. This mitigates the terminology used here to some extent.

Before we give the first result of this section, we quote the following lemma whose proof is a simple exercise in differentiation (see Collingwood [20]).

LEMMA 1. For $i = 1, 2$ let

$$\begin{aligned} \dot{x}_i &= f^i(x_i) + \sum_{j=1}^m u_j g_j^i(x_i), \quad x_i \in \mathbb{R}^{n_i}, \quad x_i(0) = \bar{x}_i, \\ \sum_i \quad y_k &= h_k^i(x_i), \quad 1 \leq k \leq p \end{aligned}$$

be two systems with the same input-output map. Let $\mathcal{H}_i, i = 1, 2$ be the observation spaces, and $\mathcal{L}_i, i = 1, 2$ the Lie algebras for systems $\sum_i, i = 1, 2$, respectively, and let $t \rightarrow x_i^u(t)$ $i = 1, 2$ denote the solutions of the respective equations subject to the same input functions $u_j(t)$. Then

(i) If \sum_2 is controllable, there is a unique linear surjection $\beta: \mathcal{H}_1 \rightarrow \mathcal{H}_2$ satisfying

$$\beta(\phi)(x_2^u(t)) = \phi(x_1^u(t))$$

for all $\phi \in \mathcal{H}_1$, and trajectories x_i^u , when both sides are defined.

(ii) If \sum_2 is observable, there is a unique Lie algebra homomorphism $\lambda: \mathcal{L}_1 \rightarrow \mathcal{L}_2$, satisfying

$$\lambda(f^1) = f^2 \quad \text{and} \quad \lambda(g_i^1) = g_i^2, \quad 1 \leq i \leq m.$$

We give a proof of the following result, originally obtained in Fliess [16], because of the more geometric flavor and its intimate relation with subsequent results, especially Theorem 4.

THEOREM 1 (Fliess [16]). *Given a controllable and observable system (1), then about any state $x_0 \in \mathbb{R}^n$, there exists a coordinate chart (U, ϕ) , $\phi(x_0) = 0$, $\phi: U \rightarrow V \subset \mathbb{R}^n$, such that in the resulting coordinates $z_1 \cdots z_n$ on V , the system's observation space contains all of the coordinate functions $z \rightarrow z_i$, $1 \leq i \leq n$.*

Proof. Since the system is observable, it satisfies the observability rank condition $d\mathcal{H}(x) = T_x^*M$, $x \in M$, and so there exists functions $\phi_1 \cdots \phi_n \in \mathcal{H}$, such that $d\phi_1(x) \cdots d\phi_n(x)$ span T_x^*M for each $x \in U$, a neighborhood of x_0 . It follows that the map $x \rightarrow \phi(x) = (\phi_1(x), \dots, \phi_n(x)) = (z_1 \cdots z_n) = z$ is a diffeomorphism from U onto some open subset $V \subset \mathbb{R}^n$. Define vector fields and functions on V by

$$\begin{aligned} F(z) &= \phi_* f(\phi^{-1}(z)), & G_i(z) &= \phi_* g_i(\phi^{-1}(z)), \\ H_i(z) &= h_i \circ \phi^{-1}(z). \end{aligned}$$

This defines a system on V

$$\begin{aligned} \dot{z} &= F(z) + \sum_{i=1}^m u_i G_i(z), & z &\in V, \quad z(0) = \phi(x_0), \\ (2) \quad y_i &= H_i(z), & 1 &\leq i \leq p, \end{aligned}$$

which has the same input-output map as system (1). Since both systems are clearly controllable on U and V , respectively, by Lemma 1 we have

$$h \circ \phi^{-1}(z(t)) = \beta(h)(z(t)) = h(x(t))$$

for $h \in \mathcal{H}$, where $t \rightarrow x(t)$ and $t \rightarrow z(t)$ are the solutions of systems (1) and (2), respectively, and defined where $z(t)$ is defined.

Denote by $\psi: V \rightarrow \mathbb{R}^n$ the analytic map $z \rightarrow (\phi_1 \circ \phi^{-1}(z), \dots, \phi_n \circ \phi^{-1}(z)) = (\beta(\phi_1)(z), \dots, \beta(\phi_n)(z))$. Now $\beta(\phi_i)(z(t)) = \phi_i(x(t))$ $1 \leq i \leq n$, so we have $\psi(z(t)) = z(t)$, when $z(t)$ is defined. Since system (2) is controllable on V , the reachable set of system (2) from $z = \phi(x_0)$ contains an open subset of V . Thus, ψ is the identity map on an open subset of V , and hence by analyticity it is the identity map on the whole of V . It follows that $\beta(\phi_i)(z) = z_i$ on V and hence $z \rightarrow z_i$ is in the observability space of system (2) for $1 \leq i \leq n$. \square

Note that if in the situation above the observation space also contains the constant functions then the algebra \mathcal{A} in the coordinates $z_1 \cdots z_n$, is simply $\mathbb{R}[z_1 \cdots z_n]$, the ring of polynomials in $z_1 \cdots z_n$. We show later that this happens globally for minimal realizations of finite Volterra series.

We now turn attention to the problem of establishing conditions under which an analytic system (1) has an input-output map described by a finite Volterra series. We use the technique of Hijab [7] and Fliess and Kupka [8], and although the result is not particularly innovative given the works above and Crouch [2], Fliess [17] and [18], we include its proof as an introduction to the structure of the observation space.

THEOREM 2. *Give a controllable system (1), then its input-output map has a finite Volterra series of length $N \geq 1$ if and only if both of the following conditions hold:*

- (i) \mathcal{H} is a finite-dimensional space;
- (ii) There exists a sequence of subspaces

$$\begin{aligned} \mathcal{H}^k &\subset \mathcal{H}, \quad 0 \leq k \leq N+1, \quad \mathcal{H}^{N+1} = \{0\}, \quad \mathcal{H}^0 = \mathcal{H}, \\ 0 &\subset \mathcal{H}^N \subset \mathcal{H}^{N-1} \subset \cdots \subset \mathcal{H}^1 \subset \mathcal{H}, \end{aligned}$$

such that

- (a) \mathcal{H}^N consists of the constant functions
- (b) $L_{g_i} \mathcal{H}^k \subset \mathcal{H}^{k+1}$, $L_f \mathcal{H}^k \subset \mathcal{H}^k$, $0 \leq k \leq N$, $1 \leq i \leq m$.

Proof. Necessity: If the system has a finite Volterra series then it is realizable by a bilinear system (Brockett [23]) in the form

$$\dot{z} = Az + \sum_{i=1}^m u_i(N_i z + b_i), \quad z(0) = 0, \quad z \in \mathbb{R}^J,$$

$$u_i = c'_i z \quad 1 \leq i \leq p.$$

Moreover, since the system (1) is controllable, we may suppose that there is an analytic map $\phi: M \rightarrow \mathbb{R}^J$, such that given any $h \in \mathcal{H}$, there exists an affine function \bar{h} on \mathbb{R}^J , such that $h = \bar{h} \circ \phi$. Since the space of affine functions on \mathbb{R}^J is finite-dimensional, we see that condition (i) is satisfied. As in Lemma 3.1 of Crouch [2] the last Volterra kernel $W_N(t, \sigma_1 \cdots \sigma_N)(x_0)$, viewed as a function $W_N(t, \sigma_1 \cdots \sigma_N)(x)$ is a constant for each $t, \sigma_1 \cdots \sigma_N$. Thus, the coefficients of the parameters $t, \sigma_1 \cdots \sigma_N$ and their powers, viewed as functions on M , are constant, and not all zero. These coefficients, which are elements of \mathcal{H} , are given by the following expressions, Fliess [18], Crouch [2]:

$$L_{ad}^{k_0}_{f^{(g_{i_1})}} L_{ad}^{k_1}_{f^{(g_{i_2})}} \cdots L_{ad}^{k_{N-1}}_{f^{(g_{i_N})}} L_f^{k_N}(h_j), \quad 1 \leq i_j \leq m, \quad 1 \leq j \leq p, \quad k_i \geq 0.$$

However, by repeated use of the formula

$$L_{ad(X)(Y)}(h) = L_X L_Y(h) - L_Y L_X(h),$$

we see that the linear span of the functions above is the same as the linear span of the functions

$$L_f^{k_0} L_{g_{i_1}} L_f^{k_1} L_{g_{i_2}} \cdots L_{g_{i_N}} L_f^{k_N}(h_j).$$

Let \mathcal{O}^r be the subspace of \mathcal{H} spanned by the functions

$$L_f^{k_0} L_{g_{i_1}} L_f^{k_1} L_{g_{i_2}} \cdots L_{g_{i_r}} L_f^{k_r}(h_j),$$

$k_i \geq 0$, $1 \leq j \leq p$, $1 \leq i_j \leq m$, and set, for $0 \leq r \leq N$,

$$\hat{\mathcal{H}}^r = \mathcal{O}^N + \mathcal{O}^{N-1} + \cdots + \mathcal{O}^r, \quad \hat{\mathcal{H}}^{N+1} = \{0\}.$$

By the arguments above $\mathcal{O}^N = \hat{\mathcal{H}}^N$ is spanned by the constant functions, and it is clear that

$$L_f \hat{\mathcal{H}}^r \subset \hat{\mathcal{H}}^r, \quad L_{g_i} \hat{\mathcal{H}}^r \subset \hat{\mathcal{H}}^{r+1}, \quad 1 \leq i \leq m, \quad 0 \leq r \leq N.$$

Thus condition (ii) of the theorem is satisfied with $\hat{\mathcal{H}}^r = \mathcal{H}^r$, $0 \leq r \leq N+1$.

Sufficiency: Select complementary subspaces \mathcal{O}^j such that $\mathcal{H}^{j+1} + \mathcal{O}^j = \mathcal{H}^j$, $0 \leq j \leq N$, $\mathcal{O}^N = \mathcal{H}^N$. Thus $L_f \mathcal{O}^j \subset \mathcal{H}^j$, $L_{g_i} \mathcal{O}^j \subset \mathcal{H}^{j+1}$. Let $\phi_1^r \cdots \phi_n^r$ be a basis of \mathcal{O}^r for $N \geq r \geq 0$, and let $z_r(t)$ denote the vector $(\phi_1^r(x(t)), \cdots, \phi_n^r(x(t)))$ where $x(t)$ is the solution of system (1). We obtain by differentiation

$$\frac{d}{dt} \phi_j^r(x(t)) = L_f \phi_j^r(x(t)) + \sum_{i=1}^m u_i L_{g_i} \phi_j^r(x(t)).$$

Thus, using the above properties, we obtain

$$(3) \quad \frac{d}{dt} z_r(t) = \sum_{j=r}^N A_j^r z_j(t) + \sum_{j=r+1}^N \sum_{i=1}^m u_i B_{ij}^r z_j(t)$$

for appropriate matrix coefficients A_j^r , B_{ij}^r with $B_{ij}^N = A_j^N = 0$. Also, since $h_i \in \mathcal{H}$ there exist vectors c_{ij} such that

$$(4) \quad h_i(x(t)) = \sum_{r=1}^N c'_{ir} z_r(t).$$

Clearly we may set $z_N(t) \equiv 1 = \phi^N$ the basis element of \mathcal{H}^N . Writing equations (3) and (4) in matrix form we obtain

$$\frac{d}{dt} \begin{bmatrix} z_N \\ z_{N-1} \\ z_{N-2} \\ \vdots \\ z_0 \end{bmatrix} = \begin{bmatrix} 0 & 0 & 0 & \cdots & 0 \\ A_N^{N-1} & A_{N-1}^{N-1} & 0 & \cdots & 0 \\ A_N^{N-2} & A_{N-1}^{N-2} & A_{N-2}^{N-2} & & 0 \\ \vdots & \vdots & & \ddots & \\ A_0^0 & & & & A_0^0 \end{bmatrix} \begin{bmatrix} z_N \\ z_{N-1} \\ z_{N-2} \\ \vdots \\ z_0 \end{bmatrix} + \sum_{i=1}^m u_i \begin{bmatrix} 0 & & & & \\ B_{iN}^{N-1} & 0 & & & \\ B_{iN}^{N-2} & B_{iN-1}^{N-2} & 0 & & \\ \vdots & \vdots & \ddots & \ddots & \\ B_{iN}^0 & & & B_{i1}^0 & 0 \end{bmatrix} \begin{bmatrix} z_N \\ z_{N-1} \\ z_{N-2} \\ \vdots \\ z_0 \end{bmatrix},$$

$$y_i(t) = \sum_{r=1}^N c'_{ir} z_r(t), \quad 1 \leq i \leq p.$$

This system evidently has a finite Volterra series of length N , for any initial state. \square

The bilinear realizations we obtain in the proof of this result are in general not minimal, and it is the structure of the minimal realizations that we consider next.

3. Graded polynomial forms for realizations of finite Volterra series. In order to define the canonical forms for systems with finite Volterra series, as constructed in Crouch [2], we must introduce some notation and results from the theory of graded vector spaces, and polynomial functions on them. All the results may be found in Goodman [4].

A graded vector space consists of a vector space \mathbb{R}^n , and a specified direct sum decomposition of \mathbb{R}^n

$$\mathbb{R}^n = \mathbb{R}^{n_1} \oplus \mathbb{R}^{n_2} \oplus \cdots \oplus \mathbb{R}^{n_N}$$

where $\sum_{i=1}^N n_i = n$. We represent a vector $x \in \mathbb{R}^n$ as the compound vector (x_1, x_2, \dots, x_N) with $x_{n_i} \in \mathbb{R}^{n_i}$. If it is necessary to refer to the components of a vector x_i , we shall denote them as x_i^j , $1 \leq j \leq n_i$.

The role of each subspace is differentiated by the introduction of a dilation $\delta_t: \mathbb{R}^n \rightarrow \mathbb{R}^n$

$$\delta_t(x_1 \cdots x_N) = (tx_1, \dots, t^N x_N).$$

We refer to N as the order of the graded vector space. We say that a polynomial function h on \mathbb{R}^n is homogeneous of degree k if $h \circ \delta_t(x) = t^k h(x)$ for each $x \in \mathbb{R}^n$. The vector space of all homogeneous polynomials of degree k will be denoted by Q^k . A vector field X on \mathbb{R}^n , with polynomial coefficients, is said to be homogeneous of degree m , $0 \leq m \leq N$, if for each $k \geq 0$ and each $h \in Q^k$, $L_X(h) \in Q^{k-m}$. We set $Q^k = 0$ for each $k < 0$. The vector space of all homogeneous vector fields of degree m will be denoted by P^m . We set

$$V^m = P^N \oplus P^{N-1} \oplus \cdots \oplus P^m, \quad C^m = Q^m \oplus Q^{m-1} \oplus \cdots \oplus Q^0.$$

A one form ω on \mathbb{R}^n , with polynomial coefficients, is said to be homogeneous of degree m , if for each k , $0 \leq k \leq N$ and each $X \in P^k$, $\omega(X) \in Q^{m-k}$. We may represent the space of all closed one forms on \mathbb{R}^n , homogeneous of degree m by dQ^m . Let Z^k be the space of constant vector fields in P^k , and W^k the space of constant one forms, homogeneous of degree k . Z^k is spanned by $\partial/\partial x_k^i$ and W^k is spanned by dx_k^i , $1 \leq i \leq n_k$. If $h \in C^m$ and $X \in V^k$ then $L_X(h) \in C^{m-k}$, which may be expressed in the form $dC^m(V^k) \subset C^{m-k}$.

The following sequence of facts are readily verifiable using the definitions above:

$$(5) \quad P^j = \bigoplus_{k=j}^N (Q^{k-j} \otimes Z^k), \quad 1 \leq j \leq N, \quad P^0 = \bigoplus_{k=1}^N (Q^k \otimes Z^k),$$

$$(6) \quad V^j = \bigoplus_{k=j}^N (C^{k-j} \otimes Z^k), \quad 1 \leq j \leq N, \quad V^0 = \bigoplus_{k=1}^N (C^k \otimes Z^k),$$

$$(7) \quad dQ^j \subset \bigoplus_{k=1}^j (Q^{j-k} \otimes W^k), \quad dC^j \subset \bigoplus_{k=1}^j (C^{j-k} \otimes W^k),$$

$$(8) \quad Q^k \otimes P^m \subset P^{m-k}, \quad Q^k \otimes Q^m \subset Q^{k+m},$$

$$(9) \quad [P^k, P^m] \subset P^{m+k}, \quad [V^k, V^m] \subset V^{m+k}.$$

Example 1. If $\mathbb{R}^3 = \mathbb{R} \oplus \mathbb{R}^2$ is a graded vector space of order $N=3$, $n_1=1$, $n_2=2$, $n_3=0$, then we may express $x \in \mathbb{R}^3$ as (x_1^1, x_2^1, x_2^2) . It follows that $x_2^2 + x_2^1 x_1^1 \in C^3$, while $x_2^2 \in Q^2$ and $x_2^1 x_1^1 \in Q^3$. $(x_1^1)^2 \partial / \partial x_1^1 + x_2^1 \partial / \partial x_2^1 \in P^0$ and $\partial / \partial x_1^1 \in P^1$.

Example 2. If $\mathbb{R}^3 = \mathbb{R} \oplus \mathbb{R} \oplus \mathbb{R}$ is a graded vector space of order $N=3$, $n_1=1$, $n_2=1$, $n_3=1$, then we may express $x \in \mathbb{R}^3$ as (x_1^1, x_2^1, x_3^1) . It follows that $x_3^1 \in Q^3$, $\partial / \partial x_1^1 + x_2^1 \partial / \partial x_3^1 \in P^1$ and $(x_1^1)^2 \partial / \partial x_2^1 + (x_2^1 + (x_1^1)^3) \partial / \partial x_3^1 \in V^0$, while $x_2^1 \partial / \partial x_3^1 \in P^1$ and $(x_1^1)^2 \partial / \partial x_2^1 + (x_1^1)^3 \partial / \partial x_3^1 \in P^0$.

Example 3. If $\mathbb{R}^2 = \mathbb{R} \oplus \mathbb{R}$ is a graded vector space of order $2=N$, $n_1=n_2=1$ we may express $x \in \mathbb{R}^2$ as $x = (x_1^1, x_2^1)$. Thus $x_2^1 \in Q^2$, $\partial / \partial x_1^1 \in P^1$ and $(x_1^1)^2 \partial / \partial x_2^1 \in P^0$.

Consider the following system on \mathbb{R}^n :

$$(10) \quad \begin{aligned} \dot{x} &= F(x) + \sum_{i=1}^m u_i G_i(x), \quad x \in \mathbb{R}^n, \quad x(0) = 0, \\ y_i &= H_i(x), \quad 1 \leq i \leq p. \end{aligned}$$

If \mathbb{R}^n is a graded vector space of (order N) we say that the system is in graded polynomial form (g.p.f.) if $F \in V^0$, $G_i \in V^1$, $1 \leq i \leq m$, and $H_i \in C^N$, $1 \leq i \leq p$, relative to the graded structure described above. With the aid of the identities (6) we may write a system (10) in g.p.f. using the coordinates (x_1, \dots, x_N) , as

$$(11) \quad \begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \\ \vdots \\ \dot{x}_N \end{bmatrix} = \begin{bmatrix} A_1 x_1 + a_1 \\ A_2 x_2 + a_2(x_1) \\ \vdots \\ A_N x_N + a_N(x_1 \cdots x_{N-1}) \end{bmatrix} + \sum_{i=1}^m u_i \begin{bmatrix} b_{i1} \\ b_{i2}(x_1) \\ \vdots \\ b_{iN}(x_1 \cdots x_{N-1}) \end{bmatrix} \begin{bmatrix} x_1(0) \\ x_2(0) \\ \vdots \\ x_N(0) \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix},$$

$$y_i = H_i(x_1 \cdots x_N).$$

Here $a_j(x_1 \cdots x_{j-1})$ represents a vector in \mathbb{R}^{n_j} with components in C^j , $b_{ij}(x_1 \cdots x_{j-1})$ represents a vector in \mathbb{R}^{n_j} with components in C^j , and $H_i(x_1 \cdots x_N)$ belong to C^N .

We claim that any system (10), in g.p.f., has an input-output map described by a finite Volterra series of length not greater than N . To show this, let $\mathcal{H}^i = \mathcal{H} \cap C^{N-i}$, $0 \leq i \leq N$ where \mathcal{H} is the observation space of the system. Since \mathcal{H} is comprised of polynomial functions in C^N , \mathcal{H} is a finite-dimensional vector space as required in (i) of Theorem 2. Since $\mathcal{F} \in V^0$, and $G_i \in V^1$, we see that the sequence of subspaces \bar{H}^i defined above satisfies the conditions (ii)(b) of Theorem 2. In case $\bar{\mathcal{H}}^N = \mathcal{H} \cap C^0$ is nonempty condition (ii)(a) is also satisfied, and the system has a Volterra series of length N . The length of the Volterra series, in the case where $\bar{\mathcal{H}}^N$ is empty, is equal to the smallest integer k such that $\bar{\mathcal{H}}^k$ contains the constant functions.

It follows that the graded polynomial form defined here is exclusively used to describe realizations of finite Volterra series. Note that in the definition relating to system (10) we have not required that $F(0)=0$, but this can be imposed very easily. Note also that there is no loss of generality in assuming that $x(0)=0$, since a change of coordinates ensuring this leaves the graded structure invariant.

We now consider the structure of the Lie algebra of a system in g.p.f. If \mathcal{R}^1 is the subspace of \mathcal{L} spanned by the vector fields $ad^k f(g_i)$, $k \geq 0$, $1 \leq i \leq m$, for a general system (1), and \mathcal{S} is the ideal of \mathcal{L} generated by \mathcal{R}^1 we make the following inductive definitions. $\mathcal{S}^1 = \mathcal{S}$, $\mathcal{S}^m = [\mathcal{S}^1, \mathcal{S}^{m-1}]$ and $\mathcal{R}^m = [\mathcal{R}^1, \mathcal{R}^m]$. In Theorem 3.2 of Crouch [2], it is shown that a minimal realization of a finite Volterra series of length N has an ideal \mathcal{S} , which is nilpotent, and $\mathcal{S}^{N+1} = \{0\}$.

It follows that $\mathcal{R}^{N+1} = \{0\}$ and for $1 \leq i \leq N$

$$\mathcal{S}^i = \mathcal{R}^N + \mathcal{R}^{N-1} + \dots + \mathcal{R}^i.$$

Assuming the system (10), defined on a graded vector space of order N , is in g.p.f. we may define the subspaces $\bar{\mathcal{S}}^i = \mathcal{S} \cap V^i$ of \mathcal{S} for $1 \leq i \leq N+1$. From the relation (9) we deduce that $\bar{\mathcal{S}}^i$ is in fact a sequence of ideals of \mathcal{S} such that $\bar{\mathcal{S}}^i / \bar{\mathcal{S}}^{i+1}$ is an abelian Lie algebra with $\bar{\mathcal{S}}^{N+1} = \{0\}$, and

$$\bar{\mathcal{S}}^N \subset \bar{\mathcal{S}}^{N-1} \subset \dots \subset \bar{\mathcal{S}}^2 \subset \bar{\mathcal{S}}^1.$$

It follows that \mathcal{S} is also a nilpotent Lie algebra (see Humphreys [24]), and in fact we have $\mathcal{S} = \mathcal{S}^1 = \bar{\mathcal{S}}^1 = \mathcal{S} \cap V^1$, $\mathcal{S}^i = (\bar{\mathcal{S}}^1)^i = (\mathcal{S} \cap V^1)^i \subset \mathcal{S}^i \cap V^i \subset \mathcal{S} \cap V^i = \bar{\mathcal{S}}^i$ for $1 \leq i \leq N$. Hence $\mathcal{S}^{N+1} \subset \bar{\mathcal{S}}^{N+1} = \{0\}$. Thus the Lie algebra of system (10) in g.p.f. has the same structure as the Lie algebra in a minimal realization of a finite Volterra series.

To state the main result of Crouch [2] we introduce another definition. We say a realization of a finite Volterra series of length N defined on a graded vector space \mathbb{R}^n of order N is in graded controllable polynomial form (g.c.p.f.) if it is in g.p.f. and

$$(12) \quad V^i(0) = \mathcal{S}^i(0), \quad 1 \leq i \leq N$$

where $V^i(0)$ and $\mathcal{S}^i(0)$ are the subspaces of $T_0\mathbb{R}^n = \mathbb{R}^n$ spanned by the corresponding Lie algebras of vector fields evaluated at $x=0$.

From this definition we see that the dimensions n_i of the subspaces \mathbb{R}^{n_i} in the graded vector space are related to the structure of the system by the relation

$$n_i = \text{Dim } \mathcal{S}^i(0) - \text{Dim } \mathcal{S}^{i+1}(0).$$

In particular, since the dimension of $\mathcal{S}^i(x)$, $x \in \mathcal{M}$ is the same for any minimal realization of a finite Volterra series; any two minimal realizations of the same finite Volterra series of length N both in g.c.p.f. must be defined on the same graded vector space $\mathbb{R}^n = \mathbb{R}^{n_1} \oplus \mathbb{R}^{n_2} \oplus \dots \oplus \mathbb{R}^{n_N}$. In particular the integers n_i , $1 \leq i \leq N$ are invariants, and coincide with those integer invariants introduced in Crouch [2].

Of course, the existence of minimal realizations of finite Volterra series in g.c.p.f. has not yet been established. This is, however, one of the main results in Crouch [2].

THEOREM 3 (Crouch [2]). *A finite Volterra series which has a realization in the form of system (1), where $f + \sum_{i=1}^m \alpha_i g_i$, $\alpha_i \in \mathbb{R}$, are complete vector fields, is also realizable by a minimal system in graded controllable polynomial form. In particular, the state space of a minimal realization is a Cartesian space \mathbb{R}^n for some $n > 0$.*

We now single out some special properties of realizations of finite Volterra series in g.p.f. which are useful in subsequent sections.

LEMMA 2. *If system (10) is a realization of a finite Volterra series of length N , which is in g.c.p.f. the Lie algebra \mathcal{S} contains vector fields of the form*

$$(13) \quad \frac{\partial}{\partial x_i^j} + \sum_{k>i}^m r_k^m(x_1 \dots x_{k-i}) \frac{\partial}{\partial x_k^m}, \quad 1 \leq i \leq N, \quad 1 \leq j \leq n_i$$

with $r_k^m \in C^{k-i}$.

Proof. Since the system is in g.c.p.f., $\mathcal{S}^i(0)/\mathcal{S}^{i+1}(0) = V^i(0)/V^{i+1}(0)$, $1 \leq i \leq N$, which by property (6) is isomorphic to \mathbb{R}^{n_i} , we may select vector fields $X_i^j \in \mathcal{S}^i$ such that their image in $\mathcal{S}^i(0)/\mathcal{S}^{i+1}(0)$ is $\partial/\partial x_i^j$ for $i \leq j \leq n_i$. Since $X_i^j \in \mathcal{S}^i \subset V^i$ it follows that it has the form given in expression (13), \square

Notice that the above result shows that we may replace the condition in (12) by the condition

$$V^i(x) = \mathcal{S}^i(x), \quad x \in \mathbb{R}^n, \quad 1 \leq i \leq N.$$

In particular, $T_x \mathbb{R}^n = \mathbb{R}^n = V^1(x) = \mathcal{S}(x)$ for $x \in \mathbb{R}^n$, so a system (10) in g.c.p.f. with $F(0) = 0$ is controllable.

LEMMA 3. *Let \sum_i , $i = 1, 2$ be two minimal realizations of the same input-output map, each in g.p.f. on a graded vector space \mathbb{R}^n of order N , and let $\Phi: \mathbb{R}^n \rightarrow \mathbb{R}^n$ be the state space isomorphism as guaranteed by Sussmann [10]. Then Φ is a polynomial mapping.*

Proof. We use the notation of Lemma 1 with $M^{n_i} = \mathbb{R}^n$, $i = 1, 2$. Thus $\Phi(x_1^u(t)) = x_2^u(t)$ for all inputs u and $t \geq 0$. Let \sum_1' be the system whose dynamics are those of \sum_1 with outputs $y_i = \Phi^i(x_1)$, $1 \leq i \leq n$, and let \sum_2' be the system whose dynamics are those of \sum_2 with outputs $y_i = x_2^i$, $1 \leq i \leq n$. Both \sum_1' and \sum_2' are therefore minimal realizations of the same input-output map. Thus, as shown in Crouch [2],

$$L_{X_1} L_{X_2} \cdots L_{X_k}(\phi) = 0$$

for all $X_i \in \mathcal{S}_2 \subset \mathcal{L}_2$, $\phi \in \mathcal{H}_2$, $k \geq N+1$, relative to \sum_2' . In particular, since \sum_1' and \sum_2' are also isomorphic,

$$L_{Y_1} L_{Y_2} \cdots L_{Y_k}(\psi) = 0$$

for all $Y_i \in \mathcal{S}_1 \subset \mathcal{L}_1$, $\psi \in \mathcal{H}_1$, $k \geq N+1$. We may now set $\psi = \Phi^i$, $1 \leq i \leq n$, so arguing as in Crouch [2], or Collingwood [20], we see that each Φ^i is a polynomial function. Thus Φ is a polynomial mapping. \square

The isomorphism Φ in Lemma 3 must also be such that Φ^{-1} is also a polynomial mapping, as is easily observed by interchanging the roles of \sum_1 and \sum_2 . Thus, the state space isomorphisms intertwining minimal realizations of finite Volterra series, in g.p.f., have a very special structure. We exploit this fact in the next section.

4. Graded observable polynomial forms for realizations of finite Volterra series. In this section we construct minimal realizations of finite Volterra series, in g.p.f. based on the structure of the observation space \mathcal{H} . We first make the principal definition. A realization of a finite Volterra series of length N , defined on a graded vector space \mathbb{R}^n of order N , is said to be in graded observable polynomial form (g.o.p.f.) if it is in g.p.f. and

$$(14) \quad dC^{N-k}(0) = d\hat{\mathcal{H}}^k(0), \quad 0 \leq k \leq N-1$$

where $dC^k(0)$ and $d\hat{\mathcal{H}}^k(0)$ are the subspaces of $T_0^* \mathbb{R}^n = \mathbb{R}^n$ spanned by the corresponding spaces of one forms evaluated at $x = 0$. $\hat{\mathcal{H}}^k$ as defined in Theorem 2, is the subspace of \mathcal{H} spanned by all functions of the form

$$L_{f^{k_0}} L_{g_{i_1}} L_{f^{k_1}} L_{g_{i_2}} \cdots L_{g_{i_r}} L_{f^{k_r}}(h_j) \quad \text{for } 1 \leq j \leq p, \quad k_i \geq 0, \quad 1 \leq i_j \leq m \text{ and } k \leq r \leq N.$$

Recalling the property (7) we see that if the graded vector space in the above definition is expressed as

$$\mathbb{R}^n = \mathbb{R}^{m_1} \oplus \mathbb{R}^{m_2} \oplus \cdots \oplus \mathbb{R}^{m_N}$$

then the dimensions m_i are related to the system by the equations

$$m_{N-i+1} = \text{Dim } d\hat{\mathcal{H}}^{i-1}(0) - \text{Dim } d\hat{\mathcal{H}}^i(0).$$

In particular, since the dimension of $d\hat{\mathcal{H}}^i(x)$, $x \in M$ is the same for any minimal realization of a finite Volterra series (see Crouch [2]); any two minimal realizations of a finite Volterra series of length N , both in g.o.c.f., must be defined on the same graded vector space. In particular, the integers m_i , $1 \leq i \leq N$ are invariants.

Our next result shows that minimal realizations of finite Volterra series in g.o.p.f. do in fact exist.

THEOREM 4. *A finite Volterra series that has a realization in the form of system (1), where $f + \sum_{i=1}^m \alpha_i g_i$, $\alpha_i \in \mathbb{R}$ are complete vector fields, is also realizable by a minimal system in graded observable polynomial form. The observation space of the system in this representation contains all the coordinate functions.*

Proof. Using standard realization theory as in Sussmann [10], we may assume that we are given a minimal realization of the finite Volterra series, with length N , represented by the system (1) in which the vector fields are complete. As in Crouch [2], we may assume that the state space is $\mathbb{R}^N = M$. By observability $d\mathcal{H}(x) = T_x^* M$ for each $x \in M$, where $x \rightarrow d\mathcal{H}(x)$ is the distribution in the cotangent bundle $d\mathcal{H}(x) = \{dh(x); h \in \mathcal{H}\}$. We select a basis for $d\mathcal{H}(x_0)$ as follows. Let $\phi_{N-1}^1 \cdots \phi_{N-1}^{m_{N-1}} \in \hat{\mathcal{H}}^{N-1} = \mathcal{O}^{N-1}$, as defined in Theorem 2, be such that $d\phi_{N-1}^1(x_0) \cdots d\phi_{N-1}^{m_{N-1}}(x_0)$ is a basis for $d\hat{\mathcal{H}}^{N-1}(x_0)$. Define the basis inductively, by completing the basis for $d\hat{\mathcal{H}}^k(x_0)$, to a basis for $d\hat{\mathcal{H}}^{k-1}(x_0)$ using elements $\phi_{k-1}^1 \cdots \phi_{k-1}^{m_{N-1-k+1}} \in \mathcal{O}^{k-1}$ such that $d\phi_{k-1}^1(x_0) \notin d\hat{\mathcal{H}}^k(x_0)$, $d\phi_{k-1}^1(x_0) \cdots d\phi_{k-1}^{m_{N-1-k+1}}(x_0)$ are linearly independent and together with $d\hat{\mathcal{H}}^k(x_0)$ span $d\hat{\mathcal{H}}^{k-1}(x_0)$.

It is shown in Proposition 4.6 of Crouch [2] that $d\hat{\mathcal{H}}^k(x)$ are constant-dimensional distributions on M for $N-1 \geq k \geq 0$. In fact, if $S \subset \text{Diff}(M)$ is the connected Lie transformation group of M with Lie Algebra \mathcal{S} , by controllability S acts transitively on M and if $\gamma \in S$, $d\phi \in d\hat{\mathcal{H}}^r$ then $\gamma^* d\phi(\gamma(x_0)) = d\phi(x_0) + \omega(x_0)$ where $\omega(x_0) \in d\hat{\mathcal{H}}^{r+1}(x_0)$. Since γ^* is an isomorphism, one deduces that $d\phi_{N-1}^i(x)$ span $d\hat{\mathcal{H}}^{N-1}(x)$ for all $x \in M$, $1 \leq i \leq m_1$. Similarly, an inductive argument shows that $d\phi_k^1(x), \dots, d\phi_k^{m_{N-k}}(x)$ span $d\hat{\mathcal{H}}^k(x)$ for each k , and in particular for $k=0$, when $d\hat{H}^0(x) = T_x^* M$.

We now differentiate these elements of \mathcal{H} along solutions of (1), to construct the desired realization. In particular

$$\frac{d}{dt} \phi_j^k(x(t)) = L_f \phi_j^k(x(t)) + \sum_{i=1}^m u_i L_{g_i} \phi_j^k(x(t)).$$

Since $\phi_j^k \in \hat{\mathcal{H}}^j$, $L_f \phi_j^k \in \hat{\mathcal{H}}^j$, $L_{g_i} \phi_j^k \in \hat{\mathcal{H}}^{j+1}$ and $d\phi_{N-1}^1(x), \dots, d\phi_j^{m_{N-j}}(x)$ span $d\hat{\mathcal{H}}^j(x)$ there exist analytic functions F_j^k, G_{ij}^k such that

$$L_f \phi_j^k(x) = F_j^k(\phi_{N-1}^1(x), \dots, \phi_j^{m_{N-j}}(x)),$$

$$L_{g_i} \phi_j^k(x) = G_{ij}^k(\phi_{N-1}^1(x), \dots, \phi_{N-1}^{m_{N-1}}(x)).$$

Moreover, since $h_i \in \mathcal{H}$ for $1 \leq i \leq p$, there exist analytic functions \hat{h}_i such that

$$h_i(x) = \hat{h}_i(\phi_{N-1}^1(x), \dots, \phi_0^{m_N}(x)).$$

We now set $\hat{x}_r(t) = (\phi_{N-r}^1(x(t)), \dots, \phi_{N-r}^{m_{N-r}}(x(t))) = (\hat{x}_r^1(t), \dots, \hat{x}_r^{m_r}(t))$, and let \hat{f}_r represent the vector with components F_{N-r}^k , $1 \leq k \leq m_r$ and \hat{g}_{ir} represent the vector with components $G_{i,N-r}^k$, $1 \leq k \leq m_r$. It now follows that we have the following representation

for system (1):

$$(15) \quad \begin{bmatrix} \dot{\hat{x}}_1 \\ \dot{\hat{x}}_2 \\ \vdots \\ \dot{\hat{x}}_N \end{bmatrix} = \begin{bmatrix} \hat{f}_1(\hat{x}_1) \\ \hat{f}_2(\hat{x}_1, \hat{x}_2) \\ \vdots \\ \hat{f}_N(\hat{x}_1, \dots, \hat{x}_N) \end{bmatrix} + \sum_{i=1}^m u_i \begin{bmatrix} \hat{g}_{i1} \\ \hat{g}_{i2}(\hat{x}_1) \\ \vdots \\ \hat{g}_{iN}(\hat{x}_1, \dots, \hat{x}_{N-1}) \end{bmatrix} \begin{bmatrix} \hat{x}_1 \in \mathbb{R}^{m_1} \\ \hat{x}_2 \in \mathbb{R}^{m_2} \\ \vdots \\ \hat{x}_N \in \mathbb{R}^{m_n} \end{bmatrix},$$

$$y_i = \hat{h}_i(\hat{x}_1, \dots, \hat{x}_N).$$

Since each component $F_j^k(\phi_{N-1}^1(x), \dots, \phi_j^{m_{N-j}}(x))$ of \hat{F}_{N-j} lies in $\hat{\mathcal{H}}^j$, and $\phi_j^k(x) \in \hat{\mathcal{H}}^j$, viewed as outputs of system (1), they give rise to Volterra series of length at most $N-j$. Thus, with respect to the graded structure defined by system (15)

$$\mathbb{R}^n = \mathbb{R}^{m_1} \oplus \mathbb{R}^{m_2} \oplus \dots \oplus \mathbb{R}^{m_n},$$

we see that $F_j^k(\hat{x}_1, \dots, \hat{x}_{N-j}) \in C^{N-j}$. Similarly, $G_{ij}^k(\hat{x}_1 \dots \hat{x}_{N-j+1}) \in C^{N-j+1}$ and $\hat{h}_i(\hat{x}_1, \dots, \hat{x}_N) \in C^N$. Thus system (15) is in g.p.f. with respect to the graded structure.

The Jacobian of the map $\Phi: M \rightarrow \mathbb{R}^n$ defined by $\Phi(x) = (\phi_{N-1}^1(x), \dots, \phi_0^{m_N}(x))$ has by construction full rank at each point $x \in M$. It follows that system (15) is a complete, controllable, and observable realization of the original Volterra series on the range of Φ . However, as in Corollary 3.8 of Crouch [2], any such realization is also minimal, so the range of Φ must be \mathbb{R}^n , and system (15) is the desired minimal realization of the Volterra series in g.p.f.

By mimicking the proof of Theorem 1, it is clear that the observation space of system (15) also contains all the coordinate functions $\hat{x} \rightarrow \hat{x}_j^i$, $1 \leq i \leq m_j$. In particular $d\hat{\mathcal{H}}^k(\Phi(x_0))$ is spanned by $d\hat{x}_1^1 \dots d\hat{x}_{N-k}^{m_{N-k}}$ and so condition (14) is satisfied, showing that the system is also in g.o.p.f. Finally, we claim that we may assume that $\Phi(x_0) = 0$ without affecting any of the above results. But this is clear since it contains the constant functions so we may modify each basis function ϕ_j^k to obtain the desired result. \square

An immediate corollary of this result is that the observation algebra of the realization (15) obtained above is equal to $\mathbb{R}[\hat{x}_1^1 \dots \hat{x}_N^{m_N}]$. However, this result is true more generally, as we now demonstrate.

LEMMA 4. *If system (15) is a realization of a finite Volterra series of length N , which is in g.o.p.f., the observation algebra \mathcal{H} contains functions of the form*

$$(16) \quad \hat{x}_i^j + S_i^j(\hat{x}_1^1 \dots \hat{x}_{i-1}^{m_{i-1}}), \quad 1 \leq i \leq N, \quad i \leq j \leq m_i$$

where $S_i^j \in C^i$.

Proof. Since the system is in g.o.p.f.

$$d\hat{\mathcal{H}}^{N-i}(0)/d\hat{\mathcal{H}}^{N-i+1}(0) = dC^i(0)/dC^{i-1}(0), \quad 1 \leq i \leq N,$$

which by property (7) is isomorphic to \mathbb{R}^{m_i} . We may select one form $d\phi_j^i \in d\hat{\mathcal{H}}^{N-i}$ such that their image in $d\hat{\mathcal{H}}^{N-i}(0)/d\hat{\mathcal{H}}^{N-i+1}(0)$ is dx_i^j for $1 \leq j \leq m_i$. But since $d\phi_j^i \in d\hat{\mathcal{H}}^{N-i} \subset dC^i$ the expression in (16) follows immediately. \square

From Lemma 4 one deduces immediately that any system in g.o.p.f. has an observation algebra equal to $\mathbb{R}[\hat{x}_1^1 \dots \hat{x}_N^{m_N}]$. This result and the remark following the proof of Theorem 4 may be viewed as a restatement of a similar result obtained in Fliess [16], although the result there is only local. Note also that Lemma 4 shows that in the definition of g.o.p.f., equation (14) may be replaced by $dC^{N-k}(x) = d\hat{\mathcal{H}}^k(x)$, $0 \leq k \leq N-1$, $x \in M$, and in particular for $k=0$ $dC^N(x) = d\hat{\mathcal{H}}^0(x) = d\mathcal{H}(x)$. Thus, a system in g.o.p.f. is always observable.

We now sharpen the result obtained above to show that any minimal system in g.p.f. has a polynomial observation algebra. Note first that if the two systems Σ_1 and

Σ_2 in Lemma 1 are minimal realizations of the same finite Volterra series, and $\mathcal{A}_1, \mathcal{A}_2$ are the observation algebras generated by the observation spaces \mathcal{H}_1 and \mathcal{H}_2 respectively, then the state space isomorphism Φ , guaranteed by Sussmann [10], induces an isomorphism of \mathcal{H}_1 and \mathcal{H}_2 and hence \mathcal{A}_1 and \mathcal{A}_2 . In the case when both systems are g.p.f., however, Lemma 3 shows that Φ has a particularly nice structure which we now exploit.

THEOREM 5. *The observation algebra of a minimal realization of a finite Volterra series in graded polynomial form is the ring of polynomials $\mathbb{R}[x_1 \cdots x_n]$ in the state coordinates $(x_1 \cdots x_n)$.*

Proof. Construct a minimal realization of the finite Volterra series in g.o.p.f. as in Theorem 4 with state $(z_1 \cdots z_n) \in \mathbb{R}^n$. Let ϕ be the state space isomorphism $\phi: \mathbb{R}^n \rightarrow \mathbb{R}^n$ $\phi(x_1 \cdots x_n) = (z_1 \cdots z_n) = z$. If $\mathcal{A} \subset \mathbb{R}[x_1 \cdots x_n]$ is the observation algebra of the original system, by the previous remarks we see that ϕ induces an isomorphism β_ϕ of \mathcal{A} onto $\mathbb{R}[z_1 \cdots z_n]$. By Lemma 3 ϕ is a polynomial mapping with a polynomial inverse. Thus, there exist $q_i \in \mathbb{R}[z_1 \cdots z_n]$ such that $q_i(z) = x_i, 1 \leq i \leq n$.

Now $(\beta_\phi)^{-1}(q_i)(x) = q_i(\phi(x)) = q_i(z) = x_i$ where $(\beta_\phi)^{-1}(q_i) \in \mathcal{A}$. Thus \mathcal{A} contains all the coordinate functions $x_1 \cdots x_n$. Since it also contains the constant functions we see that $\mathcal{A} = \mathbb{R}[x_1 \cdots x_n]$ as claimed. \square

A somewhat stronger intermediate result is given in Collingwood [20].

From Theorems 3 and 4, we see that given a finite Volterra series of length N , we may construct minimal realizations in g.c.p.f. and g.o.p.f., on graded vector spaces of order N , characterized by two sets of integers $n_1 \cdots n_N$, and $m_1 \cdots n_N$, respectively, where for $1 \leq i \leq N$

$$\begin{aligned} n_i &= \text{Dim } (\mathcal{S}^i(0)/\mathcal{S}^{i+1}(0)), \\ m_i &= \text{Dim } (d\hat{\mathcal{H}}^{N-i}(0)/d\hat{\mathcal{H}}^{N-i+1}(0)). \end{aligned}$$

One naturally asks whether these two sets of integers are identical. In general, they are not, and this is demonstrated in Examples 4 and 5 below. However, for Volterra series consisting of a single term (denoted homogeneous Volterra series in Crouch [2]), the two sets of integers are identical. Indeed, referring to § 4.4 of Crouch [2], we see that

$$\begin{aligned} m_i &= \text{Dim } (d\hat{\mathcal{H}}^{N-i}(0)/d\hat{\mathcal{H}}^{N-i+1}(0)) = \text{Dim } (d\mathcal{O}^{N-i}(0)) \\ &= \text{Dim } \mathcal{R}^i(0) = \text{Dim } (\mathcal{S}^i(0)/\mathcal{S}^{i+1}(0)) = n_i. \end{aligned}$$

In general, a system in g.p.f. which is simultaneously in g.o.p.f. and g.c.p.f. so that $n_i = m_i, 1 \leq i \leq N$, will be said to be in graded symmetric polynomial form (g.s.p.f.).

Example 4.

$$\begin{aligned} (17) \quad \dot{x}_1 &= u, & x_1(0) &= 0, \\ \dot{x}_2 &= (x_1)^2, & x_2(0) &= 0, \\ \dot{x}_3 &= x_2, & x_3(0) &= 0, \\ y &= x_3 + x_2 x_1. \end{aligned}$$

The graded structure of this system is exhibited in Example 1 with

$$(x_1, x_2, x_3) = (x_1^1, x_2^1, x_2^2) \in \mathbb{R} \oplus \mathbb{R}^2.$$

That is $N = 3, n_1 = 1, n_2 = 2, n_3 = 0$.

If $f = (x_1)^2 \partial / \partial x_2 + x_2 \partial / \partial x_3$, $g_1 = \partial / \partial x_1$ we compute $g_2 = [g_1, f] = 2x_1 \partial / \partial x_2$, $g_3 = [g_1, g_2] = 2\partial / \partial x_2$, $g_4 = [g_3, f] = 2\partial / \partial x_3$, $g_5 = [g_2, f] = 2x_1 \partial / \partial x_3$, $[g_1, g_5] = 2\partial / \partial x_3$, and

all other brackets vanish. It follows that $\mathcal{S}^3(0)=0$, $\mathcal{S}^2(0)=\text{Span}\{\partial/\partial x_3, \partial/\partial x_2, x_1\partial/\partial x_3\}=\mathbb{R}^2$, $\mathcal{S}^1(0)=\text{Span}\{\mathcal{S}^2(0), \partial/\partial x_1, x_1\partial/\partial x_2\}$, and so the system is in g.c.p.f.

If we set $h_1=x_3+x_2x_1$ we compute $h_2=L_f h_1=x_2+(x_1)^3$, $h_3=L_{g_1} h_1=x_2$, $h_4=L_f h_2=(x_1)^2$, $3h_4=L_{g_1} h_2=3(x_1)^2$, $h_4=L_f h_3=(x_1)^2$, $h_5=L_{g_1} h_4=2x_1$, $h_6=L_{g_1} h_5=2$, and all other Lie derivatives are zero. It follows that $\mathcal{H}=\text{Span}\{x_3+x_2x_1, x_2, x_1(x_1)^2, (x_1)^3, 1\}$, $\mathcal{H}^3=\text{span}\{1\}$, $\mathcal{H}^2=\text{Span}\{1, x_1\}$, $\mathcal{H}^1=\text{Span}\{1, x_1, x_2, (x_1)^2\}$. Thus, $\text{Dim } d\mathcal{H}^0(0)/d\mathcal{H}^1(0)=1$, $\text{Dim } d\mathcal{H}^1(0)/d\mathcal{H}^2(0)=1$, $\text{Dim } d\mathcal{H}^2(0)/d\mathcal{H}^3(0)=1$. Thus, system (17) is not in g.o.c.f., but is observable and hence, minimal. Note that although \mathcal{H} does not contain x_1, x_2 and x_3 , the algebra \mathcal{A} generated by \mathcal{H} does.

Example 5.

$$(18) \quad \begin{aligned} \dot{z}_1 &= u, & z_1(0) &= 0, \\ \dot{z}_2 &= (z_1)^2, & z_2(0) &= 0, \\ \dot{z}_3 &= z_2 + (z_1)^3 + z_2 u, & z_3(0) &= 0, \\ y &= z_3. \end{aligned}$$

The graded structure of this system is exhibited in Example 2 with $(z_1, z_2, z_3) = (x_1^1, x_1^2, x_1^3) \in \mathbb{R} \oplus \mathbb{R} \oplus \mathbb{R}$. That is $N=3$, $n_1=n_2=n_3=1$. Note however that there is a state space isomorphism intertwining systems (17) and (18) defined by

$$z_1 = x_1 = 1/2 h_5(x), \quad z_2 = h_3(x) = x_2, \quad z_3 = x_3 + x_2 x_1 = h_1(x).$$

Thus, systems (17) and (18) are both minimal realizations of the same finite Volterra series, and in particular, the calculations in Example 4 demonstrate that system (18) is in g.o.p.f. but not in g.c.p.f. Note that in this case, $\mathcal{H}=\text{Span}\{z_3, z_2, z_1, (z_1)^2, (z_1)^3, 1\}$, giving an example of Theorem 4.

Example 6.

$$\begin{aligned} \dot{x}_1 &= u, & x_1(0) &= 0, \\ \dot{x}_2 &= (x_1)^2, & x_2(0) &= 0, \\ y &= x_2. \end{aligned}$$

This system is minimal and has a Volterra series consisting of a single second order term. $\mathcal{S}^2(0)=\text{Span}\{\partial/\partial x_2\}$, $\mathcal{S}^1(0)=\text{Span}\{\partial/\partial x_2, x_1\partial/\partial x_2, \partial/\partial x_1\}$, $\mathcal{H}^2=\text{Span}\{1\}$, $\mathcal{H}^1=\text{Span}\{1, x_1\}$, $\mathcal{H}^0=\mathcal{H}=\text{Span}\{1, x_1, x_2, (x_1)^2\}$. Consequently, the system is in g.s.p.f., on the graded vector space of order $N=2$, $\mathbb{R}^2=\mathbb{R} \oplus \mathbb{R}$, $m_1=m_2=n_1=n_2=1$.

5. An application to algebraic estimation theory. Consider the stochastic system given in Ito form by

$$(19) \quad \begin{aligned} dx &= f(x) dt + g(x) dw, & x &\in \mathbb{R}^n, \\ dy &= h(x) dt + dv, & y &\in \mathbb{R} \end{aligned}$$

where $\{w\}$ and $\{v\}$ are independent Brownian motions and suppose that we wish to estimate some real-valued function (or statistic) $\psi(x(t))$ using the information contained in the observations process $Y(t)=\{y(s); 0 \leq s \leq t\}$. It is well known (and indeed, many excellent texts and surveys on the subject, including those of Kallianpur [25], Davis and Marcus [26] and Marcus [27], exist) that if $\hat{\psi}$ is the minimum variance, or least-squares, estimator of ψ , then there are several methods available for the calculation of $\hat{\psi}$. Here we focus our attention on only one of these algorithms, namely the recursive scheme based on the evolution equation of the unnormalized conditional density as

derived by Zakai [11], Mortensen [28] and Duncan [29]. Thus, we obtain $\hat{\psi}$ by solving the system (again in Ito form)

$$\begin{aligned} d\rho(t, x) &= \bar{F}(\rho(t, x)) dt + G(\rho(t, x)) dy, \quad \rho(0) = p_0, \\ (20) \quad \hat{\psi}(t) &= E(\psi(x(t)) | \mathcal{Y}(t)) \\ &= \int_{\mathbb{R}^n} \psi(x) \rho(t, x) dx \left[\int_{\mathbb{R}^n} \rho(t, x) dx \right]^{-1} = C_\psi(\rho), \end{aligned}$$

where \bar{F} and G are the differential operators on $C_0^\infty(\mathbb{R}^n)$ defined by

$$\begin{aligned} \bar{F}(\phi)(x) &= \frac{1}{2} \sum_{i,j=1}^n \frac{\partial^2}{\partial x_i \partial x_j} (g_i(x) g_j(x) \phi(x)) - \sum_{i=1}^n \frac{\partial}{\partial x_i} f_i(x) \phi(x), \\ G(\phi)(x) &= h(x) \phi(x), \end{aligned}$$

$f_i, g_i, 1 \leq i \leq n$ are the components of the vector fields f, g of (19) and $\mathcal{Y}(t)$ is the σ field generated by $Y(t)$. ($C_0^\infty(\mathbb{R}^n)$ is the space of smooth functions on \mathbb{R}^n with compact support.) While giving a complete solution to this filtering problem, the system described by (20) remains unsatisfactory from the point of view of applications since it has the structure of a stochastic partial differential equation. In an attempt to circumvent this problem, the concept of a finite dimensionally computable (f.d.c.) recursive statistic has been introduced. A prime objective of algebraic estimation theory is the determination of conditions under which such statistics exist.

If ψ is such a f.d.c. statistic then by definition, there is a stochastic system driven by the observations process $\{y(t)\}$ and evolving on a finite dimensional manifold, say M , which has as output the desired optimal estimate $\hat{\psi}$. For technical reasons (specifically to alleviate problems with coordinate transformations) such a system is usually described in terms of the Fisk-Stratonovich integral and thus takes the form

$$\begin{aligned} (21) \quad dz &= a(z) dt + b(z) \circ dy, \quad z(0) = z_0, \\ \hat{\psi}(t) &= c(z(t)) \end{aligned}$$

with a, b vector fields and c a real valued function on M . Now, it is reasonable, and natural, to assume that whatever data record generates the estimate, the two systems (20) and (21) should respond in the same way; thus we can think of (20) and (21) as being realizations of the same input-output map. By appealing to the Sussmann [30] and Doss [31] construction of a pathwise solution to a stochastic differential equation, and by transforming (20) into the equivalent (F-S) form we see that the underlying deterministic systems

$$\begin{aligned} (22) \quad \frac{\partial \rho}{\partial t} &= F(\rho) + uG(\rho), \quad \rho(0) = p_0, \\ \hat{\psi}(t) &= C_\psi(\rho) \end{aligned}$$

and

$$\begin{aligned} (23) \quad \frac{dz}{dt} &= a(z) + ub(z), \quad z(0) = z_0, \\ \hat{\psi} &= c(z) \end{aligned}$$

where now $F = \bar{F} - \frac{1}{2}h^2$, are also realizations of the same input-output map. Moreover, Hijab [32] has shown that (21) can be taken to be “minimal” in the sense that the corresponding system (23) is minimal. This means that we can apply the techniques used in the proof of Lemma 1 provided there is sufficient analytic structure on the state space of (22). For instance, if it contains a subspace D invariant under the (semi)-flow generated by X , for all $X \in \{F, G\}_{\text{L.A.}}$, the Lie algebra generated by F and G , and every time-analytic input produces a time-analytic output, we conclude as in Brockett [1] that there should be a homomorphism between the system Lie algebras of (22) and (23), giving a necessary condition for the existence of f.d.c. statistics. Clearly, some of the ideas used to derive this homomorphism principle are of an ad hoc nature, but the work of Hijab [33] and Michael and Chaleyat-Maurel [34], among others, has done much to place it on a more rigorous footing. Moreover, such a homomorphism can be readily found in the case that (19) is linear, and the central theme of algebraic estimation theory is its construction for more general systems.

Unfortunately, the bulk of the available evidence suggests that for most systems this approach will ultimately prove to be fruitless and, as for example, in the analysis of cubic sensor problems, preliminary Lie algebraic calculations and predictions are usually borne out by later probabilistic calculations (Hazewinkel, Marcus and Sussmann [35]). We hope to point the way towards the reasons for this paucity by applying the results of the previous sections to the system (19) under the additional assumption that the underlying deterministic system, corresponding to the equivalent (F-S) form of (19),

$$(24) \quad \begin{aligned} \dot{x} &= \hat{f}(x) + ug(x), & x(0) &= x_0, \quad x \in \mathbb{R}^n, \\ y &= h(x), \end{aligned}$$

with $\hat{f} = f - \frac{1}{2}(dg/dx)g$, is a minimal system in g.p.f. Our arguments are designed to show that there are deep connections between the estimation algebra (i.e., the Lie algebra $\Lambda \triangleq \{F, G\}_{\text{L.A.}}$) corresponding to (19) and the Weyl algebra, W_n , of all differential operators on \mathbb{R}^n with polynomial coefficients. The significance of this observation derives from Hazewinkel and Marcus' [5] result that the only homomorphisms between W_n and a Lie algebra of vector fields are trivial, suggesting, via the above arguments, that if $\Lambda \equiv W_n$, then (19) will have no f.d.c. statistics.

We shall concentrate on the following aspect of the problem. After some manipulation, it is possible to rewrite the generator F as

$$F = \frac{1}{2}L_g^{*2} + L_{\hat{f}}^* - \frac{1}{2}h^2$$

where $\phi \rightarrow L_X^*(\phi)$ is the formal adjoint (w.r.t. the L_2 inner product) of the Lie derivative operator L_X , from which it is clear that if (19) or (24) are polynomial systems on \mathbb{R}^n , $\Lambda \subset W_n$ always. However, there is the important special case, namely if (19) or (21) is linear, for which $\Lambda \neq W_n$ and it is natural to ask therefore if we can determine any other cases for which $\Lambda \neq W_n$. We show that this is equivalent to showing that Λ is not identical to an associated tensor algebra.

As in the previous sections, we denote by \mathcal{L} , \mathcal{I} , and \mathcal{H} the Lie algebra, ideal and observation space of system (24), and denote by \mathcal{L}^* the Lie algebra of differential operators generated by $L_{\hat{f}}^*$ and L_g^* . Let \mathcal{I}^* denote the ideal of \mathcal{L}^* generated by L_g^* . Finally, if F is a Lie algebra of differential operators on $C_0^\infty(\mathbb{R}^n)$, let $E(F)$ denote its enveloping algebra, which we may identify with the algebra of operators on $C_0^\infty(\mathbb{R}^n)$ comprised of repeated finite compositions of operators in F .

The following result follows immediately from these definitions and the identity $[L_X^*, h] = -L_X h$ for smooth vector fields X and smooth functions h viewed as multiplication operators on $C_0^\infty(\mathbb{R}^n)$.

LEMMA 5. $\Lambda \subset \mathbb{R}F + E(\{(\mathcal{H} + 1) \otimes (\mathcal{S}^* + 1)\}_{\text{L.A.}}) \triangleq \Omega$.

Therefore, for polynomic systems (19) we have $\Lambda \subset \Omega \subset W_n$, and so Λ can fail to equal W_n in two ways, Λ may fail to equal Ω , and Ω may fail to equal W_n . The aim of this section is to show that under appropriate conditions systems with finite Volterra series always satisfy $\Omega = W_n$. The following result, the proof of which involves a lot of tedious calculations performed in Collingwood [20], establishes that the Ito correction term $-\frac{1}{2}(dg/dx)g$ does not affect the definition of graded (symmetric) polynomial form.

LEMMA 6. *System (24) is a minimal realization of a finite Volterra series in g.p.f. (g.s.p.f.) if and only if the following system has the same property with respect to the same graded vector space.*

$$\begin{aligned}\dot{x} &= f(x) + ug(x), & x(0) &= x_0, \quad x \in \mathbb{R}^n, \\ y &= h(x).\end{aligned}$$

From this result we see that there is no ambiguity involved if we refer to the (stochastic) system (19) as being in g.p.f.

We may now state the main result of this section.

THEOREM 6. *Given a stochastic system (19) in g.p.f., such that the corresponding deterministic system (24) is minimal and in g.c.p.f., then the estimation algebra satisfies*

$$\Lambda \subset \Omega = W_n.$$

Proof. By Theorem 5 we may assume that the observation algebra of system (24) is $\mathbb{R}[x_1 \cdots x_n]$. Moreover, since system (24) is in g.p.f., $L_g^* = -L_g$ and $L_f^* = -L_f - \text{div}(f)$. Now $c = \text{div}(f) = \text{trace}(\text{diag}(A_1 \cdots A_N))$, where A_1, \dots, A_N are the matrices occurring in the representation of \hat{f} as a vector field on the graded vector space (of order N), as given in (11). Consequently, we may write $F = -L_f + \frac{1}{2}L_g^2 - \frac{1}{2}h^2 - c$, $G = h$. Since the operator consisting of multiplication by a constant commutes with Λ , we ignore the term c in the expression for F . Moreover, \mathcal{S}^* coincides with \mathcal{S} , when viewed as a space of differential operators. Since \mathcal{H} contains constant functions, we may rewrite the expression for Ω as

$$\Omega = \mathbb{R}F + E(\{\mathcal{H} \otimes (\mathcal{S} + 1)\}_{\text{L.A.}}).$$

Since we assume that system (24) is in g.c.p.f., by Lemma 2 we may assume that \mathcal{S} contains the vector fields of the form

$$\partial/\partial x_i + \sum_{j>i} r_i^j(x_1 \cdots x_{j-i})\partial/\partial x_j$$

where r_i^j are polynomials. Since Ω clearly contains $E(\mathcal{H}) = \mathbb{R}[x_1 \cdots x_n]$ and $E(\mathcal{H}) \otimes \mathcal{S}$, it also contains $\partial/\partial x_1, \dots, \partial/\partial x_n$. Thus, Ω contains the generators of W_n and hence $\Omega = W_n$ as required since $F \in W_n$ also. \square

Theorem 6 shows that the algebraic estimation problem, in the case of systems with finite Volterra series, rests on the inclusion

$$\Lambda \subset E(\{\mathcal{H} \otimes (\mathcal{S} + 1)\}_{\text{L.A.}}).$$

There seems to be considerable evidence suggesting that systems (19) in g.p.f. with

Volterra series of length greater than one satisfy $\Lambda = W_n$. Although this would imply the nonexistence of finite-dimensional filters, it may have significance for attempts to obtain approximate filters, based on a stochastic version of the approximation procedure outlined in Crouch [3].

REFERENCES

- [1] R. W. BROCKETT, *Remarks on finite-dimensional nonlinear estimation*, in *Analyses des systemes astériskues*, 75-76, 1980, pp. 45-55.
- [2] P. E. CROUCH, *Dynamical realizations of finite Volterra series*, this Journal, 19 (1981), pp. 177-202.
- [3] ———, *Solvable approximations to control systems*, this Journal, 22 (1984), pp. 40-54.
- [4] R. W. GOODMAN, *Nilpotent Lie Groups: Structure and Applications to Analysis*, Lecture Notes in Mathematics, No. 562, Springer-Verlag, Berlin-New York, 1976.
- [5] M. HAZEWINKEL AND S. I. MARCUS, *On Lie algebras and finite-dimensional filtering*, *Stochastics*, 7 (1982), pp. 29-62.
- [6] R. HERMANN AND A. J. KRENER, *Nonlinear observability and controllability*, *IEEE Trans. Automat. Control*, AC-22 (1977), pp. 728-740.
- [7] O. HIJAB, *Minimum energy estimation*, Ph.D. dissertation, Univ. California, Berkeley, 1980.
- [8] M. FLIESS AND I. KUPKA, *A finiteness criterion for nonlinear input-output differential systems*, this Journal, 21 (1983), pp. 721-728.
- [9] H. J. SUSSMANN AND V. JURDEVIC, *Controllability of nonlinear systems*, *J. Differential Equations*, 12 (1972), pp. 313-329.
- [10] H. J. SUSSMANN, *Existence and uniqueness of minimal realizations of nonlinear systems*, *Math. Systems Theory*, 10 (1977), pp. 263-284.
- [11] M. ZAKAI, *On the optimal filtering of diffusion processes*, *Z. Wahrsch. Verw. Gebiete*, 11 (1969), pp. 230-243.
- [12] I. A. P. KUPKA, *On finite Volterra series and a theorem of P. Crouch*, preprint (1983), this Journal, to appear.
- [13] J. P. GAUTHIER AND G. BORNARD, *Observability for any $u(t)$ of a class of nonlinear systems*, *IEEE Trans. Automat. Control*, AC 26 (1981), pp. 922-926.
- [14] D. A. AEYELS, *Generic observability of differentiable systems*, this Journal, 19 (1981), pp. 595-603.
- [15] H. NIJMEIJER, *Observability of a class of nonlinear systems: A geometric approach*, *Ricerche Automat.*, 12 (1981), pp. 1-19.
- [16] M. FLIESS, *Realisations des systèmes non linéaires, algèbres de Lie, filtres transitives et séries génératrices non commutatives*, *Invent. Math.*, 11 (1983), pp. 521-537.
- [17] ———, *A remark on transfer functions and realizations of homogeneous continuous-time systems*, *IEEE Trans. Automat. Control*, 24 (1979), pp. 507-508.
- [18] ———, *Functionnelles causales non linéaires et indéterminées non commutatives*, *Bull. Soc. Math. France*, 109 (1981), pp. 3-4.
- [19] P. E. CROUCH AND P. C. COLLINGWOOD, *Structure theory for realizations of finite Volterra series*, *Proc. Twente Workshop on Systems and Optimization*, in *Lecture Notes in Control and Information Sci.*, 66 (1984), pp. 44-60.
- [20] P. C. COLLINGWOOD, *Realizations of finite Volterra series and the algebraic estimation problem*, Ph.D. thesis, Univ. Warwick, Coventry, England, 1985.
- [21] Z. BARTOSIEWICZ, *On polynomial continuous time systems. Realization theory and some perspectives*, *I.M.A. J. on Mathematical Control* (1984), to appear.
- [22] ———, *Realizations of polynomial systems*, preprint (1985).
- [23] R. W. BROCKETT, *Volterra series and geometric control theory*, *Automatica—J. IFAC*, 12 (1976), pp. 167-176.
- [24] J. E. HUMPHREYS, *Introduction to Lie Algebras and Representation Theory*, *Graduate Texts in Math.*, 9, Springer, New York-Berlin, 1972.
- [25] G. KALLIANPUR, *Stochastic Filtering Theory*, Springer-Verlag, New York, 1980.
- [26] M. H. A. DAVIS AND S. I. MARCUS, *An introduction to nonlinear filtering*, in *Stochastic Systems: The Mathematics of Filtering and Applications*, M. Hazewinkel and J. C. Willems, eds., Reidel, Dordrecht, The Netherlands, 1981, pp. 565-572.
- [27] I. MARCUS, *Algebraic and geometric methods in nonlinear filtering*, this Journal, 22 (1984), pp. 817-844.
- [28] R. E. MORTENSEN, *Optimal control of continuous time stochastic systems*, Ph.D. thesis, Univ. California, Berkeley, 1966.

- [29] T. E. DUNCAN, *Probability densities for diffusion processes with applications to nonlinear filtering theory and diffusion theory*, Ph.D. thesis, Stanford Univ., Stanford, CA, 1967.
- [30] H. J. SUSSMANN, *On the gap between deterministic and stochastic ordinary differential equations*, Ann. Probab., 6 (1978), pp. 19–41.
- [31] H. DOSS, *Liens entre équations différentielles stochastiques et ordinaires*, Ann. Inst. H. Poincaré, 13 (1977), pp. 99–125.
- [32] O. HIJAB, *A realization theory for nonlinear stochastic systems*, Proc. 22nd IEEE Conf. on Dec. and Cont., San Antonio, Texas, 1983.
- [33] ———, *Finite dimensional, causal functionals of Brownian motion*, in Nonlinear Stochastic Problems, R. S. Bucy and J. M. F. Moura, eds., Proc. NATO-ASI, Reidel, Dordrecht, the Netherlands, 1982.
- [34] D. MICHAEL AND M. CHALEYAT-MAUREL, *Un théoreme de non-existence de filtre de dimension fini*, C.R. Acad. Sci. Paris, Sér. 1, 296 (1983), pp. 933–936.
- [35] M. HAZEWINKEL, S. I. MARCUS AND H. J. SUSSMANN, *Nonexistence of finite dimensional filters for conditional statistics of the cubic sensor problem*, Systems Control Lett., 3 (1983), pp. 331–340.

A CONSTRUCTIVE ALGORITHM FOR SENSITIVITY OPTIMIZATION OF PERIODIC SYSTEMS*

TRYPHON T. GEORGIOU† AND PRAMOD P. KHARGONEKAR‡

Abstract. In a recent paper (A. Feintuch, P. P. Khargonekar and A. Tannenbaum, *On the sensitivity minimization problem for linear time-varying periodic systems*, this Journal, 24 (1986), pp. 1076–1085), the problem of weighted sensitivity optimization was considered for linear, discrete-time, periodic time-varying systems. Here we present a constructive algorithm for solving this problem.

Key words. sensitivity minimization, periodic systems, Hankel matrix extension problems

AMS(MOS) subject classifications. 93B50, 30D50

1. Introduction. Zames formulated the weighted sensitivity optimization problem in his seminal paper [14]. Since then, this problem has been thoroughly investigated for linear time-invariant systems by many researchers. We refer the interested reader to the recent survey paper of Francis and Doyle [7] for a good exposition and a complete bibliography. Feintuch and Francis [5] considered this and related problems for linear time-varying systems. Our work is motivated by the recent paper [6] of Feintuch, Khargonekar and Tannenbaum where the problem of weighted sensitivity optimization for linear, discrete-time, periodic time-varying systems is considered. In [6] a formula for minimal weighted sensitivity was derived and the existence of an optimal controller was established. Motivated by possible applications to multirate sampled data systems, here we present a constructive algorithm for the computation of optimal controllers. Our algorithm is based on a simple new way of solving the one step extension problem for finite rank block Hankel matrices. The one-step extension problem for general Hankel operators has been investigated in the masterful work of Adamjan, Arov and Krein [1]. The interested reader is also referred to the recent book [12] by Power for certain related extension problems.

In [10], Khargonekar, Poolla and Tannenbaum showed that to any p -output, m -input, N -periodic, causal, linear, discrete-time system, one can associate a pN -output, mN -input causal linear *time-invariant* system with transfer function $P(z)$ such that $P(\infty)$ is (block) *lower triangular*. Indeed, lower triangularity is closely related to causality. Feintuch, Khargonekar and Tannenbaum [6] showed that the weighted sensitivity minimization problem of Zames [14] for periodic systems can be reduced to the following problem: Given a $pN \times mN$ transfer matrix $T(z)$ with no poles on the unit circle, find

$$(1.1) \quad \mu = \inf \{ \|T(z) - V(z)\|_{\infty} : V(z) \text{ is analytic in the complement of the open unit disc including } \infty \text{ and } V(\infty) \text{ is block lower triangular} \}.$$

This reduction is accomplished using coprime factorizations, Youla parametrization of all stabilizing controllers, and inner-outer factorizations. (There exist good algorithms for these factorizations, e.g., see Doyle [4], Khargonekar and Sontag [11], Vidyasagar [13], and the references cited there.)

* Received by the editors September 25, 1985; accepted for publication (in revised form) February 4, 1986. This research was supported in part by the National Science Foundation under grant ECS-8451519.

† Department of Electrical Engineering, University of Minnesota, Minneapolis, Minnesota 55455. Present address: Department of Electrical Engineering, Iowa State University, Ames, Iowa 50011.

‡ Department of Electrical Engineering, University of Minnesota, Minneapolis, Minnesota 55455.

A formula for μ in (1.1) was given by Feintuch, Khargonekar and Tannenbaum [6]. This formula shows that μ is the maximum of the norms of N Hankel operators. Our paper is devoted to a constructive algorithm to obtain $V(z)$ to solve (1.1) for the special case of rational $T(z)$. This algorithm combined with techniques for coprime and inner-outer factorizations gives a complete constructive algorithm for obtaining optimal controllers for weighted sensitivity minimization of periodic systems.

2. Main results. It has been shown by Feintuch, Khargonekar and Tannenbaum [6] that the weighted sensitivity minimization of Zames [14] for N -periodic linear time-varying finite-dimensional plants can be reduced to the following *best approximation* problem: Given a (possibly unstable) rational $pN \times mN$ matrix $T(z)$, find

$$(1.1) \quad \mu = \inf \{ \|T(z) - V(z)\|_\infty : V(z) \text{ with entries in } RH_\infty^{pN \times mN} \text{ such that } V(\infty) \text{ is lower block triangular} \}.$$

It is assumed that $T(z)$ has no poles on the unit circle. Here RH_∞ denotes the space of all rational functions with real coefficients which are analytic in the complement of the open unit disc (including infinity). Each $pN \times mN$ matrix is considered as an $N \times N$ square matrix with $p \times m$ block entries. The key constraint here is that $V(\infty)$ is required to be block lower triangular. (This corresponds to causality; see [6].) A formula for μ was given by Feintuch, Khargonekar and Tannenbaum [6] and it was also shown that a $V(z)$ achieving the minimum exists. Our main result is to give a *constructive algorithm* to obtain $V(z)$ which in turn can be used to obtain the optimal controller. We should also note that a solution to this problem will also be a key step in solving the general H^∞ -optimization problem of Doyle [4] in the setting of periodic systems.

Let

$$T(z) = \sum_{j=-\infty}^{\infty} \gamma_j z^{-j}$$

be the Fourier series expansion of $T(z)$ (which converges on an open set containing the unit circle). We are seeking a function

$$V(z) = \sum_{j=0}^{\infty} v_j z^{-j}$$

in $RH_\infty^{pN \times mN}$ such that v_0 is lower $(p \times m)$ -block triangular and $\|T - V\|_\infty$ is minimized. Our solution is to first obtain v_0 with the required constraint, and then obtain the rest of $V(z)$ using Glover's algorithm [9].

From the work of Adamjan, Arov and Krein [1] we know that

$$\inf_{\tilde{V}} \|T(z) - v_0 - \tilde{V}(z)\|_\infty$$

where $\tilde{V}(z) = \sum_{j=1}^{\infty} v_j z^{-j}$ is in $z^{-1}RH_\infty^{pN \times mN}$, is equal to the norm of the Hankel operator.

$$\Gamma_e = \begin{bmatrix} D & \gamma_{-1} & \gamma_{-2} & \cdots \\ \gamma_{-1} & \gamma_{-2} & \gamma_{-3} & \cdots \\ \gamma_{-2} & \gamma_{-3} & \gamma_{-4} & \cdots \\ \vdots & \vdots & \vdots & \ddots \end{bmatrix} \quad \text{where } D = \gamma_0 - v_0.$$

The operator Γ_e is thought of as a bounded linear operator acting between the Hilbert spaces of square summable one-sided sequences-denoted by h_2 :

$$\Gamma_e : h_2^{mN} \rightarrow h_2^{pN}.$$

This result is the matrix version of the Nehari problem. In case $T(z) - v_0$ is a rational matrix valued function, a constructive procedure to obtain a rational $\tilde{V}(z)$ minimizing $\|T(z) - v_0 - \tilde{V}(z)\|_\infty$ is given by Glover [9], and Ball and Ran [2]. Thus, we only need to consider a one-step extension problem for the finite-rank Hankel operator given by the matrix $\Gamma = [\gamma_{1-j-k}]_1^\infty$. We want to determine an "extension" D in $R^{pN \times mN}$, subject to the given constraint that the upper block triangular part of D is specified, so that $\|\Gamma_e\|$ is minimized. In the process of solving the generalized Nehari problem, Adamjan, Arov and Krein [1] have provided a solution to a one-step extension problem for block Hankel operators. However their solution does not seem to lend itself either to an easily computable scheme in the case of finite-rank Hankel operators, or to a procedure dealing with the case where D is partially specified as is required in our case.

It is a standard result in realization theory (see the book by Fuhrmann [8]) that if Γ is a finite-rank Hankel operator, then there exists a triple of matrices (F, G, H) , where F is square $n \times n$ (n being the smallest such integer possible), H is $pN \times n$, and G is $n \times mN$, such that the entries γ_{-k} admit a factorization

$$(2.1) \quad \gamma_{-k} = HF^{k-1}G, \quad k = 1, 2, \dots$$

Moreover, this induces a factorization of Γ into a product OR , where

$$R = \begin{bmatrix} G & FG & F^2G & \dots \end{bmatrix} : h_2^{mN} \rightarrow R^n : (u_i : i = 0, 1, \dots) \rightarrow \sum_{i=0}^{\infty} F^i Gu_i,$$

and

$$O = \begin{bmatrix} H \\ HF \\ HF^2 \\ \vdots \end{bmatrix} : R^n \rightarrow h_2^{pN} : x \rightarrow (HF^i x : i = 0, 1, \dots),$$

are bounded linear maps. These are the usual *reachability* and *observability* maps and, because of minimality, are surjective and injective respectively. In view of this factorization, Γ_e is given by

$$(2.2) \quad \Gamma_e = \begin{bmatrix} D & HR \\ OG & OFR \end{bmatrix}.$$

Let P, Q, Σ, Δ be defined as follows:

$$(2.3) \quad P := RR^*, \quad Q := O^*O, \quad \Sigma := P^{1/2}, \quad \Delta := Q^{1/2},$$

where $(\)^{1/2}$ denotes the "Hermitian square root of," and $(\)^*$ denotes the "adjoint of." Also define the following *finite matrix*.

$$H_e := \begin{bmatrix} D & H\Sigma \\ \Delta G & \Delta F\Sigma \end{bmatrix}.$$

We now have the following proposition.

PROPOSITION 2.4. *With the above notation $\|\Gamma_e\| = \|H_e\|$.*

Proof. Define

$$U := \begin{bmatrix} I_{mN} & 0 \\ 0 & R^*\Sigma^{-1} \end{bmatrix} : R^{mN} + R^n \rightarrow R^{mN} + h_2^{mN},$$

$$V := \begin{bmatrix} I_{pN} & 0 \\ 0 & \Delta^{-1}O^* \end{bmatrix} : R^{pN} + h_2^{pN} \rightarrow R^{pN} + R^n,$$

where I_k denotes the $k \times k$ identity matrix. (That the indicated inverses exist follows easily from the fact that \mathbf{R} is surjective and $\mathbf{0}$ is injective.) It is easily seen that

$$\pi_{\mathbf{R}^*} := \mathbf{R}^*(\mathbf{R}\mathbf{R}^*)^{-1}\mathbf{R} : h_2^{mN} \rightarrow h_2^{mN}$$

is the orthogonal projection onto the range of \mathbf{R}^* , and that

$$\pi_{\mathbf{0}} := \mathbf{0}(\mathbf{0}^*\mathbf{0})^{-1}\mathbf{0}^* : h_2^{pN} \rightarrow h_2^{pN}$$

is the orthogonal projection onto the range of $\mathbf{0}$. It is now straightforward to verify that

$$(2.5a) \quad \mathbf{U}\mathbf{U}^* = I_{mN} + \pi_{\mathbf{R}^*},$$

$$(2.5b) \quad \mathbf{U}^*\mathbf{U} = I_{mN+n},$$

$$(2.5c) \quad \mathbf{V}^*\mathbf{V} = I_{pN} + \pi_{\mathbf{0}},$$

$$(2.5d) \quad \mathbf{V}\mathbf{V}^* = I_{pN+n}.$$

We now prove that $\Gamma_e = \mathbf{V}^*H_e\mathbf{U}^*$. Note first that $H_e = \mathbf{V}\Gamma_e\mathbf{U}$. Then

$$\begin{aligned} \mathbf{V}^*H_e\mathbf{U}^* &= \mathbf{V}^*\mathbf{V}\Gamma_e\mathbf{U}\mathbf{U}^* \\ &= \begin{bmatrix} I_{pN} & 0 \\ 0 & \pi_{\mathbf{0}} \end{bmatrix} \begin{bmatrix} D & H\mathbf{R} \\ \mathbf{0}G & \mathbf{0}F\mathbf{R} \end{bmatrix} \begin{bmatrix} I_{mN} & 0 \\ 0 & \pi_{\mathbf{R}^*} \end{bmatrix} \\ &= \begin{bmatrix} D & H\mathbf{R}\pi_{\mathbf{R}^*} \\ \pi_{\mathbf{0}}\mathbf{0}G & \pi_{\mathbf{0}}\mathbf{0}F\mathbf{R}\pi_{\mathbf{R}^*} \end{bmatrix}. \end{aligned}$$

But $\pi_{\mathbf{0}}\mathbf{0} = \mathbf{0}$, and $\mathbf{R}\pi_{\mathbf{R}^*} = \mathbf{R}[\mathbf{R}^*(\mathbf{R}\mathbf{R}^*)^{-1}\mathbf{R}] = \mathbf{R}$. Consequently,

$$\mathbf{V}^*H_e\mathbf{U}^* = \begin{bmatrix} D & H\mathbf{R} \\ \mathbf{0}G & \mathbf{0}F\mathbf{R} \end{bmatrix} = \Gamma_e.$$

Finally, from (2.5b) and (2.5d) it follows that

$$\|\Gamma_e\| = \|H_e\|,$$

and this completes the proof. \square

Thus, our original problem has now become: *Obtain D subject to the original constraints, so that it minimizes $\|H_e\|$.* Note that the entries of H_e are now finite matrices, and moreover, the quantities Σ, Δ can be obtained from (F, G, H) by first computing P and Q as solutions to the following Lyapunov equations (see [9]):

$$(2.6a) \quad P = FPF^* + GG^*,$$

$$(2.6b) \quad Q = F^*QF + H^*H,$$

and then taking the Hermitian square roots of P and Q (see (2.3)). Below, we focus on how to explicitly compute such a D , given F, G, H, Σ, Δ .

Therefore, our problem has now been reduced to obtain a $pN \times mN$ matrix D whose $(p \times m)$ -block entries above the main diagonal are completely specified by γ_0 , and which minimizes

$$\|H_e\| = \left\| \begin{bmatrix} D & H\Sigma \\ \Delta G & \Delta F\Sigma \end{bmatrix} \right\|.$$

Let D be represented by

$$D = \begin{bmatrix} x_{11} & d_{12} & d_{13} & \cdots & d_{1N} \\ x_{21} & x_{22} & d_{23} & \cdots & d_{2N} \\ \vdots & \vdots & \vdots & & \vdots \\ x_{N1} & x_{N2} & x_{N3} & \cdots & x_{NN} \end{bmatrix},$$

where the entries indicated are $p \times m$ matrices, the d 's being specified by γ_0 and the x 's representing the entries to be filled in. Define

$$\begin{aligned}\hat{H}_e &:= \begin{bmatrix} \Delta G & \Delta F \Sigma \\ D & H \Sigma \end{bmatrix}, \\ L_k &:= \begin{bmatrix} I_{n+kp} & 0 \\ 0 & 0 \end{bmatrix}, \\ R_k &:= \begin{bmatrix} 0 & 0 \\ 0 & I_{n+km} \end{bmatrix}.\end{aligned}$$

We now have the following proposition.

PROPOSITION 2.7. $\mu = \max \{ \|L_1 \hat{H}_e R_{N-1}\|, \dots, \|L_N \hat{H}_e R_0\| \}$.

The matrix $L_K \hat{H}_e R_{N-k}$ is the *top right* $(n+km) \times (n+Nm-km)$ submatrix of \hat{H}_e . Note that $L_k \hat{H}_e R_{N-k}$, for $k=0, 1, \dots, N$ are all the maximal rectangular submatrices of \hat{H}_e whose entries do not depend on the variables x . (It will be seen that the norm of $L_0 \hat{H}_e R_N$ is equal to the norm of $L_N \hat{H}_e R_0$, and this is why only one of the two appears in the expression of the above proposition.) The proof is based on the following very important result—see the book by Power [12] and also [3] for a proof of this result.

LEMMA 2.8 (Parrott, Davis–Kahan–Weinberger). *Consider the block matrix*

$$M(X) = \begin{bmatrix} C & A \\ X & B \end{bmatrix},$$

where A, B, C, X are matrices of compatible dimensions. Then

$$\alpha = \inf_X \|M(X)\| = \max \left\{ \|(C \ A)\|, \left\| \begin{pmatrix} A \\ B \end{pmatrix} \right\| \right\}.$$

Moreover, the above infimum is attained by the choice

$$(2.9) \quad X = -BA^*(\alpha^2 I - AA^*)^{-1}C.$$

In case the indicated inverse does not exist, then this should be interpreted as a pseudo-inverse. Also, in [3], one can find a description of all possible choices for X that attain this infimum.

Proof of Proposition 2.7. By the results of Adamjan, Arov and Krein [1], as we mentioned earlier, it follows that

$$\mu = \inf_{x_{ik}} \|\Gamma_e\|.$$

From Proposition 2.4 we now have that

$$\mu = \inf_{x_{ik}} \|H_e\| = \inf_{x_{ik}} \|\hat{H}_e\|,$$

where the last equality follows from the unitary equivalence of H_e and \hat{H}_e . By repeated application of Lemma 2.8 we now have that

$$\begin{aligned}\mu &= \max \{ \|(L_N \hat{H}_e R_0)\|, \inf_{\substack{x_{j1}, \dots, x_{jj} \\ 1 \leq j \leq N-1}} \|(L_{N-1} \hat{H}_e R_N)\| \} \\ &\quad \vdots \\ &= \max \{ \|(L_N \hat{H}_e R_0)\|, \dots, \|L_2 \hat{H}_e R_{N-2}\|, \inf_{x_{11}} \|L_1 \hat{H}_e R_N\| \} \\ &= \max \{ \|(L_N \hat{H}_e R_0)\|, \|(L_{N-1} \hat{H}_e R_1)\|, \dots, \|(L_0 \hat{H}_e R_N)\| \}.\end{aligned}$$

A final point to be noted is that the first and the last term of the above expression are equal. We have

$$\|(L_N \hat{H}_e R_0)\| = \left\| \begin{pmatrix} \Delta F \Sigma \\ H \Sigma \end{pmatrix} \right\| = \left\| \begin{pmatrix} H \Sigma \\ \Delta F \Sigma \end{pmatrix} \right\|.$$

But

$$\mathbf{V}^* \begin{bmatrix} H \Sigma \\ \Delta F \Sigma \end{bmatrix} \Sigma^{-1} \mathbf{R} = \begin{bmatrix} H \mathbf{R} \\ 0 F \mathbf{R} \end{bmatrix} = \Gamma,$$

and

$$\mathbf{V} \mathbf{V}^* = I_{n+pN} \quad \text{and} \quad \Sigma^{-1} \mathbf{R} \mathbf{R}^* \Sigma = I_n.$$

Consequently, $\|L_N \hat{H}_e R_0\| = \|\Gamma\|$. It follows similarly that $\|L_0 \hat{H}_e R_N\| = \|\Gamma\|$, and this completes the proof. \square

It can be readily shown that the matrices

$$L_k \hat{H}_e R_{N-k}, \quad k = 1, \dots, N,$$

are directly related to the operators $A_k, k = 1, \dots, N$ respectively, of Feintuch, Khargonekar and Tannenbaum [6], and that in fact they have equal norms. Thus, in Proposition 2.7, we have a formula for the optimal sensitivity μ , that is equivalent to the one given in [6]. But, in addition to that, the sequence of steps (2.9) in the proof of the Proposition 2.7 leads to the following procedure to obtain a $V(z)$ that solves the problem (1.1).

ALGORITHM.

1. Obtain a minimal realization (F, G, H) of the negative Fourier coefficients $(\gamma_k: k \leq -1)$ of $T(z)$; i.e., (F, G, H) satisfy (2.1).

2. Obtain P, Q by solving Lyapunov equations (2.6a) and (2.6b) and find their Hermitian square roots Σ, Δ , respectively.

3₁. Select x_{11} using (2.9) to minimize $\|L_1 \hat{H}_e R_N\|$. Note that in $L_1 \hat{H}_e R_N$, the only variable is x_{11} and the rest of $L_1 \hat{H}_e R_N$ is completely specified.

3₂. Select (x_{21}, x_{22}) to minimize $\|L_2 \hat{H}_e R_N\|$. Note that $L_2 \hat{H}_e R_N$ contains x_{11}, x_{21}, x_{22} as the only variables. In this step x_{11} obtained in step 3₁ is used. Again (2.9) is used to select (x_{21}, x_{22}) .

3_j. Select $(x_{j1}, x_{j2}, \dots, x_{jj})$ to minimize $\|L_j \hat{H}_e R_N\|$ for $j = 3, 4, \dots, N$, where the entries x_{k1}, \dots, x_{kk} , for $1 \leq k \leq j-1$ have already been determined at the previous steps. This is done by applying (2.9) to the corresponding submatrices of $L_j \hat{H}_e R_N$.

4. Let $v_0 = \gamma_0 - D$. Now using techniques of Glover [9] obtain

$$\tilde{V}(z) = \sum_{i=1}^{\infty} v_i z^{-i}$$

such that

$$\mu = \|T(z) - v_0 - \tilde{V}(z)\|_{\infty}.$$

For this last step, see also Ball and Ran [2]. Then $V(z) = v_0 + \tilde{V}(z)$ is a solution to (1.1).

We would like to note that only the elements of D are calculated by recursively applying (2.9) $(N-1)$ times. After obtaining D , the procedure of Glover [9] or Ball and Ran [2] can be used to obtain $V(z)$.

REFERENCES

- [1] V. M. ADAMJAN, D. Z. AROV AND M. G. KREIN, *Infinite Hankel block matrices and related extension problems*, AMS Trans., (2) 111 (1978), pp. 133-156.
- [2] J. A. BALL AND A. C. M. RAN, *Optimal Hankel norm reductions and Wiener-Hopf factorization I: The canonical case*, Virginia Polytechnic Inst. and State Univ., 1985, preprint.
- [3] C. DAVIS, W. M. KAHAN AND H. F. WEINBERGER, *Norm preserving dilations and their applications to optimal error bounds*, SIAM J. Numer. Anal., 10 (1982), pp. 445-469.
- [4] J. C. DOYLE, *Synthesis of robust controllers and filters*, Proc. 22nd IEEE Conf. on Decision and Control, 1983, pp. 109-114.
- [5] A. FEINTUCH AND B. A. FRANCIS, *Uniformly optimal control of linear feedback systems*, Automatica, 21 (1985), pp. 563-574.
- [6] A. FEINTUCH, P. P. KHARGONEKAR AND A. TANNENBAUM, *On the sensitivity minimization problem for linear time-varying periodic systems*, this Journal, 24 (1986), pp. 1076-1085.
- [7] B. A. FRANCIS AND J. C. DOYLE, *Linear control theory with an H_∞ -optimality criterion*, to appear.
- [8] P. A. FUHRMANN, *Linear Systems and Operators in Hilbert Space*, McGraw-Hill, New York, 1981.
- [9] K. GLOVER, *All optimal Hankel-norm approximations of linear multivariable systems and their L^∞ -error bounds*, Internat. J. Control, 39 (1984), pp. 1115-1193.
- [10] P. P. KHARGONEKAR, K. POOLLA AND A. TANNENBAUM, *Robust control of linear time-invariant plants using periodic compensation*, IEEE Trans. Automat. Control, AC-30 (1986), pp. 1088-1096.
- [11] P. P. KHARGONEKAR AND E. D. SONTAG, *On the relation between stable matrix fraction factorizations and regulable realizations of linear systems over rings*, IEEE Trans. Automat. Control, AC-27 (1982), pp. 627-638.
- [12] S. C. POWER, *Hankel Operators on Hilbert Space*, Pitman, Boston, MA, 1982.
- [13] M. VIDYASAGAR, *Control Systems Synthesis-A Factorization Approach*, The MIT Press, Cambridge, MA, 1985.
- [14] G. ZAMES, *Feedback and optimal sensitivity*, IEEE Trans. Automat. Control, AC-26 (1981), pp. 301-320.

THE MAXIMUM PRINCIPLE FOR OPTIMAL CONTROL OF DIFFUSIONS WITH PARTIAL INFORMATION*

U. G. HAUSSMANN†

Abstract. We derive necessary conditions for the optimal control of a system that satisfies an Ito equation with control entering the drift term. The control is a function of a noise corrupted observation of the state, and the cost is the expectation of an integral and a final term. The robust form of the Zakai equation from nonlinear filtering is used to compute the variation of the cost due to a strong or "needle" variation of a control. This gives rise to an explicit formula for the adjoint process, much as in the case with complete information.

Key words. maximum principle, optimal control, partial information, stochastic differential equation, Zakai equation, nonlinear filtering

AMS(MOS) subject classification. 49B60

1. Introduction. Necessary conditions, satisfied by solutions of the problem,

$$(1.1) \quad \min \{J(u): u \in \mathcal{U}\},$$

$$(1.2) \quad J(u) = E \left\{ \int_0^T l(t, X_t, u(t, Y)) dt + c(X_T) \right\},$$

$$(1.3) \quad dX_t = f(t, X_t, u(t, Y)) dt + \sigma(t, X_t) dw_t, \quad X_0 = x_0,$$

$$(1.4) \quad dY_t = h(t, X_t) dt + d\tilde{w}_t, \quad Y_0 = 0,$$

have recently been given by Bensoussan [3]. The control u may depend only on Y , not X . In addition to the uniform ellipticity of $\sigma(t, x)\sigma(t, x)'$ (here σ' is the transpose of σ), he assumed much differentiability and boundedness of the functions l, c, f, σ and h as well as convexity and compactness of the set of control points U . In this work we relax most of these hypotheses thus including the linear regulator. The more difficult problem when constraints are included will be attacked elsewhere. We point out also that the method (of strong variations) cannot be extended to allow σ to depend on the control. Bensoussan [3] used the (stochastic) Zakai equation of nonlinear filtering to define the state of the separated problem and then he applied the method of weak variations to obtain the necessary conditions. The improvement here stems from using weak solutions of the robust (nonstochastic) form of the Zakai equation (Hausmann [7]), from using strong variations, and from finding an explicit representation for the adjoint process.

An attempt on this problem was made by Kushner [9], Hausmann [5], Elliott [4], Arkin and Saksonov [1], where in posing the control problem, the control u is taken to be adapted to a filtration which may be smaller than that generated by the joint process (X, Y) . Unfortunately in the work of Kushner [9] and of Arkin and Saksonov [1] this filtration must be specified a priori, i.e. independently of the control u , and hence *cannot* be the one generated by Y . This difficulty does not arise in Hausmann [5] or Elliott [4], but in those works no representation of the adjoint process is given, so again the result is not a viable maximum principle. Additionally

* Received by the editors September 18, 1985; accepted for publication (in revised form) February 7, 1986. This work was supported by the Natural Sciences and Engineering Research Council of Canada under grant A8051 and was carried out in part while the author visited the Université de Paris IX, Paris, France.

† Mathematics Department, University of British Columbia, Vancouver, Canada V6T 1Y4.

in these last two works σ must be nonsingular, and in the work of Elliott no state constraints are admitted and the method is inherently incapable of treating them.

In § 2 we give a precise formulation of the problem and we summarize the results from filtering theory which will be used in the sequel. In § 3 we compute the variation in J due to a variation in u , and in § 4 we apply this calculation to derive the maximum principle. We also relate this result to that of Bensoussan [3], and we interpret the adjoint process in terms of the value function. Finally as an example we consider the linear regulator.

The results established here were reported in Haussmann [6].

2. Formulation of the problem. We say that a function $f: \mathcal{X} \times \mathcal{Y} \rightarrow \mathcal{Z}$, where $\mathcal{X}, \mathcal{Y}, \mathcal{Z}$ are subsets of Euclidean spaces, is *locally uniformly continuously differentiable in x* if $f(\cdot, y)$ is continuously differentiable for each y , and for (x, y) in a compact set C , the modulus of continuity of $x \rightarrow f_x(x, y)$ depends only on C .

The following hypotheses are assumed. U is a Borel set in some Euclidean space, and \mathcal{B}^n is the family of Borel sets of \mathbf{R}^n .

(A₁) $f: [0, T] \times \mathbf{R}^n \times U \rightarrow \mathbf{R}^n$ is Borel measurable, continuous in u for each (t, x) , locally uniformly continuously differentiable in x , and for some constant K_1 ,

$$(1 + |x| + |u|)^{-1} |f(t, x, u)| + |f_x(t, x, u)| \leq K_1;$$

(A₂) $\sigma: [0, T] \times \mathbf{R}^n \rightarrow \mathbf{R}^n \otimes \mathbf{R}^m$ is Borel measurable, locally uniformly continuously differentiable in x , and for some K_2 ,

$$|\sigma(t, x)| + |\sigma_x(t, x)| \leq K_2;$$

(A₃) $h: [0, T] \times \mathbf{R}^n \rightarrow \mathbf{R}^d$ is Borel measurable, locally uniformly continuously differentiable in x , and for some K_3 ,

$$(1 + |x|)^{-1} |h(t, x)| + |h_x(t, x)| \leq K_3;$$

(A₄) $l: [0, T] \times \mathbf{R}^n \times U \rightarrow \mathbf{R}$ is Borel measurable, continuous in u for each (t, x) , locally uniformly continuously differentiable in x , and for some constants $q \geq 1, K_4$,

$$|l(t, x, u)| + |l_x(t, x, u)| \leq K_4(1 + |x|^q + |u|^q);$$

(A₅) $c: \mathbf{R}^n \rightarrow \mathbf{R}$ is continuously differentiable and for some K_5 ,

$$|c(x)| + |c_x(x)| \leq K_5(1 + |x|^q);$$

(A₆) P_0 is a probability measure on $(\mathbf{R}^n, \mathcal{B}^n)$ such that for some $\bar{q} > q + 1$,

$$\int |x|^{\bar{q}} P_0(dx) < \infty.$$

We use the following notation:

c_x is the vector $(c_{x_i}) = (\partial c / \partial x_i)$,

h^i, σ^{ij} are the components of h, σ ,

h_x is the matrix $(h_x)^{ij} = h_{x_j}^i$,

σ_x is the tensor $(\sigma_x)^{ijk} = \sigma_{x_k}^{ij}$,

$\mathcal{C}(0, T; \mathbf{R}^k)$ is the space of continuous functions $[0, T] \rightarrow \mathbf{R}^k$,

$\{\mathcal{G}_t^k\}$ is the canonical Borel filtration on $\mathcal{C}(0, T; \mathbf{R}^k)$,

$\|y\|_t = \sup \{|y(s)|: s \leq t\}$.

An *admissible control* is a function u

$$u: [0, T] \times \mathcal{C}(0, T; \mathbf{R}^d) \rightarrow U$$

which is Borel measurable, $\{\mathcal{G}_t^d\}$ adapted, and for which there exists a constant K_u such that $|u(t, y)| \leq K_u(1 + \|y\|_t)$. We write \mathcal{U} for the set of all admissible controls.

For the underlying probability space we take the canonical space of (x_0, w, Y) where (w, Y) is a standard Brownian motion on $\mathbf{R}^m \times \mathbf{R}^d$ and x_0 is an independent random variable with distribution P_0 . Hence

$$\Omega = \mathbf{R}^n \times \mathcal{C}(0, T; \mathbf{R}^m) \times \mathcal{C}(0, T; \mathbf{R}^d),$$

$$\mathcal{F} = \mathcal{B}^n \otimes \mathcal{G}_T^m \otimes \mathcal{G}_T^d,$$

$$P = P_0 \times P_w^m \times P_w^d,$$

where P_w^k is the Wiener measure on $\mathcal{C}(0, T; \mathbf{R}^k)$. The generic element ω of Ω is denoted by

$$\omega = (\xi, \zeta, \eta).$$

For $u \in \mathcal{U}$, $x \in \mathbf{R}^n$, $s \in [0, T]$, let X^u denote the unique strong solution of

$$(2.1) \quad \begin{aligned} dX_t &= f(t, X_t, u(t, Y)) dt + \sigma(t, X_t) dw_t, \quad t \geq 0, \\ X_0 &\sim P_0 \end{aligned}$$

and let X^{usx} denote the unique strong solution of

$$(2.1)' \quad \begin{aligned} dX_t &= f(t, X_t, u(t, Y)) dt + \sigma(t, X_t) dw_t, \quad t \geq s, \\ X_s &= x. \end{aligned}$$

Moreover, for $\eta \in \mathcal{C}(0, T; \mathbf{R}^d)$, let $X^{u\eta sx}$ be the unique strong solution of

$$(2.2) \quad \begin{aligned} dX_t &= f(t, X_t, u(t, \eta)) dt + \sigma(t, X_t) dw_t, \quad t \geq s, \\ X_s &= x. \end{aligned}$$

Note that $X_t^u(\omega) = X_t^{u0\xi}(\omega)$ P -a.s., and $X_t^{usx}(\omega) = X_t^{u\eta sx}(\omega)$ P -a.s.

From Girsanov's theorem it now follows that if

$$(2.3) \quad Z_t^u = \exp \left\{ \int_0^t h(r, X_r^u) \cdot dY_r - \frac{1}{2} \int_0^t |h(r, X_r^u)|^2 dr \right\},$$

and if $dP^u := Z_T^u dP$, then (X^u, Y) is a solution of (1.3), (1.4) on $(\Omega, \mathcal{F}, P^u)$, i.e., is a weak solution of (1.3), (1.4)—which moreover is unique in law. In addition,

$$J(u) = E^u \left\{ \int_0^T l(t, X_t^u, u(t, Y)) dt + c(X_T^u) \right\},$$

i.e.,

$$(2.4) \quad J(u) = E \left\{ Z_T^u \left[\int_0^T l(t, X_t^u, u(t, Y)) dt + c(X_T^u) \right] \right\}$$

where E^u denotes expectation on $(\Omega, \mathcal{F}, P^u)$. So the problem becomes

$$(2.5) \quad \min \{J(u): u \in \mathcal{U}\}$$

where J is defined by (2.4), (2.3) and (2.1).

Later we shall also require the density

$$Z_t^{usx} := \exp \left\{ \int_s^t h(r, X_r^{usx}) \cdot dy_r - \frac{1}{2} \int_s^t |h(r, X_r^{usx})|^2 dr \right\}$$

on (Ω, \mathcal{F}, P) . Again note that

$$Z_t^u(\omega) = Z_t^{u0\xi}(\omega) \quad P\text{-a.s.}$$

and that $Z_t^{u0\xi}(\omega)$ is in fact constant as a function of ξ P -a.s. (n.b. $\omega = (\xi, \zeta, \eta)$).

In what follows we shall use some concepts from differential equations for which we now introduce some notation. Let $H = L^2(R^n)$ with inner product $\langle v, g \rangle$. Define

$$H^1 := \{v \in H : \partial_i v \in H, i = 1, \dots, n\}$$

where ∂_i is the distributional derivative with respect to x_i (and ∂_0 with respect to t). Let H^{-1} be the dual of H^1 with the pairing (v, g) , $v \in H^{-1}$, $g \in H^1$. The norm in H is written as $|\cdot|_H$ and in H^1 as $\|\cdot\|$, i.e.,

$$\|v\|^2 = |v|_H^2 + \sum_{i=1}^n |\partial_i v|_H^2.$$

Then define

$$W(0, T) := \left\{ v \in L^2(0, T; H^1) : \frac{dv}{dt} \in L^2(0, T; H^{-1}) \right\}.$$

Note that dv/dt is the distributional derivative of $t \rightarrow v(t) \in H^1$ —for details cf. Bensoussan [2].

Let us use the convention that repeated indices are to be summed. Then define $a(t, x) = \sigma(t, x)\sigma(t, x)'$, where $'$ denotes transpose,

$$a_t^{ij} : x \rightarrow a^{ij}(t, x),$$

$$b_t^{u\eta i} : x \rightarrow f^i(t, x, u(t, \eta)) - \frac{1}{2} a^{ij}(t, x) \partial_j [\eta(t) \cdot h(t, x)] - \frac{1}{2} \partial_j a^{ij}(t, x),$$

$$h_t : x \rightarrow h(t, x),$$

$$\partial_0 h_t : x \rightarrow \partial_0 h(t, x) \quad (\text{when this is a function}),$$

$$F_t^{\mu\eta} : x \rightarrow l(t, x, u(t, \eta)) \exp [\eta(t) \cdot h(t, x)],$$

$$\eta_t = \eta(t).$$

It will be useful to impose temporarily an additional hypothesis.

(H)(i) $h \in C^{1,2}((0, T) \times \mathbf{R}^n)$ and there exist constants K_0, α with $\alpha > 0$ such that for all $t \in [0, T]$, $x \in \mathbf{R}^n$, $u \in U$

$$|f(t, x, u)| + |\sigma(t, x)| + |h(t, x)| + |\partial_0 h(t, x)| + \sum |\partial_i h(t, x)| + \sum |\partial_i \partial_j h(t, x)| \leq K_0,$$

$$a(t, x) \leq \alpha I,$$

$$|c|_H < \infty$$

for each $u \in \mathcal{U}$ and any compact $C \subset \mathcal{C}(0, T; \mathbf{R}^d)$

$$\sup_{\eta \in C} \int_0^T |l(t, \cdot, u(t, \eta))|_H^2 dt < \infty.$$

(H)(ii) P_0 has a density p_0 with compact support.

Now we can define, for each $t \in [0, T)$, $u \in \mathcal{U}$, $\eta \in \mathcal{C}(0, T; \mathbf{R}^d)$ a bilinear form on H^1 :

$$\begin{aligned} \mathcal{A}_t^{u\eta}(\mu, v) = & -\frac{1}{2}\langle a_t^{ij}\partial_i\mu, \partial_jv \rangle + \langle b_t^{u\eta i}\partial_i\mu, v \rangle + \frac{1}{2}\langle \partial_i(\eta_i \cdot h_t) a_t^{ij}\mu, \partial_jv \rangle \\ & - \langle [\partial_i(\eta_i \cdot h_t) b_t^{u\eta i} + \eta_i \cdot \partial_0 h_t + \frac{1}{2}|h_t|^2]\mu, v \rangle. \end{aligned}$$

The following results are established in Haussmann [7]. Let $\mathcal{F}_t^y = \{\phi, \mathbf{R}^n \times \mathcal{C}(0, T; \mathbf{R}^m)\} \otimes \mathcal{G}_t^d$ be the σ -algebra generated by $\{Y_s: 0 \leq s \leq t\}$, and let G be a measurable function

$$G: [0, T] \times \mathbf{R}^n \times \mathcal{C}(0, T; \mathbf{R}^d) \rightarrow \mathbf{R}$$

such that for any compact set C

$$(2.6) \quad \sup_{\eta \in C} \int_0^T \int_{\mathbf{R}^n} |G(t, x, \eta)|^2 dx dt < \infty.$$

PROPOSITION 2.1. Assume (A_1) – (A_6) and (H) . Let $u \in \mathcal{U}$ and write $G_t^\eta(x) = G(t, x, \eta)$. Then

(a) for each $\eta \in \mathcal{C}(0, T; \mathbf{R}^d)$ there exists a unique $\mu^{u\eta} \in W(0, t)$ such that

$$(2.7) \quad \left(\frac{d\mu_s^{u\eta}}{ds}, v \right) + \mathcal{A}_s^{u\eta}(\mu_s^{u\eta}, v) + \langle G_s^\eta \exp(\eta_s \cdot h_s), v \rangle = 0, \quad v \in H^1, \quad 0 \leq s \leq t,$$

$$\mu_t^{u\eta} = c \exp(\eta_t \cdot h_t);$$

(b) for almost all x

$$\mu_s^{uY(\omega)}(x) = E \left\{ c(X_t^{usx}) Z_t^{usx} + \int_s^t G_r^{Y(\omega)}(X_r^{usx}) Z_r^{usx} dr \middle| \mathcal{F}_T^Y \right\} (\omega) \exp[Y_s(\omega) \cdot h_s(x)];$$

(c) for each $\eta \in \mathcal{C}(0, T; \mathbf{R}^d)$ there exists a unique $\psi^{u\eta} \in W(0, T)$ such that

$$(2.8) \quad \left(\frac{d\psi_t^{u\eta}}{dt}, v \right) - \mathcal{A}_t^{u\eta}(v, \psi_t^{u\eta}) = 0 \quad \forall v \in H^1, \quad t \geq 0,$$

$$\psi_0^{u\eta} = p_0;$$

(d) $\rho_t^{uY(\omega)}(\cdot) := \psi_t^{uY(\omega)}(\cdot) \exp[Y_t(\omega) \cdot h_t(\cdot)]$ is an unnormalised conditional density of $X_t^u | \mathcal{F}_t^Y$.

Proof. In Haussmann [7, Cor. 2.1, Cor. 3.2 and Thm. 3.4] the result is established under the added hypothesis that

$$(2.9) \quad \sup_{\eta \in C} |G(t, x, \eta)| \leq K(1 + |x|^q).$$

Note that Corollary 3.1 as stated is false: in fact ν must be continuous and this hypothesis must be added throughout. Since we have $\partial_i \partial_j h$ bounded, then we can avoid (2.9) as follows. We take $c = 0$, since the quoted result covers the case $c \neq 0$, $G = 0$. Recalling that $\omega = (\xi, \zeta, \eta)$ and $Y(\omega) = \eta$ a.s. we have from Haussmann [7] that

$$\begin{aligned} \mu_s^{u\eta}(x) &= E \left\{ \int_s^t G_r^\eta(X_r^{usx}) Z_r^{usx} dr \middle| \mathcal{F}_T^Y \right\} (\omega) e^{\eta_s \cdot h_s(x)} \quad \text{a.s.} \\ &= E^\eta \int_s^t G_r^\eta(X_r^{u\eta sx}) e^{\delta_r(\eta)} dr \end{aligned}$$

provided G is also bounded. Here E^η is expectation under P^η , a measure obtained from P by a Girsanov transformation, and $X^{u\eta sx}$ satisfies

$$(2.10) \quad \begin{aligned} dX &= [f(t, X, u(t, \eta)) - a(t, X) \nabla_x (\eta_t \cdot h_t(X))] dt + \sigma(t, X) d\bar{w}, \quad t \geq s, \\ X_s &= x, \end{aligned}$$

where \bar{w} , Y are independent Brownian motions under P^η . Moreover,

$$(2.11) \quad \begin{aligned} \delta_r(\eta) &= \eta_r \cdot h_r(X_r^{u\eta sx}) - \int_s^r [\eta_\theta \cdot \partial_\theta h_\theta(X_\theta^{u\eta sx}) + \tilde{L}_\theta^\eta(\eta_\theta \cdot h_\theta)(X_\theta^{u\eta sx}) \\ &\quad + \frac{1}{2} |h_\theta(X_\theta^{u\eta sx})|^2] d\theta, \\ \tilde{L}_\theta^\eta \phi(x) &= \frac{1}{2} a_\theta^{ij}(x) \partial_i \partial_j \phi(x) + f(\theta, x, u(\theta, \eta)) \cdot \nabla \phi(x) - \frac{1}{2} |\sigma(\theta, x)' \nabla \phi(x)|^2, \end{aligned}$$

so that

$$\sup_{\eta \in C} \sup_{u, r, x} |\delta_r(\eta)| \leq K < \infty,$$

with K depending only on C and K_0 , cf. (H).

Suppose now that G is not bounded. Let

$$G_m(t, x, \eta) = \begin{cases} G(t, x, \eta) & \text{if } |G(t, x, \eta)| \leq m, \\ G(t, x, \eta) \frac{m}{|G(t, x, \eta)|} & \text{otherwise.} \end{cases}$$

Then G_m is bounded, $|G_m| \uparrow |G|$. If we write $\mu_s^{m\eta}(x)$ for $\mu_s^{u\eta}(x)$ when G_m is used, then as in Haussmann [7, Thm. 3.2] it follows that $|\mu_s^{m\eta} - \mu_s^{u\eta}|_H \rightarrow 0$ for each η in $\mathcal{C}(0, T; \mathbb{R}^d)$. Let $\rho \in H$ be a probability density. Then

$$\langle \mu_s^{m\eta}, \rho \rangle \rightarrow \langle \mu_s^{u\eta}, \rho \rangle$$

and

$$\langle \mu_s^{m\eta}, \rho \rangle = E^\eta \int_s^t G_{mr}^\eta(X_r^{u\eta sp}) e^{\delta_r(\eta, \rho)} dr$$

where $\delta_r(\eta, \rho)$ is defined by (2.11) with $X^{u\eta sx}$ replaced by $X^{u\eta sp}$ where $X^{\mu\eta sp}$ is the unique solution of (2.10) with initial distribution

$$X_s \sim \rho(x) dx.$$

Since $|G_m^\eta(\cdot)| \uparrow |G^\eta(\cdot)|$ a.e. ($dt dx$), and since $X^{u\eta sp}$ under P^η has a density $p_r^\eta(x)$ in $L^2(s, T; H^1)$, then $|G_{mr}^\eta(X_r^{u\eta sp}) e^{\delta_r(\eta, \rho)}| \leq |G_r^\eta(X_r^{u\eta sp})| e^K$ and recalling (2.6) we have

$$E^\eta \int_s^t |G_r^\eta(X_r^{u\eta sp})| dr = \int \int |G_r^\eta(x) p_r^\eta(x) dx dr| < \infty.$$

Hence, by the above and the dominated convergence theorem,

$$\begin{aligned} \langle \mu_s^{u\eta}, \rho \rangle &= \lim_m \langle \mu_s^{m\eta}, \rho \rangle \\ &= \lim_m E^\eta \int_s^t G_{mr}^\eta(X_r^{u\eta sp}) e^{\delta_r(\eta, \rho)} dr \\ &= E^\eta \int_s^t G_r^\eta(X_r^{u\eta sp}) e^{\delta_r(\eta, \rho)} dr \\ &= \left\langle E^\eta \int_s^t G_r^\eta(X_r^{u\eta sx}) e^{\delta_r(\eta)} dr, \rho \right\rangle \end{aligned}$$

$$= \left\langle E \left\{ \int_s^t G_r^\eta(X_r^{usx}) Z_r^{usx} dr \middle| \mathcal{F}_T^Y \right\} (\omega) e^{\eta_s \cdot h_s(x)}, \rho \right\rangle.$$

Since the probability densities form a total set in H then the result follows.

We add that, from the usual energy estimates, it follows that for $i = 1, \dots, n$

$$(2.12) \quad \int_0^t |\partial_i \mu_s^{u\eta}|^2 ds \leq K \left\{ |c e^{\eta_i \cdot h_i}|_H^2 + \int_0^t \left| G_s^\eta e^{\eta_s \cdot h_s} \right|_H^2 ds \right\}.$$

3. Strong variations. Let \hat{u} be a solution of (2.5). For s, ε fixed such that $0 \leq s \leq s + \varepsilon \leq T$, a *strong variation* (corresponding to (s, ε)) is an element $u \in \mathcal{U}$ such that

$$u(t, \eta) = \hat{u}(t, \eta), \quad t \notin [s, s + \varepsilon].$$

In this section we study the resultant perturbation of J , i.e., $J(u) - J(\hat{u})$.

We begin with some definitions. Let $\hat{\Phi}_t^{sx}$ be the fundamental matrix solution of

$$(3.1) \quad dz_t = f_x(t, \hat{X}_t^{sx}, \hat{u}(t, Y)) z_t dt + \sigma_x^{(i)}(t, \hat{X}_t^{sx}) z_t dw^i, \quad t \geq s$$

where \hat{X}_t^{sx} is the solution of (2.1)' with $u = \hat{u}$ and $\sigma^{(i)}$ is the i th column of σ . Hence each column of $\hat{\Phi}_t^{sx}$ satisfies (3.1) and $\hat{\Phi}_s^{sx} = I$ a.s. Assume (H) and let $\hat{\mu}^\eta$ be the function $\mu^{u\eta}$ of Proposition 2.1(a) with $u = \hat{u}$, with $G(t, x, \eta) = l(t, x, \hat{u}(t, \eta))$ and with terminal condition at $t = T$. With $\hat{Z}_t = Z_t^{\hat{u}}$ and $\hat{X}_t = X_t^{\hat{u}}$ it follows from Proposition 2.1(b) and (2.4) that

$$(3.2) \quad \begin{aligned} J(\hat{u}) &= EE \left\{ c(\hat{X}_T) \hat{Z}_T + \int_0^T l(t, \hat{X}_t, \hat{u}(t, Y)) \hat{Z}_t dt \middle| \mathcal{F}_T^Y \vee \hat{X}_0 \right\} \\ &= E \hat{\mu}_0^Y(\hat{X}_0) \\ &= E \langle \hat{\mu}_0^Y, p_0 \rangle. \end{aligned}$$

We now set $\hat{v}_t^\eta = \hat{\mu}_t^\eta \exp(-\eta_t \cdot h_t)$. Observe that $\hat{\mu}_t^\eta$, hence \hat{v}_t^η , depends on the past of η through \hat{u} and on the future through the dynamics (2.7). Our aim is to compute $E\{\partial \hat{v}_s^Y(x) | \mathcal{F}_s^Y\}$ where ∂v is the vector with components $\partial_i v$ for $v \in H^1$.

LEMMA 3.1. Assume (A₁)–(A₆) and (H). For each s and almost all x

$$(3.3) \quad \begin{aligned} E\{\partial \hat{v}_s^Y(x) | \mathcal{F}_s^Y\} &= E \left\{ \left[c_x(\hat{X}_T^{sx}) \hat{\Phi}_T^{sx} + \int_s^T l_x(t, \hat{X}_t^{sx}, \hat{u}(t, Y)) \hat{\Phi}_t^{sx} dt \right. \right. \\ &\quad \left. \left. + \left(c(\hat{X}_T^{sx}) + \int_s^T l(t, \hat{X}_t^{sx}, \hat{u}(t, Y)) dt \right) \right. \right. \\ &\quad \left. \left. \cdot \int_s^T h_x^i(t, \hat{X}_t^{sx}) \hat{\Phi}_t^{sx} (dY_t^i - h^i(t, \hat{X}_t^{sx}) dt) \right] \hat{Z}_T^{sx} \middle| \mathcal{F}_s^Y \right\} \quad \text{a.s.} \end{aligned}$$

Note that the right side of (3.3) is defined for all x and equals $\hat{q}_s(x)$, where (n.b. $\hat{E} = E^{\hat{u}}$)

$$(3.4) \quad \begin{aligned} \hat{q}_s(x) &= \hat{E} \left\{ c_x(\hat{X}_T^{sx}) \hat{\Phi}_T^{sx} + \int_s^T l_x(t, \hat{X}_t^{sx}, \hat{u}(t, Y)) \hat{\Phi}_t^{sx} dt \right. \\ &\quad \left. + \left[c(\hat{X}_T^{sx}) + \int_s^T l(t, \hat{X}_t^{sx}, \hat{u}(t, Y)) dt \right] \int_s^T h_x^i(t, \hat{X}_t^{sx}) \hat{\Phi}_t^{sx} d\tilde{w}_t^i \middle| \mathcal{F}_s^Y \right\}. \end{aligned}$$

Proof. From Proposition 2.1(b) we have for almost all x

$$\hat{v}_s^Y(x) = E \left\{ c(\hat{X}_T^{sx}) \hat{Z}_T^{sx} + \int_s^T l(t, \hat{X}_t^{sx}, \hat{u}(t, Y)) \hat{Z}_t^{sx} dt \middle| \mathcal{F}_T^Y \right\} \quad \text{P-a.s.,}$$

so that

$$E\{\hat{v}_s^Y(x) | \mathcal{F}_s^Y\} = E\left\{\left[c(\hat{X}_T^{sx}) + \int_s^T l(t, \hat{X}_t^{sx}, \hat{u}(t, Y)) dt\right] \hat{Z}_T^{sx} \middle| \mathcal{F}_s^Y\right\} \quad P\text{-a.s.}$$

Now fix χ in \mathbf{R}^n and let z_t^{sx} be the solution of (3.1) with $z_s^{sx} = \chi$. A standard computation, linearization of (2.2), shows that for any compact set C

$$\sup_{\omega} E\{\|\hat{X}^{s(x+\chi)} - \hat{X}^{sx}\|_T^2 | \mathcal{F}_T^Y\} = O(|\chi|^2)$$

and

$$(3.5) \quad \sup_{\{\omega: \eta \in C\}} \sup_{s \leq t \leq T} E\{|\hat{X}_t^{s(x+\chi)} - \hat{X}_t^{sx} - z_t^{sx}|^2 | \mathcal{F}_T^Y\} = o(|\chi|^2).$$

Set

$$\bar{Z}_t = \hat{Z}_t^{s(x+\chi)} - \hat{Z}_t^{sx} \left[1 + \int_s^t h_x^i(r, \hat{X}_r^{sx}) \cdot z_r^{sx} dY_r^i - \int_s^t h^i(r, \hat{X}_r^{sx}) h_x^i(r, \hat{X}_r^{sx}) \cdot z_r^{sx} dr \right].$$

Then

$$\begin{aligned} d\bar{Z}_t &= \bar{Z}_t h(t, \hat{X}_t^{s(x+\chi)}) \cdot dY_t + \hat{Z}_t^{sx} [h^i(t, \hat{X}_t^{s(x+\chi)}) - h^i(t, \hat{X}_t^{sx}) - h_x^i(t, \hat{X}_t^{sx}) \cdot z_t^{sx}] dY_t^i \\ &\quad + \hat{Z}_t^{sx} \left[\int_s^t h_x^i(r, \hat{X}_r^{sx}) \cdot z_r^{sx} dY_r^i - \int_s^t h^i(r, \hat{X}_r^{sx}) h_x^i(r, \hat{X}_r^{sx}) \cdot z_r^{sx} dr \right] \\ &\quad \cdot [h(t, \hat{X}_t^{s(x+\chi)}) - h(t, \hat{X}_t^{sx})] \cdot dY_t^i. \end{aligned}$$

Now (3.5), Gronwall's lemma and the fact that for all $q < \infty$, $E\{\|\hat{Z}^{sx}\|_T^q\} < \infty$ by the Burkholder inequality (since h is bounded), imply that

$$\sup_t E|\bar{Z}_t|^2 = o(|\chi|^2).$$

It follows that

$$\begin{aligned} E\{\hat{v}_s^Y(x+\chi) | \mathcal{F}_s^Y\} &= E\{\hat{v}_s^Y(x) | \mathcal{F}_s^Y\} \\ &\quad + E\left\{\left[c_x(\hat{X}_T^{sx}) z_T^{sx} + \int_s^T l_x(t, \hat{X}_t^{sx}, \hat{u}(t, Y)) \cdot z_t^{sx} dt \right. \right. \\ &\quad \left. \left. + \left[c(\hat{X}_T^{sx}) + \int_s^T l(t, \hat{X}_t^{sx}, \hat{u}(t, Y)) dt\right] \right. \right. \\ &\quad \left. \left. \cdot \left[\int_s^T h_x^i(t, \hat{X}_t^{sx}) \cdot z_t^{sx} dY_t^i \right. \right. \right. \\ &\quad \left. \left. \left. - \int_s^T h^i(t, \hat{X}_t^{sx}) h_x^i(t, \hat{X}_t^{sx}) \cdot z_t^{sx} dt\right] \right] \hat{Z}_T^{sx} \middle| \mathcal{F}_s^Y\right\} + R \quad \text{a.s.} \end{aligned}$$

where $E|R| = o(|\chi|)$. But

$$z_t^{sx} = \hat{\Phi}_t^{sx} \chi.$$

So for almost all x , $E\{\hat{v}_s^Y(x) | \mathcal{F}_s^Y\}$ is differentiable in x with derivative $\hat{q}_s(x)$.

It remains to show that

$$\partial E\{\hat{v}_s^Y(x) | \mathcal{F}_s^Y\} = E\{\partial \hat{v}_s^Y(x) | \mathcal{F}_s^Y\} \quad \text{a.s.}$$

Since \hat{v}_s^Y is in H^1 for almost all s , then $\partial \hat{v}_s^Y$ is in $L^2(\mathbf{R})$, i.e., for each i

$$\lim_{\chi \rightarrow 0} \|\chi^{-1}[\hat{v}_s^Y(x + \chi e_i) - \hat{v}_s^Y(x)] - \partial_i \hat{v}_s^Y(x)\|_{L^2(\mathbf{R}^n)} = 0 \quad \text{a.s.}$$

from whence it follows that $\partial \hat{v}_s^Y(x)$ is (x, ω) measurable, and for a suitable subsequence $\{\chi_n\}$ converging to zero we have for almost all x

$$\chi_n^{-1}[\hat{v}_s^Y(x + \chi_n e_i) - \hat{v}_s^Y(x)] \rightarrow \partial_i \hat{v}_s^Y(x) \quad \text{a.s.}$$

But calculations of the kind used above show that, for some $p > 1$,

$$\sup_x E\{|\chi^{-1}[\hat{v}_s^Y(x + \chi e_i) - \hat{v}_s^Y(x)]|^p | \mathcal{F}_s^Y\} < \infty \quad \text{a.s.,}$$

i.e., we have uniform integrability. Since the conditional expectation $E\{\cdot | \mathcal{F}_s^Y\}$ is in this case integration with respect to a measure $P_y(\cdot)$ for y in $\mathcal{C}(0, s; \mathbf{R}^d)$, then Lebesgue's theorem shows that, for almost all s and almost all x ,

$$\lim E\{\chi^{-1}[\hat{v}_s^Y(x + \chi e_i) - \hat{v}_s^Y(x)] | \mathcal{F}_s^Y\} = E\{\partial_i \hat{v}_s^Y(x) | \mathcal{F}_s^Y\} \quad \text{a.s.}$$

The result follows.

Although the above result is only given for $u = \hat{u}$, it holds for any u in \mathcal{U} . We write $\nu_s^{uY}(x)$, $q_s^u(x)$, Φ_t^{usx} correspondingly. Now without assuming (H) $q_s^u(x)$ is still well defined by

$$(3.4)' \quad \begin{aligned} q_s^u(x) = E^u \left\{ c_x(X_T^{usx}) \Phi_T^{usx} + \int_s^T l_x(t, X_t^{usx}, u(t, Y)) \Phi_t^{usx} dt \right. \\ \left. + \left[c(X_T^{usx}) + \int_s^T l(t, X_t^{usx}, u(t, Y)) dt \right] \int_s^T h_x^i(t, X_t^{usx}) \Phi_t^{usx} d\tilde{w}_t^i \middle| \mathcal{F}_s^Y \right\}. \end{aligned}$$

In preparation for the main result we establish two lemmata.

LEMMA 3.2. Assume (A₁)–(A₆). Then there exists a constant K_6 such that for all u in \mathcal{U}

$$|q_s^u(x)| \leq K_6(1 + |x|^q + \|y\|_s^q).$$

Moreover K_6 depends only on K_1, \dots, K_5, K_u, q and T .

Proof. This follows readily from (3.4)'.

With \bar{u} in \mathcal{U} define

$$(3.6) \quad u_{\varepsilon s}(t, y) = \begin{cases} \bar{u}(t, y) & \text{if } s \leq t < s + \varepsilon, \\ \hat{u}(t, y) & \text{otherwise.} \end{cases}$$

Then $u_{\varepsilon s}$ is a strong variation. Wherever the superscript $u_{\varepsilon s}$ appears we replace it by ε .

LEMMA 3.3. Assume (A₁)–(A₆). For any $M < \infty$,

$$\lim_{\varepsilon \rightarrow 0} \sup_{0 \leq s \leq T} \sup_{\|y\|_s \leq M} 1_{\{|x| \leq M\}}(\omega) |q_s^\varepsilon(x) - \hat{q}_s(x)| = 0 \quad \text{a.s.}$$

Proof. Let $\bar{X}_t = X_t^{\varepsilon sx} - \hat{X}_t^{sx}$, and suppress the superscript sx . Then

$$\begin{aligned} d\bar{X}_t = & [f(t, X_t^\varepsilon, u^\varepsilon(t, Y)) - f(t, \hat{X}_t, u^\varepsilon(t, Y))] dt \\ & + [f(t, \hat{X}_t, \bar{u}(t, Y)) - f(t, \hat{X}_t, \hat{u}(t, Y))] 1_{[s, s+\varepsilon)}(t) dt \\ & + [\sigma(t, X_t^\varepsilon) - \sigma(t, \hat{X}_t)] dw_t \end{aligned}$$

so that, by (A₁), (A₂) and Gronwall's inequality, for any $p < \infty$

$$(3.7) \quad \begin{aligned} E\{\|\bar{X}\|_T^p | \mathcal{F}_s^Y\} & \leq KE \left\{ \varepsilon^{p-1} \int_s^{s+\varepsilon} |f(t, \hat{X}_t, \bar{u}(t, Y)) - f(t, \hat{X}_t, \hat{u}(t, Y))|^p dt \middle| \mathcal{F}_s^Y \right\} \\ & \leq \varepsilon^p K(1 + |x|^p + \|Y\|_s^p) \quad \text{a.s.} \end{aligned}$$

where the last inequality follows from (A₁) and

$$(3.8) \quad E\{\|\hat{X}\|_T^p | \mathcal{F}_s^Y\} \leq K(1 + |x|^p + \|Y\|_s^p) \quad \text{a.s.}$$

and where the constant K may change from line to line, but is independent of s , ω , ε (it may depend on p , K_1 , K_2 , \dots , K_5 , q , $\bar{K}(=K_{\bar{u}})$, $\hat{K}(=K_{\hat{u}})$). We also have for any $p < \infty$

$$\sup_{\varepsilon, s, x, \omega} E\{\|\Phi^\varepsilon\|_T^p | \mathcal{F}_s^Y\} < \infty.$$

Let $\bar{\Phi}_t = \Phi_t^\varepsilon - \hat{\Phi}_t$ and let $\sigma^{(i)}$ be the i th column of σ . Then

$$\begin{aligned} d\bar{\Phi}_t &= f_x(t, X_t^\varepsilon, u_t^\varepsilon)\bar{\Phi}_t dt + \sigma_x(t, X_t^\varepsilon)\bar{\Phi}_t dw_t \\ &\quad + [f_x(t, X_t^\varepsilon, \hat{u}) - f_x(t, \hat{X}_t, \hat{u})]\hat{\Phi}_t dt + [\sigma_x^{(i)}(t, X_t^\varepsilon) - \sigma_x^{(i)}(t, \hat{X}_t)]\hat{\Phi}_t dw_t^i \\ &\quad + [f_x(t, X_t^\varepsilon, u^\varepsilon) - f_x(t, X_t^\varepsilon, \hat{u})]\hat{\Phi}_t dt. \end{aligned}$$

Proceeding as in (3.7) it follows that

$$\begin{aligned} (3.9) \quad E\{\|\bar{\Phi}\|_T^p | \mathcal{F}_s^Y\} &\leq K[\varepsilon^p + \sqrt{I_1} + \sqrt{I_2}], \\ I_1 &= E\left\{\int_s^T |f_x(t, X_t^\varepsilon, \hat{u}) - f_x(t, \hat{X}_t, \hat{u})|^{2p} dt \middle| \mathcal{F}_s^Y\right\}, \\ I_2 &= E\left\{\int_s^T |\sigma_x(t, X_t^\varepsilon) - \sigma_x(t, \hat{X}_t)|^{2p} dt \middle| \mathcal{F}_s^Y\right\}. \end{aligned}$$

Consider I_1 . For any $N < \infty$ set

$$A_{\varepsilon N} = \{\omega: \|X^\varepsilon\|_T > N\} \cup \{\omega: \|\hat{X}\|_T > N\} \cup \{\omega: \|Y\|_T > N\};$$

then by (3.7), (3.8)

$$P\{A_{\varepsilon N} | \mathcal{F}_s^Y\} \leq K(1 + |x|^2 + \|Y\|_s^2)/N^2.$$

Moreover by the local uniformly continuous differentiability of f , we have that for ω not in $A_{\varepsilon N}$, $\delta > 0$, there exists η_N such that

$$|f_x(t, X_t^\varepsilon, \hat{u}) - f_x(t, \hat{X}_t, \hat{u})|^{2p} < \delta T^{-1}$$

if $|X_t^\varepsilon - \hat{X}_t| < \eta_N$. Let $B_{\varepsilon N} = \{\omega: \|X^\varepsilon - \hat{X}\|_T \geq \eta_N\}$. By (3.7)

$$P\{B_{\varepsilon N} | \mathcal{F}_s^Y\} \leq \varepsilon^2 K(1 + |x|^2 + \|Y\|_s^2)/\eta_N^2.$$

From the uniform bound on f_x it follows that

$$(3.10) \quad I_1 \leq K\left(\frac{1}{N^2} + \frac{\varepsilon^2}{\eta_N^2}\right)(1 + |x|^2 + \|Y\|_s^2) + \delta.$$

Since I_2 satisfies a similar inequality, then it follows from (3.9) that

$$(3.11) \quad \limsup_{\varepsilon \rightarrow 0} \sup_s \sup_{|x| \leq M} 1_{\{\|Y\|_s \leq M\}} E\{\|\Phi^\varepsilon - \hat{\Phi}\|_T^p | \mathcal{F}_s^Y\} = 0.$$

The boundedness and convergence of X^ε , Φ^ε , and an argument of the kind used for I_1 but applied to

$$E\left\{\int_s^T |l_x(t, X_t^\varepsilon, \hat{u}) - l_x(t, \hat{X}_t, \hat{u})|^{2p} dt \middle| \mathcal{F}_s^Y\right\},$$

now show that for any $p < \infty$

$$\sup_s \sup_{|x| \leq M} 1_{\{\|Y\|_s \leq M\}} E \left\{ \left| c_x(X_T^\varepsilon) \Phi_T^\varepsilon - c_x(\hat{X}_T) \hat{\Phi}_T + \int_s^T [l_x(t, X_t^\varepsilon, u^\varepsilon(t, Y)) \Phi_t^\varepsilon - l_x(t, \hat{X}_t, \hat{u}(t, Y)) \hat{\Phi}_t] dt \right|^p \middle| \mathcal{F}_s^Y \right\}$$

converges to 0 with ε . Similarly

$$\sup_s \sup_{|x| \leq M} 1_{\{\|Y\|_s \leq M\}} E \left\{ \left| c(X_T^\varepsilon) - c(\hat{X}_T) + \int_s^T [l(t, X_t^\varepsilon, u^\varepsilon(t, Y)) - l(t, \hat{X}_t, \hat{u}(t, Y))] dt \right|^p \middle| \mathcal{F}_s^Y \right\}$$

also converges to 0 with ε . The terms

$$\int_s^T [h_x^i(t, X_t^\varepsilon) \Phi_t^\varepsilon - h_x^i(t, \hat{X}_t) \hat{\Phi}_t] dY_t^i, \\ \int_s^T [h_x^i(t, X_t^\varepsilon) \Phi_t^\varepsilon h^i(t, X_t^\varepsilon) - h_x^i(t, \hat{X}_t) \hat{\Phi}_t h^i(t, \hat{X}_t)] dt$$

are treated by the same methods.

We turn now to Z_T^ε . Write ζ^ε for $\log Z_T^\varepsilon$. Then

$$P\{|Z_T^\varepsilon - \hat{Z}_T| > \delta \mid \mathcal{F}_s^Y\} \leq P\{\hat{\zeta}^\varepsilon > N \mid \mathcal{F}_s^Y\} + P\{|\exp(\zeta^\varepsilon - \hat{\zeta}^\varepsilon) - 1| > \delta e^{-N} \mid \mathcal{F}_s^Y\} \\ \leq N^{-2} E \left\{ \int_s^T |h(t, \hat{X}_t)|^2 dt \middle| \mathcal{F}_s^Y \right\} + P\{|\hat{\zeta}^\varepsilon - \zeta| > \delta_N \mid \mathcal{F}_s^Y\}$$

where $\delta_N = \log(1 + \delta e^{-N})$. But

$$P\{|\zeta^\varepsilon - \hat{\zeta}^\varepsilon| > \delta_N \mid \mathcal{F}_s^Y\} \leq P\{\zeta^\varepsilon - \hat{\zeta}^\varepsilon > \delta_N \mid \mathcal{F}_s^Y\} + P\{\hat{\zeta}^\varepsilon - \zeta^\varepsilon > \delta_N \mid \mathcal{F}_s^Y\} \\ \leq 2\delta_N^{-2} E \left\{ \int_s^T |h(t, X_t^\varepsilon) - h(t, \hat{X}_t)|^2 dt \middle| \mathcal{F}_s^Y \right\}.$$

Hence $Z_T^\varepsilon \rightarrow \hat{Z}_T$ in measure uniformly in s, x for $|x| \leq M, \|Y\|_s \leq M$. The linear bound on h implies that for some $p'' > 1$

$$\sup_{\varepsilon, s} \sup_{|x| \leq M} 1_{\{\|Y\|_s \leq M\}} E\{|Z_T^\varepsilon|^{p''} \mid \mathcal{F}_s^Y\} < \infty.$$

Hence for $1 < p' < p''$

$$\lim_{\varepsilon \rightarrow 0} \sup_s \sup_{|x| \leq M} 1_{\{\|Y\|_s \leq M\}} E\{|Z_T^\varepsilon - Z_T|^{p'} \mid \mathcal{F}_s^Y\} = 0.$$

Repeated application of Hölder's inequality now proves the lemma (cf. (3.3), (3.4), (3.4)').

Let us emphasize that the above continuity is independent of, hence uniform with respect to, (H). Similar but somewhat easier calculations establish

COROLLARY 3.1. $\hat{q}_s(\cdot)$ is a.s. continuous.

We are now in a position to present the main result of this section. Define

$$g_t^u(x) = l(t, x, u(t, Y)) + \hat{q}_t(x) \cdot f(t, x, u(t, Y)).$$

THEOREM 3.1. Assume (A₁)–(A₆) and (H). Then

$$J(u_{\varepsilon s}) - J(\hat{u}) = \hat{E} \int_s^{s+\varepsilon} [g_t^{\bar{u}}(\hat{X}_t) - \hat{g}_t(\hat{X}_t)] dt + o(\varepsilon).$$

Proof. With $u = u_{\varepsilon s}$ define $\bar{\mu}_t^{\varepsilon\eta} = \mu_t^{\varepsilon\eta} - \hat{\mu}_t^\eta$. Recall that $\mu_t^{\varepsilon\eta}$ is the solution of (2.7) on $[0, T]$ with $G_t^\eta(x) = l(t, x, u_{\varepsilon s}(t, \eta))$. Then from (2.7) for $v \in H^1$

$$\bar{\mu}_t^{\varepsilon\eta} = 0 \quad \text{if } t \geq s + \varepsilon,$$

$$(3.12) \quad \left(\frac{d\bar{\mu}^{\varepsilon\eta}}{dt}, v \right) + \hat{\mathcal{A}}_t^\eta(\bar{\mu}_t^{\varepsilon\eta}, v) + (\mathcal{A}_t^{\bar{u}\eta} - \hat{\mathcal{A}}_t^\eta)(\mu_t^{\varepsilon\eta}, v) \\ + \langle [l(t, \cdot, \bar{u}(t, \eta)) - l(t, \cdot, \hat{u}(t, \eta))] e^{\eta_i \cdot h_i}, v \rangle = 0 \quad \text{if } s < t \leq s + \varepsilon,$$

$$(3.13) \quad \left(\frac{d\bar{\mu}^{\varepsilon\eta}}{dt}, v \right) + \hat{\mathcal{A}}_t^\eta(\bar{\mu}_t^{\varepsilon\eta}, v) = 0 \quad \text{if } t \leq s.$$

If we write Δf for $f(t, x, \hat{u}(t, \eta)) - f(t, x, u(t, \eta))$, then

$$(\mathcal{A}_t^{\bar{u}\eta} - \hat{\mathcal{A}}_t^\eta)(\mu_t^{\varepsilon\eta}, v) = \langle \Delta f^i [\partial_i \mu_t^{\varepsilon\eta} - \partial_i(\eta_i \cdot h_i) \mu_t^{\varepsilon\eta}], v \rangle \\ = \langle \Delta f^i \partial_i \mu_t^{\varepsilon\eta} e^{\eta_i \cdot h_i}, v \rangle.$$

Recall that $u = u_{\varepsilon s}$ and for $\phi \in \mathcal{U}$ set

$$\tilde{g}_t^{\phi\eta}(x) = l(t, x, \phi(t, \eta)) + \partial \nu_t^{\varepsilon\eta}(x) \cdot f(t, x, \phi(t, \eta)), \\ \tilde{g}_t^\phi(x) = E\{\tilde{g}_t^{\phi Y}(x) | \mathcal{F}_t^Y\} = l(t, x, \phi(t, Y)) + q_t^\varepsilon(x) \cdot f(t, x, \phi(t, Y))$$

and write $\Delta \tilde{g}_t^\eta(x) = \tilde{g}_t^{u\eta}(x) - \tilde{g}_t^{\hat{u}\eta}(x)$, $\Delta \tilde{g}_t(x) = \bar{g}_t^u(x) - \bar{g}_t^{\hat{u}}(x)$. Then (3.12) can be written as

$$(3.12)' \quad \left(\frac{d\bar{\mu}^{\varepsilon\eta}}{dt}, v \right) + \hat{\mathcal{A}}_t^\eta(\bar{\mu}_t^{\varepsilon\eta}, v) + \langle \Delta \tilde{g}_t^\eta e^{\eta_i \cdot h_i}, v \rangle = 0 \quad \text{if } s < t \leq s + \varepsilon.$$

We now want to apply Proposition 2.1(b) to (3.12)' with $t = s + \varepsilon$, $c = 0$, and $G_t^\eta = \Delta \tilde{g}_t^\eta$. Since $\hat{\mu}_t^\eta$ satisfies (2.12), since f is bounded and since c, l satisfy (A₄) and (H), then $\Delta \tilde{g}_t^\eta$ satisfies (2.6), and hence

$$(3.14) \quad J(u_{\varepsilon s}) - J(\hat{u}) = E\langle \bar{\mu}_0^{\varepsilon Y}, p_0 \rangle \\ = E\langle \bar{\mu}_s^{\varepsilon Y}, \hat{\rho}_s^Y e^{-Y_s \cdot h_s} \rangle \\ = E\left\langle E\left\{ \int_s^{s+\varepsilon} \Delta \tilde{g}_t^Y(\hat{X}_t^{sx}) \hat{Z}_t^{sx} dt \middle| \mathcal{F}_T^Y \right\}, \hat{\rho}_s^Y \right\rangle \\ = E\left\langle E\left\{ \int_s^{s+\varepsilon} \Delta \tilde{g}_t(\hat{X}_t^{sx}) dt \hat{Z}_T^{sx} \middle| \mathcal{F}_s^Y \right\}, \hat{\rho}_s^Y \right\rangle \\ = E\left\{ \int_s^{s+\varepsilon} \Delta \tilde{g}_t(\hat{X}_t) dt \hat{Z}_T \right\}$$

where the first equality comes from Proposition 2.1(b), the second from the fact that (2.8), satisfied by $\hat{\psi} = \hat{\rho} e^{-Y \cdot h}$ with $u = \hat{u}$, and (3.13) are adjoint, hence the solutions

have constant inner product, and the third comes from (3.12)' and Proposition 2.1(b), plus the fact that $\bar{\mu}_{s+\varepsilon}^{\varepsilon\eta} = 0$. For the fourth note that $\hat{\rho}_s^Y$ is \mathcal{F}_s^Y measurable, and

$$\begin{aligned} E\{\Delta\tilde{g}_t^Y(\hat{X}_t^{sx})\hat{Z}_t^{sx}|\mathcal{F}_t^Y\} &= E\{\Delta\tilde{g}_t^Y(\hat{X}_t^{sx})\hat{Z}_t^{sx}|\mathcal{F}_s^Y\} \\ &= E\{E\{\Delta\tilde{g}_t^Y(\hat{X}_t^{sx})|\mathcal{F}_t\}\hat{Z}_t^{sx}|\mathcal{F}_s^Y\} \\ &= E\{E\{\Delta\tilde{g}_t(\hat{X}_t^{sx})|\mathcal{F}_t\}\hat{Z}_t^{sx}|\mathcal{F}_s^Y\} \\ &= E\{\Delta\tilde{g}_t(\hat{X}_t^{sx})\hat{Z}_t^{sx}|\mathcal{F}_s^Y\} \\ &= E\{\Delta\tilde{g}_t(\hat{X}_t^{sx})\hat{Z}_T^{sx}|\mathcal{F}_s^Y\} \end{aligned}$$

since $\Delta\tilde{g}_t(\hat{X}_t^{sx})$ is \mathcal{F}_t measurable. The last equality follows from

$$\begin{aligned} (3.15) \quad & \langle E\{\Delta\tilde{g}_t(\hat{X}_t^{sx})\hat{Z}_T^{sx}|\mathcal{F}_s^Y\}, \hat{\rho}_s^Y \rangle = E\{\Delta\tilde{g}_t(\hat{X}_t)\hat{Z}_T\hat{Z}_s^{-1}|\mathcal{F}_s^Y, \hat{X}_s = x\}, \hat{\rho}_s^Y\} \\ &= E\{E\{\Delta\tilde{g}_t(\hat{X}_t)\hat{Z}_T\hat{Z}_s^{-1}|\mathcal{F}_s^Y \vee \hat{X}_s\}\hat{Z}_s|\mathcal{F}_s^Y\} \\ &= E\{E\{\Delta\tilde{g}_t(\hat{X}_t)\hat{Z}_T\hat{Z}_s^{-1}|\mathcal{F}_s^Y \vee \mathcal{F}_s^{\hat{X}}\}\hat{Z}_s|\mathcal{F}_s^Y\} \\ &= E\{\Delta\tilde{g}_t(\hat{X}_t)\hat{Z}_T|\mathcal{F}_s^Y\} \end{aligned}$$

since \hat{X}_s given \mathcal{F}_s^Y is conditionally Markovian and since \hat{Z}_s is $\mathcal{F}_s^Y \vee \mathcal{F}_s^{\hat{X}}$ measurable.

As the final step we shall replace q_t^ε by \hat{q}_t on the right side of (3.14). Note that

$$(3.16) \quad E\{\Delta\tilde{g}_t(\hat{X}_t)\hat{Z}_T\} = E\{1_{\{\|\hat{X}\|_T \leq M\}}1_{\{\|Y\|_T \leq M\}}\Delta\tilde{g}_t(\hat{X}_t)\hat{Z}_T\} + R_t^M$$

where

$$\begin{aligned} |R_t^M| &\leq \int_{A_M} K(1 + \|\hat{X}\|_T^{q+1} + \|Y\|_T^q) d\hat{P}, \\ A_M &= \{\omega: \|\hat{X}\|_T > M\} \cup \{\omega: \|Y\|_T > M\}. \end{aligned}$$

Hypotheses (A₁)–(A₆) imply that the integrand on the right above is integrable and that $\hat{P}(A_M) \rightarrow 0$ so that $\|R^M\|_T \rightarrow 0$ uniformly with respect to the constants in (H). Now Lemmas 3.2, 3.3 and (3.16) imply that

$$E\{\Delta\tilde{g}_t(\hat{X}_t)\hat{Z}_T\} = E\{\Delta\tilde{g}_t(\hat{X}_t)\hat{Z}_T\} + o(1),$$

so the theorem follows.

We emphasize that $o(\varepsilon)$ in the theorem is *uniform* in the data satisfying (A₁)–(A₅) with the same K_1, \dots, K_5 and in P_0 in $\mathcal{P}(K) := \{P = \int |x|^q dP \leq K\}$. We use this remark to eliminate (H). Let us also define the Hamiltonian and the adjoint process p . Recall that ' is transpose and $\hat{E} = E^{\hat{a}}$.

$$(3.17) \quad H(t, x, u, p) = p \cdot f(t, x, u) - l(t, x, u),$$

$$\begin{aligned} (3.18) \quad & p'_s(\omega) = -\hat{q}_s^{Y(\omega)}(\hat{X}_s(\omega)) \\ &= -\hat{E}\left\{c_x(\hat{X}_T)\hat{\Phi}_T^s + \int_s^T l_x(t, \hat{X}_t, \hat{u}(t, Y))\hat{\Phi}_t^s dt \right. \\ & \quad \left. + \left[c(\hat{X}_T) + \int_s^T l(t, \hat{X}_t, \hat{u}(t, Y)) dt\right] \int_s^T h_x^i(t, \hat{X}_t)\hat{\Phi}_t^s d\tilde{w}_t^i \middle| \mathcal{F}_s\right\} \end{aligned}$$

where $\hat{\Phi}_t^s = \hat{\Phi}_t^{s\hat{X}}$ is the fundamental matrix solution of

$$(3.19) \quad dz_t = f_x(t, \hat{X}_t, \hat{u}(t, Y))z_t dt + \sigma_x^{(i)}(t, \hat{X}_t)z_t d\tilde{w}_t^i, \quad t \geq s,$$

with $\hat{\Phi}_s^s = I$. Note that in (3.18) conditioning on \mathcal{F}_s is equivalent to conditioning on $\mathcal{F}_s^Y \vee \hat{X}_s$.

COROLLARY 3.2. Assume (A₁)–(A₆). Then

$$(3.20) \quad J(u_{\varepsilon s}) - J(\hat{u}) = -\hat{E} \int_s^{s+\varepsilon} [H(t, \hat{X}_t, \bar{u}(t, Y), p_t) - H(t, \hat{X}_t, \hat{u}(t, Y), p_t)] dt + o(\varepsilon).$$

Proof. We may assume that P_0 has compact support.¹ Choose $f^n, \bar{\sigma}^n, \bar{h}^n, c^n, l^n$ which satisfy (A₁)–(A₆) with the same constants K_1, \dots, K_5, q , such that

$$f^n(t, x, u) = f(t, x, u) \quad \text{if } |x| \leq n, \quad |u| \leq n,$$

$$l^n(t, x, u) = l(t, x, u) \quad \text{if } |x| \leq n, \quad |u| \leq n,$$

$$c^n(x) = c(x) \quad \text{if } |x| \leq n,$$

$$\bar{h}^n(t, x) = \begin{cases} h(t, x) & \text{if } |x| \leq n, \quad 0 \leq s \leq T, \\ 0 & \text{if } s \notin [0, T]. \end{cases}$$

Let $\beta_n(t)$ be a smooth positive function, support in $\{|t| \leq 1/n\}$, such that $\int \beta_n dt = 1$ and let $\bar{\beta}_n(x) = \prod_{i=1}^n \beta_n(x_i)$. Let

$$h^n(t, x) = \int_{-\infty}^{\infty} \int_{\mathbf{R}^n} \beta_n(t-s) \bar{\beta}_n(x-\xi) \bar{h}^n(s, \xi) d\xi ds,$$

$$dX_t^{nu} = f^n(t, X_t^{nu}, u(t, Y)) dt + \sigma^n(t, X_t^{nu}) d\bar{w}_t + n^{-1} d\bar{w}_t$$

where the initial Ω has been enlarged to $\mathbf{R}^n \times \mathcal{C}(0, T, \mathbf{R}^{2m}) \times \mathcal{C}(0, T; \mathbf{R}^d)$ and \bar{w} is an independent Brownian motion. Using the locally uniform differentiability of h , we can show that for any compact C in \mathbf{R}^n , $\sup_{x \in C} |h^n(\cdot, x) - h(\cdot, x)| \rightarrow 0$ in $L^1(dt)$, hence in measure. Similarly for h_x^n . Also for any u in $\mathcal{U} f^n(t, x, u(t, \eta)) \rightarrow f(t, x, u(t, \eta))$ for each (t, x, η) as $n \rightarrow \infty$. Then f^n , etc. can be chosen so that (H) is satisfied with $\alpha = n^{-2}$.

If we write $\hat{X}_t^n, \hat{\Phi}_t^{sxn}, \hat{Z}_t^{sxn}$ correspondingly and if (H)(ii) is satisfied, then by Theorem 3.1

$$(3.21) \quad J^n(u_{\varepsilon s}) - J^n(\hat{u}) = -E \left\{ \hat{Z}_T^n \int_s^{s+\varepsilon} [H^n(t, \hat{X}_t^n, \bar{u}(t, Y), p_t^n) - H^n(t, \hat{X}_t^n, \hat{u}(t, Y), p_t^n)] dt \right\} + o(\varepsilon).$$

¹ Given P_0 in $\mathcal{P}(K)$, set $P_0^N(\cdot) = k_N P_0(\cdot \cap \{|x| \leq N\})$ where k_N is a normalization constant, $k_N \downarrow 1$. If $J_x(u)$ is the cost corresponding to $X_0 = x$, it can be shown that

$$J^N(u) - J(u) = \int (k_N 1_{\{|x| \leq N\}} - 1) J_x(u) P_0(dx) \rightarrow 0$$

as $N \rightarrow \infty$, uniformly in P in $\mathcal{P}(K)$ and uniformly in the data satisfying (A₁), \dots , (A₅) with the same constants K_1, \dots, K_5 . The term

$$\hat{E} \int_s^{s+\varepsilon} H(t, \hat{X}_t, u(t, y), p_t) dt,$$

cf. (3.20), is treated similarly.

As remarked above, $o(\varepsilon)$ will be uniform in n , so to prove the result we must pass to the limit in (3.21). Since (A_1) – (A_6) are satisfied uniformly in n , then for any $p < \infty$

$$E\{\|\hat{X}^n - \hat{X}\|_T^p + \|X^{nu} - X^u\|_T^p\} \rightarrow 0,$$

and for any compact C in \mathbf{R}^n

$$(3.22) \quad \sup_{x \in C} E\{\|\hat{X}^{nsx} - \hat{X}^{sx}\|_T^p | \mathcal{F}_s^Y\} \rightarrow 0 \quad \text{a.s.},$$

$$(3.23) \quad \sup_n \sup_{x \in C} E\{\|\hat{X}^{nsx}\|_T^p | \mathcal{F}_s^Y\} < \infty \quad \text{a.s.}$$

We shall need similar results for most of the other data. Observe that by (3.23) and (A_3)

$$(3.24) \quad \sup_{x \in C} \sup_n E\left\{\int_s^T |h^n(t, \hat{X}_t^{nsx})|^q dt \middle| \mathcal{F}_s^Y\right\} < \infty \quad \text{a.s.}$$

Moreover if Q^y stands for the product measure $dt \times dP_y$, where P_y is the conditional of $P | \mathcal{F}_s^Y$ (which in this case is a probability measure for each $y \in \mathcal{C}(0, s; \mathbf{R}^d)$), then for any $\varepsilon > 0$, $\delta > 0$

$$\begin{aligned} \sup_{x \in C} Q^y\{ |h^n(t, \hat{X}_t^{nsx}) - h(t, \hat{X}_t^{sx})| > \varepsilon \} &\leq 2\delta T + Q^y\left\{ \sup_{x \in C'} |h^n(t, x) - h(t, x)| > \varepsilon \right\} \\ &\leq (2T + 1)\delta, \end{aligned}$$

for n sufficiently large by the uniform convergence in measure of $h^n \rightarrow h$. Here we take n large enough that by (3.22)

$$\sup_{x \in C} P_y\{\|\hat{X}^{nsx} - \hat{X}^{sx}\|_T > 1\} < \delta,$$

and we take $C' = \{x: |x| \leq M\}$, where M is so large that

$$\sup_{x \in C} P_y\{\|\hat{X}^{sx}\|_T > M - 1\} < \delta.$$

Hence

$$(3.25) \quad \sup_{x \in C} E\left\{\int_s^T |h^n(t, \hat{X}_t^{nsx}) - h(t, \hat{X}_t^{sx})|^p dt \middle| \mathcal{F}_s^Y\right\} \rightarrow 0 \quad \text{a.s.}$$

Since (A_1) , (A_2) , (A_3) hold uniformly in n , then there exist $p'' > p' > 1$ such that

$$(3.26) \quad \sup_{x \in C} \sup_n E\{(\hat{Z}_T^{nsx})^{p''} | \mathcal{F}_s^Y\} < \infty \quad \text{a.s.},$$

and

$$(3.27) \quad \sup_{x \in C} E\{|\hat{Z}_T^{nsx} - \hat{Z}_T^{sx}|^{p'} | \mathcal{F}_s^Y\} \rightarrow 0 \quad \text{a.s.}$$

The result (3.27) follows from (3.26) and (3.25) with $p = 2$. The same arguments also show that

$$\sup_n E(Z_T^{nu})^{p''} < \infty,$$

$$E|Z_T^{nu} - Z_T^u|^{p'} \rightarrow 0,$$

whence it follows easily that the left side of (3.21) converges to that of (3.20).

From (3.24), (3.25), (3.26), (3.27) and similar results involving $h_x^n, c^n, c_x^n, l^n, l_x^n, \Phi^{nsx}$ it can be shown that $\hat{q}_s^n(x) \rightarrow \hat{q}_s(x)$ a.s. uniformly in x in C . Corollary 3.1 now implies that for each s $\hat{q}_s^n(\hat{X}_s^n) \rightarrow \hat{q}_s(\hat{X}_s)$ in probability. Lemma 3.2 then implies that the integral on the right side of (3.21) converges to that of (3.20).

Let us finally eliminate (H)(ii). The above convergence is uniform in P_0 in $\mathcal{P}(K)$ with support in a fixed compact set. Now set $p_0^m(x) = \int \beta_m(|x-y|)P_0(dy)$. Since P_0 has support in $\{|x| \leq N\}$ then $\text{supp } p_0^m \subset \{|x| \leq N+1\}$ and hence (H) is satisfied. Moreover by Fubini's theorem

$$J^m(u) = \int J_x(u) \int \beta_m(|x-y|)P_0(dy) dx \rightarrow \int J_y(u)P_0(dy) = J(u)$$

since $J_x(u)$ is continuous in x . Since also for u in \mathcal{U}

$$\hat{E} \left\{ \int_s^{s+\varepsilon} H(t, \hat{X}_t, u(t, Y), p_t) dt \mid \hat{X}_0 = x \right\}$$

is continuous in x , then the result follows and the corollary is established.

4. The main result. We are now in a position to derive a maximum principle.

THEOREM 4.1. Assume (A₁)–(A₆). If $\hat{u} \in \mathcal{U}$ is a solution of (1.1)–(1.4), then for almost all t

$$\hat{E}\{H(t, \hat{X}_t, \hat{u}(t, Y), p_t) \mid \mathcal{F}_t^Y\} = \max_{u \in U} \hat{E}\{H(t, \hat{X}_t, u, p_t) \mid \mathcal{F}_t^Y\} \quad \text{a.s.}$$

Proof. We begin by defining a countable class of variations. We write $\tilde{\mathcal{U}}_s$ for $L^1(\mathcal{C}(0, T; \mathbf{R}^d), \mathcal{G}_s^d, P_w^d; U)$, the integrable functions mapping $(\mathcal{C}(0, T; \mathbf{R}^d), \mathcal{G}_s^d)$ into U ; then $\tilde{\mathcal{U}}_s$ is separable. Let \mathcal{V} be a countable, dense subset of $\tilde{\mathcal{U}}_T$ such that each element is a bounded function, and define the measurable map

$$i_s: (\mathcal{C}(0, T; \mathbf{R}^d), \mathcal{G}_s^d) \rightarrow (\mathcal{C}(0, T; \mathbf{R}^d), \mathcal{G}_T^d)$$

by $(i_s \eta)(t) = \eta(t \wedge s)$ where $t \wedge s = \min\{t, s\}$. Now $\mathcal{V} \circ i_s$ is a countable dense subset of $\tilde{\mathcal{U}}_s$ (cf. Haussmann [8]), and

$$\bar{\mathcal{U}} := \{u \in \mathcal{U}: u(t, \eta) = v \circ i_t(\eta), v \in \mathcal{V}\}$$

is a countable subset of \mathcal{U} . We shall only consider strong perturbations of the form $u_{\varepsilon s}$ with \bar{u} in $\bar{\mathcal{U}}$, cf. (3.6). Since $\hat{E}|H(t, \hat{X}_t, \bar{u}(t, Y), p_t)| < \infty$ then there is a null set $N(\bar{u})$ such that for $s \notin N(\bar{u})$

$$(4.1) \quad \frac{d}{ds} \int_0^s \hat{E}H(t, \hat{X}_t, \bar{u}(t, Y), p_t) dt = \hat{E}H(s, \hat{X}_s, \bar{u}(s, Y), p_s).$$

We set

$$N = \bigcup_{u \in \bar{\mathcal{U}}} N(\bar{u}) \cup N(\hat{u});$$

then N is a null set.

From Corollary 3.2 we have

$$(4.2) \quad 0 \geq -J(u_{\varepsilon s}) + J(\hat{u}) = \int_s^{s+\varepsilon} [H(t, \hat{X}_t, \bar{u}(t, Y), p_t) - H(t, \hat{X}_t, \hat{u}(t, Y), p_t)] dt + o(\varepsilon),$$

so that when we divide (4.2) by $\varepsilon > 0$ and take the limit as $\varepsilon \rightarrow 0$, then according to (4.1) we have for $s \notin N$

$$0 \geq \hat{E}\{H(s, \hat{X}_s, u(s, Y), p_s) - H(s, \hat{X}_s, \hat{u}(s, Y), p_s)\}$$

for all u in $\mathcal{V} \circ i_s$, hence for all u in $\tilde{\mathcal{U}}_s$ by the denseness of $\mathcal{V} \circ i_s$ in $\tilde{\mathcal{U}}_s$ and by the continuity in u of f and l . It follows as usual (cf. Kushner [9]) that the inequality is preserved when $\hat{E}\{\cdot\}$ is replaced by $\hat{E}\{\cdot | \mathcal{F}_s^Y\}$. This establishes the theorem.

Let us investigate, in a nonrigorous way, the adjoint process. As we saw in the proof of Lemma 3.1 $E\{\hat{p}_s^Y(x) | \mathcal{F}_s^Y\} := V_s^Y(x)$ is the expected cost to go given the past observations and the present state x . Now observe that (cf. (3.18))

$$p'_s = -\hat{q}_s^Y(\hat{X}_s) = -\nabla V_s^Y(\hat{X}_s) \quad \text{a.s.},$$

so as usual the adjoint process is the negative of the gradient of the value function (given the past observations).

On the other hand, if we consider the separated problem as was done in Bensoussan [3], we have

$$\begin{aligned} & \min \{ \mathcal{J}(u) : u \in \mathcal{U} \}, \\ & \mathcal{J}(u) = E_w^d \left\{ \int_0^T \langle l(t, \cdot, u(t, Y)), \rho_t^u \rangle dt + \langle c, \rho_T^u \rangle \right\}, \\ (4.3) \quad & d\rho_t^u = L_t^{u*} \rho_t^u dt + \rho_t^u h_t \cdot dY_t, \quad \rho_0^u = p_0, \end{aligned}$$

where p_0 is the density of P_0 (assumed to exist) and L_t^{u*} is the formal adjoint of

$$L_t^u = \frac{1}{2} a^{ij}(t, \cdot) \partial_i \partial_j + f^i(t, \cdot, u(t, Y(\omega))) \partial_i.$$

Now the optimal cost to go, given the present state ρ and past observation Y , is

$$\tilde{V}_s^Y(\rho) = E_w^d \left\{ \int_s^T \langle l(t, \cdot, \hat{u}(t, Y)), \hat{\rho}_t \rangle dt + \langle c, \hat{\rho}_T \rangle \middle| \mathcal{G}_s^d \right\}$$

where $\hat{\rho}_t$ satisfies (4.3) for $t \geq s$ with u replaced by \hat{u} and with initial condition $\hat{\rho}_s = \rho$. As is readily verified, $\tilde{V}_s^Y(\rho) = \langle V_s^Y, \rho \rangle$ so that the gradient of the value function \tilde{V}_s^Y is (represented by) V_s^Y . But for the separated problem we now define the Hamiltonian

$$\tilde{H}(t, \rho, u, \tilde{p}) = -\langle l(t, \cdot, u), \rho \rangle + \tilde{p}(L_t^{u*} \rho)$$

(\tilde{p} is a functional) so that

$$\begin{aligned} \tilde{H}(t, \rho, u, -\nabla \tilde{V}_t^Y) &= -\langle l(t, \cdot, u), \rho \rangle - \langle V_t^Y, L_t^{u*} \rho \rangle \\ &= -\langle l(t, \cdot, u) + L_t^u V_t^Y, \rho \rangle \\ &= \langle H(t, \cdot, u, p_t), \rho \rangle - \phi_t^Y \end{aligned}$$

where

$$\phi_t^Y = -\frac{1}{2} \langle a_t^{ij} \partial_i \partial_j V_t^Y, \rho \rangle$$

is independent of u . Since

$$\hat{E}\{H(t, \hat{X}_t, u(t, Y), p_t) | \mathcal{F}_t^Y\} = \langle H(t, \cdot, u(t, Y), p_t), \hat{\rho}_t \rangle,$$

then it follows from Theorem 4.1 that

$$\tilde{H}(t, \hat{\rho}_t, \hat{u}(t, Y), \tilde{p}_t) = \max_{u \in U} \tilde{H}(t, \hat{\rho}_t, u, \tilde{p}_t),$$

with $\tilde{p}_t = -\nabla \tilde{V}_t^Y = -\langle V_t^Y, \cdot \rangle$, i.e., the maximum principle holds for the separated problem and the adjoint is the negative of the gradient of the value function.

As an example let us now consider the linear regulator:

$$\begin{aligned} \min \{J(u): u \in \mathcal{U}\}, \\ J(u) = E \left\{ \int_0^T [X_t' M(t) X_t + u_t' N(t) u_t] dt + X_T' D X_T \right\}, \\ dX_t = [A(t) X_t + B(t) u_t] dt + \sigma(t) dw_t, \quad X_0 = x_0, \\ dY_t = H(t) X_t dt + d\tilde{w}_t, \quad Y_0 = 0 \end{aligned}$$

where \mathcal{U} is the set of admissible controls (cf. § 2, with $U = \mathbb{R}^m$). As usual A, B, σ, H, M, N are bounded matrix valued functions of t , and $N(t), M(t), D$ are symmetric with $N(t)$ positive definite and the other two semidefinite. In order that the optimal control, which is a stochastic integral with respect to dY , be in \mathcal{U} we shall assume $H(\cdot)$ to have bounded variation.

Here the Hamiltonian is given as

$$H(t, x, u, p) = p'(A(t)x + B(t)u) - (x'M(t)x + u'N(t)u),$$

so that according to Theorem 4.1, if \hat{u} is optimal, then

$$(4.4) \quad \hat{u}(t, Y) = \frac{1}{2} N(t)^{-1} B(t)' \hat{E}(p_t | \mathcal{F}_t^Y)$$

where

$$(4.5) \quad \begin{aligned} p_t = -2 \left[\Phi(T, t)' D \hat{X}_T + \int_t^T \Phi(s, t)' M(s) \hat{X}_s ds \right] \\ - \left[\hat{X}_T' D \hat{X}_T + \int_t^T (\hat{X}_s' M(s) \hat{X}_s + \hat{u}_s' N(s) \hat{u}_s) ds \right] \int_t^T \Phi(t, s)' H(s)' d\tilde{w} \end{aligned}$$

with

$$\frac{d\Phi}{ds}(s, t) = A(s)\Phi(s, t), \quad \Phi(t, t) = I.$$

We must now give a more explicit representation of \hat{u} , i.e., we must put $\hat{E}(p_t | \mathcal{F}_t^Y)$ into a more usable form. We allow ourselves to be inspired by the linear regulator with complete observation to guess that \hat{u} is linear in the state \hat{X}_t , or at least in $m_t = \hat{E}(\hat{X}_t | \mathcal{F}_t^Y)$, i.e.,

$$(4.6) \quad \hat{u}(t, Y) = K(t) m_t,$$

$$(4.7) \quad dm_t = [A(t) + B(t)K(t)]m_t dt + R(t)H(t)'[dY_t - H(t)m_t dt],$$

$$(4.8) \quad \frac{dR}{dt} - A(t)R - RA(t)' + RH(t)'H(t)R - \sigma(t)\sigma(t)' = 0.$$

If we assume that x_0 is normally distributed, $N(m_0, R_0)$, then the parameters m_0, R_0 give the initial conditions for m_t, R_t . If K is bounded, then it follows that

$$\begin{aligned} m_t &= \phi(t, 0)m_0 + \int_0^t \phi(t, s)R_s H_s' dY_s \\ &= \phi(t, 0)m_0 + R_t H_t' Y_t - \int_0^t \psi(t, s)Y_s ds \end{aligned}$$

for some locally bounded functions ϕ, ψ , because H is absolutely continuous and ϕ, R are differentiable. Hence \hat{u} as given by (4.6) is in \mathcal{U} .

Let $\tilde{\Phi}(s, t)$ be the fundamental matrix solution of

$$\frac{dx}{dt} = (A + BK)x$$

and let $\Psi(s, t)$ be that of

$$\frac{dx}{dt} = \tilde{A}x$$

where

$$\tilde{A} = \begin{bmatrix} A + BK - RH'H & RH'H \\ BK & A \end{bmatrix}.$$

It follows that

$$(4.9) \quad \begin{aligned} d \begin{pmatrix} m \\ \hat{X} \end{pmatrix} &= \tilde{A} \begin{pmatrix} m \\ \hat{X} \end{pmatrix} dt + \begin{pmatrix} RH' & 0 \\ 0 & \sigma \end{pmatrix} d \begin{pmatrix} \tilde{w} \\ w \end{pmatrix}, \\ \begin{pmatrix} m_s \\ \hat{X}_s \end{pmatrix} &= \Psi(s, t) \begin{pmatrix} m_t \\ \hat{X}_t \end{pmatrix} + \int_t^s \Psi(s, \theta) \begin{pmatrix} R_\theta H'(\theta)' & 0 \\ 0 & \sigma(\theta) \end{pmatrix} d \begin{pmatrix} \tilde{w} \\ w_\theta \end{pmatrix}. \end{aligned}$$

Now using (4.9) we can compute

$$\begin{aligned} \hat{E}(p_t | \mathcal{F}_t) &= -2 \left\{ \Phi(T, t)' D(0 \ I) \Psi(T, t) + \int_t^T \Phi(s, t)' M_s(0 \ I) \Psi(s, t) ds \right. \\ &\quad + \int_t^T \Phi(s, t)' H_s' H_s R_s(I \ 0) \Psi(T, s)' ds \begin{pmatrix} 0 \\ I \end{pmatrix} D(0 \ I) \Psi(T, t) \\ &\quad + \int_t^T \int_t^s \Phi(\theta, t)' H_\theta' H_\theta R_\theta(I \ 0) \Psi(s, \theta)' d\theta \begin{pmatrix} 0 \\ I \end{pmatrix} M_s(0 \ I) \Psi(s, t) ds \\ &\quad + \int_t^T \int_t^s \Phi(\theta, t)' H_\theta' H_\theta R_\theta(I \ 0) \Psi(s, \theta)' d\theta \\ &\quad \left. \cdot \begin{pmatrix} I \\ 0 \end{pmatrix} K_s' N_s K_s(0 \ I) \Psi(s, t) ds \right\} \begin{pmatrix} m_t \\ \hat{X}_t \end{pmatrix}. \end{aligned}$$

Since

$$\tilde{A} \begin{pmatrix} I \\ I \end{pmatrix} = \begin{pmatrix} I \\ I \end{pmatrix} (A + BK), \quad \Psi(s, t) \begin{pmatrix} I \\ I \end{pmatrix} = \begin{pmatrix} I \\ I \end{pmatrix} \tilde{\Phi}(s, t),$$

we have

$$(4.10) \quad \begin{aligned} \hat{E}(p_t | \mathcal{F}_t^Y) &= -2 \left\{ \Phi(T, t)' D \tilde{\Phi}(T, t) + \int_t^T \Phi(s, t)' M(s) \tilde{\Phi}(s, t) ds \right. \\ &\quad \left. + \int_t^T \Phi(s, t)' H(s)' H(s) R_s(I \ 0) Q_s \begin{pmatrix} I \\ I \end{pmatrix} \tilde{\Phi}(s, t) ds \right\} m_t \end{aligned}$$

where

$$\begin{aligned} Q_s &= \Psi(T, s)' \begin{pmatrix} 0 \\ I \end{pmatrix} D \Psi(T, s) + \int_s^T \Psi(\theta, s)' \begin{pmatrix} 0 \\ I \end{pmatrix} M(\theta) (0 \ I) \Psi(\theta, s) ds \\ &\quad + \int_s^T \Psi(\theta, s)' \begin{pmatrix} I \\ 0 \end{pmatrix} K(\theta)' N(\theta) K(\theta) (I \ 0) \Psi(\theta, s) ds, \end{aligned}$$

so that Q satisfies

$$\frac{d}{ds}Q_s = -\tilde{A}(s)'Q_s - Q_s\tilde{A}(s) - \begin{pmatrix} K(s)'N(s)K(s) & 0 \\ 0 & M(s) \end{pmatrix},$$

$$Q_T = \begin{pmatrix} 0 & 0 \\ 0 & D \end{pmatrix}.$$

Let us write

$$Q_s \begin{pmatrix} I \\ I \end{pmatrix} = q_s;$$

then

$$\frac{dq}{ds} = -\tilde{A}'(s)q - q(A(s) + B(s)K(s)) - \begin{pmatrix} K(s)'N(s)K(s) \\ M(s) \end{pmatrix},$$

and if $q = (q_1', q_2')'$ then

$$\frac{dq_1}{ds} = -(A + BK - RH'H)'q_1 - K'B'q_2 - q_1(A + BK) - K'NK,$$

$$\frac{dq_2}{ds} = -H'HRq_1 - A'q_2 - q_2(A + BK) - M,$$

$$q_1(T) = 0, \quad q_2(T) = D.$$

We observe that if

$$(4.11) \quad B'(t)q_2(t) = -N(t)K(t)$$

then $q_1(\cdot) = 0$, and if $q_2(s)$ is symmetric then

$$\frac{dq_2}{ds} = -A(s)'q_2 - q_2A(s) + K(s)'N(s)K(s) - M(s), \quad q_2(T) = D.$$

Let us write this solution as $P(s)$. Then the second term in (4.10) is zero since $q_1 = 0$, and

$$\hat{E}(p_t | \mathcal{F}_t^Y) = -2P(t)m_t,$$

so (4.4) holds if $K(t) = -N(t)^{-1}B(t)'P(t)$. Moreover, this K is bounded on $[0, T]$ and (4.11) does hold! Hence the control

$$\hat{u}(t, Y) = -N(t)B(t)'P(t)m_t$$

does satisfy the necessary conditions for optimality.

REFERENCES

- [1] V. I. ARKIN AND M. T. SAKSONOV, *Necessary optimality conditions for stochastic differential equations*, Soviet Math. Dokl., 20 (1979), pp. 1-5.
- [2] A. BENSOUSSAN, *Stochastic Control by Functional Analysis Methods*, North-Holland, Amsterdam, 1982.
- [3] ———, *Maximum principle and dynamic programming approaches of the optimal control of partially observed diffusions*, Stochastics, 9 (1983), pp. 169-222.
- [4] R. J. ELLIOTT, *The optimal control of a stochastic system*, this Journal, 15 (1977), pp. 756-778.
- [5] U. G. HAUSSMANN, *General necessary conditions for optimal control of stochastic systems*, Math. Programming Stud., 6 (1976), pp. 30-48.

- [6] ———, *The maximum principle for optimal control of diffusions with partial information*, Proc. International Conference on Stochastic Optimization, Kiev, 1984.
- [7] ———, *L'équation de Zakai et le problème séparé du contrôle optimal stochastique*, Séminaire de Probabilités XIX, Lecture Notes in Math. 1123, Springer, Berlin, 1985, pp. 37–62.
- [8] ———, *A Stochastic Maximum Principle for Optimal Control of Diffusions*, Longman, London, 1986.
- [9] H. J. KUSHNER, *Necessary conditions for continuous parameter stochastic optimization problems*, this Journal, 10 (1972), pp. 550–565.

OPTIMAL HANKEL NORM MODEL REDUCTIONS AND WIENER-HOPF FACTORIZATION I: THE CANONICAL CASE*

JOSEPH A. BALL† AND ANDRÉ C. M. RAN‡

Abstract. We consider the problem for discrete time systems of approximating a given stable rational matrix function $K(z) = \sum_{j=1}^{\infty} K_j z^{-j}$ of McMillan degree n by a function $\hat{K}(z) + H(z)$, where \hat{K} has McMillan degree $l < n$ and H is antistable; we include the case $l=0$ in which we then take $\hat{K}=0$. The minimum possible L^∞ -norm (on the unit circle) of the error $\|K - \hat{K} - H\|_{L^\infty}$, or equivalently the minimum possible spectral norm of the induced Hankel matrix $\|\mathcal{H}_{K-\hat{K}}\|$, is known to be equal to the $(l+1)$ st singular value $\sigma_{l+1}(K)$ of the Hankel matrix $\mathcal{H}_K = [K_{i+j-1}]_{i,j}$. Assume $\sigma_{l+1}(K) < \sigma_l(K)$ and choose the number σ to satisfy $\sigma_{l+1}(K) < \sigma < \sigma_l(K)$; if $l=0$, the condition is simply $\sigma_1(K) = \|\mathcal{H}_K\| < \sigma$. We give an explicit linear fractional map parametrization of the class of all functions $\hat{K} + H$ as above which satisfy $\|K - \hat{K} - H\|_{L^\infty} \leq \sigma$. The coefficients of the linear fractional map are completely determined by the matrices A, B, C in a realization $K(z) = C(zI - A)^{-1}B$ for $K(z)$ and the observability and controllability gramians for the discrete time system (A, B, C) . The analogous results for continuous time systems are derived by a linear fractional change of variable; in this way we recover some recent results of Glover. The basic idea is to use the Grassmannian approach of Ball and Helton to reduce the problem to one of spectral factorization; this in turn can be solved by the geometric factorization principle of Bart, Gohberg, Kaashoek and van Dooren. Known applications include sensitivity minimization in H^∞ control theory and model reduction for linear systems.

Key words. McMillan degree, Lyapunov equation, controllability and observability gramian, singular values, shift invariant subspace representation

AMS(MOS) subject classifications. 93B25, 93B40, 93C35

1. Introduction. We suppose that $K(z)$ is a given strictly proper stable rational $p \times q$ matrix function of McMillan degree n which is analytic on the unit circle. Here we mean "stable" in the sense of discrete time systems, so all poles of K are assumed to be in the open unit disk. We assume that we know a realization

$$K(z) = C(zI_n - A)^{-1}B$$

for K ; here A, B and C are matrices of sizes $n \times n$, $n \times q$ and $p \times n$, respectively, and the spectrum $\sigma(A)$ of A is in the open unit disk.

The model reduction problem that comes up in many engineering applications is that of approximating K by a stable rational matrix function \hat{K} of McMillan degree $l < n$. It is usually desired (see [10], [8] and the references there) to obtain the best approximation in the sense of minimizing the Hankel norm $\|K - \hat{K}\|_{\mathcal{H}}$, where $\|F\|_{\mathcal{H}}$ is the spectral norm of the block Hankel matrix $\mathcal{H}_F = [F_{i+j-1}]_{i,j}$, $j = 1, 2, \dots$ associated with F (where $F(z) = \sum_{i=1}^{\infty} F_i z^{-i}$). Here the block Hankel matrix is considered as an operator from l_q^2 to l_p^2 (where l_r^2 is the space of norm-square-summable \mathbb{C}^r -valued sequences). It is known (see [1] for the scalar case and [10] and [3] for the matrix case) that

$$(1.1) \quad \inf \{ \|K - \hat{K}\|_{\mathcal{H}} : \hat{K} \text{ stable with McMillan degree } \leq l \} = \sigma_{l+1}(K),$$

* Received by the editors May 1, 1985; accepted for publication (in revised form) February 13, 1986.

† Department of Mathematics, College of Arts and Sciences, Virginia Polytechnic Institute and State University, Blacksburg, Virginia 24061-4097. The work of this author was partially supported by a grant from the National Science Foundation.

‡ Subfaculteit der Wiskunde en Informatica, Vrije Universiteit, 1007 MC Amsterdam, the Netherlands. The work of this author was supported by a grant from the Niels Stensen Stichting at Amsterdam.

where $\sigma_1(K) \geq \sigma_2(K) \geq \dots \geq \sigma_n(K)$ are the singular values of \mathcal{H}_K . The singular values of \mathcal{H}_K are alternatively characterized as the square roots of the eigenvalues of the product $\hat{P}\hat{Q}$ of the controllability gramian

$$\hat{P} = \sum_{k=0}^{\infty} A^k B B^* A^{*k}$$

and the observability gramian

$$\hat{Q} = \sum_{k=0}^{\infty} A^{*k} C^* C A^k$$

for the discrete time system (A, B, C) (see [8]). The gramians \hat{P} and \hat{Q} can also be defined as the unique solutions of the Lyapunov equations

$$(1.2) \quad \hat{P} - A\hat{P}A^* = BB^*,$$

$$(1.3) \quad \hat{Q} - A^*\hat{Q}A = C^*C.$$

The approximation result (1.1) has an equivalent formulation in the frequency domain. To describe this we now introduce some notation and terminology for function spaces on the unit circle. For F any function defined on the unit circle, let $\|F\|_{L^\infty} = \sup \{\|F(z)\| : |z| = 1\}$. The space $L_{p \times q}^\infty$ is the set of all measurable $p \times q$ matrix functions F with $\|F\|_{L^\infty} < \infty$. The subspace $H_{p \times q}^\infty$ is the set of all $F \in L_{p \times q}^\infty$ with bounded analytic continuation to the unit disk (the antistable functions).

Denote by $H_{p \times q}^\infty(l)$ the (nonconvex) subset of $L_{p \times q}^\infty$ consisting of all functions H of the form $\hat{K} + H_1$, where \hat{K} is a stable rational $p \times q$ matrix function of McMillan degree at most l and $H_1 \in H_{p \times q}^\infty$. Then an equivalent frequency domain version of (1.1) is

$$(1.1') \quad \inf \{\|K - H\|_{L^\infty} : H \in H_{p \times q}^\infty(l)\} = \sigma_{l+1}(K).$$

For $l = 0$ this amounts to the matrix version of Nehari's theorem [11]; this special case is also relevant to the sensitivity minimization problem in H^∞ control theory (see, e.g. [7] and the references there).

Various authors (see [10], [8] and the references there) have derived algorithms for computing an $H_0 \in H_{p \times q}^\infty(l)$ for which the infimum in (1.1') is achieved, that is, for which

$$(1.4) \quad \|K - H_0\|_{L^\infty} = \sigma_{l+1}(K).$$

In [3] a (degenerate) linear fractional map is produced which parametrizes the set of all $F = K - H_0 \in K + H_{p \times q}^\infty(l)$ that satisfy (1.4). Computation of this map and a starting point for many of the algorithms mentioned above involves a knowledge of the singular value decomposition of \mathcal{H}_K . Glover [8] was the first to obtain a complete parametrization of the set of all solutions directly from the matrices A, B, C , but in the continuous time setting.

In this paper we obtain a result analogous to that of Glover's for the discrete time as well as continuous time setting, but for a slightly modified (and easier) problem. From (1.1') it is clear that for a given tolerance $\sigma > 0$, the smallest integer l for which there is an $F \in K + H_{p \times q}^\infty(l)$ such that $\|F\|_{L^\infty} \leq \sigma$ is the first l for which $\sigma \geq \sigma_{l+1}(K)$. In this paper we assume that a tolerance σ is chosen so that $\sigma_l(K) > \sigma > \sigma_{l+1}(K)$ if $l \geq 1$, or $\sigma > \|\mathcal{H}_K\| = \sigma_1(K)$ if $l = 0$. We then give explicit formulas (in terms of the matrices A, B, C and the controllability and observability gramians \hat{P} and \hat{Q} given by (1.2) and (1.3)) for a linear fractional map which can be used to parametrize the

set of all $F \in K + H_{p \times q}^\infty(I)$ satisfying $\|F\|_{L^\infty} \leq \sigma$. The existence of such a map was already established in [3] for a general $\sigma > 0$; the form of the map is somewhat simpler when one assumes $\sigma_l(K) > \sigma > \sigma_{l+1}(K)$ (or $\sigma > \sigma_1(K)$). The precise result for discrete time is as follows.

THEOREM 1. *Suppose $K(z) = C(zI_n - A)^{-1}B$ is a rational $p \times q$ matrix function with all poles in the open unit disk and of McMillan degree n . Let $\sigma_1(K) \geq \sigma_2(K) \geq \dots \geq \sigma_n(K)$ denote the Hankel singular values of K and choose the positive number σ and the integer l so that $\sigma_l(K) > \sigma > \sigma_{l+1}(K)$ if $l \geq 1$, or $\sigma > \sigma_1(K)$ if $l = 0$. Then there is a rational $(p+q) \times (p+q)$ matrix function*

$$\theta(z) = \begin{bmatrix} \theta_{11}(z) & \theta_{12}(z) \\ \theta_{21}(z) & \theta_{22}(z) \end{bmatrix}$$

such that any matrix function $F = K - H$ where $H \in H_{p \times q}^\infty(I)$ and $\|F\|_{L^\infty} \leq \sigma$ has a representation

$$(1.5) \quad F(z) = (\theta_{11}(z)G(z) + \theta_{12}(z))(\theta_{21}(z)G(z) + \theta_{22}(z))^{-1}$$

for a matrix function $G \in H_{p \times q}^\infty$ such that $\|G\|_{L^\infty} \leq 1$; the function F uniquely determines the function G . Conversely, if $G \in H_{p \times q}^\infty$ with $\|G\|_{L^\infty} \leq 1$, then (1.5) defines a function F with $\|F\|_{L^\infty} \leq \sigma$ of the form $F = K - H$ for a matrix function $H \in H_{p \times q}^\infty(I)$. Moreover, F is rational if and only if G is rational and $\sigma^{-1}F(z)$ is isometric for $|z| = 1$ if and only if $G(z)$ is isometric for $|z| = 1$.

Assume that the matrix A is invertible and has spectrum in the unit disk. Then the matrix function $\theta(z)$ can be given explicitly in the following way. Let \hat{P} and \hat{Q} be the controllability and observability gramians given by (1.2) and (1.3). Set $Z = (I - \sigma^{-2}\hat{Q}\hat{P})^{-1}$ and let c be the $(p+q) \times (p+q)$ Hermitian matrix

$$c = \begin{bmatrix} I_p - \sigma^{-2}CA^{-1}\hat{P}ZA^{*-1}C^* & -CA^{-1}Z^*B \\ -B^*ZA^{*-1}C^* & -\sigma^2I_q - B^*Z\hat{Q}B \end{bmatrix}.$$

Then it develops that c has p positive and q negative eigenvalues, so there is a $(p+q) \times (p+q)$ matrix

$$e = \begin{bmatrix} e_{11} & e_{12} \\ e_{21} & e_{22} \end{bmatrix}$$

such that

$$c^{-1} = e \begin{bmatrix} I_p & 0 \\ 0 & -I_q \end{bmatrix} e^*.$$

Then the matrix function $\theta(z)$ can be taken to be

$$\theta(z) = \{I + \underline{C}(zI - A)^{-1}\underline{B}\}e$$

where

$$\underline{A} = \begin{bmatrix} A & -\sigma^{-2}BB^*A^{*-1} \\ 0 & A^{*-1} \end{bmatrix}, \quad \underline{B} = \begin{bmatrix} 0 & B \\ ZA^{*-1}C^* & Z\hat{Q}B \end{bmatrix}$$

and

$$\underline{C} = \begin{bmatrix} C & \sigma^{-2}C\hat{P} \\ 0 & -\sigma^{-2}B^*A^{*-1} \end{bmatrix}.$$

Equivalently, the block entries $\theta_{ij}(z)$ ($1 \leq i, j \leq 2$) of $\theta(z)$ are expressed as

$$(1.6) \quad \theta_{11}(z) = e_{11} + C(zI_n - A)^{-1}(\sigma^{-2}\hat{P}ZA^{*-1}C^*e_{11} + Z^*Be_{21}),$$

$$(1.7) \quad \theta_{12}(z) = e_{12} + C(zI_n - A)^{-1}(\sigma^{-2}\hat{P}ZA^{*-1}C^*e_{12} + Z^*Be_{22}),$$

$$(1.8) \quad \theta_{21}(z) = e_{21} - \sigma^{-2}B^*A^{*-1}(zI_n - A^{*-1})^{-1}(ZA^{*-1}C^*e_{11} + Z\hat{Q}Be_{21}),$$

and

$$(1.9) \quad \theta_{22}(z) = e_{22} - \sigma^{-2}B^*A^{*-1}(zI_n - A^{*-1})^{-1}(ZA^{*-1}C^*e_{12} + Z\hat{Q}Be_{22}).$$

Our basic approach to the proof of Theorem 1 is to use the Grassmannian and invariant subspace approach from [3] to reduce the computation of the matrix θ to the computation of a signed spectral factorization. The signed spectral factorization problem is as follows. Since we are given a realization for the function $K(z)$, we easily derive a realization $Y_-(z) = D_- + C_-(zI - A_-)^{-1}B_-$ for the rational matrix function

$$Y_-(z) = \begin{bmatrix} I_p & K(z) \\ 0 & \sigma I_q \end{bmatrix}.$$

The problem is to find explicitly a rational matrix function $X_+(z) = D_+ + C_+(zI - A_+)^{-1}B_+$ which is analytic and invertible on the closed unit disk such that

$$X_+^*(z) \begin{bmatrix} I_p & 0 \\ 0 & -I_q \end{bmatrix} X_+(z) = Y_-^*(z) \begin{bmatrix} I_p & 0 \\ 0 & -I_q \end{bmatrix} Y_-(z)$$

where in general $F^*(z) := F(\bar{z}^{-1})^*$. In short, the problem is that of computing a signed spectral factorization from a known signed antispectral factorization of the matrix function

$$W(z) := Y_-^*(z) \begin{bmatrix} I_p & 0 \\ 0 & -I_q \end{bmatrix} Y_-(z).$$

This problem can be solved by the geometric state space approach to Wiener-Hopf factorization developed by Bart, Gohberg and Kaashoek [5] (see also [6]). In this context the controllability and observability gramians are identical (up to trivial scalar factors) to the angle operators needed to describe the key subspaces in the construction of [5]. The hypotheses that the matrix A is invertible is needed only to insure that the matrix function θ is analytic and invertible at ∞ . Similar formulas hold if A is not invertible where in this case θ also has a polynomial part; in the continuous time case this issue does not arise. We do not pursue this point here.

In § 2 of the paper we prove Theorem 1 by the method sketched above. In § 3 we derive the analogous results for continuous time systems by a linear fractional change of variable. It is also possible to get the formulas for the continuous time case directly by the same reduction to a signed spectral factorization problem as was done in § 2 for the discrete time case, but with the left half plane playing the role of the unit disk; this is the approach in [4].

As mentioned above, Theorem 1 parallels results of Glover [8] for continuous time systems. In practice, one is interested only in the stable part \hat{K} of an optimal L^∞ -approximant $H = \hat{K} + H_1$ ($H_1 \in H_{p \times q}^\infty$) from $H_{p \times q}^\infty(I)$. Then \hat{K} is an optimal approximant to K in the Hankel norm but not in the L^∞ -norm. The paper of Glover goes on to give estimates of the L^∞ -norm of $K - \hat{K}$ in terms of the Hankel singular values $\sigma_j(K)$ for $j \geq l + 1$. We have nothing to add here to this aspect of the problem.

Also as mentioned above, the paper of Glover actually handles (for continuous time systems) the more complicated situation where one chooses $\sigma = \sigma_{l+1}(K)$. This case can also be handled by our approach, but involves more involved invariant subspace representations (see [2]) and factorization problems (see [9]). We plan to deal with this topic in a future report.

Finally we remark that our results apply equally well to nonrational functions $K(z)$ which have a realization $K(z) = C(zI - A)^{-1}B$ for bounded operators A, B, C with the spectrum of A in the open unit disk (discrete time) or left half plane (continuous time).

2. The model reduction problem. Our starting point to the model reduction problem discussed in the Introduction is the invariant subspace approach to the problem found in [3]. We summarize the results from there which we need here as follows. We denote by H_{p+q}^2 the Hardy space of \mathbb{C}^{p+q} -valued norm square integrable functions on the unit circle with vanishing negative Fourier coefficients.

THEOREM 2.1 (see [3]). *Let $K = \sum_{j=1}^{\infty} K_j z^{-j}$ be a rational $p \times q$ matrix function with no poles on the unit circle. Let $\sigma_1(K) \geq \sigma_2(K) \geq \cdots \geq \sigma_n(K)$ be the singular values of the Hankel matrix operator $\mathcal{H}_K: l_q^2 \rightarrow l_p^2$ with block matrix representation*

$$[\mathcal{H}_K] = [K_{i+j-1}]_{i,j=1,2,\dots}$$

Then

$$\inf \{ \|K - H\|_{L^\infty} : H \in H_{p \times q}^\infty(l) \} = \sigma_{l+1}(K).$$

Moreover, if the numbers σ and l are chosen so that $\sigma_l(K) > \sigma > \sigma_{l+1}(K)$ if $l \geq 1$ ($\sigma > \sigma_1(K)$ if $l = 0$), then there is a rational $(p+q) \times (p+q)$ matrix function

$$\theta(z) = \begin{bmatrix} \theta_{11}(z) & \theta_{12}(z) \\ \theta_{21}(z) & \theta_{22}(z) \end{bmatrix}$$

such that any function $F = K - H$ where $H \in H_{p \times q}^\infty(l)$ and $\|F\|_{L^\infty} \leq \sigma$ must have a representation

$$(2.1) \quad F(z) = (\theta_{11}(z)G(z) + \theta_{12}(z))(\theta_{21}(z)G(z) + \theta_{22}(z))^{-1}$$

for a matrix function $G \in H_{p \times q}^\infty$ such that $\|G\|_{L^\infty} \leq 1$. Conversely, if $G \in H_{p \times q}^\infty$ with $\|G\|_{L^\infty} \leq 1$, then (2.1) defines a function F with $\|F\|_{L^\infty} \leq \sigma$ of the form $F = K - H$ for a matrix function $H \in H_{p \times q}^\infty(l)$. Moreover, F is rational if and only if G is rational and $\sigma^{-1}F(z)$ is isometric for $|z|=1$ if and only if $G(z)$ is isometric for $|z|=1$. The matrix function $\theta(z)$ can be chosen to be any matrix function satisfying the two conditions

$$(2.2) \quad \theta^*(z) \begin{bmatrix} I_p & 0 \\ 0 & -\sigma^2 I_q \end{bmatrix} \theta(z) = \begin{bmatrix} I_p & 0 \\ 0 & -I_q \end{bmatrix}$$

and

$$(2.3) \quad \theta H_{p+q}^2 = \begin{bmatrix} I_p & K \\ 0 & I_q \end{bmatrix} H_{p+q}^2.$$

Thus the proof of Theorem 1 in the Introduction will be complete once we show how to construct a rational matrix function which satisfies (2.2) and (2.3). The following lemma reduces the problem to a signed spectral factorization problem.

LEMMA 2.2. Suppose K is a rational $p \times q$ matrix function. Then a rational $(p+q) \times (p+q)$ matrix function θ exists which satisfies (2.2) and (2.3) if and only if the matrix function

$$(2.4) \quad W(z) = \begin{bmatrix} I_p & 0 \\ K^*(z) & I_q \end{bmatrix} \begin{bmatrix} I_p & 0 \\ 0 & -\sigma^2 I_q \end{bmatrix} \begin{bmatrix} I_p & K(z) \\ 0 & I_q \end{bmatrix}$$

has a signed spectral factorization

$$(2.5) \quad W(z) = X^*(z) \begin{bmatrix} I_p & 0 \\ 0 & -I_q \end{bmatrix} X(z)$$

(where X is analytic and invertible on the closed unit disk \bar{D}). If this is the case, then the rational matrix function θ defined by

$$(2.6) \quad \theta(z) = \begin{bmatrix} I_p & K(z) \\ 0 & I_q \end{bmatrix} X(z)^{-1}$$

satisfies (2.2) and (2.3).

Proof. If X is any rational $(p+q) \times (p+q)$ matrix function and θ is defined by (2.6), one easily checks that the factorization (2.2) is equivalent to the factorization (2.5), and that the subspace condition (2.3) is equivalent to the subspace condition

$$X^{-1}H_{p+q}^2 = H_{p+q}^2.$$

But this holds for a rational matrix function X if and only if X is analytic and invertible on the closed unit disk \bar{D} . In this way we see that (2.5) is a signed spectral factorization for W if and only if θ satisfies (2.2) and (2.3), where θ and W are related via (2.4) and (2.6). \square

The next lemma describes $X(z)^{-1}$ in case $K(z) = C(zI - A)^{-1}B$ in terms of A , B and C .

LEMMA 2.3. Suppose $K(z) = C(zI - A)^{-1}B$ where $\sigma(A) \subset D \setminus \{0\}$ and $W(z)$ is given by (2.4). Let P and Q be the unique solutions of the Lyapunov equations

$$(2.7) \quad A(\sigma^2 P)A^* - (\sigma^2 P) = BB^*$$

and

$$(2.8) \quad A^*QA - Q = C^*C.$$

Then $W(z)$ has a signed spectral factorization if and only if the matrix $I - QP$ is invertible.

When this is the case, the factor $X(z)$ for a signed spectral factorization

$$W(z) = X^*(z) \begin{bmatrix} I_p & 0 \\ 0 & -I_q \end{bmatrix} X(z)$$

is computed as follows. Set $Z = (I - QP)^{-1}$ and let c be the $(p+q) \times (p+q)$ matrix

$$(2.9) \quad c = \begin{bmatrix} I + CA^{-1}PZA^{*-1}C^* & -CA^{-1}Z^*B \\ -B^*ZA^{-1*}C^* & -\sigma^2 I + B^*ZQB \end{bmatrix}.$$

Then c is Hermitian with p positive and q negative eigenvalues, and so has a factorization

$$(2.10) \quad c = d^* \begin{bmatrix} I_p & 0 \\ 0 & -I_q \end{bmatrix} d$$

for an invertible $(p+q) \times (p+q)$ matrix d . Then the spectral factor $X(z)$ for $W(z)$ in this case is given by

$$(2.11) \quad X(z) = d \left\{ \begin{bmatrix} I_p & 0 \\ 0 & I_q \end{bmatrix} + \begin{bmatrix} CP \\ \sigma^{-2} B^* A^{*-1} \end{bmatrix} Z(zI - A^{*-1})^{-1} [A^{*-1} C^*, -QB] \right\}$$

with inverse given by

$$(2.12) \quad X(z)^{-1} = \left\{ \begin{bmatrix} I_p & 0 \\ 0 & I_q \end{bmatrix} - \begin{bmatrix} CP \\ \sigma^{-2} B^* A^{*-1} \end{bmatrix} (zI - A^{*-1})^{-1} Z [A^{*-1} C^*, -QB] \right\} d^{-1}.$$

Proof. First we find a realization for $W(z)$ using (2.4). Obviously

$$\begin{bmatrix} I_p & K(z) \\ 0 & I_q \end{bmatrix} = \begin{bmatrix} I_p & 0 \\ 0 & I_q \end{bmatrix} + \begin{bmatrix} C \\ 0 \end{bmatrix} (zI - A)^{-1} [0, B],$$

and hence

$$\begin{bmatrix} I_p & 0 \\ K^*(z) & I_q \end{bmatrix} = \begin{bmatrix} I_p & 0 \\ -B^* A^{*-1} C^* & I_q \end{bmatrix} - \begin{bmatrix} 0 \\ B^* A^{*-1} \end{bmatrix} (zI - A^{*-1})^{-1} [A^{*-1} C^*, 0].$$

Multiplying realizations as on page 6 of [5] we arrive at the following realization for $W(z)$:

$$W(z) = \tilde{D} + \tilde{C}(zI - \tilde{A})^{-1} \tilde{B}$$

where

$$(2.13) \quad \begin{aligned} \tilde{A} &= \begin{bmatrix} A^{*-1} & A^{*-1} C^* C \\ 0 & A \end{bmatrix}, & \tilde{B} &= \begin{bmatrix} A^{*-1} C^* & 0 \\ 0 & B \end{bmatrix}, \\ \tilde{C} &= \begin{bmatrix} 0 & C \\ -B^* A^{*-1} & -B^* A^{*-1} C^* C \end{bmatrix}, & \tilde{D} &= \begin{bmatrix} I_p & 0 \\ -B^* A^{*-1} C^* & -\sigma^2 I_q \end{bmatrix}. \end{aligned}$$

If we set $\tilde{A}^x = \tilde{A} - \tilde{B} \tilde{D}^{-1} \tilde{C}$, then

$$(2.14) \quad \tilde{A}^x = \begin{bmatrix} A^{*-1} & 0 \\ -\sigma^{-2} B B^* A^{*-1} & A \end{bmatrix}.$$

Let \mathcal{M} be the spectral subspace of \tilde{A} corresponding to the open unit disk, and \mathcal{M}^x the spectral subspace of \tilde{A}^x corresponding to the exterior of the unit disk. Then, according to Theorem 1.5 in [5], there exists a Wiener-Hopf factorization of $\tilde{D}^{-1} W(z)$ of the form

$$(2.15) \quad \tilde{D}^{-1} W(z) = X_-(z) X_+(z)$$

where $X_+(z)$ is analytic and invertible on the closed unit and $X_-(z)$ is analytic and invertible on the exterior of the unit disk, if and only if \mathbb{C}^{2n} has the direct sum decomposition

$$(2.16) \quad \mathbb{C}^{2n} = \mathcal{M} \dot{+} \mathcal{M}^x.$$

Moreover the factors $X_-(z)$ and $X_+(z)$ can be given explicitly as

$$X_-(z) = I_{p+q} + \tilde{D}^{-1} \tilde{C}(zI - \tilde{A})^{-1} (I - \Pi) \tilde{B}$$

and

$$(2.17) \quad X_+(z) = I_{p+q} + \tilde{D}^{-1} \tilde{C} \Pi (zI - \tilde{A})^{-1} \tilde{B}$$

where Π is the projection onto \mathcal{M}^x along \mathcal{M} . From the formula

$$[I + C(zI - A)^{-1} B]^{-1} = I - C(zI - A + BC)^{-1} B,$$

we see from this that

$$(2.18) \quad \begin{aligned} X_+(z)^{-1} &= I_{p+q} - \tilde{D}^{-1} \tilde{C} \Pi (zI - \tilde{A} + \tilde{B} \tilde{C} \Pi)^{-1} \tilde{B} \\ &= I_{p+q} - \tilde{D}^{-1} \tilde{C} (zI - \tilde{A}^x)^{-1} \Pi \tilde{B}. \end{aligned}$$

We next analyze what the direct sum condition (2.15) means for our particular case. From the upper triangular form of \tilde{A} in (2.13) and the hypothesis $\sigma(A) \subset D$ (so $\sigma(A^{*-1}) \subset \mathbb{C} \setminus \bar{D}$), we see that the spectral subspace for \tilde{A} for the exterior of the closed unit disk $\mathbb{C} \setminus \bar{D}$ is $\text{Im} \begin{bmatrix} I \\ 0 \end{bmatrix}$. The spectral subspace \mathcal{M} for A corresponding to D is then completely determined by the following two conditions:

- (i) $\mathcal{M} \dot{+} \text{Im} \begin{bmatrix} I \\ 0 \end{bmatrix} = \mathbb{C}^{2n}$, and
- (ii) $\tilde{A} \mathcal{M} \subset \mathcal{M}$.

The first condition forces \mathcal{M} to have the form $\mathcal{M} = \text{Im} \begin{bmatrix} Q \\ I \end{bmatrix}$ for some $n \times n$ matrix. The second condition then means that for each $x \in \mathbb{C}^n$ there exists a $y \in \mathbb{C}^n$ such that

$$\begin{bmatrix} A^{*-1} & C^*C \\ 0 & A \end{bmatrix} \begin{bmatrix} Q \\ I \end{bmatrix} x = \begin{bmatrix} Q \\ I \end{bmatrix} y.$$

From the second row we get $y = Ax$. From the first row we then get

$$A^{*-1}Qx + A^{*-1}C^*Cx = QAx.$$

Since this identity must hold for all $x \in \mathbb{C}^n$, we see that Q must satisfy the Lyapunov equation (2.8); note that the solution of (2.8) is unique since A and A^{*-1} have disjoint spectra.

In a similar way, from the lower triangular form of \tilde{A}^x in (2.14), we see that the spectral subspace for \tilde{A}^x corresponding to D is $\text{Im} \begin{bmatrix} 0 \\ I \end{bmatrix}$. Then the spectral subspace \mathcal{M}^x for \tilde{A}^x corresponding to $\mathbb{C} \setminus \bar{D}$ must satisfy

- (i') $\mathcal{M}^x \dot{+} \text{Im} \begin{bmatrix} 0 \\ I \end{bmatrix} = \mathbb{C}^{2n}$, and
- (ii') $\tilde{A}^x \mathcal{M}^x \subset \mathcal{M}^x$.

This leads to the conclusion that \mathcal{M}^x has the form $\mathcal{M}^x = \text{Im} \begin{bmatrix} I \\ P \end{bmatrix}$ where P is the unique solution of the Lyapunov equation (2.7). The direct sum condition (2.16) is easily seen to be equivalent to the invertibility of the $2n \times 2n$ matrix $\begin{bmatrix} I & Q \\ P & I \end{bmatrix}$. From the factorization

$$\begin{bmatrix} I & Q \\ P & I \end{bmatrix} = \begin{bmatrix} I & Q \\ 0 & I \end{bmatrix} \begin{bmatrix} I - QP & 0 \\ 0 & I \end{bmatrix} \begin{bmatrix} I & 0 \\ P & I \end{bmatrix},$$

we see that this in turn is equivalent to the invertibility of $I - QP$. We now conclude that the Wiener-Hopf factorization (2.15) exists if and only if $I - QP$ is invertible, where P and Q are the unique solutions of the respective Lyapunov equations (2.7) and (2.8).

To convert the Wiener-Hopf factorization (2.15) to the desired signed spectral factorization (2.5), we use the symmetry $W = W^*$ of W together with the uniqueness of Wiener-Hopf factorization. More precisely, if the normalization

$$X_-(\infty) = \begin{bmatrix} I_p & 0 \\ 0 & I_q \end{bmatrix} = X_+(\infty)$$

is imposed, then the factors $X_+(z)$ and $X_-(z)$ in the Wiener-Hopf factorization are uniquely determined; indeed if $\tilde{D}^{-1}W(z) = X'_-(z)X'_+(z)$ were another such factorization, then $X'_-(z)^{-1}X_-(z) = X'_+(z)X_+(z)^{-1}$ would be analytic on the whole complex plane with value I_{p+q} at ∞ , and hence would be identically equal to I_{p+q} by Liouville's Theorem. For our case, $W(z) = \tilde{D}X_-(z)X_+(z)$ where $W(z) = W^*(z)$, and hence we also have $W(z) = X_+^*(z)X_-^*(z)\tilde{D}^*$. This yields the factorization

$$\tilde{D}^{-1}W(z) = \tilde{D}^{-1}X_+^*(z)X_-^*(z)\tilde{D}^*.$$

Evaluating at ∞ yields $I = \tilde{D}^{-1}X_+(0)^*X_-(0)^*\tilde{D}^*$, so $X_+(0)^{*^{-1}}\tilde{D} = [\tilde{D}^{*-1}X_-(0)^{*^{-1}}]^{-1}$. Then we rewrite the above factorization as

$$(2.19) \quad \tilde{D}^{-1}W(z) = X'_-(z)X'_+(z)$$

where

$$X'_-(z) = \tilde{D}^{-1}X_+^*(z)X_+(0)^{*^{-1}}\tilde{D}$$

and

$$X'_+(z) = \tilde{D}^{*-1}X_-(0)^{*^{-1}}X_-^*(z)\tilde{D}^*.$$

Note that X'_- is analytic and invertible on $\mathbb{C} \setminus D$ with $X'_-(\infty) = I$ and similarly X'_+ is analytic and invertible on \bar{D} with $X'_+(\infty) = I$. Thus the factorization (2.19) is another normalized Wiener-Hopf factorization for $\tilde{D}^{-1}W(z)$. By the uniqueness mentioned above we necessarily have $X'_-(z) = X_-(z)$. Thus the factorization (2.15) has the form

$$(2.20) \quad W(z) = X_+^*(z)X_+(0)^{*^{-1}}\tilde{D}X_+(z).$$

If we evaluate this expression on the unit circle, we see that the matrix $c := X_+(0)^{*^{-1}}\tilde{D}$ is Hermitian with p positive and q negative eigenvalues and hence has a signed Cholesky factorization

$$c = d^* \begin{bmatrix} I_p & 0 \\ 0 & -I_q \end{bmatrix} d$$

for some matrix d . If we set

$$(2.21) \quad X(z) = dX_+(z),$$

then

$$W(z) = X^*(z) \begin{bmatrix} I_p & 0 \\ 0 & -I_q \end{bmatrix} X(z)$$

is our desired signed spectral factorization.

It remains only to compute the explicit formulas for c , $X(z)$ and $X(z)^{-1}$ in the statement of the lemma. A straightforward computation shows that the projection Π of \mathbb{C}^{2n} onto

$$\mathcal{M}^x = \text{Im} \begin{bmatrix} I \\ P \end{bmatrix} \quad \text{along} \quad \mathcal{M} = \text{Im} \begin{bmatrix} Q \\ I \end{bmatrix}$$

is given by

$$\Pi = \begin{bmatrix} I \\ P \end{bmatrix} Z[I, -Q]$$

where we have set $Z = (I - QP)^{-1}$.

We identify $\text{Im } \Pi$ with \mathbb{C}^n via the map $S: \mathbb{C}^n \rightarrow \text{Im } \Pi$ given by

$$S = \begin{bmatrix} I \\ P \end{bmatrix},$$

with inverse $S^{-1}: \text{Im } \Pi \rightarrow \mathbb{C}^n$ given by

$$S^{-1} = [I \ 0] | \text{Im } \Pi.$$

Rewrite (2.17) and (2.18) as

$$(2.22) \quad X_+(z) = I_{p+q} + \tilde{D}^{-1} \tilde{C} \Pi S (zI - S^{-1} \Pi \tilde{A} \Pi S)^{-1} S^{-1} \tilde{B}$$

and

$$(2.23) \quad X_+(z)^{-1} = I_{p+q} - \tilde{D}^{-1} \tilde{C} S (zI - S^{-1} \Pi \tilde{A}^x \Pi S)^{-1} S^{-1} \Pi \tilde{B}.$$

Use (2.13) and (2.8) to compute

$$\begin{aligned} S^{-1} \Pi \tilde{A} \Pi S &= S^{-1} \Pi \tilde{A} S = Z[I, -Q] \begin{bmatrix} A^{*-1} & A^{*-1} C^* C \\ 0 & A \end{bmatrix} \begin{bmatrix} I \\ P \end{bmatrix} \\ &= Z[A^{*-1}, A^{*-1} C^* C - QA] \begin{bmatrix} I \\ P \end{bmatrix} \\ &= ZA^{*-1} + Z[QA - A^{*-1} Q]P - ZQAP \\ &= ZA^{*-1} Z^{-1}. \end{aligned}$$

From (2.13) we get

$$(2.24) \quad \begin{aligned} \tilde{D}^{-1} \tilde{C} S &= \begin{bmatrix} I_p & 0 \\ -\sigma^{-2} B^* A^{*-1} C^* & -\sigma^{-2} I_q \end{bmatrix} \begin{bmatrix} CP \\ -B^* A^{*-1} - B^* A^{*-1} C^* CP \end{bmatrix} \\ &= \begin{bmatrix} CP \\ \sigma^{-2} B^* A^{*-1} \end{bmatrix} \end{aligned}$$

and

$$(2.25) \quad \begin{aligned} S^{-1} \Pi \tilde{B} &= Z[I, -Q] \begin{bmatrix} A^{*-1} C^* & 0 \\ 0 & B \end{bmatrix} \\ &= Z[A^{*-1} C^*, -QB]. \end{aligned}$$

From (2.14) we get

$$(2.26) \quad \begin{aligned} S^{-1} \Pi \tilde{A}^x S &= S^{-1} \tilde{A}^x S \\ &= [I \ 0] \begin{bmatrix} A^{*-1} & 0 \\ -\sigma^{-2} B B^* A^{*-1} & A \end{bmatrix} \begin{bmatrix} I \\ P \end{bmatrix} \\ &= A^{*-1}. \end{aligned}$$

From (2.23)

$$\begin{aligned} X_+(0)^{-1} &= I_{p+q} + \tilde{D}^{-1} \tilde{C} S (S^{-1} \Pi \tilde{A}^x \Pi S)^{-1} S^{-1} \Pi \tilde{B} \\ &= I_{p+q} + \begin{bmatrix} CP \\ \sigma^{-2} B^* A^{*-1} \end{bmatrix} A^* Z[A^{*-1} C^*, -QB]. \end{aligned}$$

We compute

$$\begin{aligned}
 c &= X_+(0)^{*^{-1}} \tilde{D} \\
 &= \left\{ I_{p+q} + \begin{bmatrix} CA^{-1} \\ -B^*Q \end{bmatrix} Z^* A [PC^*, \sigma^{-2}A^{-1}B] \right\} \begin{bmatrix} I_p & 0 \\ -B^*A^{*-1}C^* & -\sigma^2I_q \end{bmatrix} \\
 &= \begin{bmatrix} I_p & 0 \\ -B^*A^{*-1}C^* & -\sigma^2I_q \end{bmatrix} \\
 &\quad + \begin{bmatrix} CA^{-1} \\ -B^*Q \end{bmatrix} Z^* A [PC^* - \sigma^{-2}A^{-1}BB^*A^{*-1}C^*, -A^{-1}B].
 \end{aligned}$$

Use (2.7) to then get

$$c = \begin{bmatrix} I_p & 0 \\ -B^*A^{*-1}C^* & -\sigma^2I_q \end{bmatrix} + \begin{bmatrix} CA^{-1} \\ -B^*Q \end{bmatrix} Z^* A [A^{-1}PA^{*-1}C^*, -A^{-1}B].$$

Write

$$c = \begin{bmatrix} c_{11} & c_{12} \\ c_{21} & c_{22} \end{bmatrix}$$

and get from this

$$\begin{aligned}
 c_{11} &= I_p + CA^{-1}Z^*PA^{*-1}C^*, \\
 c_{12} &= -CA^{-1}Z^*B, \\
 c_{21} &= -B^*A^{*-1}C^* - B^*QZ^*PA^{*-1}C^* \\
 &= -B^*[I + ZQP]A^{*-1}C^* \\
 &= -B^*ZA^{*-1}C^*
 \end{aligned}$$

and

$$c_{22} = -\sigma^2I_q + B^*QZ^*B = -\sigma^2I_q + B^*ZQB,$$

which agrees with (2.9). Finally, formulas (2.11) and (2.12) follow by plugging the expressions (2.24), (2.25), (2.26), (2.27) into (2.21), (2.22), (2.23), and the lemma follows. \square

We are now in position to complete the proof of Theorem 1 in the Introduction.

Proof of Theorem 1. By Theorem 2.1 and Lemma 2.2, the matrix function

$$\theta(z) = \begin{bmatrix} I_p & K(z) \\ 0 & I_q \end{bmatrix} X(z)^{-1}$$

satisfies all the requirements of Theorem 1 if the factorization (2.5) is a signed spectral factorization for the matrix function $W(z)$ defined by (2.4). We assume that $K(z) = C(zI - A)^{-1}B$ where the $n \times n$ matrix A has spectrum in D . Suppose that P and Q are the unique solutions of the Lyapunov equations (2.7) and (2.8). Then $\hat{P} := -\sigma^2P$ and $\hat{Q} := -Q$ solve (1.2) and (1.3), and hence are the controllability and observability gramians for the system associated with the above realization of $K(z)$. Therefore $I_n - QP = I_n - \sigma^{-2}\hat{Q}\hat{P}$ is invertible if and only if σ is not one of the Hankel singular values $\sigma_j(K)$ of K . Then by Lemma 2.3, $W(z)$ has a signed spectral factorization as in (2.5) where $X(z)^{-1}$ is given by (2.12). We may write

$$\begin{bmatrix} I_p & K(z) \\ 0 & I_q \end{bmatrix}$$

in realization form as

$$\begin{bmatrix} I_p & K(z) \\ 0 & I_q \end{bmatrix} = \begin{bmatrix} I_p & 0 \\ 0 & I_q \end{bmatrix} + \begin{bmatrix} C \\ 0 \end{bmatrix} (zI_n - A)^{-1} [0, B].$$

We then multiply transfer functions as, e.g., [5, p. 6] to obtain

$$\theta(z) = \begin{bmatrix} I_p & K(z) \\ 0 & I_q \end{bmatrix} X(z)^{-1} = \{I + \underline{C}(zI_{2n} - \underline{A})^{-1} \underline{B}\} d^{-1}$$

where

$$\underline{A} = \begin{bmatrix} A & -\sigma^{-2}BB^*A^{*-1} \\ 0 & A^{*-1} \end{bmatrix}, \quad \underline{B} = \begin{bmatrix} 0 & B \\ ZA^{*-1}C^* & Z\hat{Q}B \end{bmatrix}$$

and

$$\underline{C} = \begin{bmatrix} C & \sigma^{-2}C\hat{P} \\ 0 & -\sigma^{-2}B^*A^{*-1} \end{bmatrix}.$$

We may use the Lyapunov equation (1.2) to rewrite \underline{A} as

$$\underline{A} = \begin{bmatrix} A & \sigma^{-2}(A\hat{P} - \hat{P}A^{*-1}) \\ 0 & A^{*-1} \end{bmatrix}$$

from which we get

$$(zI_{2n} - \underline{A})^{-1} = \begin{bmatrix} (zI_n - A)^{-1} & \sigma^{-2}(zI_n - A)^{-1}(A\hat{P} - \hat{P}A^{*-1})(zI_n - A^{*-1})^{-1} \\ 0 & (zI_n - A^{*-1})^{-1} \end{bmatrix}.$$

We compute first

$$(zI_{2n} - \underline{A})^{-1} \underline{B} = \begin{bmatrix} x_{11} & x_{12} \\ x_{21} & x_{22} \end{bmatrix}$$

where

$$\begin{aligned} x_{11} &= \sigma^{-2}(zI_n - A)^{-1}(A\hat{P} - \hat{P}A^{*-1})(zI_n - A^{*-1})^{-1}ZA^{*-1}C^*, \\ x_{12} &= (zI_n - A)^{-1}B + \sigma^{-2}(zI_n - A)^{-1}(A\hat{P} - \hat{P}A^{*-1})(zI_n - A^{*-1})^{-1}Z\hat{Q}B, \\ x_{21} &= (zI_n - A^{*-1})^{-1}ZA^{*-1}C^*, \\ x_{22} &= (zI_n - A^{*-1})^{-1}Z\hat{Q}B. \end{aligned}$$

Now we compute

$$\underline{C}(zI_{2n} - \underline{A})^{-1} \underline{B} = \begin{bmatrix} y_{11} & y_{12} \\ y_{21} & y_{22} \end{bmatrix}$$

where

$$\begin{aligned} y_{11} &= Cx_{11} + \sigma^{-2}C\hat{P}x_{21} \\ &= \sigma^{-2}C(zI_n - A)^{-1}(A\hat{P} - \hat{P}A^{*-1})(zI_n - A^{*-1})^{-1}ZA^{*-1}C^* \\ &\quad + \sigma^{-2}C\hat{P}(zI_n - A^{*-1})^{-1}ZA^{*-1}C^* \\ &= \sigma^{-2}C(zI_n - A)^{-1}\{A\hat{P} - \hat{P}A^{*-1} + (zI_n - A)\hat{P}\}(zI_n - A^{*-1})^{-1}ZA^{*-1}C^* \\ &= \sigma^{-2}C(zI_n - A)^{-1}\hat{P}ZA^{*-1}C^* \end{aligned}$$

and

$$\begin{aligned}
 y_{12} &= Cx_{12} + \sigma^{-2}C\hat{P}x_{22} \\
 &= C(zI_n - A)^{-1}B + \sigma^{-2}C(zI_n - A)^{-1}(A\hat{P} - \hat{P}A^{*-1})(zI_n - A^{*-1})^{-1}Z\hat{Q}B \\
 &\quad + \sigma^{-2}C\hat{P}(zI_n - A^{*-1})^{-1}Z\hat{Q}B \\
 &= C(zI_n - A)^{-1}B + \sigma^{-2}C(zI_n - A)^{-1}\{A\hat{P} - \hat{P}A^{*-1} + (zI_n - A)\hat{P}\}(zI_n - A^{*-1})^{-1}Z\hat{Q}B \\
 &= C(zI_n - A)^{-1}B + \sigma^{-2}C(zI_n - A)^{-1}\hat{P}Z\hat{Q}B \\
 &= C(zI_n - A)^{-1}(I + \sigma^{-2}\hat{P}Z\hat{Q})B.
 \end{aligned}$$

Noting that $I + \sigma^{-2}\hat{P}Z\hat{Q} = I + PZQ = I + PQZ^* = ((I - PQ) + PQ)Z^* = Z^*$, we obtain that

$$y_{12} = C(zI_n - A)^{-1}Z^*B.$$

Further,

$$\begin{aligned}
 y_{21} &= -\sigma^{-2}B^*A^{*-1}x_{21} \\
 &= -\sigma^{-2}B^*A^{*-1}(zI_n - A^{*-1})^{-1}ZA^{*-1}C^*,
 \end{aligned}$$

and finally,

$$\begin{aligned}
 y_{22} &= -\sigma^{-2}B^*A^{*-1}x_{22} \\
 &= -\sigma^{-2}B^*A^{*-1}(zI_n - A^{*-1})^{-1}Z\hat{Q}B.
 \end{aligned}$$

If we now write d^{-1} in block form

$$d^{-1} = \begin{bmatrix} e_{11} & e_{12} \\ e_{21} & e_{22} \end{bmatrix},$$

then

$$\theta(z) = \begin{bmatrix} \theta_{11}(z) & \theta_{12}(z) \\ \theta_{21}(z) & \theta_{22}(z) \end{bmatrix}$$

where

$$\begin{aligned}
 \theta_{11} &= e_{11} + y_{11}e_{11} + y_{12}e_{21}, & \theta_{12} &= e_{12} + y_{11}e_{12} + y_{12}e_{22}, \\
 \theta_{21} &= e_{21} + y_{21}e_{11} + y_{22}e_{21}, & \theta_{22} &= e_{22} + y_{21}e_{12} + y_{22}e_{22}.
 \end{aligned}$$

Substitution of the above expressions for y_{ij} into these expressions for θ_{ij} now yield formulas (1.6)–(1.9). This completes the proof of Theorem 1. \square

3. The continuous time case. In this section we consider the continuous time analogue of the above results; this is the setting of Glover's original paper [8]. These results were obtained in [4] directly by reducing to a signed spectral factorization problem as in § 1 but with the left half plane in place of the unit disk; here we simply reduce to the discrete time case by a linear fractional change of variable. We suppose that we are given a stable rational $p \times q$ matrix function of McMillan degree n which is analytic on the imaginary $j\omega$ axis with $G(\infty) = 0$; for the present continuous time setting, “stable” means that G has no poles in the right half plane $\{s: \operatorname{Re} s > 0\}$. Thus $G(s)$ may be realized as

$$G(s) = C(sI - A)^{-1}B$$

where A , B , C are matrices of sizes $n \times n$, $n \times q$ and $p \times n$, respectively, and where the spectrum of A lies in the open left half plane.

The symbol $H_{p \times q}^{\infty+}$ now denotes the class of $p \times q$ matrix functions analytic in the right half plane and uniformly bounded there; $H_{p \times q}^{\infty-}$ is the analogous class for the left half plane. For l a nonnegative integer, we let $H_{p \times q}^{\infty-}(l)$ denote the class of functions $\hat{G} + F$ where \hat{G} has McMillan degree l and $F \in H_{p \times q}^{\infty-}$. We now let $L_{p \times q}^{\infty}$ denote the class of measurable uniformly bounded functions K on the imaginary axis with norm

$$\|K\|_{L^{\infty}} = \sup_{-\infty < w < \infty} \|K(jw)\|.$$

We also introduce the Hardy spaces H_q^{2-} for the left half plane and H_q^{2+} for the right half plane of \mathbb{C}^q -valued functions and the Lebesgue space L_q^2 of \mathbb{C}^q -valued functions which are norm square integrable on the jw axis.

The problem to be considered in this section is that of approximating a given stable rational matrix function $G(s) = C(sI - A)^{-1}B$ by functions $K(s) \in H_{p \times q}^{\infty-}(l)$ in L^{∞} -norm. From the earlier work of [1], [10] and [3] translated to the continuous time setting by a linear fractional change of variable, it is known that

$$\inf \{\|G - K\|_{L^{\infty}} : K \in H_{p \times q}^{\infty-}(l)\} = \sigma_{l+1}(G)$$

where $\sigma_{l+1}(G)$ is the $(l+1)$ st Hankel singular value of G . The Hankel singular values $\sigma_1(G) \geq \sigma_2(G) \geq \dots \geq \sigma_n(G) > 0$ can be defined for the continuous time case as the square roots of the eigenvalues $\lambda_i(\hat{P}\hat{Q})$ of the product of the controllability gramian

$$\hat{P} := \int_0^{\infty} \exp(At)BB^* \exp(A^*t) dt$$

and the observability gramian

$$\hat{Q} := \int_0^{\infty} \exp(A^*t)C^*C \exp(At) dt.$$

Equivalently, they can be defined as the singular values $\sigma_i(\mathcal{H}_G)$ of the Hankel operator $\mathcal{H}_G : L_q^2(0, \infty) \rightarrow L_q^2(0, \infty)$ defined by

$$(3.1) \quad \mathcal{H}_G : v(t) \rightarrow \int_0^{\infty} C \exp(A(t+\tau))Bv(\tau) d\tau$$

(see [8]). With σ and l chosen so that $\sigma_l(G) > \sigma \geq \sigma_{l+1}(G)$ (or $\sigma \geq \sigma_1(G)$ if $l=0$), there is again a linear fractional map parametrization for the set of all $K \in H_{p \times q}^{\infty-}(l)$ satisfying $\|G - K\|_{L^{\infty}} \leq \sigma$. The form of the linear fractional map simplifies considerably if $\sigma \neq \sigma_{l+1}(G)$. Glover [8] was the first to give formulas for the coefficients of the linear fractional map directly in terms of the matrices A, B, C appearing in the realization of G . The result for the present continuous time setting is as follows.

THEOREM 3.1. *Suppose $G(s) = C(sI - A)^{-1}B$ is a stable rational $p \times q$ matrix function of McMillan degree n , and that the number σ and the integer l are chosen to satisfy $\sigma_l(G) > \sigma > \sigma_{l+1}(G)$ or $\sigma > \sigma_1(G)$ if $l=0$. Then there is a rational $(p+q) \times (p+q)$ matrix function*

$$\theta(s) = \begin{bmatrix} \theta_{11}(s) & \theta_{12}(s) \\ \theta_{21}(s) & \theta_{22}(s) \end{bmatrix}$$

such that a matrix function $\hat{F} \in L_{p \times q}^{\infty}$ is of the form $\hat{F} = G - K$ for some $K \in H_{p \times q}^{\infty-}(l)$ with $\|F\|_{L^{\infty}} \leq \sigma$ if and only if

$$\hat{F}(s) = (\theta_{11}(s)H(s) + \theta_{12}(s))(\theta_{21}(s)H(s) + \theta_{22}(s))^{-1}$$

for some $H \in H_{p \times q}^{\infty-}$ with $\|H\|_{L^{\infty}} \leq 1$. The function \hat{F} uniquely determines the function H .

Moreover, \hat{F} is rational if and only if the corresponding H is rational, and $(1/\sigma)\hat{F}(jw)$ is isometric for all real w if and only if $H(jw)$ is isometric for all real w . The matrix function $\theta(s)$ is given by

$$\theta(s) = \begin{bmatrix} I_p & 0 \\ 0 & \sigma^{-1}I_q \end{bmatrix} + \underline{C}(sI - \underline{A})^{-1}\underline{B}$$

where

$$(3.2) \quad \underline{A} = \begin{bmatrix} A & -\sigma^{-2}BB^* \\ 0 & -A^* \end{bmatrix},$$

$$(3.3) \quad \underline{B} = \begin{bmatrix} 0 & \sigma^{-1}B \\ ZC^* & \sigma^{-1}Z\hat{Q}B \end{bmatrix},$$

$$(3.4) \quad \underline{C} = \begin{bmatrix} C & \sigma^{-2}C\hat{P} \\ 0 & -\sigma^{-2}B^* \end{bmatrix}.$$

Equivalently, the block entries $\theta_{ij}(s)$ ($i, j = 1, 2$) are given by

$$(3.5) \quad \theta_{11}(s) = I_p + \sigma^{-2}C(sI - A)^{-1}\hat{P}ZC^*,$$

$$(3.6) \quad \theta_{12}(s) = \sigma^{-1}C(sI - A)^{-1}Z^*B,$$

$$(3.7) \quad \theta_{21}(s) = -\sigma^{-2}B^*(sI + A^*)^{-1}ZC^*,$$

$$(3.8) \quad \theta_{22}(s) = \sigma^{-1}I_q - \sigma^{-3}B^*(sI + A^*)^{-1}Z\hat{Q}B.$$

Here \hat{P} and \hat{Q} are the controllability and observability gramian, respectively, associated with the realization $G(s) = C(sI - A)^{-1}B$, and $Z = (I - \sigma^{-2}\hat{Q}\hat{P})^{-1}$.

The controllability and observability gramians \hat{P} and \hat{Q} alternatively arise as the unique solutions of the Lyapunov equations

$$(3.9) \quad A\hat{P} + \hat{P}A^* = -BB^*,$$

$$(3.10) \quad A^*\hat{Q} + \hat{Q}A = -C^*C,$$

(see [8]).

The linear fractional map in Glover's solution [8] has a different form from that given in Theorem 3.1. Any linear fractional map of the form

$$H \rightarrow (\theta_{11}H + \theta_{12})(\theta_{21}H + \theta_{22})^{-1}$$

can be represented in the equivalent form

$$(3.11) \quad H \rightarrow H_{11} + H_{12}H(I - H_{22}H)^{-1}H_{21}$$

where

$$\begin{bmatrix} H_{11} & H_{12} \\ H_{21} & H_{22} \end{bmatrix} = \begin{bmatrix} \theta_{12}\theta_{22}^{-1} & \theta_{11} - \theta_{12}\theta_{22}^{-1}\theta_{21} \\ \theta_{22}^{-1} & -\theta_{22}^{-1}\theta_{21} \end{bmatrix}.$$

When this is done for the case where θ_{11} , θ_{12} , θ_{21} , θ_{22} are given by (3.5)–(3.8), one obtains by a straightforward but tedious computation

$$(3.12) \quad \begin{aligned} H_{11}(s) &= \sigma^{-2}C\hat{P}(sI + Z(A^* + \sigma^{-2}\hat{Q}A\hat{P}))^{-1}Z\hat{Q}B + C(sI - A)^{-1}B, \\ H_{12}(s) &= I + \sigma^{-2}C\hat{P}(sI + Z(A^* + \sigma^{-2}\hat{Q}A\hat{P}))^{-1}ZC^*, \\ H_{21}(s) &= \sigma I + \sigma^{-1}B^*(sI + Z(A^* + \sigma^{-2}\hat{Q}A\hat{P}))^{-1}Z\hat{Q}B, \\ H_{22}(s) &= \sigma^{-1}B^*(sI + Z(A^* + \sigma^{-2}\hat{Q}A\hat{P}))^{-1}ZC^*. \end{aligned}$$

The parametrization of the errors $G - \hat{G} - F$ given by (3.11) and (3.12) agrees with Glover's parametrization of the approximants $\hat{G} + F$ ([8, Cor. 8.6], interpreted for the case $\sigma_{l+1}(G) < \sigma < \sigma_l(G)$ or $\sigma_1(G) < \sigma$ if $l = 0$).

Proof of Theorem 3.1. As is well known and is easily checked, the map $s \rightarrow z(s) = (1+s)/(1-s)$ maps the left half plane conformally onto the unit disk with inverse $z \rightarrow s(z) = (z-1)/(z+1)$. Thus, if $F = F(s)$ is in the Hardy space $H_{p \times q}^\infty(l)$ for the left half plane, then $\tilde{F}(z) := F(s(z)) = F((z-1)/(z+1))$ is in the corresponding space $H_{p \times q}^\infty(l)$ for the unit disk, and conversely, if $K(z) \in H_{p \times q}^\infty(l)$ for the unit disk, then $K'(s) = K(z(s)) = K((1+s)/(1-s))$ is in the space $H_{p \times q}^\infty(l)$ for the left half plane.

Now suppose $G(s) = C(sI - A)^{-1}B$ is a rational matrix function as in Theorem 3.1. In addition to $\sigma(A) \subset \{s: \operatorname{Re} s < 0\}$, we assume $-1 \notin \sigma(A)$; this hypothesis can be avoided by using a different linear fractional change of variables. Then we set $\tilde{G}(z) = G(s(z))$ and note that $\tilde{G}(z) = \tilde{D} + \tilde{C}(zI - \tilde{A})\tilde{B}$, where

$$(3.13) \quad \tilde{D} = C(I - A)^{-1}B,$$

$$(3.14) \quad \tilde{C} = \sqrt{2}C(I - A)^{-1},$$

$$(3.15) \quad \tilde{B} = \sqrt{2}(I - A)^{-1}B,$$

$$(3.16) \quad \tilde{A} = (I + A)(I - A)^{-1}.$$

The extra assumption that $-1 \notin \sigma(A)$ implies that \tilde{A} is invertible; also $\sigma(\tilde{A}) \subset D$ since A has spectrum in the left half plane. Thus we can apply the formulas in Theorem 1 to $\tilde{G}(z)$ in place of $K(z)$. Assume that the number σ is such that $\sigma_{l+1}(\tilde{G}) < \sigma < \sigma_l(\tilde{G})$ (or simply $\sigma_1(\tilde{G}) < \sigma$ if $l = 0$). We shall see later that the singular values $\sigma_i(\tilde{G})$ for the discrete time Hankel matrix $\mathcal{H}_{\tilde{G}}$ are identical with the singular values $\sigma_i(G)$ for the continuous time Hankel operator \mathcal{H}_G given by (3.1). Let

$$\tilde{\theta}(z) = \begin{bmatrix} \tilde{\theta}_{11}(z) & \tilde{\theta}_{12}(z) \\ \tilde{\theta}_{21}(z) & \tilde{\theta}_{22}(z) \end{bmatrix}$$

be the matrix function associated with \tilde{G} and σ as in Theorem 1. Thus

$$(3.17) \quad \tilde{\theta}(z) = \{I_{p+q} + \tilde{C}(zI - \tilde{A})^{-1}\tilde{B}\}\tilde{\theta}(\infty)$$

where

$$(3.18) \quad \tilde{A} = \begin{bmatrix} \tilde{A} & -\sigma^{-2}\tilde{B}\tilde{B}^*\tilde{A}^{*-1} \\ 0 & \tilde{A}^{*-1} \end{bmatrix},$$

$$(3.19) \quad \tilde{B} = \begin{bmatrix} 0 & \tilde{B} \\ \tilde{Z}\tilde{A}^{*-1}\tilde{C}^* & \tilde{Z}\tilde{Q}\tilde{B} \end{bmatrix},$$

$$(3.20) \quad \tilde{C} = \begin{bmatrix} \tilde{C} & \sigma^{-2}\tilde{C}\tilde{P} \\ 0 & -\sigma^{-2}\tilde{B}^*\tilde{A}^{*-1} \end{bmatrix}.$$

Here \tilde{P} and \tilde{Q} are solutions of the discrete time Lyapunov equations

$$(3.21) \quad \tilde{P} - \tilde{A}\tilde{P}\tilde{A}^* = \tilde{B}\tilde{B}^*,$$

$$(3.22) \quad \tilde{Q} - \tilde{A}^*\tilde{Q}\tilde{A} = \tilde{C}^*\tilde{C}$$

and $\tilde{Z} = (I - \sigma^{-2}\tilde{Q}\tilde{P})^{-1}$. The result of Theorem 1 then is that \tilde{F} satisfies

$$\sup \{\|\tilde{F}(z)\|: |z| = 1\} \leq \sigma$$

and

$$\tilde{F} \in \tilde{G} + H_{p \times q}^\infty(l)$$

if and only if \tilde{F} has the form

$$\tilde{F}(z) = T_{\tilde{\theta}(z)}(\tilde{H}(z))$$

for some $\tilde{H} \in H_{p \times q}^\infty$ with $\|\tilde{H}\|_\infty \leq 1$ (sup norm on the unit circle), where in general we define

$$T_\theta(X) = (\theta_{11}X + \theta_{12})(\theta_{21}X + \theta_{22})^{-1}$$

if

$$\theta = \begin{bmatrix} \theta_{11} & \theta_{12} \\ \theta_{21} & \theta_{22} \end{bmatrix}.$$

If we make the change of variable $z \rightarrow z(s)$ and define

$$\theta(s) = \tilde{\theta}(z(s)), \quad G(s) = \tilde{G}(z(s)) \quad \text{and} \quad F(s) = \tilde{F}(z(s)),$$

we see from this that $\theta(s) = \tilde{\theta}(z(s))$ is as desired in Theorem 3.1. It remains only to use (3.13)–(3.22) to compute $\theta(s)$ explicitly in terms of the original A , B , C and continuous time controllability and observability gramians \hat{P} and \hat{Q} .

We first convert the discrete time gramians \tilde{P} , \tilde{Q} to the continuous time gramians \hat{P} , \hat{Q} . From (3.15) and (3.16), the Lyapunov equation (3.21) is equivalent to

$$\tilde{P} - (I + A)(I - A)^{-1}\tilde{P}(I - A^*)^{-1}(I + A^*) = 2(I - A)^{-1}BB^*(I - A^*)^{-1}$$

or

$$(I - A)\tilde{P}(I - A^*) - (I + A)\tilde{P}(I + A)^* = 2BB^*.$$

This collapses to

$$A^*\tilde{P} + \tilde{P}A^* = -B^*B.$$

From (3.9) we see that the discrete time controllability gramian \tilde{P} is identical to the continuous time controllability gramian \hat{P} . Similarly, using (3.14) and (3.16) we see that (3.22) collapses to

$$A^*\tilde{Q} + \tilde{Q}A = -C^*C,$$

so $\tilde{Q} = \hat{Q}$. Thus $\tilde{Z} = (I - \sigma^{-2}\tilde{Q}\tilde{P})^{-1} = (I - \sigma^{-2}\hat{Q}\hat{P})^{-1} = Z$ as well.

The matrix function $\theta(s)$ is determined by the conditions in Theorem 3.1 only up to a constant J -unitary right factor ($J = \begin{bmatrix} I_p & 0 \\ 0 & -I_q \end{bmatrix}$). By using this freedom we can arrange that

$$\theta(\infty) = \begin{bmatrix} I_p & 0 \\ 0 & \sigma^{-1}I_q \end{bmatrix}.$$

Thus with this normalization $\theta(s)$ is given by

$$\theta(s) = \tilde{\theta}(z(s))\tilde{\theta}(z(\infty))^{-1} \begin{bmatrix} I_p & 0 \\ 0 & \sigma^{-1}I_q \end{bmatrix}$$

where

$$\tilde{\theta}(z(s)) = \left[I_{p+q} + \tilde{C} \left(\frac{1+s}{1-s} I - \tilde{A} \right)^{-1} \tilde{B} \right] \tilde{\theta}(\infty).$$

Thus

$$\tilde{\theta}(z(\infty)) = [I_{p+q} - \tilde{C}(I + \tilde{A})^{-1}\tilde{B}]\tilde{\theta}(\infty).$$

Compute next

$$\begin{aligned}
 \tilde{\theta}(z(s)) - \tilde{\theta}(z(\infty)) &= \tilde{C} \left[\left(\frac{1+s}{1-s} I - \tilde{A} \right)^{-1} + (I + \tilde{A})^{-1} \right] \tilde{B} \tilde{\theta}(\infty) \\
 &= \tilde{C} \left(\frac{1+s}{1-s} I - \tilde{A} \right)^{-1} \left[I + \tilde{A} + \frac{1+s}{1-s} I - \tilde{A} \right] (I + \tilde{A})^{-1} \tilde{B} \tilde{\theta}(\infty) \\
 &= \frac{2}{1-s} \tilde{C} \left(\frac{1+s}{1-s} I - \tilde{A} \right)^{-1} (I + \tilde{A})^{-1} \tilde{B} \tilde{\theta}(\infty) \\
 &= 2 \tilde{C} (s(I + \tilde{A}) + (I - \tilde{A}))^{-1} (I + \tilde{A})^{-1} \tilde{B} \tilde{\theta}(\infty) \\
 &\quad \cdot 2 \tilde{C} (I + \tilde{A})^{-1} (sI - (\tilde{A} - I)(\tilde{A} + I)^{-1})^{-1} (I + \tilde{A})^{-1} \tilde{B} \tilde{\theta}(\infty).
 \end{aligned}$$

Thus

$$\begin{aligned}
 \theta(s) &= \begin{bmatrix} I_p & 0 \\ 0 & \sigma^{-1} I_q \end{bmatrix} + 2 \tilde{C} (I + \tilde{A})^{-1} (sI - (\tilde{A} - I)(\tilde{A} + I)^{-1})^{-1} (I + \tilde{A})^{-1} \\
 &\quad \cdot \tilde{B} [I_{p+q} - \tilde{C} (I + \tilde{A})^{-1} \tilde{B}]^{-1} \begin{bmatrix} I_p & 0 \\ 0 & \sigma^{-1} I_q \end{bmatrix}.
 \end{aligned}$$

We thus see that

$$\theta(s) = \begin{bmatrix} I_p & 0 \\ 0 & \sigma^{-1} I_q \end{bmatrix} + \underline{C} (sI - \underline{A})^{-1} \underline{B},$$

where

$$(3.23) \quad \underline{A} = (\tilde{A} - I)(\tilde{A} + I)^{-1},$$

$$(3.24) \quad \underline{B} = \sqrt{2} (I + \tilde{A})^{-1} \tilde{B} [I_{p+q} - \tilde{C} (I + \tilde{A})^{-1} \tilde{B}]^{-1} \begin{bmatrix} I_p & 0 \\ 0 & \sigma^{-1} I_q \end{bmatrix}$$

and

$$(3.25) \quad \underline{C} = \sqrt{2} \tilde{C} (I + \tilde{A})^{-1}.$$

The computation of \underline{A} is aided by noting from the definition (3.15) of \tilde{A} and the Lyapunov equation (3.18) that \tilde{A} can be block-diagonalized.

$$(3.26) \quad \tilde{A} = \begin{bmatrix} \tilde{A} & \sigma^{-2} \tilde{A} \hat{P} - \sigma^{-2} \hat{P} \tilde{A}^{*-1} \\ 0 & \tilde{A}^{*-1} \end{bmatrix} = \begin{bmatrix} I & -\sigma^{-2} \hat{P} \\ 0 & I \end{bmatrix} \begin{bmatrix} \tilde{A} & 0 \\ 0 & \tilde{A}^{*-1} \end{bmatrix} \begin{bmatrix} I & \sigma^{-2} \hat{P} \\ 0 & I \end{bmatrix}.$$

From (3.16) it is immediate that

$$(\tilde{A} - I)(\tilde{A} + I)^{-1} = \underline{A}$$

and

$$(\tilde{A}^{*-1} - I)(\tilde{A}^{*-1} + I)^{-1} = -\underline{A}^*.$$

Thus from this, (3.23) and (3.26) we get

$$\begin{aligned}
 \underline{A} &= \begin{bmatrix} I & -\sigma^{-2} \hat{P} \\ 0 & I \end{bmatrix} \begin{bmatrix} \underline{A} & 0 \\ 0 & -\underline{A}^* \end{bmatrix} \begin{bmatrix} I & \sigma^{-2} \hat{P} \\ 0 & I \end{bmatrix} \\
 &= \begin{bmatrix} \underline{A} & \sigma^{-2} \underline{A} \hat{P} + \sigma^{-2} \hat{P} \underline{A}^* \\ 0 & -\underline{A}^* \end{bmatrix} \\
 &= \begin{bmatrix} \underline{A} & -\sigma^{-2} \underline{B} \underline{B}^* \\ 0 & -\underline{A}^* \end{bmatrix}
 \end{aligned}$$

where we used (3.9) for the last step. This verifies the formula (3.2) for \underline{A} .

We next compute \underline{B} from (3.24). To do this we must compute

$$\tilde{B}[I_{p+q} - \tilde{C}(I + \tilde{A})^{-1}\tilde{B}]^{-1} = [I - \tilde{B}\tilde{C}(I + \tilde{A})^{-1}]^{-1}\tilde{B} = (I + \tilde{A})[I + \tilde{A} - \tilde{B}\tilde{C}]^{-1}\tilde{B}.$$

From (3.19) and (3.20) we get

$$\tilde{B}\tilde{C} = \begin{bmatrix} 0 & -\sigma^{-2}\tilde{B}\tilde{B}^*\tilde{A}^{*-1} \\ Z\tilde{A}^{*-1}\tilde{C}^*\tilde{C} & \sigma^{-2}Z\tilde{A}^{*-1}\tilde{C}^*\tilde{C}\hat{P} - \sigma^{-2}Z\hat{Q}\tilde{B}\tilde{B}^*\tilde{A}^{*-1} \end{bmatrix}.$$

Thus, from 3.18

$$\tilde{A} - \tilde{B}\tilde{C} = \begin{bmatrix} \tilde{A} & 0 \\ -Z\tilde{A}^{*-1}\tilde{C}^*\tilde{C} & x_{22} \end{bmatrix}$$

where (by using (3.21) and (3.22))

$$\begin{aligned} x_{22} &= \tilde{A}^{*-1} - \sigma^{-2}Z\tilde{A}^{*-1}\tilde{C}^*\tilde{C}\hat{P} + \sigma^{-2}Z\hat{Q}\tilde{B}\tilde{B}^*\tilde{A}^{*-1} \\ &= Z[(I - \sigma^{-2}\hat{Q}\hat{P})\tilde{A}^{*-1} + \sigma^{-2}(\hat{Q}\tilde{A} - \tilde{A}^{*-1}\hat{Q})\hat{P} + \sigma^{-2}\hat{Q}(\hat{P}\tilde{A}^{*-1} - \tilde{A}\hat{P})] \\ &= Z[\tilde{A}^{*-1} - \sigma^{-2}\tilde{A}^{*-1}\hat{Q}\hat{P}] \\ &= Z\tilde{A}^{*-1}Z^{-1}. \end{aligned}$$

Therefore

$$\begin{aligned} \tilde{A} - \tilde{B}\tilde{C} &= \begin{bmatrix} \tilde{A} & 0 \\ -Z\tilde{A}^{*-1}\tilde{C}^*\tilde{C} & Z\tilde{A}^{*-1}Z^{-1} \end{bmatrix} \\ &= \begin{bmatrix} I & 0 \\ 0 & Z \end{bmatrix} \begin{bmatrix} \tilde{A} & 0 \\ -\tilde{A}^{*-1}\tilde{C}^*\tilde{C} & \tilde{A}^{*-1} \end{bmatrix} \begin{bmatrix} I & 0 \\ 0 & Z^{-1} \end{bmatrix} \\ &= \begin{bmatrix} I & 0 \\ 0 & Z \end{bmatrix} \begin{bmatrix} \tilde{A} & 0 \\ \hat{Q}\tilde{A} - \tilde{A}^{*-1}\hat{Q} & \tilde{A}^{*-1} \end{bmatrix} \begin{bmatrix} I & 0 \\ 0 & Z^{-1} \end{bmatrix} \\ &= \begin{bmatrix} I & 0 \\ 0 & Z \end{bmatrix} \begin{bmatrix} I & 0 \\ \hat{Q} & I \end{bmatrix} \begin{bmatrix} \tilde{A} & 0 \\ 0 & \tilde{A}^{*-1} \end{bmatrix} \begin{bmatrix} I & 0 \\ -\hat{Q} & I \end{bmatrix} \begin{bmatrix} I & 0 \\ 0 & Z^{-1} \end{bmatrix} \end{aligned}$$

and

$$I + \tilde{A} - \tilde{B}\tilde{C} = \begin{bmatrix} I & 0 \\ Z\hat{Q} & Z \end{bmatrix} \begin{bmatrix} I + \tilde{A} & 0 \\ 0 & I + \tilde{A}^{*-1} \end{bmatrix} \begin{bmatrix} I & 0 \\ -\hat{Q} & Z^{-1} \end{bmatrix},$$

so

$$(I + \tilde{A} - \tilde{B}\tilde{C})^{-1} = \begin{bmatrix} I & 0 \\ Z\hat{Q} & Z \end{bmatrix} \begin{bmatrix} (I + \tilde{A})^{-1} & 0 \\ 0 & (I + \tilde{A}^{*-1})^{-1} \end{bmatrix} \begin{bmatrix} I & 0 \\ -\hat{Q} & Z^{-1} \end{bmatrix}.$$

We finally conclude that

$$\begin{aligned} \tilde{B}[I_{p+q} - \tilde{C}(I + \tilde{A})^{-1}\tilde{B}]^{-1} \\ = (I + \tilde{A}) \begin{bmatrix} I & 0 \\ Z\hat{Q} & Z \end{bmatrix} \begin{bmatrix} (I + \tilde{A})^{-1} & 0 \\ 0 & (I + \tilde{A}^{*-1})^{-1} \end{bmatrix} \begin{bmatrix} I & 0 \\ -\hat{Q} & Z^{-1} \end{bmatrix} \tilde{B}. \end{aligned}$$

From (3.24), plugging in (3.16) we therefore get

$$\underline{B} = \sqrt{2} \begin{bmatrix} I & 0 \\ Z\hat{Q} & Z \end{bmatrix} \begin{bmatrix} \frac{1}{2}(I - A) & 0 \\ 0 & \frac{1}{2}(I + A^*) \end{bmatrix} \begin{bmatrix} I & 0 \\ -\hat{Q} & Z^{-1} \end{bmatrix} \tilde{B} \begin{bmatrix} I_p & 0 \\ 0 & \sigma^{-1}I_q \end{bmatrix}.$$

Substituting (3.19) for \tilde{B} and (3.14)–(3.16) for \tilde{A} , \tilde{B} , \tilde{C} , we get

$$\begin{aligned} \underline{B} &= \sqrt{2} \begin{bmatrix} \frac{1}{2}(I-A) & 0 \\ \frac{1}{2}Z[\hat{Q}(I-A) - (I+A^*)\hat{Q}] & \frac{1}{2}Z(I+A^*)Z^{-1} \end{bmatrix} \\ &\quad \cdot \begin{bmatrix} 0 & \sqrt{2}\sigma^{-1}(I-A)^{-1}B \\ \sqrt{2}Z(I+A^*)^{-1}C^* & \sqrt{2}\sigma^{-1}Z\hat{Q}(I-A)^{-1}B \end{bmatrix} \\ &= \begin{bmatrix} 0 & \sigma^{-1}B \\ ZC^* & \sigma^{-1}Z\hat{Q}B \end{bmatrix}, \end{aligned}$$

which verifies (3.3). To verify formula (3.4) for \underline{C} , plug (3.20) and (3.26) into (3.25) to get

$$\underline{C} = \sqrt{2} \begin{bmatrix} \tilde{C} & \sigma^{-2}\tilde{C}\hat{P} \\ 0 & -\sigma^{-2}\tilde{B}^*\tilde{A}^{*-1} \end{bmatrix} \begin{bmatrix} I & -\sigma^{-2}\hat{P} \\ 0 & I \end{bmatrix} \begin{bmatrix} (I+\tilde{A})^{-1} & 0 \\ 0 & (I+\tilde{A}^{*-1})^{-1} \end{bmatrix} \begin{bmatrix} I & \sigma^{-2}\hat{P} \\ 0 & I \end{bmatrix}.$$

Now substituting (3.14)–(3.16), we get

$$\begin{aligned} \underline{C} &= \begin{bmatrix} 2C(I-A)^{-1} & 2\sigma^{-2}C(I-A)^{-1}\hat{P} \\ 0 & -2\sigma^{-2}B^*(I+A^*)^{-1} \end{bmatrix} \\ &\quad \cdot \begin{bmatrix} I & -\sigma^{-2}\hat{P} \\ 0 & I \end{bmatrix} \begin{bmatrix} \frac{1}{2}(I-A) & 0 \\ 0 & \frac{1}{2}(I+A^*) \end{bmatrix} \begin{bmatrix} I & \sigma^{-2}\hat{P} \\ 0 & I \end{bmatrix} \\ &= \begin{bmatrix} C(I-A)^{-1} & \sigma^{-2}C(I-A)^{-1}\hat{P} \\ 0 & -\sigma^{-2}B^*(I+A^*)^{-1} \end{bmatrix} \begin{bmatrix} I-A & -\sigma^{-2}(A\hat{P}+\hat{P}A^*) \\ 0 & I+A^* \end{bmatrix} \\ &= \begin{bmatrix} C & \sigma^{-2}C(I-A)^{-1}[-A\hat{P}-\hat{P}A^*+\hat{P}(I+A^*)] \\ 0 & -\sigma^{-2}B^* \end{bmatrix} \\ &= \begin{bmatrix} C & \sigma^{-2}C\hat{P} \\ 0 & -\sigma^{-2}B^* \end{bmatrix}. \end{aligned}$$

This verifies formula (3.4).

It remains only to check that (3.5)–(3.8) follow from (3.2)–(3.4). But this is easily seen by using the diagonalization of \underline{A} derived from the Lyapunov equation (3.9)

$$\begin{aligned} \underline{A} &= \begin{bmatrix} A & \sigma^{-2}A\hat{P}+\sigma^{-2}\hat{P}A^* \\ 0 & -A^* \end{bmatrix} \\ &= \begin{bmatrix} I & -\sigma^{-2}\hat{P} \\ 0 & I \end{bmatrix} \begin{bmatrix} A & 0 \\ 0 & -A^* \end{bmatrix} \begin{bmatrix} I & \sigma^{-2}\hat{P} \\ 0 & I \end{bmatrix} \end{aligned}$$

and then multiplying out. We leave the details to the reader. \square

REFERENCES

- [1] V. M. ADAMJAN, D. Z. AROV AND M. G. KREIN, *Analytic properties of Schmidt pairs for a Hankel operator and the generalized Schur–Takagi problem*, Math. USSR-Sb., 15 (1971), pp. 31–73.
- [2] J. A. BALL, *Invariant subspace representations, unitary interpolants, and factorization indices*, in Topics in Operator Theory Systems and Networks, H. Dym and I. Gohberg, eds., OT12 Birkhäuser, Basel, 1984.
- [3] J. A. BALL AND J. W. HELTON, *A Beurling–Lax theorem for the Lie group $U(m, n)$ which contains most classical interpolation theory*, J. Operator Theory, 9 (1983), pp. 107–142.
- [4] J. A. BALL AND A. C. M. RAN, *Hankel norm approximation of a rational matrix function in terms of its realization*, in Modelling, Identification and Robust Control, C. I. Byrnes and A. Lindquist, eds., North-Holland, Amsterdam, 1986.

- [5] H. BART, I. GOHBERG AND M. A. KAASHOEK, *Minimal Factorization of Matrix and Operator Functions*, OT1 Birkhäuser, Basel, 1979.
- [6] H. BART, I. GOHBERG, M. A. KAASHOEK AND P. VAN DOOREN, *Factorization of transfer functions*, this Journal, 18 (1980), pp. 675–696.
- [7] B. A. FRANCIS, J. W. HELTON AND G. ZAMES, *H^∞ optimal feedback controllers for linear multivariate systems*, IEEE Trans. Automat. Control, AC-29 (1984), pp. 888–900.
- [8] K. GLOVER, *All optimal Hankel-norm approximations of linear multivariable systems and their L^∞ -error bounds*, Internat. J. Control, 39 (1984), pp. 1115–1193.
- [9] M. A. KAASHOEK AND A. C. M. RAN, *Symmetric Wiener–Hopf Factorization of Selfadjoint Rational Matrix Functions and Realization*, OT21 Birkhäuser, Basel, to appear.
- [10] S.-Y. KUNG AND D. W. LIN, *Optimal Hankel norm model reductions: multivariable systems*, IEEE Trans. Automat. Control, 26 (1981), pp. 832–852.
- [11] Z. NEHARI, *On bounded bilinear forms*, Ann. of Math. (2), 65 (1957), pp. 153–162.

ITERATIVE METHODS FOR LARGE CONVEX QUADRATIC PROGRAMS: A SURVEY*

Y. Y. LIN† AND J.-S. PANG‡

Abstract. In this paper, we give a state-of-the-art review of many iterative methods for solving large convex quadratic programs. We attempt to classify several of the more basic methods in two categories, within each of which a unified iterative scheme will be introduced and its convergence analyzed. Hybrid iterative methods (such as the proximal point algorithm and a diagonalization scheme) that make use of the more basic schemes will also be described. The results of an extensive computer experimentation which is aimed at comparing the relative performance of the various methods will be reported and discussed. Finally, several important topics which require future research will be highlighted.

Key words. iterative methods, quadratic programs, linear complementarity problem, matrix splitting, duality

AMS(MOS) subject classifications. 90C20, 90C33

1. Introduction.

1.1. Foreword. Convex quadratic programming is an old and important topic in mathematical programming. Its countless applications appear in diverse areas of engineering, mathematical, physical, social and management sciences. In addition, quadratic programs are central to many algorithms for solving nonlinear programming problems.

Over the years, a large number of methods have been developed for solving convex quadratic programs. These methods can be divided into two categories: finite methods and iterative methods. Finite methods solve a given program by some kind of pivoting procedures and terminate in finite time. On the other hand, iterative methods generate an infinite sequence which converges to a limit point that solves the program. In general, finite methods are effective for small-to-medium sized problems, but tend to become less efficient and uneconomical as the problem size increases. This is due to two reasons. One is the fact that round-off errors accumulate very rapidly and often lead to numerical difficulties. The other is the huge computer storage required which places a severe limit on the size of the problem being solved. Iterative methods, on the other hand are immune from these two handicaps because they (i) are self-correcting and (ii) operate on the input data only. As a result, they are capable of preserving any data sparsity and thus are particularly attractive for solving large-scale sparse problems.

Our principal objective in this paper is twofold: (i) to undertake a comprehensive survey of the most commonly used and (in our view) most promising iterative methods for solving large-scale sparse convex quadratic programs (we refer to [44] for a survey of finite methods for solving general convex quadratic programs); and (ii) to perform and report the numerical results of an extensive computer experimentation comparing the relative performance of the methods. Based on the results obtained, we shall make some general comments on each individual method with regard to its strength and drawbacks.

1.2. Review of literature. Among the earliest iterative methods proposed for solving convex quadratic programs is the one by Hildreth [27]. In present-day language,

* Received by the editors September 25, 1985, and in revised form March 4, 1986. This research was based on work supported by the National Science Foundation under grant ECS-8407240.

† School of Management, The University of Texas at Dallas, Richardson, Texas 75083-0688.

Hildreth's procedure may be described as the projected (point) Gauss-Seidel method applied to an equivalent linear complementarity formulation of the given quadratic program. The relaxation method proposed by Agmon [2], Motzkin and Schoenberg [37] for solving a system of linear inequalities is related to Hildreth's procedure and actually appears before the latter.

Hildreth's algorithm was published in the 1950s. It was not until 1971 that Cryer [17] proposed the first successive overrelaxation (SOR) method for minimizing a strictly convex quadratic function over the nonnegative orthant. Subsequently, Cottle, Golub and Sacher [15] extended Cryer's method to a block version; Cottle and Goheen [14] further extended Cryer's method to include box constraints. In [8], Cea and Glowinski proposed a related block SOR scheme. (Many more references on similar SOR methods may be found in the cited articles.) In [31], Mangasarian proposed a general iterative scheme for solving the symmetric linear complementarity problem and improved on the convergence results obtained previously by Cryer, Cottle and others. Subsequently, Mangasarian [32], [34] discussed how a SOR method can be applied to solve separable strictly convex quadratic programs and linear programs. Based on a penalty function approach, Han and Mangasarian [24] proposed an SOR scheme for solving a nonseparable strictly convex quadratic program. Most recently, Lin and Cryer [30] proposed an alternating direction implicit algorithm for the solution of (symmetric) linear complementarity problems arising from free boundary problems. The numerical results reported in [30] show that the alternating direction implicit algorithm is significantly faster than the SOR algorithms.

Motivated by the works of Mangasarian [31], Ahn [3], Aganagic [1] and others, Pang [43] introduced a fundamental iterative scheme for solving the linear complementarity problem. The scheme is based on the classical notion of matrix splittings in numerical analysis. Specialization of the scheme to a strictly convex quadratic program was discussed in [45], [46] where some necessary and sufficient conditions for convergence were established.

Separately, Lent and Censor [28] extended Hildreth's original algorithm and developed an almost cyclic SOR algorithm for a strictly convex quadratic program. The accompanying paper by Herman and Lent [26] described an application of the Lent-Censor extension of Hildreth's algorithm. Another recent application of Hildreth's original algorithm was documented in Bachem and Korte [6]. In [28], the authors pointed out the connection between Hildreth's algorithm and the relaxation method of Agmon [2], Motzkin and Schoenberg [37]. Lent and Censor classified the extended Hildreth algorithm as a row-action method [9]. The term "row-action" refers to the notion that the constraints of the (quadratic) program are acted upon one row at a time.

More recently, Cottle et al. [13] described a Lagrangian relaxation algorithm for solving a constrained matrix problem formulated as a strictly convex quadratic program over the transportation polytope. (See also [38].) An earlier paper by Cottle and Pang [16] discussed the solution of the same problem by an SOR scheme. (See also Cottle [12].) In [46], a convergence result on the Lagrangian relaxation scheme of Cottle et al. was established.

The iterative methods mentioned above can be classified as either (i) LCP (i.e. linear complementarity problem) based, or (ii) dual based. The LCP-based iterative methods solve a given quadratic program via an equivalent linear complementarity formulation. These methods have their origin in the solution of systems of linear equations [40], [53]. The fundamental scheme proposed in Pang [43] is a unification of many of the LCP-based methods. The dual-based iterative methods, on the other

hand, solve a given program via its (Lagrangian) dual program. A typical dual method operates by maximizing the dual function using a certain periodic basis or gradient-type ascent iterative procedure. For certain types of problems, the LCP and the dual formulations are the same.

Many of the iterative methods for solving a general convex quadratic program require a strictly convex objective function. An iterative scheme, proposed recently by Shiau [52], is an exception. Shiau's scheme was designed to solve a linear complementarity problem and proved to be convergent in the case of an asymmetric positive semidefinite matrix. In particular, it is applicable to the linear complementarity problem arising from a convex quadratic program. No computational result concerning this application is available, however. (Incidentally, Shiau's scheme is rather similar to the Frank-Wolfe 1957 procedure [21] for quadratic programs.) A third approach to deal with the convex, but not strictly convex case is to rely on the proximal point algorithm [50], [51]. Ha [23] discussed this idea in his Ph.D. dissertation and showed how it leads to certain decomposition methods for block structured convex quadratic programs. The main idea of the proximal point approach is to first strongly convexify the (quadratic) objective function before applying an inner iterative scheme to the strongly convexified subprograms. Specialized to a linear objective function, the proximal point approach provides an effective iterative procedure for solving large, sparse linear programs which may prove to be a practical alternative to the well-known simplex method and/or many of its variants. Mangasarian's iterative method for linear programs [34] is closely related to the proximal point approach (see Cheng's dissertation [10] for more details).

In recent years, there have been several proposals [18], [39], [54] which attempt to incorporate an iterative scheme in a finite method for solving convex quadratic programs with box constraints. The central idea is to use an iterative method for systems of linear equations (like SOR or conjugate gradient) to carry out the fundamental step of equation solving in a pivoting or active-set method. Under exact arithmetic and suitable stopping rule for the inner scheme, the overall method remains finitely convergent. At present, it is not clear how such an iterative active-set method (as it is called in [39]) can be applied to large-scale problems with general linear constraints. (It should be pointed out that although in theory, any strictly convex quadratic program can be transformed into one with box constraints, the transformation is not suitable for large programs with sparse data. This will be explained in more detail later.)

1.3. Organization. The remainder of this paper is divided into 5 sections. Section 2 contains 3 subsections. In § 2.1, we review a fundamental iterative scheme for solving the symmetric linear complementarity problem. In § 2.2, we state several convergence results of the fundamental scheme. Section 2.3 deals with the issue of error bounds of an approximate solution to the linear complementarity problem. In § 3.1, we discuss the specialization of the iterative scheme of § 2.1 to a separable strictly convex quadratic program and point out how separability is crucial. Section 3.2 deals with the nonseparable case. There, the Han-Mangasarian approach as well as a diagonalization algorithm is described. In § 3.3, we explain how the proximal point algorithm can be applied to a convex quadratic program. Section 4 is concerned with the dual-based iterative methods for solving a strictly convex quadratic program. In § 4.1, we define the dual function and summarize several of its key properties. In § 4.2, we introduce a unified dual ascent method. Section 4.3 explains how the method specializes in two important cases. In § 4.4, we state two convergence results for the dual ascent method of § 4.2.

In § 4.5, we refine the analysis of § 4.3 and discuss two related extensions. In the fifth section, we report the computational experience we have gathered with the various methods surveyed in the previous sections. We divide our discussion into the separable (§ 5.1) and the nonseparable (§ 5.2) case. Numerical results are presented from which observations are drawn. Finally, in the sixth and last section, we point out several issues that are not treated in this survey but which are important and deserve attention.

2. Iterative methods for symmetric LCP's.

2.1. A fundamental scheme. The close connection between a convex quadratic program and a linear complementarity problem is well recognized. Indeed, many iterative methods for (strictly) convex quadratic programs can be derived from a certain basic algorithm for solving the LCP with a symmetric matrix. In this section, we shall explain this fundamental iterative scheme for the LCP and present various convergence results related to it.

Consider the general linear complementarity problem of finding a vector x in R^n such that

$$(2.1.1) \quad w = q + Mx \geq 0, \quad x \geq 0 \quad \text{and} \quad x^T w = 0$$

where q and M are a given n -vector and an n by n matrix, respectively. We shall assume that the matrix M is *symmetric*. Under this symmetry assumption, the LCP (2.1.1) becomes the Karush-Kuhn-Tucker optimality conditions of the quadratic program over the nonnegative orthant:

$$(2.1.2) \quad \begin{aligned} &\text{minimize} \quad f(x) = q^T x + \frac{1}{2} x^T M x \\ &\text{subject to} \quad x \geq 0. \end{aligned}$$

If in addition, M is positive semidefinite, then the quadratic objective function $f(x)$ is convex. In this case, the two problems (2.1.1) and (2.1.2) are completely equivalent. As we shall see, the objective function $f(x)$ plays a crucial role in the convergence results of the basic iterative scheme for solving the LCP (2.1.1).

It is well known that many iterative methods for systems of linear equations can be uniformly described in terms of matrix splittings [40], [53]. Based on the same notion, we introduce a fundamental iterative scheme for solving the LCP (2.1.1). Specifically, let

$$M = B + C$$

be a splitting of the matrix M .

BASIC ALGORITHM. Let x^0 be an arbitrary nonnegative vector. In general, given $x^k \geq 0$, we generate x^{k+1} (which is assumed to exist) as a solution to the linear complementarity subproblem

$$(2.1.3) \quad w = q + Cx^k + Bx \geq 0, \quad x \geq 0 \quad \text{and} \quad x^T w = 0.$$

Terminate if some stopping rule is satisfied.

The above algorithm is very general. In practice, the matrix B should be chosen so that each subproblem (2.1.3) has at least one solution which is easily computable.

An example of such a choice is the standard (point) SOR-splitting where

$$(2.1.4) \quad B = L + D/\omega \quad \text{and} \quad C = U + (1 - 1/\omega)D$$

with D , L and U being the diagonal, strictly lower triangular and strictly upper triangular parts of M , respectively, and with ω being a given scalar satisfying $0 < \omega < 2$. If M has positive diagonals, then under the splitting (2.1.4), the basic algorithm generates the SOR sequence $\{x^k\}$ where

$$(2.1.5) \quad x_i^{k+1} = \max \left\{ 0, x_i^k - \left(\frac{\omega}{M_{ii}} \right) \left(q_i + \sum_{j < i} M_{ij} x_j^{k+1} + \sum_{j \geq i} M_{ij} x_j^k \right) \right\}, \quad i = 1, \dots, n.$$

The latter iteration (2.1.5) is precisely Cryer's 1971 method [17]. In a similar fashion, if M is partitioned in blocks (M_{ij}) and if D , L and U are accordingly defined, then the splitting (2.1.4) leads to a block SOR scheme in which each subproblem (2.1.3) decomposes into smaller sub-subproblems which, typically, can be solved much more easily than the original LCP (2.1.1).

2.2. Convergence results. In this subsection, we give several convergence results pertaining to the basic algorithm. All these results depend crucially on the symmetry assumption of M . (There are a number of results available for the asymmetric LCP [43], [45]. However, they do not seem directly applicable to convex quadratic programs.) Proofs will be omitted but can be found in the references.

We denote the splitting $M = B + C$ by the pair (B, C) . We say that the splitting (B, C) is *regular* if $B - C$ is positive definite. Notice that the SOR splitting (2.1.4) is regular if M is symmetric with positive diagonals and if $0 < \omega < 2$. We shall assume that the matrix B is such that each linear complementarity subproblem (2.1.3) has a solution. In the terminology of linear complementarity theory [41], such a B is referred to as a *Q-matrix*. If B is a *Q-matrix*, the splitting (B, C) is called a *Q-splitting*. Throughout the analysis that follows, except for the implicit understanding that each iterate x^{k+1} is well defined and easily computable, we leave open the question of how the computation of x^{k+1} should be carried out.

The first result states that under the mere assumption of symmetry on M , any accumulation point of a sequence $\{x^k\}$ generated by the basic algorithm with a regular *Q-splitting* (B, C) is a solution to the LCP (2.1.1).

THEOREM 2.2.1. *Let M be a symmetric matrix and (B, C) a regular *Q-splitting* of M . Then, for any $x^0 \geq 0$, any accumulation point of the sequence $\{x^k\}$ produced by the basic algorithm solves (2.1.1).*

Notice that Theorem 2.2.1 does not assert the existence of an accumulation point. For such a point to exist, it is sufficient for the sequence $\{x^k\}$ to be bounded. Conditions ensuring such boundedness consequence are given in the next result.

THEOREM 2.2.2. *Let M , B and C be as given in Theorem 2.2.1. If the two conditions below hold:*

- (A) *the quadratic function $f(x) = q^T x + \frac{1}{2} x^T M x$ is bounded below for $x \geq 0$;*
- (B) *the homogeneous LCP $[x \geq 0, Mx \geq 0 \text{ and } x^T M x = 0]$ has $x = 0$ as the unique solution,*

then any sequence $\{x^k\}$ generated by the basic algorithm is bounded.

It is well known from quadratic programming theory that if a quadratic objective function is bounded below on a polyhedral set, then it achieves its minimum there (see [19], [21]). In particular, assumption (A) implies that the quadratic program (2.1.2) has an optimal solution which must necessarily solve the LCP (2.1.1) by the

symmetry of M . Thus, the existence of a solution to (2.1.1) is implicit in condition (A). Moreover, (A) holds if and only if the matrix M is copositive (i.e. $x^T Mx \geq 0$ for all $x \geq 0$) and the implication below holds:

$$[x \geq 0 \text{ and } x^T Mx = 0] \text{ implies } q^T x \geq 0$$

(see [19] for a proof). In general, the existence of a solution to the LCP (2.1.1) does not imply (A).

Condition (B), on the other hand, is related to the boundedness of the solution set to the LCP (2.1.1). Indeed, it is not difficult to show that (B) holds if and only if for all vectors q , the LCP (2.1.1) has a (possibly empty) bounded solution set. Obviously, if M is strictly copositive (i.e. $x^T Mx > 0$ for all $x \geq 0$), then condition (B) holds trivially. Moreover, according to the discussion above, condition (A) also holds for all vectors q .

We say that a Q -splitting (B, C) of M is *weakly convergent* if for all vectors q and all initial $x^0 \geq 0$, any sequence $\{x^k\}$ generated by the basic algorithm contains at least one accumulation point; moreover, any such point is a solution to the LCP (2.1.1). (Notice that this notion of weak convergence is a global one in the sense that it applies to all constant vectors q .) We have just proven that if M is a symmetric strictly copositive matrix, then any regular Q -splitting of M is weakly convergent. As a matter of fact, the converse of this statement is also true and is made precise in the next theorem.

THEOREM 2.2.3. *Let M be a symmetric matrix. If M is strictly copositive, then any regular Q -splitting of M is weakly convergent. Conversely, if M has a regular Q -splitting that is weakly convergent, then M is strictly copositive.*

Specializing Theorem 2.2.3 to an SOR method, we obtain the following.

COROLLARY 2.2.1. *Let M be symmetric with positive diagonals. Then the SOR splitting (2.1.4) of M is weakly convergent for all $0 < \omega < 2$ if and only if M is strictly copositive.*

Remark. A similar result can be stated for a block SOR method.

Corollary 2.2.1 resembles the Ostrowski–Reich theorem in numerical analysis [40], [53]. The latter theorem states that for the system of linear equations $[Mx = p]$, if M is symmetric with positive diagonals, then the sequence $\{x^k\}$ generated by the (point) SOR method is convergent for any x^0 and all $0 < \omega < 2$ if and only if M is positive definite. There are several differences however. An obvious one is the notion of convergence involved. In the case of a system of linear equations, the entire sequence of iterates converges; whereas in the case of the LCP, the convergence is in terms of subsequences. (Incidentally, the latter kind of convergence is very common in nonlinear programming algorithms.) In [45], Pang raised the question of whether for the LCP (2.1.1), the convergence of the SOR sequence (cf. (2.1.5)) can be characterized by the positive definiteness of M , just like the Ostrowski–Reich result. It is easy to see that positive definiteness of M implies the convergence of the SOR sequence. Unfortunately, the converse is false. To explain this, we quote a convergence result which is a much simplified version of Theorem 2.1 in [46].

THEOREM 2.2.4. *Let M be a symmetric nondegenerate matrix (i.e. all principal minors of M are nonzero). Suppose that M has positive diagonal entries. Then the following statements are equivalent:*

- (i) *For any $x^0 \geq 0$, the SOR sequence $\{x^k\}$ (2.1.5) is convergent for all $0 < \omega < 2$ and all vectors q ;*
- (ii) *M is strictly copositive;*
- (iii) *M is copositive.*

Since there are obviously symmetric nondegenerate copositive matrices that are not positive definite, it follows that the convergence of the SOR sequence (2.1.5) does not necessarily imply the positive definiteness of M .

In several important applications of the LCP (2.1.1) to convex quadratic programs, the matrix M is, in addition to being symmetric, positive semidefinite but not definite. In this case, condition (B) of Theorem 2.2.2 will fail to hold. Thus, Theorem 2.2.2 becomes inapplicable. However, we have the following convergence result.

THEOREM 2.2.5. *Let M be a symmetric positive semidefinite matrix and let (B, C) be a regular Q -splitting of M . Then, for any initial vector $x^0 \geq 0$, the sequence $\{x^k\}$ generated by the basic algorithm is uniquely defined. Moreover, the following three statements are equivalent:*

- (A) *the quadratic function $f(x) = q^T x + \frac{1}{2} x^T M x$ is bounded below for $x \geq 0$;*
- (B) *the LCP (2.1.1) has a solution;*
- (C) *for any initial vector $x^0 \geq 0$, the sequence $\{Mx^k\}$ converges to some vector Mz and z solves the LCP (2.1.1).*

Notice that Theorem 2.2.5 does not assert the convergence of the sequence $\{x^k\}$. As a matter of fact, it remains an open question to determine if $\{x^k\}$ is bounded under the assumptions of Theorem 2.2.5. (More specifically, the question is: Let M , B and C be as given in Theorem 2.2.5. Does condition (A) or (B) imply that the sequence $\{x^k\}$ is bounded?) The next result shows that $\{x^k\}$ is indeed bounded if a Slater constraint qualification holds for the LCP (2.1.1). This result was first proved by Mangasarian [31] for his iterative scheme under a slightly weaker assumption on the matrix M . Basically, the additional Slater condition (together with the properties of M) ensures that the (quadratic) function $f(x)$ has bounded level sets. From this, the boundedness of $\{x^k\}$ follows easily.

THEOREM 2.2.6. *Let M be a symmetric positive semidefinite matrix. Let (B, C) be a regular Q -splitting of M . If there exists a vector x so that*

$$(2.2.1) \quad q + Mx > 0,$$

then for any $x^0 \geq 0$, the (uniquely defined) sequence $\{x^k\}$ generated by the basic algorithm is bounded, and thus has an accumulation point. Moreover, any such point solves the LCP (2.1.1) and the sequence $\{Mx^k\}$ converges.

Remark. The constraint qualification (2.2.1) implies assumption (A) of Theorem 2.2.5 but not conversely. It should also be pointed out that (2.2.1) bears little relationship to assumption (B) in Theorem 2.2.2. In particular, Theorem 2.2.6 is not a special case of Theorem 2.2.2.

2.3. A posteriori error bounds. In practical implementation, an iterative method terminates when a suitable stopping rule is satisfied. A commonly used rule for the LCP (2.1.1) is the following one:

$$(2.3.1) \quad \|\min(x^k, q + Mx^k)\| \leq \varepsilon$$

where $\|\cdot\|$ denotes a vector norm, x^k is the iterate being tested and ε is a given tolerance. The rule is justified on the observation that a vector x^* solves (2.1.1) if and only if

$$\|\min(x^*, q + Mx^*)\| = 0.$$

A question naturally arises, namely, if x^k is the (approximate) solution obtained at the termination of an iterative method under the rule (2.3.1), how close is x^k to an exact solution x^* of (2.1.1)? In other words, can one bound the error $\|x^k - x^*\|$ in terms of the quantity $\|\min(x^k, q + Mx^k)\|$? The result below gives an affirmative answer

to this question under a positive definiteness assumption on M . For a proof, see [47] where similar results for more general problems are established.

THEOREM 2.3.1. *Let M be a symmetric positive definite matrix with least eigenvalue α and let x^* be the (unique) solution to (2.1.1). Let x be an arbitrary vector. Then the following two error bounds hold:*

$$(2.3.2) \quad \|x - x^*\|_2 \leq ((\|M\|_2 + 1)/\alpha) \|\min(x, q + Mx)\|_2$$

and

$$(2.3.3) \quad \frac{\|x - x^*\|_2}{\|x^*\|_2} \leq \text{cond}(M)(\|M\|_2 + 1) \frac{\|\min(x, q + Mx)\|_2}{\|(-q)_+\|_2}$$

where $\text{cond}(M)$ denotes the condition number of M and $(-q)_+ = \max(0, -q)$. (It is assumed that $(-q)_+ \neq 0$ in (2.3.3).)

Remark. If $(-q)_+ = 0$, then $x^* = 0$ is the unique solution to (2.1.1).

The product $\text{cond}(M)(\|M\|_2 + 1)$ in (2.3.3) plays the same role as the condition number in a system of linear equations. Indeed, for the system $[Mx = p]$ where M is nonsingular, the bound

$$(2.3.4) \quad \frac{\|x - M^{-1}p\|}{\|M^{-1}p\|} \leq \text{cond}(M) \frac{\|p - Mx\|}{\|p\|}$$

is well known [40]. Thus, in (2.3.3), the relative error $\|x - x^*\|_2/\|x^*\|_2$ is bounded in terms of the relative residual $\|\min(x, q + Mx)\|_2/\|(-q)_+\|_2$ just like (2.3.4). Similarly, in (2.3.2), the absolute error $\|x - x^*\|_2$ is bounded in terms of the absolute residual $\|\min(x, q + Mx)\|_2$.

In [36], Mangasarian and Shiau have derived error estimates similar to (2.3.2) and (2.3.3) for the LCP (2.1.1) with a positive semidefinite matrix M . Among other things, they give an example to show that in this case, the quantity $\|\min(x, q + Mx)\|_2$ cannot be used as a measure of error. Instead, an alternative residual is needed. However, the bounding constants in the Mangasarian-Shiau error expressions do not appear as easily computable as those in Theorem 2.3.1. The reader is referred to Mangasarian [33] for related results.

3. LCP-based iterative methods for QP's.

3.1. The separable case. Consider the strictly convex quadratic program (QP)

$$(3.1.1) \quad \begin{aligned} &\text{minimize} && f(x) = q^T x + \frac{1}{2} x^T D x \\ &\text{subject to} && Ax \geq b, \quad Cx = d \end{aligned}$$

where the matrix D is symmetric and positive definite. Throughout the discussion, we shall assume that the program (3.1.1) is feasible. By the assumed property on D , it follows that (3.1.1) has a unique optimal solution which we denote by x^* . The Karush-Kuhn-Tucker optimality conditions for (3.1.1) are

$$(3.1.2a) \quad 0 = q + Dx - A^T y - C^T z,$$

$$(3.1.2b) \quad u = -b + Ax \geq 0, \quad y \geq 0, \quad u^T y = 0,$$

$$(3.1.2c) \quad 0 = -d + Cx,$$

which define a “mixed” LCP with matrix M and vector p given by

$$(3.1.3) \quad M = \begin{bmatrix} D & -A^T & -C^T \\ A & 0 & 0 \\ C & 0 & 0 \end{bmatrix} \quad \text{and} \quad p = \begin{bmatrix} q \\ -b \\ -d \end{bmatrix}.$$

The word "mixed" here refers to the equations in (3.1.2a) and (3.1.2c) and the related fact that the variables x and z are unrestricted in sign. For simplicity, we shall call (3.1.2) an LCP with data (3.1.3) and leave out the word "mixed."

In theory, one could apply the basic algorithm of § 2 to the LCP (3.1.2). However, there are serious considerations. One is the fact that the matrix in (3.1.3) is not symmetric. This immediately invalidates all the convergence results of § 2. (The results for the asymmetric LCP are too restrictive to be useful here, unfortunately.) Second, the zero blocks in (3.1.3) make an algorithm like the point SOR method inapplicable. As a result, one is led to seek for alternative LCP formulation(s).

One such formulation can be derived as follows. From (3.1.2a), we solve for x in terms of y and z , obtaining

$$(3.1.4) \quad x = -D^{-1}q + D^{-1}A^T y + D^{-1}C^T z.$$

Substitution of (3.1.4) into (3.1.2b) and (3.1.2c) yields

$$(3.1.5a) \quad u = -b - AD^{-1}q + AD^{-1}A^T y + AD^{-1}C^T z \geq 0, \quad y \geq 0, \quad u^T y = 0,$$

$$(3.1.5b) \quad 0 = -d - CD^{-1}q + CD^{-1}A^T y + CD^{-1}C^T z,$$

which form a ("mixed") LCP defined by the matrix

$$(3.1.6) \quad \begin{bmatrix} AD^{-1}A^T & AD^{-1}C^T \\ CD^{-1}A^T & CD^{-1}C^T \end{bmatrix}.$$

The latter matrix is obviously symmetric. Moreover, it is positive semidefinite but in general not positive definite. Therefore the basic algorithm of § 2 and its convergence results (Theorem 2.2.5 in particular) can be applied to (3.1.5). By backward substitution, (3.1.4) will then produce an (approximate) solution to the original QP (3.1.1). As a special case, one obtains an SOR scheme for solving (3.1.1).

There is a major drawback in the formulation (3.1.5); namely, it involves the inverse of the matrix D . For large-scale problems with sparse data, this inverse could easily destroy the practicality of the approach. Consequently, the formulation (3.1.5) is recommended for QP's with a separable objective function (i.e. with a diagonal D) or for programs where the creation of the matrix (3.1.6) will not cause computer storage difficulties. Observe that in the separable case, the computation of x from y and z in (3.1.4) is trivial.

Two practical points should be mentioned with regard to the implementation of an SOR scheme to (3.1.5). First, (3.1.5b) should be treated as an equation and the variables z unrestricted. In particular, no projection on the nonnegative orthant is required for the z variables (cf. (2.1.5)). Second, for large-scale sparse problems, it is not advisable to form the matrix (3.1.6) explicitly. This is because any sparsity structure could easily be destroyed in the formation of (3.1.6). As explained in [32], it is possible to implement the SOR scheme in such a way that any sparsity or structural properties of the data can be maintained and taken advantage of.

In what follows, we give two convergence results for the basic algorithm of § 2 applied to the LCP (3.1.5). The first result (Theorem 3.1.1) is a specialization of Corollary 2.2.1 and concerns an SOR method. Its proof can be found in [45]. The second result (Theorem 3.1.2) follows from Theorem 2.2.5 and applies to an arbitrary regular Q -splitting of the matrix (3.1.6). Its proof can be found in [46].

THEOREM 3.1.1. *Let D be a symmetric positive definite matrix. Suppose that the matrix A is nonvacuous. Then the (point) SOR splitting of the matrix (3.1.6) is weakly convergent for all $0 < \omega < 2$ if and only if the matrix C has linearly independent rows and there exists a vector x such that $Ax > 0$ and $Cx = 0$.*

THEOREM 3.1.2. *Let D be a symmetric positive definite matrix. Let (G, H) be any regular Q -splitting of the matrix (3.1.6). For any initial (y^0, z^0) with $y^0 \geq 0$, the (uniquely defined) sequence $\{(y^k, z^k)\}$ generated by the basic algorithm applied to (3.1.5) induces a corresponding sequence of iterates $\{x^k\}$ via (3.1.4); namely,*

$$x^k = -D^{-1}q + D^{-1}A^T y^k + D^{-1}C^T z^k \quad \text{for all } k.$$

The following two statements are equivalent:

(D) *for any (y^0, z^0) with $y^0 \geq 0$, the induced sequence $\{x^k\}$ converges to the unique solution of the program (3.1.1);*

(E) *the program (3.1.1) is feasible, or equivalently, solvable.*

As in Theorem 2.2.5, Theorem 3.1.2 does not assert the convergence of the sequence $\{(y^k, z^k)\}$. In order for the latter sequence to be bounded and thus to have the subsequential convergence property, it is sufficient for the matrix C to have linearly independent rows and for a Slater constraint qualification to hold; see [31], [32] for details (cf. also Theorems 2.2.6 and 3.1.1).

3.2. The nonseparable case. As we have pointed out, the formulation (3.1.5) is useful mainly in the separable case. In the sequel, we describe two approaches used to deal with the nonseparable case. One approach is due to Han and Mangasarian [24] who derive it using an exact penalty function theory. The other approach transforms the nonseparable problem into a sequence of separable ones to which the methodology of § 3.1 is applicable.

Consider the quadratic program (3.1.1) and its Karush–Kuhn–Tucker conditions (3.1.2). Observe that if γ is such that $\gamma D - I$ is nonsingular then (3.1.2) is equivalent to

$$(3.2.1a) \quad 0 = (\gamma D - I)q + (\gamma D - I)Dx - (\gamma D - I)A^T y - (\gamma D - I)C^T z,$$

$$(3.2.1b) \quad u = -b - \gamma Aq - A(\gamma D - I)x + \gamma A A^T y + \gamma A C^T z \geq 0, \quad y \geq 0, \quad u^T y = 0,$$

$$(3.2.1c) \quad 0 = -d - \gamma Cq - C(\gamma D - I)x + \gamma C A^T y + \gamma C C^T z.$$

The latter conditions (3.2.1a,b,c) define a mixed LCP with matrix

$$(3.2.2) \quad \begin{bmatrix} (\gamma D - I)D & -(\gamma D - I)A^T & -(\gamma D - I)C^T \\ -A(\gamma D - I) & \gamma A A^T & \gamma A C^T \\ -C(\gamma D - I) & \gamma C A^T & \gamma C C^T \end{bmatrix},$$

which is obviously symmetric. Han and Mangasarian [24] proposed the application of the point SOR method to the formulation (3.2.1). They showed that for $\gamma \geq 1/\rho$ where $\rho > 0$ is the least eigenvalue of D , the matrix (3.2.2) is symmetric positive semidefinite and that if $\gamma > 1/\rho$, the problem (3.2.1) has a solution $(\bar{x}(\gamma), \bar{y}(\gamma), \bar{z}(\gamma))$ such that $\bar{x}(\gamma) = x^*$ where x^* is the unique solution of the program (3.1.1). We remark that if $\gamma > 1/\rho$ and if the matrix $\begin{pmatrix} A \\ C \end{pmatrix}$ has linearly independent rows, then the matrix (3.2.2) is positive definite.

The following gives a convergence result for the (point) SOR method applied to (3.2.1). Its proof can be found in [46].

THEOREM 3.2.1. *Let D be a symmetric positive definite matrix with least eigenvalue $\rho > 0$. Suppose that $\begin{pmatrix} A \\ C \end{pmatrix}$ has no vanishing rows. Fix $\gamma > 1/\rho$. Then for any $\omega \in (0, 2)$ and any initial vector (x^0, y^0, z^0) with $y^0 \geq 0$, the sequence of iterates $\{(x^k, y^k, z^k)\}$ generated by the (point) SOR method applied to (3.2.1) is uniquely defined. Moreover, the following two statements are equivalent:*

(F) for any $\omega \in (0, 2)$ and any initial vector (x^0, y^0, z^0) with $y^0 \geq 0$, the sequence $\{(x^k, y^k, z^k)\}$ is such that $\{x^k\}$ converges to the unique solution x^* of (3.1.1) and that $\{A^T y^k + C^T z^k\}$ converges to some vector $A^T y^* + C^T z^*$ where (x^*, y^*, z^*) solves (3.2.1);

(E) same as the one in Theorem 3.1.2.

As in Theorem 3.1.2, Theorem 3.2.1 does not assert the convergence of the sequence $\{(y^k, z^k)\}$. In order for this to happen, a linear independence property of the matrix $\begin{pmatrix} A \\ C \end{pmatrix}$ as well as a Slater constraint qualification need to hold. See [24] for details.

With regard to the practical implementation of the (point) SOR method applied to (3.2.1), we point out that the inverse of D is absent in the formulation (3.2.1). Moreover, as in the case of (3.1.5), it is not recommended, for large-scale sparse problems, to form the matrix (3.2.2) explicitly. Indeed, the method should (and can) be implemented to take full advantage of any sparsity structure that the data might have.

The other approach used to deal with the program (3.1.1) with a nonseparable D is based on a scheme which, in recent years, has been recognized as an effective algorithm for solving variational inequalities (Ahn and Hogan [4], Pang and Chan [48]) as well as convex programs (Feijou and Meyer [20]). The central idea is to transform (3.1.1) into a sequence of separable quadratic programs. Specifically, let us write

$$(3.2.3) \quad D = G + H$$

where G and H denote, respectively, the diagonal and off-diagonal parts of D . The following is a detailed description of the method that we have termed the diagonalization algorithm.

DIAGONALIZATION ALGORITHM. Let x^0 be a feasible solution to (3.1.1). In general, given x^k feasible, solve the (separable) subprogram

$$(3.2.4) \quad \begin{aligned} &\text{minimize} && \frac{1}{2}x^T Gx + (q + Hx^k)^T x \\ &\text{subject to} && \text{same constraints as (3.1.1)} \end{aligned}$$

and let $x^{k+1/2}$ denote the unique optimal solution. Define $d^k = x^{k+1/2} - x^k$. Perform the one-dimensional search over θ :

$$(3.2.5) \quad \begin{aligned} &\text{minimize} && f(x^k + \theta d^k) \\ &\text{subject to} && x^k + \theta d^k \text{ feasible to (3.1.1)} \end{aligned}$$

and let θ^k be the unique minimizer. Set $x^{k+1} = x^k + \theta^k d^k$. Terminate if some appropriate stopping rule is satisfied.

It should be pointed out that since the objective function $f(x)$ is quadratic, the one-dimensional search problem (3.2.5) is trivial and can be carried out exactly. The vector $x^{k+1/2}$ can be obtained by an SOR scheme or any appropriate method. The following result establishes the convergence of the above algorithm.

THEOREM 3.2.2. *Let D be a symmetric positive definite matrix. Suppose that the feasible set of the program (3.1.1) is nonempty and compact. Then for any initial x^0 which is feasible to (3.1.1), the sequence $\{x^k\}$ generated by the diagonalization algorithm converges to the unique optimal solution x^* of (3.1.1).*

A detailed proof of Theorem 3.2.2 can be found in [20] where the algorithm was stated for general convex objective functions. (See also [29].) In what follows, we sketch the essential ideas underlying the proof. To begin, let F denote the feasible set of (3.1.1). The sequences $\{x^k\}$, $\{x^{k+1/2}\}$ and $\{x^{k+1}\}$ are contained in the compact set F and hence have convergent subsequences $\{x^k\}_K$, $\{x^{k+1/2}\}_K$ and $\{x^{k+1}\}_K$ with limits \bar{x} , \tilde{x} and x' respectively. Since $f(x^{k+1}) \leq f(x^k)$ for all k , the sequence $\{f(x^k)\}$ is

monotonically nonincreasing and bounded below by $f(x^*)$. Thus $\{f(x^k)\}$ converges and $f(\bar{x}) = f(x')$. The rest of the proof consists of showing that \bar{x} is in fact the optimal solution of (3.1.1), and this is done by contradiction. Once this is established, the desired convergence of $\{x^k\}$ follows because the program (3.1.1) has a unique optimal solution by the positive definiteness of D . See the cited references for more details.

Remark. According to [20], a weaker version of Theorem 3.2.2 holds if D is symmetric and positive semidefinite. In this case, the sequence of objective values $\{f(x^k)\}$ converges to the optimum value $f(x^*)$ and the sequence of gradient vectors $\{\nabla f(x^k)\}$ converges to $\nabla f(x^*)$. However, one can not conclude that $\{x^k\}$ converges to x^* .

Notice that Theorem 3.2.2 requires a compactness assumption on the feasible set of the program (3.1.1). The next result shows that if the splitting (3.2.3) is regular, i.e., if $G - H$ is positive definite, then a modified version of the diagonalization algorithm is convergent, regardless of whether the feasible region is compact.

THEOREM 3.2.3. *Let D be a symmetric positive definite matrix. Suppose that the program (3.1.1) is feasible and that $G - H$ is positive definite. For a given x^0 feasible to (3.1.1), let $\{x^k\}$ be such that x^{k+1} is the (unique) solution of the subprogram (3.2.4). Then the sequence $\{f(x^k)\}$ is monotonically nonincreasing and $\{x^k\}$ converges to the unique solution x^* of (3.1.1).*

We sketch the proof of Theorem 3.2.3. First of all, the positive definiteness of $G - H$ implies that for all k ,

$$f(x^{k+1}) - f(x^k) \leq -\frac{1}{2}\alpha \|x^{k+1} - x^k\|_2^2 \leq 0$$

where α is some positive number less than the least eigenvalue of $G - H$. Thus, the sequence $\{f(x^k)\}$ is monotonically nonincreasing and bounded below by $f(x^*)$. Thus $\{f(x^k)\}$ converges. Consequently, $\|x^{k+1} - x^k\|_2 \rightarrow 0$. By a similar manipulation, we may deduce

$$f(x^*) - f(x^k) \leq -\frac{1}{2}\alpha \|x^* - x^k\|_2^2.$$

Thus, the sequence $\{x^k\}$ is bounded. From here on, the desired convergence of $\{x^k\}$ follows from a routine argument.

It is not difficult to show that a necessary and sufficient condition for $G - H$ to be positive definite is that $\|G^{-1/2}DG^{-1/2}\|_2 < 1$. Moreover it has been shown (see [48] e.g.) that if D is symmetric and strictly diagonally dominant, then the latter norm condition holds. Consequently, for such a D , the modified diagonalization algorithm without the one-dimensional search (3.2.5) produces a sequence convergent to the desired solution of (3.1.1).

3.3. The convex case. In this subsection, we discuss the proximal point algorithm applied to the quadratic program (3.1.1), where the matrix D is assumed to be symmetric and positive semidefinite. In general, the proximal point algorithm can solve a variational inequality problem with a monotone operator [50] as well as a general convex program [51]. The central idea of the algorithm is to create a sequence of strongly convexified subprograms, which are solved by certain (iterative) methods like the ones discussed in the previous sections. An attractive feature of this approach is that the subproblems become numerically more stable and error bounds (such as those in § 2.3) can be derived.

Specifically, given an iterate x^k (which may not be feasible to (3.1.1)), let x^{k+1} be an “approximate” solution to the quadratic subprogram

$$\begin{aligned} (3.3.1) \quad & \text{minimize} \quad \frac{1}{2}x^T(D + c_k I)x + (q - c_k x^k)^T x \\ & \text{subject to} \quad \text{same constraints as (3.1.1)} \end{aligned}$$

where c_k is a given positive scalar. Notice that (3.3.1) has a strongly convex objective function. (Strong convexity is equivalent to strict convexity for a quadratic function.) Also, the iterate x^{k+1} is not required to be feasible to (3.1.1). (For example, if (3.3.1) is solved by an iterative scheme like SOR and if x^{k+1} is the iterate obtained at termination, then x^{k+1} is in general, not a feasible vector.) In [50], Rockafellar proposed two criteria under which the sequence $\{x^k\}$ will converge. These two rules are

$$(3.3.2) \quad \|x^{k+1} - z^{k+1}\| \leq \varepsilon_k, \quad \varepsilon_k > 0, \quad \sum_{k=0}^{\infty} \varepsilon_k < \infty$$

and

$$(3.3.3) \quad \|x^{k+1} - z^{k+1}\| \leq \delta_k \|x^{k+1} - x^k\|, \quad \delta_k > 0, \quad \sum_{k=0}^{\infty} \delta_k < \infty,$$

where z^{k+1} denotes the exact solution of (3.3.1) and $\{\varepsilon_k\}$ and $\{\delta_k\}$ are two given sequences of tolerance. Observe that if $\{x^k\}$ is bounded, then (3.3.3) implies (3.3.2).

The following gives a convergence result of the above proximal point algorithm. Its proof can be found in [50], where a much more general version was established.

THEOREM 3.3.1. *Let D be a symmetric positive semidefinite matrix. Let $\{x^k\}$ be a sequence of vectors generated by the proximal point algorithm under the rule (3.3.2). Suppose that the sequence $\{c_k\}$ is bounded. Then $\{x^k\}$ is bounded if and only if the program (3.1.1) is solvable. If $\{x^k\}$ is bounded, then it converges to a vector x^* that solves (3.1.1). Finally, suppose that D is symmetric positive definite, that $\{c_k\}$ is non-increasing and converges to zero and that the rule (3.3.3) holds. Then, if $\{x^k\}$ is bounded, the convergence of $\{x^k\}$ to x^* is superlinear, i.e.,*

$$\lim_{k \rightarrow \infty} \frac{\|x^{k+1} - x^*\|}{\|x^k - x^*\|} = 0.$$

When the proximal point algorithm is used in conjunction with an inner iterative scheme to solve (3.3.1), a question naturally arises, namely, how can one terminate the inner iterations so that the rule (3.3.2) or (3.3.3) will be satisfied? To phrase this question in slightly more general terms, consider a strictly convex QP in the form (3.1.1) where D is symmetric and positive definite. (Note that (3.3.1) is strictly convex.) The question becomes: Given an $\varepsilon < 0$, if x is an iterate obtained in an iterative scheme for solving (3.1.1), how can one ensure that $\|x - x^*\| \leq \varepsilon$ with x^* being the optimum solution of (3.1.1)? The answer to this question relies on the kind of a posteriori error bounds described in § 2.3. Specifically, we need two things. First, we need a suitable quantity similar to $\|\min(x, q + Mx)\|$ for the LCP (2.1.1) in order to stop the iterations. Second, we need to be able to bound the error $\|x - x^*\|$ in terms of the chosen quantity (just like (2.3.2) or (2.3.3)). Based on the equivalent formulation (3.1.4) and (3.1.5), we propose the following measure. Let y and z be two arbitrary vectors and let x be defined from y and z according to (3.1.4). It seems reasonable to use the quantity

$$r(y, z) = \|(\min(y, -b - AD^{-1}q + AD^{-1}A^T y + AD^{-1}C^T z), \\ -d - CD^{-1}q + CD^{-1}A^T y + CD^{-1}C^T z)\|$$

as a measure of error. Indeed, the vector x defined by (3.1.4) solves (3.1.1) if and only if $r(y, z) = 0$. Since most iterative methods for solving (3.1.1) do produce the vectors y and z , we may use something like $r(y, z) \leq \delta$ as a stopping rule for the iterations. The following result (drawn from [47]) shows how $\|x - x^*\|$ is bounded by $r(y, z)$.

THEOREM 3.3.2. *Let D be a symmetric positive definite matrix. Suppose that the matrix $E = \begin{pmatrix} A \\ C \end{pmatrix}$ has linearly independent rows. If x is the vector defined by (3.1.4) for an arbitrary (y, z) , then*

$$(3.3.4) \quad \|x - x^*\|_2 \leq \text{cond}(D)\beta \|ED^{-1}E^T + I\|_2 r(y, z),$$

$$(3.3.5) \quad \frac{\|x - x^*\|_2}{\|x^*\|_2} \leq \text{cond}(D)\beta \|ED^{-1}E^T + I\|_2 \|E\|_2 \frac{r(y, z)}{\|e\|_2},$$

where β is the inverse of the least singular value of the matrix E and e is the vector $\begin{pmatrix} b \\ d \end{pmatrix}$ which is assumed to be nonzero.

Remark. The linear independence assumption of the matrix E ensures that the constant β is positive.

Referring to the comments made at the end of § 2.3, we point out that the linear independence assumption on the matrix E is essential for the bounds (3.3.4) and (3.3.5) to hold. Indeed, without this assumption, the quantity $r(y, z)$ can not be used as a measure of error. In this case, an alternative measure is needed. We refer to [36] for more details.

4. Dual-based iterative methods.

4.1. Background. Consider a strictly convex quadratic program of the form

$$(4.1.1) \quad \begin{aligned} &\text{minimize} \quad f(x) = q^T x + \frac{1}{2} x^T D x \\ &\text{subject to} \quad Ax = b, \quad x \in X \end{aligned}$$

where D is a symmetric positive definite matrix and X is a polyhedral set. Throughout the discussion, we assume that the QP (4.1.1) is feasible. As a result, (4.1.1) has a unique optimal solution which we denote by x^* . The constraints of (4.1.1) have the following features. The set X contains the “easy-to-handle” constraints (e.g. simple bounds) or is itself the Cartesian product of several sets of such nature. The equation “ $Ax = b$ ” consist of the “complicating” constraints. Notice that an implicit assumption in the formulation (4.1.1) is that all the complicating constraints are equalities. (The set X may contain both inequalities and equalities, however.) A consequence of the formulation (4.1.1) is that its (Lagrangian) dual program becomes unconstrained:

$$(4.1.2) \quad \text{maximize } d(y): y \text{ unrestricted}$$

where $d(y)$ is the (Lagrangian) dual function

$$(4.1.3) \quad d(y) = \min_{x \in X} q^T x + \frac{1}{2} x^T D x + y^T (b - Ax).$$

(If there are inequalities among the complicating constraints, the dual program (4.1.2) will have some variables restricted to be nonnegative. This case will be discussed in more detail later.)

In the following result, we summarize several key properties of the dual function and some basic relationships between the primal program (4.1.1) and its dual (4.1.2). See [49] for proofs.

PROPOSITION 4.1.1. *Let D be a symmetric positive definite matrix. Suppose that the program (4.1.1) is feasible. Then the following statements hold:*

(i) *For each y , there exists a unique $x(y)$ in X such that*

$$d(y) = q^T x(y) + \frac{1}{2} x(y)^T D x(y) + y^T (b - Ax(y));$$

(ii) *The dual function $d(y)$ is continuously differentiable and*

$$\nabla d(y) = b - Ax(y);$$

(iii) If y^* is a vector of optimal Lagrange multipliers associated with the equalities $Ax = b$, then y^* solves the dual program (4.1.2). Moreover, such a y^* must exist and

$$d(y^*) = f(x^*)$$

(in particular, there is no duality gap);

(iv) Conversely, if \bar{y} solves the dual program (4.1.2), then $x(\bar{y}) = x^*$.

Remark. For a given y , the vector $x(y)$ as defined in Proposition 4.1.1(i) is in general not primal feasible. It is so if and only if it is optimal.

4.2. A unified dual ascent method. A dual ascent method solves the primal program (4.1.1) by generating a sequence of dual vectors $\{y^k\}$ (which induces the sequence of primal vectors $\{x(y^k)\}$) through the maximization of the dual function $d(y)$. The generation of $\{y^k\}$ is as follows. The initial y^0 is arbitrary. In general, given y^k , choose a search direction δ^k . Define $y^{k+1} = y^k + \theta^k \delta^k$ where θ^k is such that

$$d(y^{k+1}) = \max_{\theta} d(y^k + \theta \delta^k).$$

Since

$$d(y^k + \theta \delta^k) = \min_{x \in X} f(x) + (y^k)^T(b - Ax) + \theta(\delta^k)^T(b - Ax),$$

it follows that the search for θ^k can be achieved by solving the subprogram

$$(4.2.1) \quad \begin{aligned} &\text{minimize} && f(x) + (y^k)^T(b - Ax) \\ &\text{subject to} && (\delta^k)^T(b - Ax) = 0 \text{ and } x \in X \end{aligned}$$

and by letting θ^k be an optimal Lagrange multiplier of the constraint $(\delta^k)^T(b - Ax) = 0$. The subprogram (4.2.1) is clearly feasible (and thus solvable). Indeed, the unique minimizer of (4.2.1) is the vector $x(y^{k+1})$.

The constraint $(\delta^k)^T(b - Ax) = 0$ in (4.2.1) represents an aggregation of the complicating constraints $Ax = b$. Thus, one may interpret the above dual ascent method as solving a sequence of simplified subproblems each of which has the same objective function as the given program (4.1.1), but modified by a Lagrangian term involving the complicating constraints and their corresponding multipliers, and also having all the complicating constraints aggregated into a single one by a certain vector δ^k . The following description summarizes how the dual ascent method is implemented in practice.

DUAL ALGORITHM. Let y^0 be arbitrary. In general, given y^k , generate δ^k . Solve the subproblem (4.2.1). Let $x(y^{k+1})$ be the (unique) minimizer and let θ^k be an optimal Lagrange multiplier of the aggregated constraint $(\delta^k)^T(b - Ax) = 0$. Set $y^{k+1} = y^k + \theta^k \delta^k$. Terminate if some stopping rule is satisfied.

Observe that $x(y^{k+1})$ is obtained as a by-product of the method and no extra effort is required for its computation.

There are many choices for the aggregation vector δ^k . The following are two large families of such choices: (i) Periodic basis ascent—there is a basis of vectors $\{v^1, \dots, v^m\}$ (where m is the number of constraints in $Ax = b$) such that (a) $\delta^k \in \{v^1, \dots, v^m\}$ for each k and (b) there exists an integer l ($\geq m$) such that for all $k \geq 0$, $\{v^1, \dots, v^m\} \subseteq \{\delta^{k+1}, \dots, \delta^{k+l}\}$ and (ii) Gradient-type ascent— $\delta^k = H^k(b - Ax(y^k))$ where H^k is some symmetric matrix. (Recall that $\nabla d(y^k) = b - Ax(y^k)$ from Proposition 4.1.1(ii).) Periodic basis ascent methods include the cyclic coordinate ascent method [13], [38] which has $l = m$ and the basis $\{v^1, \dots, v^m\}$ consisting of the coordinate

vectors. The latter method amounts to the relaxation of the complicating constraints one at a time. Gradient-type ascent methods include the steepest ascent method which has H^k equal to the identity matrix, many quasi-Newton methods [22], [23] as well as the conjugate gradient method with restart that has

$$(4.2.2) \quad H^k = \begin{cases} I & \text{if } k \equiv 0 \pmod{N}, \\ I + \delta^{k-1}(\nabla d(y^k) - \nabla d(y^{k-1})^T) / \nabla d(y^{k-1})^T \nabla d(y^{k-1}) & \text{otherwise,} \end{cases}$$

where N is some positive integer not exceeding m . Under (4.2.2), the vector δ^k is given by

$$(4.2.3) \quad \delta^k = \begin{cases} \nabla d(y^k) & \text{if } k \equiv 0 \pmod{N}, \\ \nabla d(y^k) + \frac{\nabla d(y^k)^T (\nabla d(y^k) - \nabla d(y^{k-1}))}{\nabla d(y^{k-1})^T \nabla d(y^{k-1})} \delta^{k-1} & \text{otherwise,} \end{cases}$$

which is the Polak-Ribiere-Polyak conjugate gradient formula [5]. The integer N denotes the number of iterations after which the method is restarted with the steepest ascent direction.

4.3. Two special cases. Before discussing the convergence of the dual algorithm, we discuss two special instances in which the algorithm is expected to be particularly attractive. One is the case where the program (4.1.1) is separable and the set X consists of simple upper and lower bounds. In this case, each subproblem (4.2.1) is a singly constrained strictly convex separable QP with simple bounds on the variables. As such, it can be solved by a very simple one-dimensional search routine which we briefly outline below (see [13], [7], [25] for more details). Consider a typical subprogram written in the form

$$(4.3.1) \quad \begin{aligned} &\text{minimize} \quad \sum_{i=1}^n q_i x_i + \frac{1}{2} \sum_{i=1}^n c_i x_i^2 \\ &\text{subject to} \quad \sum_{i=1}^n p_i x_i = t, \quad a_i \leq x_i \leq b_i \text{ all } i \end{aligned}$$

where $c_i > 0$ all i . The (Lagrangian) dual program of (4.3.1) is

$$(4.3.2) \quad \text{maximize } d(r): r \text{ unrestricted}$$

where

$$(4.3.3) \quad d(r) = \min_{\substack{a_i \leq x_i \leq b_i \\ \text{all } i}} \left\{ \sum_{i=1}^n (q_i - r p_i) x_i + \frac{1}{2} \sum_{i=1}^n c_i x_i^2 \right\} + r t.$$

For each fixed r , the unique $\{x_i(r)\}$ minimizing (4.3.3) is given by

$$x_i(r) = \min \{b_i, \max \{a_i, -(q_i - r p_i) / c_i\}\}.$$

According to the remark following Proposition 4.1.1, the maximization of $d(r)$ is equivalent to the search of a r^* such that $\sum_{i=1}^n p_i x_i(r^*) = t$. Since each $x_i(r)$ is a one-dimensional piecewise linear function in r , the function $h(r) = \sum_{i=1}^n p_i x_i(r) - t$ is piecewise linear in r . Thus, the search for the desired r^* is easy to carry out. The so-obtained $\{x_i(r^*)\}$ is then the solution to (4.3.1). In the cited references, several strategies to enhance the practical efficiency of the search process are described.

The other instance in which the dual algorithm is potentially attractive is when the objective function $f(x)$ in (4.1.1) is partially separable (i.e. when the matrix D is

block diagonally structured) and the set X is the Cartesian product of lower-dimensional sets. Specifically, assume that the matrix D is block diagonal with diagonal blocks D_i of order n_i by n_i ($i = 1, \dots, m$) and X is equal to $\prod_{i=1}^m X_i$ where each X_i is in R^{n_i} . Under this setting, each subproblem (4.2.1) may be written in the form

$$(4.3.4) \quad \begin{aligned} & \text{minimize} \quad \sum_{i=1}^m g_i^T x_i + \frac{1}{2} \sum_{i=1}^m x_i^T D_i x_i \\ & \text{subject to} \quad \sum_{i=1}^m p_i^T x_i = t, \quad x_i \in X_i \quad (i = 1, \dots, m). \end{aligned}$$

The dual program of (4.3.4) is of the form (4.3.2) with

$$(4.3.5) \quad d(r) = \min \left\{ \sum_{i=1}^m (g_i - r p_i)^T x_i + \frac{1}{2} \sum_{i=1}^m x_i^T D_i x_i : x_i \in X_i \text{ all } i \right\}.$$

For a fixed r , the minimization in (4.3.5) decomposes into m subminimizations ($i = 1, \dots, m$)

$$(4.3.6) \quad \begin{aligned} & \text{minimize} \quad (g_i - r p_i)^T x_i + \frac{1}{2} x_i^T D_i x_i \\ & \text{subject to} \quad x_i \in X_i. \end{aligned}$$

With r as the parameter, the latter problem (4.3.6) becomes a parametric quadratic program which can be solved for example, by the parametric principal pivoting algorithm (see [11], [42]). This parametric algorithm will compute the (unique) solution $x_i(r)$ to (4.3.6) for each value of r . As before, the maximization of $d(r)$ can be achieved by finding a suitable r^* such that $\sum_{i=1}^m p_i^T x_i(r^*) = t$. In this fashion, one obtains a decomposition scheme for solving (4.3.4) which by its definition, occurs as a typical subproblem in the dual algorithm for solving the original QP (4.1.1). We refer to [22], [23] for other uses of the dual approach as a decomposition strategy for large convex QP's.

4.4. Convergence results. In this subsection, we discuss the convergence of the dual algorithm introduced in § 4.2. First of all, observe that if $\{y^k\}$ is the sequence of (dual) vectors generated by the algorithm, then $d(y^{k+1}) \geq d(y^k)$. Thus, the sequence $\{y^k\}$ lies in the level set

$$\{y: d(y) \geq d(y^0)\}.$$

The following result characterizes the boundedness of such a set.

PROPOSITION 4.4.1. *Let D be a symmetric positive definite matrix. Assume that the program (4.1.1) is feasible. Then the dual function $d(y)$ in (4.1.3) has bounded level sets if and only if there exists a neighborhood N of the vector b such that for each vector b' in N , the perturbed system*

$$Ax = b', \quad x \in X$$

is consistent. In particular, if $d(y)$ has bounded level sets, then A must have linearly independent rows.

The proof of Proposition 4.4.1 follows from Corollary 14.2.2 in [49].

In general, if the dual function is assumed to have bounded level sets, then the proof of convergence of the dual algorithm is rather routine. In the next result, we establish the convergence of the dual cyclic coordinate ascent and the dual steepest ascent methods under no such bounded-level-set assumption. Its proof can be found in Lin [29].

THEOREM 4.4.1. *Let D be a symmetric positive definite matrix. Assume that the program (4.1.1) is feasible. Suppose that the sequence of aggregation vectors $\{\delta^k\}$ is chosen according to either a periodic basis ascent method or a gradient-type ascent method where the matrices $\{H^k\}$ are uniformly positive definite on the range space of A , i.e. there exist constants $\alpha > \beta > 0$ such that for all integers k and vectors y*

$$(4.4.1) \quad \alpha \|Ay\|_2^2 \geq y^T A^T H^k A y \geq \beta \|Ay\|_2^2.$$

Then the sequence of primal vectors $\{x(y^k)\}$ converges to the unique solution x^ of (4.1.1). In particular, the conclusion holds for the dual cyclic coordinate ascent method and the dual steepest ascent method.*

Remark. Since the matrix A is not assumed to have full row rank, the condition (4.4.1) is in general less stringent than the requirement that the $\{H^k\}$ are uniformly positive definite on the entire space.

A detailed proof of Theorem 4.4.1 can be found in Lin [29]. In what follows, we sketch the main ideas underlying the proof. First of all, write $x^k = x(y^k)$. It can then be shown that

$$d(y^{k+1}) - d(y^k) \geq \frac{1}{2}\gamma \|x^{k+1} - x^k\|_2^2$$

where γ is some positive number smaller than the least eigenvalue of D . Thus, the sequence $\{d(y^k)\}$ is monotonically nondecreasing and bounded above by $d(y^*)$, where y^* is a maximizer of the dual function $d(y)$. Consequently, $\|x^{k+1} - x^k\| \rightarrow 0$. By a similar manipulation, we may obtain

$$d(y^*) - d(y^k) \geq \frac{1}{2}\gamma \|x^* - x^k\|_2^2,$$

from which the boundedness of $\{x^k\}$ follows. The rest of the proof consists of verifying that any accumulation point of $\{x^k\}$ is in fact an optimal solution of (4.1.1). Since (4.1.1) has a unique optimal solution x^* , the desired convergence of $\{x^k\}$ follows.

There are several versions of the conjugate gradient method for unconstrained optimization [5]; each depends on the particular formula used to define the search direction δ^k . We choose the Polak-Ribiere-Polyak formula (4.2.3) because, reportedly, it seems to yield better performance than other formulas. Considered as a gradient-type ascent method, the H^k matrix (cf. (4.2.2)) defining (4.2.3) does not seem to satisfy condition (4.4.1) easily. For this reason, we state a separate convergence result for the dual conjugate gradient method with restart.

THEOREM 4.4.2. *Let D be a symmetric positive definite matrix. Assume that the program (4.1.1) is feasible. Let the sequence of aggregation vectors $\{\delta^k\}$ be chosen according to formula (4.2.3). Then the sequence of primal vectors $\{x(y^k)\}$ produced by the dual algorithm converges to the unique solution x^* of (4.1.1).*

The proof of Theorem 4.4.2 resembles that of Theorem 4.4.1. The only difference lies in the way to establish the (primal) feasibility of an accumulation point of $\{x^k\}$. The details can be found in Lin [29].

4.5. Two extensions. In § 4.3, we have explained how the separability of the objective function together with the box structure of the set X can facilitate the solution of the subproblems (4.2.1), and thereby increase the attraction of the dual algorithm. In the case of a nonseparable objective (and with the same box structured X), the diagonalization algorithm described in § 3.2 can be used to reduce the program (4.1.1) to a sequence of separable ones to which the specialized dual algorithm is applicable. In what follows, we describe an alternate approach to reduce a nonseparable strictly convex quadratic program to a separable strictly convex quadratic program.

Consider the program (4.1.1) where D is symmetric positive definite with least eigenvalue $\rho > 0$. We may write

$$D = \delta I + G^T G$$

where $0 < \delta < \rho$ and G is a nonsingular matrix. Then, obviously, (4.1.1) is equivalent to

$$(4.5.1) \quad \begin{aligned} &\text{minimize} && q^T x + \frac{1}{2} \delta x^T x + \frac{1}{2} y^T y \\ &\text{subject to} && Ax = b, \quad Gx = y, \quad x \in X. \end{aligned}$$

The latter program (4.5.1) has a strictly convex separable objective. (Incidentally, the equivalence between (4.5.1) and (4.1.1) is well known and is valid even if $\delta = 0$. For $\delta = 0$, the objective in (4.5.1) is not strictly convex, however. This lack of strict convexity could invalidate the convergence results of the iterative methods for solving (4.5.1).)

The formulation (4.5.1) depends on two things: the knowledge of the least eigenvalue of D (or at least a lower bound) and the factorization of the matrix $D - \delta I$. (Recall that the Han-Mangasarian approach for nonseparable QP's also requires a similar quantity γ which is like the inverse of the δ above.) If both δ and the factorization are readily available, then (4.5.1) could turn out to be a useful formulation as some of the iterative schemes for the separable case can be applied. If only δ is known, it is still possible to implement a certain dual periodic basis ascent method as well as the dual conjugate gradient method with restart without the explicit knowledge of the matrix G . See [29] for details of how this is done. Finally, if an extensive amount of effort is required for the computation of δ , then it is not advisable to use the formulation (4.5.1).

To end our discussion on the dual methods, consider a general convex quadratic program:

$$(4.5.2) \quad \begin{aligned} &\text{minimize} && f(x) = q^T x + \frac{1}{2} x^T D x \\ &\text{subject to} && Ax = b, \quad Cx \geq d, \quad x \in X \end{aligned}$$

where D is symmetric positive semidefinite. The equalities $Ax = b$ as well as the inequalities $Cx \geq d$ express the complicating constraints. By adding slack variables to the inequality constraints, we obtain

$$(4.5.3) \quad \begin{aligned} &\text{minimize} && f(x) = q^T x + \frac{1}{2} x^T D x \\ &\text{subject to} && Ax = b, \quad Cx - v = d, \quad x \in X, \quad v \geq 0. \end{aligned}$$

Notice that in terms of the variables (x, v) jointly, the objective function in (4.5.3) is not strictly convex (even if D is positive definite). As a result of this lack of strict convexity, there is no guarantee that a straightforward application of a dual algorithm is necessarily convergent. In order to solve (4.5.3) by a dual approach, one may apply the proximal point algorithm described in § 3.3 to generate a sequence of strongly convex subprograms of the form

$$\begin{aligned} &\text{minimize} && (q - c_k x^k)^T x - c_k (v^k)^T v + \frac{1}{2} x^T (D + c_k I) x + \frac{1}{2} c_k v^T v \quad (c_k > 0) \\ &\text{subject to} && \text{same constraints as (4.5.3)} \end{aligned}$$

and then solve each subprogram by an appropriate dual algorithm. An exception arises when D is positive definite. In this case, the dual cyclic coordinate method (which corresponds to the relaxation of the complicating constraints one at a time) can be applied directly to (4.5.2) despite the presence of the inequalities $Cx \geq d$. (In particular,

the use of the proximal point algorithm becomes unnecessary.) Indeed, let us assume that D is symmetric positive definite. The dual of (4.5.2) is

$$(4.5.4a) \quad \text{maximize } d(y, z): y \text{ unrestricted and } z \geq 0$$

where

$$(4.5.4b) \quad d(y, z) = \min_{x \in X} \{q^T x + \frac{1}{2}x^T D x - y^T(b - Ax) - z^T(-d + Cx)\}.$$

Due to the simple restriction on the z -variables (and no restriction on the y -variables), it is apparent how the cyclic coordinate ascent method can be applied to maximize $d(y, z)$. In practice, the resulting method can be implemented in the primal space (cf. the description of the dual algorithm in § 4.2) and it produces a sequence of primal variables $\{x^k\}$ converging to the (unique) solution of (4.5.2).

The formulation (4.5.4) also helps to explain the technical difficulty involved in a direct application of a dual method (other than the dual cyclic coordinate ascent method) to (4.5.2) and consequently, why the formulation (4.5.3) is needed. The bottleneck is the presence of the nonnegativity restriction on the z -variables. In this regard, [55] might be useful. This is a topic that requires further study.

5. Computational study.

5.1. Separable case. In the last three sections, we have described many iterative methods for solving large convex quadratic programs. From a practical point of view, it is important to know how these methods perform and compare to one another. In this section, we report the numerical results of an extensive computer experimentation with the methods. Data of the test problems were randomly generated and all the computations were performed on an IBM 4381 computer at the University of Texas at Dallas. The computer codes were written in Fortran double precision.

Four sets of experiments were performed. In the first set of experiments, we consider a separable strictly convex QP of the form (4.1.1) with X being the nonnegative orthant and D a positive diagonal matrix. Three methods were tested: (i) an SOR method applied to the following equivalent mixed symmetric LCP formulation:

$$\begin{aligned} x &= -D^{-1}q + D^{-1}y + D^{-1}A^T z \geq 0, \quad y \geq 0, \quad y^T x = 0, \\ 0 &= -b - AD^{-1}q + AD^{-1}y + AD^{-1}A^T z, \end{aligned}$$

(ii) the dual cyclic coordinate ascent method and (iii) the dual conjugate gradient ascent method with restart. The following termination criteria were used in all 3 methods:

$$(5.1.1) \quad \max(\|\min(x^k, y^k)\|_\infty, \|-b + Ax^k\|_\infty) \leq 10^{-6}.$$

In the SOR method, a sequence $\{(y^k, z^k)\}$ is generated; this then defines the iterates $\{x^k\}$ according to the equation

$$x^k = -D^{-1}q + D^{-1}y^k + D^{-1}A^T z^k.$$

The so-generated sequence $\{x^k\}$ is then tested for termination under the rule (5.1.1). In the two dual methods, a sequence $\{(x^k, z^k)\}$ is generated with the property that x^k is the unique solution to the subproblem

$$\begin{aligned} &\text{minimize } q^T x + \frac{1}{2}x^T D x + (z^k)^T(b - Ax) \\ &\text{subject to } x \geq 0. \end{aligned}$$

Thus, with $y^k = q + Dx^k - A^T z^k$, it holds that $\min(x^k, y^k) = 0$. Hence, the rule (5.1.1) reduces to

$$(5.1.2) \quad \|-b + Ax^k\|_{\infty} \leq 10^{-6},$$

which is simply a test of primal feasibility of the iterate x^k .

We ran the methods on 2 sets of test problems, one with $n = 500$ and the other $n = 1000$ (n is the dimension of the x -vector). In the first set of problems ($n = 500$), three values of m (the number of rows in the matrix A) were chosen: these are $m = 40, 80, 160$. For each value of m , different densities of the matrix A were tested (5%, 8%, 11% and 14%). In the second set of problems ($n = 1000$), m was set to be 50, 100 and 200 and the densities of A were 2%, 4%, 6%, 8% and 10%. The results are summarized in Tables 1 and 2. As it is well known, the SOR method is very sensitive to the relaxation parameter. For each individual problem, we ran the method with different values of the parameter (ranging from 1.0 to 1.9 with an increment of 0.1 each run) and picked out the best ω value. The column for the SOR method in Tables 1 and 2 gives the result pertaining to the best ω value. In the two dual methods, we have used the one-dimensional search technique described in § 4.3 to solve each singly constrained subproblem. The specific implementation of the search routine follows that explained in [13].

The results of Tables 1 and 2 pertain to problems where the matrix A has no specific structure and is not extremely sparse. We have compared the two dual methods (under the stopping rule (5.1.2)) on some constrained matrix problems of the type discussed in [13]. The latter problems are of the form (5.1.1) where A is the node-arc incidence matrix of a bipartite graph. Such a matrix A has the following characteristics: (i) all entries are 0 or 1, (ii) it has many times more columns than rows and (iii) it is extremely sparse. For example, if the two groups of nodes in the bipartite graph each have M nodes, then the matrix A is of order $2M \times M^2$ and its density is $1/M$. So for M equal to 50 or more, the density of A will be less than 2%. We should point out

TABLE 1
Separable with $Ax = b$.

Density	ω	LCP-SOR iterations	CPU	Dual Cyclic Coord. iterations	CPU	Dual Cong. Gradient iterations	CPU
40 × 500							
5%	1.4	29	1.84	22	1.88	25	1.80
8%	1.4	31	2.27	18	2.13	22	1.98
11%	1.3	28	2.49	21	2.69	23	2.22
14%	1.3	32	3.07	19	3.03	21	2.32
80 × 500							
5%	1.5	48	3.84	31	3.92	38	3.36
8%	1.4	47	4.91	36	5.44	33	3.72
11%	1.4	44	5.74	35	7.05	32	4.22
14%	1.4	52	7.60	36	8.63	34	4.85
160 × 500							
5%	1.5	89	10.04	97	15.51	73	7.93
8%	1.5	94	14.49	101	23.38	67	9.50
11%	1.5	91	18.09	110	35.10	65	11.21
14%	1.5	95	22.67	107	42.02	62	12.75

TABLE 2
Separable with $Ax = b$.

Density	ω	LCP-SOR		Dual Cyclic Coord.		Dual Conj. Gradient	
		iterations	CPU	iterations	CPU	iterations	CPU
50 × 1000							
4%	1.3	24	3.70	14	3.40	21	3.75
6%	1.3	24	4.21	14	4.07	20	4.10
8%	1.3	27	4.93	14	4.61	20	4.36
10%	1.3	23	5.10	14	5.05	18	4.55
100 × 1000							
2%	1.5	42	6.05	25	5.72	40	6.44
4%	1.4	34	7.00	24	7.34	31	6.87
6%	1.4	36	8.61	25	9.18	26	7.32
8%	1.4	35	9.82	24	10.49	27	8.16
10%	1.3	33	10.88	23	11.73	24	8.30
200 × 1000							
2%	1.5	59	12.28	44	13.10	57	12.03
4%	1.4	51	15.82	50	19.95	43	13.53
6%	1.4	54	20.26	46	25.74	40	15.47
8%	1.4	52	23.82	44	31.41	38	16.60
10%	1.4	50	26.69	50	38.45	37	18.44

that for this set of test problems, the fact that all nonzero entries of A are 1's can be fully taken advantage of by the dual cyclic coordinate method but does not seem to have a large impact in the dual conjugate gradient method. The results are contained in Table 3.

From the results in Tables 1, 2 and 3, the following observations are apparent:

(i) For problems with very sparse matrix A , the dual cyclic coordinate ascent method stands out as the clear winner;

(ii) For problems with not-so-sparse matrix A (say with density of 5% or more), the dual conjugate gradient method tends to be most preferable, followed by the LCP-SOR method and the dual cyclic coordinate method; the distinction between the latter two methods does not seem very significant;

(iii) For problems with a matrix A of even higher density (say 10% or more), the relative attractiveness of the dual cyclic coordinate method tends to decrease rather rapidly;

(iv) For very large problems with extremely sparse matrix A (like the last two problems in Table 3), the dual conjugate gradient method becomes highly inefficient.

TABLE 3
Constrained matrix problem.

Density	Dual Cyclic Coord.		Dual Conj. Gradient	
	iterations	CPU	iterations	CPU
30 × 30	9	1.61	61	4.97
40 × 40	11	3.18	34	5.91
50 × 50	11	4.75	73	17.55
60 × 60	11	6.64	36	18.03
80 × 80	11	7.47	38	48.63
100 × 100	9	9.98	59	133.00

We should point out that in terms of computer storage requirement, the 3 methods are roughly equal.

Our next set of experiments is concerned with a separable strictly convex quadratic program of the form

$$(5.1.3) \quad \begin{aligned} &\text{minimize} && q^T x + \frac{1}{2} x^T D x \\ &\text{subject to} && Ax \geq b \quad \text{and} \quad x \geq 0 \end{aligned}$$

where D is a positive diagonal matrix. Two methods are tested: (i) an LCP-SOR method and (ii) the dual cyclic coordinate method. The termination criteria used is similar to (5.1.1). The results are summarized in Table 4. As in Tables 1 and 2, the columns in Table 4 pertaining to the LCP-SOR method report the best output among a set of ten different test values of the relaxation parameter.

TABLE 4
Separable with $Ax \geq b$.

Density	ω	LCP-SOR		Dual Cyclic Coord.	
		iterations	CPU	iterations	CPU
40 × 500					
5%	1.4	22	1.67	10	1.35
8%	1.4	26	2.10	14	1.58
11%	1.3	20	2.11	11	1.61
14%	1.4	24	2.53	10	1.60
80 × 400					
5%	1.4	29	2.99	16	2.39
8%	1.4	32	3.85	20	3.05
11%	1.3	29	4.42	16	3.32
14%	1.3	24	4.54	14	3.12
160 × 500					
5%	1.4	36	5.83	27	4.98
8%	1.4	51	9.11	33	6.80
11%	1.4	38	9.35	28	7.69
14%	1.4	42	11.90	35	10.38

We have also tested the proximal point algorithm used in conjunction with the dual conjugate gradient algorithm as explained in § 4.4. The results are not as good as either method in Table 4. Furthermore, there seems to be some erratic behavior in the convergence of the proximal point algorithm when the density of the matrix A exceeds 10%. We do not fully understand the reason for this.

We observe that the two methods in Table 4 are fairly compatible in all cases, with the dual cyclic coordinate method slightly outperforming the LCP-SOR algorithm. In the case of a low-density matrix A , this observation is consistent with that drawn earlier from Tables 1, 2 and 3. However, in the case of a less-sparse matrix A , the reason why the dual cyclic coordinate method still performs somewhat better is not completely clear to us.

5.2. Nonseparable case. Our third set of experiments is concerned with a nonseparable strictly convex quadratic program of the form (4.1.1) where the set X is the nonnegative orthant and the matrix D is symmetric diagonally dominant. Four methods

were tested: (i) the Han–Mangasarian penalty function approach, (ii) the modified (no-line-search) diagonalization algorithm with the dual conjugate gradient algorithm as an inner routine, (iii) the same diagonalization algorithm with the dual cyclic coordinate algorithm as an inner routine and (iv) the dual conjugate gradient algorithm applied to the formulation (4.5.1). The density of the matrix D was set at 4.2% and its order at 500×500 . The termination criteria used in all 4 methods is the satisfaction of the Karush–Kuhn–Tucker optimality conditions within a prescribed tolerance. More specifically, if x^k is an approximate primal solution and if λ^k is an approximate multiplier for the constraint $Ax = b$, then (x^k, λ^k) is considered acceptable if

$$(5.2.1) \quad \varepsilon_k = \max (\| \min (q + Dx^k - A^T \lambda^k, x^k) \|_\infty, \| b - Ax^k \|_\infty) \leq 10^{-5}.$$

In the two diagonalization methods tested, another termination criterion was used to stop the inner iterations. Since a dual method was used as an inner scheme, satisfactory primal feasibility was chosen as the (inner) stopping rule. Specifically, if x^k is the current iterate that defines the subproblem (cf. (3.2.4))

$$(5.2.2) \quad \begin{aligned} & \text{minimize} \quad \frac{1}{2} x^T G x + (q + H x^k)^T x \\ & \text{subject to} \quad \text{same constraints as (4.1.1)} \end{aligned}$$

where G and H are, respectively, the diagonal and off-diagonal parts of D , then a vector x^{k+1} obtained by the dual algorithm of § 4.2 applied to (5.2.2) is considered an acceptable solution if

$$(5.2.3) \quad \| b - A x^{k+1} \|_\infty \leq \varepsilon_k^2$$

where ε_k is defined in (5.2.1). We term (5.2.3) a progressive termination rule in the sense that the accuracy in the solution of each subprogram (5.2.2) depends on the quality of the iterate x^k : the closer x^k is to satisfying the overall stopping rule (5.2.1) for the outer iterations, the more accurately the corresponding subproblem (5.2.2) is solved. Presumably, the advantage of the rule (5.2.3) is that when x^k is far from satisfying (5.2.1), some unnecessary inner iterations can be saved.

The results of this set of experiments are reported in Table 5. The entries there require explanation. First of all, in the Han–Mangasarian method, there are two parameters which will affect its performance. One is the γ value (cf. (3.2.1)) that defines the LCP formulation to which an SOR method is applied. The other is the relaxation parameter in the SOR scheme. We have generated the matrix D so that its least eigenvalue is no less than one. Different values of γ (which is required to be greater than the reciprocal of the least eigenvalue of D) were tested, and for each γ , different ω were experimented. The results in Table 5 pertain to the best γ and ω values obtained. In the dual conjugate gradient method applied to (4.5.1), there is the parameter δ which affects the performance of the method. Again, different values of δ were tested and the best results were reported in Table 5. Finally, the columns (# of iterations) in the two diagonalization methods report the numbers of outer as well as total inner iterations.

From the experience we have gathered on this set of problems, the following conclusions can be drawn:

(i) The diagonalization method seems to be the clear winner among the methods tested. Depending on the density of the A matrix, either the dual cyclic coordinate or the dual conjugate gradient method may be used as an inner routine to solve the subproblems;

TABLE 5
Nonseparable with $Ax = b$.
 Density of D : 4.2%.

Density	Han-Mangasarian ($\gamma = 0.6$)			Diagonalization Dual Conj. Gradient		Diagonalization Dual Cyclic Coord.		Dual Conj. Gradient to (4.5.1), ($\delta = 1.9$)	
	ω	iterations	CPU	iterations	CPU	iterations	CPU	iterations	CPU
40 × 500									
5%	1.4	82	19.07	12/87	9.92	12/86	8.90	82	13.54
8%	1.4	89	21.25	12/90	10.97	13/101	11.27	89	14.98
11%	1.3	87	21.94	12/84	12.00	12/93	12.12	93	16.21
14%	1.4	88	23.45	12/85	12.41	13/112	15.76	108	18.94
80 × 500									
5%	1.3	99	24.06	12/112	12.39	12/135	14.48	111	18.20
8%	1.3	97	25.86	12/104	13.94	12/135	18.89	114	20.26
11%	1.3	94	27.59	11/105	16.49	12/146	24.67	125	23.40
14%	1.4	103	32.06	12/104	17.26	12/156	29.96	130	25.82
160 × 500									
5%	1.4	137	37.88	11/149	19.98	11/283	39.15	134	25.69
8%	1.4	143	45.63	11/148	24.52	11/275	58.48	150	32.02
11%	1.4	148	54.04	10/149	28.83	11/327	87.97	160	37.45
14%	1.4	161	66.13	11/145	32.23	11/285	91.51	184	46.69

(ii) Both the Han-Mangasarian approach and the dual approach applied to the (4.5.1) formulation are very sensitive to the respective parameters γ and δ . Although no formal connection can yet be established, our feeling is that these two parameters relate like reciprocals of one another. As far as performance is concerned, we conjecture that the closer γ is to the reciprocal of the least eigenvalue of D , the better the Han-Mangasarian approach will become. A similar conjecture can be made with respect to δ ;

(iii) The density of the matrix D is not expected to have a significant influence on the above two conclusions.

Our last set of experiments is concerned with a nonseparable strictly convex QP of the form (5.1.3), where the matrix D is symmetric and diagonally dominant. The specification of D is the same as in the previous set of problems. Two methods were tested: (i) the Han-Mangasarian penalty function approach and (ii) the modified (no-line-search) diagonalization scheme with the dual cyclic coordinate method as an inner subroutine. The results are summarized in Table 6. Again, the diagonalization algorithm consistently outperforms the Han-Mangasarian approach.

5.3. Additional runs. Responding to the comments of a referee, we have performed some additional experiments whose outputs are summarized in Tables 7 and 8. Table 7 solves a separable strictly convex quadratic program of the form (4.1.1) and extends Tables 1 and 2 to problems of much larger size. The objective of this set of runs is to (further) demonstrate the ability of the three basic methods for solving large problems. Table 8 solves the same type of quadratic programs as Tables 1, 2 and 7. It tests the three methods under different stopping accuracies. (Tables 1, 2 and 7 are all obtained under the termination rules (5.1.1) and (5.1.2) with an $\varepsilon = 10^{-6}$ accuracy.) The referee believes that the LCP-SOR method will outperform the other two methods when ε is large. The results in Table 8 do not support this belief. In our view, it is the sparsity

TABLE 6
Nonseparable with $Ax \geq b$.
Density of D: 4.2%.

Density	Han-Mangasarian ($\gamma = 0.6$)			Diagonalization	
	ω	iterations	CPU	Dual Cyclic Coord. iterations	CPU
40 × 500					
5%	1.3	72	17.56	13/74	7.98
8%	1.4	76	19.06	13/73	8.70
11%	1.3	74	19.58	13/70	9.40
14%	1.4	77	21.15	13/65	9.33
80 × 500					
5%	1.3	74	19.95	13/104	10.66
8%	1.3	81	23.30	13/105	13.09
11%	1.4	83	25.62	12/104	15.22
14%	1.3	80	27.07	13/90	14.77
160 × 500					
5%	1.3	87	27.35	13/159	18.97
8%	1.3	99	34.89	13/138	22.79
11%	1.3	88	36.50	12/147	28.47
14%	1.3	98	43.55	12/147	35.22

TABLE 7
Separable with $Ax = b$.

Density	ω	LCP-SOR		Dual Cyclic Coord.		Dual Conj. Gradient	
		iterations	CPU	iterations	CPU	iterations	CPU
500 × 5000							
0.6%	1.4	54	99	25	90	38	102
1%	1.4	37	102	23	97	31	105
800 × 5000							
0.6%	1.4	64	161	41	156	47	162
1000 × 5000							
0.6%	1.4	68	203	56	209	54	203

TABLE 8
Separable with $Ax = b$.

ε	ω	LCP-SOR		Dual Cyclic Coord.		Dual Conj. Gradient	
		iterations	CPU	iterations	CPU	iterations	CPU
200 × 1000 (4% density)							
10^{-2}	1.4	25	10.95	23	11.73	22	9.93
10^{-3}	1.4	31	11.89	31	13.52	27	10.54
10^{-4}	1.4	38	13.19	34	15.24	32	11.20
10^{-5}	1.4	45	14.13	40	17.30	37	11.88
1000 × 5000 (0.6% density)							
10^{-2}	1.4	32	175	25	172	26	183
10^{-3}	1.4	41	182	33	180	33	188
10^{-4}	1.4	49	186	42	189	38	190
10^{-5}	1.4	59	193	48	197	46	194

of the data that determines the relative efficiency of the respective methods, regardless of the termination accuracy.

6. Conclusion. In this paper, we have reviewed a wide variety of iterative methods for solving large convex quadratic programs and reported our computational experience with many of them. Although the problems solved in the experiments are all randomly generated and may not be considered as most general convex QP's, they are broad enough to be indicative of the relative performance of the various methods. In particular, the conclusions drawn from the tables in § 5 are expected to be valid for a typical QP arising from practice. (Recall that those conclusions are all relative to the methods being tested.)

There are a number of issues that we have either barely touched upon or completely ignored. These issues are important and deserve further investigation. Among them, we name the following four: (i) the effect of conditioning, (ii) the convex but not strictly convex program, (iii) acceleration of the reported methods and (iv) the issue of parallelism. In what follows, we briefly discuss each one of these points.

The problem of conditioning has always been an important issue in a traditional numerical method (such as the SOR algorithm) for solving systems of linear equations. To the best of our knowledge, very little research has been done to treat ill-conditioned quadratic programs by iterative methods. The thesis [54] discusses how certain preconditioning schemes can be incorporated in an iterative active-set method for solving box-constrained strictly convex QP's. (See also [39].) The two papers [36], [47] discuss how error bounds can be obtained in terms of a suitably defined condition number of a strictly convex QP and/or LCP (cf. Theorems 2.3.1 and 3.3.2). Except for these few references, we are not aware of other work done on this important topic. At this juncture, we should mention an experience of ours having to do with an attempt to solve a symmetric LCP with the following (positive definite) matrix

$$\begin{bmatrix} 6 & -4 & 1 & & & & \\ -4 & 6 & -4 & 1 & & & \\ 1 & -4 & 6 & -4 & 1 & & \\ & \ddots & \ddots & \ddots & \ddots & \ddots & \\ & & 1 & -4 & 6 & -4 & 1 \\ & & & 1 & -4 & 6 & -4 \\ & & & & 1 & -4 & 6 \end{bmatrix}$$

by an SOR scheme. (This LCP arises from a least-square concave regression problem in statistics.) For a problem of size 50×50 , the SOR scheme fails to terminate under the rule (2.3.1) with $\varepsilon = 10^{-5}$ in 8,000 iterations. We regard this as a failure for the method. (As expected, the above matrix is very poorly conditioned.)

From a theoretical as well as practical point of view, the issue of applying an iterative method to a convex, but not strictly convex, QP is far from being satisfactorily resolved. At the present time, the proximal point algorithm and the recent iterative scheme of Shiau [52] are the only two applicable approaches. However, very little experience is available on their practical performance. (In a private communication, Olvi Mangasarian informed us that he has some encouraging results with the proximal point algorithm for solving very large linear programs.) We feel that extensive studies are urgently needed to resolve this important issue.

Although very little is formally documented, it is generally believed that many commonly used iterative methods, including those reviewed in this paper, have a linear rate of convergence. The following question naturally arises: Is it possible to accelerate

some of these existing methods or to develop methods with better rate of convergence? The recent paper [30] represents the only contribution that is known to us. We feel that this question is important and requires further study.

Finally, as parallel computers are becoming more and more readily available, it is natural to think for parallel iterative schemes for QP's. The recent paper [35] represents the first contribution on this subject. We expect more activities to follow.

REFERENCES

- [1] M. AGANAGIC, *Iterative methods for linear complementarity problems*, Technical Report SOL 78-10, Systems Optimization Laboratory, Department of Operations Research, Stanford University, 1978.
- [2] S. AGMON, *The relaxation method for linear inequalities*, Canad. J. Math., 6 (1954), pp. 382-392.
- [3] B. H. AHN, *Computation of asymmetric linear complementarity problem by iterative methods*, J. Optim. Theory Appl., 33 (1981), pp. 175-185.
- [4] B. H. AHN AND W. W. HOGAN, *On convergence of the PiES algorithm for computing equilibria*, Oper. Res., 30 (1982), pp. 281-300.
- [5] M. AVRIEL, *Nonlinear Programming: Analysis and Methods*, Prentice-Hall, Englewood Cliffs, NJ, 1976.
- [6] A. BACHEM AND B. KORTE, *An algorithm for quadratic optimization over transportation polytopes*, Z. Angew. Math. Mech., 58 (1978), pp. T459-T461.
- [7] P. BRUCKER, *An $O(n)$ algorithm for quadratic knapsack problems*, Oper. Res. Lett., 3 (1984), pp. 163-166.
- [8] J. CEA AND R. GLOWINSKI, *Sur des methodes d'optimisation par relaxation*, Revue Francaise d'Automatique, Informatique et Recherche Operationelle, R-3 (1973), pp. 5-32.
- [9] Y. CENSOR, *Row-action methods for huge and sparse systems and their applications*, SIAM Rev., 23 (1981), pp. 444-466.
- [10] Y. C. CHENG, *Iterative methods for solving linear complementarity and linear programming problems*, Ph.D. dissertation, Department of Computer Sciences, University of Wisconsin-Madison, 1981.
- [11] R. W. COTTLE, *Monotone solution of parametric linear complementarity problem*, Math. Programming, 3 (1972), pp. 210-224.
- [12] ———, *Application of a block successive overrelaxation method to a class of constrained matrix problems*, in Math. Programming, R. W. Cottle, M. L. Kelmanson and B. Korte, eds., North-Holland, Amsterdam, 1984, pp. 89-103.
- [13] R. W. COTTLE, S. G. DUVAL AND K. ZIKAN, *A Lagrangian relaxation algorithm for the constrained matrix problem*, Naval Res. Logist. Quart., to appear.
- [14] R. W. COTTLE AND M. S. GOHEEN, *A special class of large quadratic programs*, in Nonlinear Programming, 3, O. L. Mangasarian, R. R. Meyer and S. M. Robinson, eds., Academic Press, New York, 1978, pp. 361-390.
- [15] R. W. COTTLE, G. H. GOLUB AND R. S. SACHER, *On the solution of large structured linear complementarity problems: The block partitioned case*, Appl. Math. Optim., 4 (1978), pp. 347-363.
- [16] R. W. COTTLE AND J. S. PANG, *On the convergence of a block successive overrelaxation method for a class of linear complementarity problems*, Math. Programming Stud., 17 (1982), pp. 126-138.
- [17] C. W. CRYER, *The solution of a quadratic programming problem using systematic overrelaxation*, SIAM J. Control, 9 (1971), pp. 385-392.
- [18] R. S. DEMBO AND U. TULOWITZKI, *On the minimization of quadratic functions subject to box constraints*, Working Paper Series B No. 71, Yale University, 1983.
- [19] B. C. EAVES, *On quadratic programming*, Management Sci., 17 (1971), pp. 698-711.
- [20] B. FEIJOU AND R. R. MEYER, *Piecewise-linear approximation methods for nonseparable convex programming*, Technical Report # 521, Department of Computer Sciences, University of Wisconsin-Madison, December 1984.
- [21] M. FRANK AND P. WOLFE, *An algorithm for quadratic programming*, Naval Res. Logist. Quart., 3 (1957), pp. 95-110.
- [22] C. D. HA, *Decomposition methods for structured convex programming*, Ph.D. dissertation, Department of Industrial Engineering, University of Wisconsin-Madison, 1980.
- [23] ———, *An algorithm for structured large-scale quadratic programming problems*, Technical Report # 2276, Mathematics Research Center, University of Wisconsin-Madison, September 1981.
- [24] S. P. HAN AND O. L. MANGASARIAN, *A dual differentiable exact penalty function*, Math. Programming, 25 (1983), pp. 293-306.
- [25] R. HELGASON, J. KENNINGTON AND H. LALL, *A polynomially bounded algorithm for a single constrained quadratic program*, Math. Programming, 18 (1980), pp. 338-343.

- [26] G. T. HERMAN AND A. LENT, *A family of iterative quadratic optimization algorithms for pairs of inequalities with application in diagnostic radiology*, Math. Programming Stud., 9 (1978), pp. 15–29.
- [27] C. HILDRETH, *A quadratic programming procedure*, Naval Res. Logist. Quart., 4 (1957), pp. 79–85, Erratum, *ibid.*, p. 361.
- [28] A. LENT AND T. CENSOR, *Extensions of Hildreth's row action method for quadratic programming*, this Journal, 18 (1980), pp. 444–454.
- [29] Y. Y. LIN, *Iterative methods for large convex quadratic programs*, Ph.D. dissertation, School of Management, The University of Texas at Dallas, December 1985.
- [30] Y. LIN AND C. W. CRYER, *An alternating direction implicit algorithm for the solution of linear complementarity problems arising from free boundary problems*, Appl. Math. Optim., 13 (1985), pp. 1–17.
- [31] O. L. MANGASARIAN, *Solution of symmetric linear complementarity problems by iterative methods*, J. Optim. Theory Appl., 22 (1977), pp. 465–485.
- [32] ———, *Sparsity-preserving SOR algorithms for separable quadratic and linear programming problems*, Comput. Oper. Res.
- [33] ———, *A Condition Number of Linear Inequalities and Linear Programs*, in Methods of Operations Research, G. Bamberg and O. Optiz, eds., Proceedings of the 6th Symposium Operations Research, Augsburg 7–9, September 1981, Verlagsgruppe Athenaum, 1981, pp. 3–15.
- [34] ———, *Iterative solution of linear programs*, SIAM J. Numer. Anal., 18 (1981), pp. 606–614.
- [35] O. L. MANGASARIAN AND R. DE LEONE, *A Parallel Successive Overrelaxation (SOR) Algorithm for Linear Programming*, paper presented at the 12th International Symposium on Mathematical Programming held at the Massachusetts Institute of Technology, August 1985.
- [36] O. L. MANGASARIAN AND T. H. SHIAU, *Error Bounds for Monotone Linear Complementarity Problems*, Technical Report 606, Department of Computer Sciences, University of Wisconsin-Madison, July 1985.
- [37] T. MOTZKIN AND I. J. SCHOENBERG, *The relaxation method for linear inequalities*, Canad. J. Math., 6 (1954), pp. 393–404.
- [38] A. OHUCHI AND I. KAJI, *Lagrangian dual coordinatewise maximization for network and transportation problem with quadratic costs*, Networks, 14 (1984), pp. 515–530.
- [39] D. P. O'LEARY, *A generalized conjugate gradient algorithm for solving a class of quadratic programming problems*, Linear Algebra Appl., 34 (1980), pp. 371–399.
- [40] J. M. ORTEGA, *Numerical Analysis: A Second Course*, Academic Press, New York, 1972.
- [41] J.-S. PANG, *On Q-matrices*, Math. Programming, 17 (1979), pp. 243–247.
- [42] ———, *A new and efficient algorithm for a class of portfolio selection problems*, Oper. Res., 28 (1980), pp. 754–767.
- [43] ———, *On the convergence of a basic iterative method for the implicit complementarity problem*, J. Optim. Theory Appl., 37 (1982), pp. 149–162.
- [44] ———, *Methods for quadratic programming: a survey*, Computers and Chemical Engineering, 7 (1983), pp. 583–594.
- [45] ———, *Necessary and sufficient conditions for the convergence of iterative methods for the linear complementarity problem*, J. Optim. Theory Appl., 42 (1984), pp. 1–18.
- [46] ———, *More results on the convergence of iterative methods for the symmetric linear complementarity problems*, J. Optim. Theory Appl., 47 (1985).
- [47] ———, *A posteriori error bounds for the linearly-constrained variational inequality problem*, manuscript, School of Management, The University of Texas at Dallas, June 1985.
- [48] J.-S. PANG AND D. CHAN, *Iterative methods for variational and complementarity problems*, Math. Programming, 24 (1982), pp. 284–313.
- [49] R. T. ROCKAFELLAR, *Convex Analysis*, Princeton University Press, Princeton, NJ, 1970.
- [50] ———, *Monotone operators and the proximal point algorithm*, this Journal, 14 (1976), pp. 877–898.
- [51] ———, *Augmented Lagrangians and application of the proximal point algorithm in convex programming*, Math. Oper. Res., 1 (1976), pp. 97–116.
- [52] T. H. SHIAU, *An iterative scheme for linear complementarity problems*, Technical Report #2737, Mathematics Research Center, University of Wisconsin-Madison, August 1984.
- [53] R. VARGA, *Matrix Iterative Analysis*, Prentice-Hall, Englewood Cliffs, NJ, 1968.
- [54] E. K. YANG, *A class of methods for solving large convex quadratic programs subject to box constraints*, Ph.D. dissertation, Curriculum in Operations Research, University of North Carolina, Chapel Hill, forthcoming.
- [55] D. P. BERTSEKAS, *Projected Newton methods for optimization problems with simple constraints*, this Journal, 20 (1982), pp. 221–246.

LOCAL CONTROLLABILITY AND INFINITESIMAL GENERATORS OF SEMIGROUPS OF SET-VALUED MAPS*

HALINA FRANKOWSKA†

Abstract. We study the semigroup properties of reachable sets of a differential inclusion

$$x' \in F(x)$$

using the derivative of the set-valued map which associates to each initial state the set of solutions. The results are applied to the local controllability problem.

Key words. differential inclusion, local controllability, reachable set, derivative of solution with respect to initial condition, derivative of a set-valued map, generalized tangent cone, Clarke's tangent cone

AMS(MOS) subject classifications. 49A34, 49A50, 93B05

1. Introduction. In this paper we study the problem of local controllability of a system governed by a differential inclusion

$$(1) \quad x' \in F(x)$$

where F is a set-valued map from \mathbb{R}^n into the subsets of \mathbb{R}^n .

A particular case of (1) is the parametrized system (also called "control system")

$$(2) \quad x' = f(x, u(t)), \quad u(t) \in U \text{ is measurable}$$

where U is a given set; then F is defined by

$$F(x) = \{f(x, u) : u \in U\}.$$

Let $\xi \in \mathbb{R}^n$, $T > 0$ be given. Denote by $S_T(\xi)$ the set of solutions to (1) starting at ξ and defined on the time interval $[0, T]$. The reachable set to (1) at time T from ξ is denoted by $R(T, \xi)$, i.e.

$$R(T, \xi) = \{x(T) : x \in S_T(\xi)\}.$$

The system (1) is called *locally controllable* around ξ at time $T > 0$ if

$$(3) \quad \xi \in \text{Int } R(T, \xi).$$

Under quite general assumptions (boundness and upper semicontinuity of F) the necessary condition for local controllability of (1) at ξ for small $T > 0$ is

$$0 \in \overline{\text{co}} F(\xi)$$

(the closed convex hull of $F(\xi)$), i.e., ξ has to be a weak equilibrium of the map F .

The purpose of this paper is to provide a sufficient condition for (3) when ξ is an equilibrium of F , i.e., when $0 \in F(\xi)$.

We use the techniques of nonsmooth analysis and differential inclusions to answer this question in the following way:

We consider an adequate concept of tangent cone $C_{S_T(\xi)}^{U,V}(\xi)$ to the set of solutions $S_T(\xi)$ at the constant trajectory ξ and prove a kind of an Open Mapping Principle, which states that (3) follows from the "surjectivity" condition

$$(4) \quad \{w(T) : w \in C_{S_T(\xi)}^{U,V}(\xi)\} = \mathbb{R}^n$$

(we specify in § 3 the definition of $C_{S_T(\xi)}^{U,V}(\xi)$ and prove an abstract version of (4)).

How can (4) be verified? In the case when ξ is an equilibrium we proceed in the following way:

* Received by the editors July 10, 1984, and in revised form October 2, 1985.

† CEREMADE, Université de Paris-Dauphine, 75775 Paris, France.

Under a Lipschitzianity assumption, the differential inclusion (1) may be replaced by a "linear" approximation (along the solutions) around the equilibrium and we prove that if the linearized system is locally controllable at zero so does the initial system.

We have to explain now what do we mean by approximating a set-valued map. Set-valued analogues of linear operators are closed convex processes, i.e., set-valued maps whose graph are closed convex cones. We consider the set-valued derivative $CF(\xi, 0)$ associated with Clarke's tangent cone to graph of F at¹ $(\xi, 0)$. The set-valued mapping $CF(\xi, 0)$ is a closed convex process. We prove that if F is Lipschitzian, then (4) holds when the reachable set to the inclusion

$$(5) \quad w' \in \text{co } F(\xi) + CF(\xi, 0)w, \quad w(0) = 0,$$

at time T contains zero in its interior.

We proceed by investigating the analogies with the single-valued smooth case when the differential equation

$$x' = f(x), \quad x(0) = \xi,$$

has a unique solution, which can be written $r(t, \xi)$. The maps $r(t, \xi)$ form a semigroup in the sense that

$$r(0, \xi) = \xi, \quad r(t+s, \xi) = r(t, r(s, \xi)),$$

and f is the *infinitesimal generator* of this semigroup in the sense that

$$f(\xi) = \frac{\partial}{\partial t} r(t, \xi) \Big|_{t=0}.$$

Furthermore, the derivative of $r(t, \cdot)$ with respect to the initial condition is given by the formula

$$\frac{\partial}{\partial \xi} r(t, \xi)(\eta) = w(t)$$

where w is the solution to the linearized differential equation

$$w'(t) = f'(r(t, \xi))w(t), \quad w(0) = \eta.$$

This still holds true in the set-valued case (we have to replace the derivatives by intermediate derivatives for the statements below to hold true). We shall prove essentially that:

- (a) The set-valued map F is the infinitesimal generator to the semigroup $R(t, \cdot)$ and $F(\xi)$ is the derivative of $t \rightarrow R(t, \xi)$ at zero in the direction 1.
- (b) The "derivative" of $\xi \rightarrow S_T(\xi)$ at a solution $x(\cdot)$ of (1) in the direction η is the set of solutions of

$$(6) \quad w'(t) \in dF(x(t), x'(t))(w(t)), \quad w(0) = \eta.$$

This is a property which is the key to solving the local controllability problem in a straightforward way. This motivates not only the title of this paper but also provides some new properties of the reachable sets and the sets of solutions to differential inclusions studied by many authors (see, for example, Aubin and Cellina [2], Hermes [17], Clarke [10], Haddad [16], Castaing and Valadier [8], Olech [21]). (See Aubin and Cellina [2] for an exhaustive bibliography on the subject.)

¹ Recall that the graph of the derivative of a differential map f at point a is the tangent space to the graph of f at the point $(a, f(a))$. Hence a natural way to extend the concept of derivative of a set-valued map F at a point $(a, b) \in \text{graph}(F)$ is to use the set-valued map $DF(a, b)$ whose graph is a tangent cone to graph (F) at (a, b) . (In this paper we use Clarke's intermediate tangent cone: see § 2 for a precise definition.)

The main difficulty we overcome in the paper is that we do not assume that the set-valued map F has convex values. For this reason the reachable sets of (1) are not necessarily closed.

The semigroup of closed reachable sets generated by control systems and, more generally, by dynamical systems without uniqueness were studied by Bushaw [7], Roxin [25], [26], Kloeden [18], [19]. These authors considered the topological properties of such semigroups.

We investigate the differentiability properties (a), (b) of the set-valued maps R and S_T in § 2. In § 3 we prove an Open Mapping Principle and derive from it a sufficient condition for the local controllability of (1). This result is applied in § 4 to prove that the local controllability of the linearized system (5) implies the local controllability of (1). The proofs of several results of this section are given in § 6. We provide some applications in § 5.

2. Differentiability of solutions with respect to initial conditions. Let $F: \mathbb{R}^n \rightrightarrows \mathbb{R}^n$ be a set-valued map. A function $x \in W^{1,1}(0, T)$ (Sobolev space) is called a solution to the differential inclusion

$$(1) \quad x' \in F(x)$$

if and only if $x'(t) \in F(x(t))$ almost everywhere in $[0, T]$.

Remark. Let us assume that F can be parametrized in the following way for all $x \in \mathbb{R}^n$

$$F(x) = \{f(x, u): u \in U\}$$

where U is a compact metric space and $f: \mathbb{R}^n \times U \rightarrow \mathbb{R}^n$ is a continuous function. Let $x \in W^{1,1}(0, T)$ be a solution to the differential inclusion (1). By a Filippov result (see Aubin and Cellina [2, p. 91]) there exists a measurable control u (a Lebesgue measurable function $u: [0, T] \rightarrow U$) such that $x'(t) = f(x(t), u(t))$ a.e. in $[0, T]$. So we can replace (1) by the control system

$$(2) \quad x'(t) = f(x(t), u(t)), \quad u(t) \in U \text{ is a measurable control.}$$

This can be done for example when F is continuous in the Hausdorff metric and has nonempty convex compact images (see Aubin and Cellina [2, p. 73]).

We recall that, in general, a set-valued map F cannot be parametrized in a way keeping the regularity of F . For instance, when F has compact images and is Lipschitzian in the Hausdorff metric, in general, there does not exist a parametrization of F as described above with f Lipschitzian in the first variable. It seems to be still an open question even in the case when F has convex images. A partial answer to it is given in Le Donne and Marchi [20], where F is parametrized in a Lipschitzian way with the set of parameters U being a compact (nonmetrizable) space.

The control system with feedback

$$x' = f(x, u), \quad u \in U(x)$$

can be reduced to the differential inclusion (1) by setting $F(x) = \{f(x, u): u \in U(x)\}$. Observe that when f is Lipschitzian and U has compact images and is Lipschitzian in the Hausdorff metric then the set-valued map F also has compact images and is Lipschitzian in the Hausdorff metric.

We denote by B the closed unit ball.

Several results of this paper are based on the following:

THEOREM 2.0 (Filippov). *Let V be a subset of \mathbb{R}^n and let $F: V \rightrightarrows \mathbb{R}^n$ be a Lipschitzian set-valued map with closed nonempty images. Then for all $\varepsilon > 0$ there exists a $K > 0$ such*

that for all $y \in W^{1,1}(0, T)$ satisfying $y([0, T]) + \varepsilon B \subset V$ and

$$\rho_F(y) := \int_0^T \text{dist}(y'(t), F(y(t))) dt \leq \varepsilon/K$$

there exists a solution x of (1) satisfying

$$x(0) = y(0), \quad \|x' - y'\|_{L^1(0,T)} \leq K\rho_F(y).$$

For the proof see Filippov [12], Aubin and Cellina [2, p. 120].

We denote by $S_T(\xi)$ the set of all solutions to the differential inclusion (1) defined on the time interval $[0, T]$ and starting at ξ .

Existence theorems imply the nonemptiness of $S_T(\xi)$ under several combinations of assumptions (see, for instance, Aubin and Cellina [2, Chaps. 2, 4], Antosiewicz and Cellina [1], Bressan [6], Olech [21]).

DEFINITION 2.1. The set

$$R(t, \xi) := S_T(\xi)(t) = \{x(t) : x \in S_T(\xi)\}$$

where $t \in [0, T]$ is called the *reachable set* at time t from the initial condition ξ .

The reachable sets satisfy the semigroup property. Namely if $R(t, \xi) \neq \emptyset$ for all $t \in [0, T]$ then

$$(2.2) \quad \begin{aligned} (i) \quad & R(0, \xi) = \xi, \\ (ii) \quad & R(t+s, \xi) = R(t, R(s, \xi)) \quad \text{if } t+s \leq T. \end{aligned}$$

We wish next to define the derivative of the map $t \rightarrow R(t, \xi)$ at $t=0$ in the direction 1. By analogy with the single-valued case, this derivative will be called the infinitesimal generator of the semigroup.

DEFINITION 2.3. Let X, Y be Banach spaces and $G: X \rightrightarrows Y$ be a set-valued map, Lipschitzian at $x, y \in G(x)$. By $dG(x, y)$ we denote the set-valued map from X into Y defined by

$$v \in dG(x, y)(u) \text{ if and only if } \lim_{h \rightarrow 0^+} \text{dist} \left(v, \frac{G(x+hu) - y}{h} \right) = 0.$$

The map $dG(x, y)$ is called the intermediate derivative of G at (x, y) .

Remark. We assume in Definition 2.3 that whenever $G(x+hu) = \emptyset$ then $\text{dist}(v, (G(x+hu) - y)/h) = +\infty$. When G is a single-valued function, Fréchet differentiable at x_0 , then the intermediate derivative of G at the point $(x_0, G(x_0)) \in \text{graph } G$ is equal to the derivative of G .

Remark. The graph of the map $dG(x, y)$ is the so-called intermediate tangent cone to $\text{graph}(G)$ at (x, y) (see Frankowska [13], where the definition of intermediate tangent cone to an arbitrary set is given).

DEFINITION 2.4. Let $R(t, \cdot)$ be a family of set-valued maps satisfying the semigroup properties (2.2), Lipschitzian in t . We say that the map

$$\xi \rightarrow d_t R(0, \xi)(1)$$

from \mathbb{R}^n into itself is the *infinitesimal generator* of the semigroup R .

As for the single-valued case the question arises whether a given set-valued map $F: \mathbb{R}^n \rightrightarrows \mathbb{R}^n$ is the infinitesimal generator of the semigroup R . The answer is positive when F is continuous bounded with compact convex values.

THEOREM 2.5. *Let $F: \mathbb{R}^n \rightrightarrows \mathbb{R}^n$ be a bounded set-valued map with compact values and $\xi \in \text{Int Dom}(F)$. Then*

- (i) *If F is upper semicontinuous at ξ , then $d_t R(0, \xi)(1) \subset \text{co } F(\xi)$.*
- (ii) *If F is continuous on a neighborhood of ξ , then $F(\xi) \subset d_t R(0, \xi)(1)$.*
- (iii) *If F is locally Lipschitzian on a neighborhood of ξ , then $d_t R(0, \xi) = \text{co } F(\xi)$.*

Proof. If $v \in d_t R(0, \xi)(1)$ for all $h > 0$ there exists $x_h \in S_h(\xi)$ such that $x_h(h) \in R(h, \xi)$ and $\lim_{h \rightarrow 0+} (x_h(h) - \xi)/h = v$.

Since F is bounded for some $M > 0$ and all $h > 0$ we have

$$|x_h(t) - \xi| \leq tM \quad \text{for } t \in [0, h].$$

By the upper semicontinuity of F and the above inequality, for all $\varepsilon > 0$ there exists $h > 0$ such that

$$F(x_h(t)) \subset F(\xi) + \varepsilon B \quad \text{for all } t \in [0, h].$$

Thus

$$\frac{1}{h}(x_h(h) - \xi) = \frac{1}{h} \int_0^h x'_h(t) dt \in \frac{1}{h} \int_0^h (\text{co } F(\xi) + \varepsilon B) dt.$$

Since $\text{co } F(\xi) + \varepsilon B$ is a closed convex subset the mean-value theorem (see for instance Aubin and Cellina [2, p. 21]) implies that

$$\frac{x_h(h) - \xi}{h} \in \text{co } F(\xi) + \varepsilon B$$

and thus that $v \in \text{co } F(\xi) + \varepsilon B$. Since ε is arbitrary we have proved (i).

(ii) By a theorem of Filippov (see Filippov [12] and Aubin and Cellina [2, p. 112]) for all $\xi \in \text{Int Dom}(F)$ and $v \in F(\xi)$ there exists $T > 0$ and $x \in S_T(\xi)$ satisfying $x'(0) = v$. Therefore, the sequence $(x(h) - \xi)/h$ converges to v when $h \rightarrow 0+$. Hence, $v \in d_t R(0, \xi)(1)$.

(iii) By the Filippov-Ważewski relaxation theorem (see, for example, Aubin and Cellina [2, p. 124] or Clarke [10, p. 117]) for all small $T > 0$, $W^{1,1}(0, T)$ -solutions of (1) are dense in the set of $W^{1,1}(0, T)$ -solutions of the relaxed inclusion in the metric of uniform convergence. This and inclusions (i), (ii) end the proof. \square

COROLLARY 2.6. *If F is continuous, with compact values, $F(\xi)$ is convex and $\xi \in \text{Int Dom}(F)$ then*

$$d_t R(0, \xi)(1) = F(\xi).$$

As in the case of ordinary differential equations, we need to study the differentiability of the solution map with respect to initial conditions.

Consider the solution map $S_T: \mathbb{R}^n \rightrightarrows W^{1,1}(0, T)$ and the intermediate derivative $dS_T(\xi, z)(\eta)$ of S_T at point (ξ, z) in the direction η .

THEOREM 2.7. *Assume that F has closed graph and choose $z \in S_T(\xi)$. If the map F is Lipschitzian on an open neighborhood of $z([0, T])$, then*

$$\{w \in W^{1,1}(0, T): w'(t) \in dF(z(t), z'(t))w(t), w(0) = \eta\} \subset dS_T(\xi, z)(\eta).$$

Moreover if for almost all $t \in [0, T]$ and for all $u \in \mathbb{R}^n$

$$(2.8) \quad dF(z(t), z'(t))(u) = \left\{ v: \liminf_{h \rightarrow 0+} \text{dist} \left(v, \frac{F(z(t) + hu) - z'(t)}{h} \right) = 0 \right\}$$

(for example when $\text{graph}(F)$ is convex) then we have an equality in the last inclusion.

To prove the theorem we need the following

LEMMA 2.9. *Let F be a set-valued map of closed graph and $L^1(0, T)$ -functions*

$$[0, T] \ni t \rightarrow \alpha_i(t) = (x_i(t), y_i(t)) \in \text{graph } F, \quad i = 1, 2, \dots,$$

$$[0, T] \ni t \rightarrow \phi(t) = (u(t), v(t)) \in \mathbb{R}^n \times \mathbb{R}^n,$$

be given. We assume that for a sequence $h_i \rightarrow 0+$ the following holds true

$$(2.10) \quad \lim_{i \rightarrow \infty} \text{dist} \left(v(t), \frac{F(x_i(t) + h_i u(t)) - y_i(t)}{h_i} \right) = 0 \quad \text{a.e. in } [0, T].$$

If F is Lipschitzian around $\bigcup_i x_i([0, T])$ then there exists a sequence $v_i \in L^1(0, T; \mathbb{R}^n)$ converging to v in $L^1(0, T)$ such that for all large i

$$y_i(t) + h_i v_i(t) \in F(x_i(t) + h_i u(t)) \quad \text{a.e. in } [0, T].$$

Proof. Let L denote the Lipschitz constant of F . For all natural number k set

$$V_{i,k}(t) = \frac{F(x_i(t) + h_i u(t)) - y_i(t)}{h_i} \cap \left(v(t) + \frac{1}{k} B \right),$$

$$U_{i,k}(t) = \frac{F(x_i(t) + h_i u(t)) - y_i(t)}{h_i} \cap L \|u(t)\| B,$$

$$M_{i,k} = \{t \in [0, T]: V_{i,k}(t) \neq \emptyset\},$$

$$W_{i,k}(t) = \begin{cases} V_{i,k}(t) & \text{if } t \in M_{i,k}, \\ U_{i,k}(t) & \text{otherwise.} \end{cases}$$

Since the graph of F is closed and F is L -Lipschitzian around $\bigcup_i x_i([0, T])$ for all k and all large i the set-valued map $[0, T] \ni t \rightarrow W_{i,k}(t)$ has nonempty closed values and is measurable. Moreover, for all k and almost all $t \in [0, T]$, $V_{i,k}(t) \neq \emptyset$ when i is sufficiently large.

Fix k . By a measurable selection theorem (see for instance, Wagner [30]) for all large i there exists a measurable selection $v_{i,k}(t) \in W_{i,k}(t)$.

From the definition of $W_{i,k}(t)$ we obtain

$$\|v_{i,k}(t) - v(t)\| \leq 1/k \quad \text{if } t \in M_{i,k},$$

$$\|v_{i,k}(t)\| \leq L \|u(t)\| \quad \text{otherwise.}$$

Observe that for all large i

$$y_i(t) + h_i v_{i,k}(t) \in F(x_i(t) + h_i u(t)) \quad \text{a.e.,}$$

$$\|v_{i,k}(t)\| \leq \|v(t)\| + \frac{1}{k} + L \|u(t)\| \quad \text{a.e.}$$

Let $i_0 = 0$. Using the induction arguments and (2.10) we define for all $k \geq 1$ numbers $i_k > i_{k-1}$ such that for all $i > i_k$ the Lebesgue measure

$$\mu(M_{i,k+1}) \geq T - 1/k$$

and set for all $i_{k-1} < i \leq i_k$, $t \in [0, T]$

$$v_i(t) = v_{i,k}(t).$$

It is clear that v_i satisfy all the requirements of our lemma. \square

Proof of Theorem 2.7. If $w \in dS_T(\xi, z)(\eta)$ then for all $h > 0$ there exist $w_h \in W^{1,1}(0, T)$ such that

$$z + hw_h \in S_T(\xi + h\eta),$$

$$\lim_{h \rightarrow 0+} w_h = w \quad \text{in } W^{1,1}(0, T).$$

Thus $w(0) = \eta$ and

$$(2.11) \quad z'(t) + hw'_h(t) \in F(z(t) + hw_h(t)) \quad \text{a.e.},$$

$$(2.12) \quad \lim_{h \rightarrow 0+} w_h = w \quad \text{in } C([0, T]).$$

$$(2.13) \quad \text{For a subsequence } \{h_i\} \text{ converging to zero } \lim_{i \rightarrow \infty} w'_{h_i}(t) = w'(t) \quad \text{a.e.}$$

Let L denote the Lipschitz constant of F . Then by (2.11) for almost all t and for all small $h > 0$ we have

$$\begin{aligned} \text{dist} \left(w'(t), \frac{F(z(t) + hw(t)) - z'(t)}{h} \right) &\leq \text{dist} \left(w'_h(t), \frac{F(z(t) + hw_h(t)) - z'(t)}{h} \right) \\ &\quad + \|w'_h(t) - w'(t)\| + L\|w_h(t) - w(t)\| \\ &= \|w'_h(t) - w'(t)\| + L\|w_h(t) - w(t)\|. \end{aligned}$$

This and (2.12), (2.13) imply that under assumption (2.8)

$$w'(t) \in dF(z(t), z'(t))w(t) \quad \text{a.e.}$$

i.e.,

$$dS_T(\xi, z)(\eta) \subset \{w \in W^{1,1}(0, T): w'(t) \in dF(z(t), z'(t))w(t), w(0) = \eta\}.$$

To prove the opposite inclusion we shall show that if

$$\begin{aligned} w'(t) &\in dF(z(t), z'(t))w(t) \quad \text{for almost all } t \in [0, T], \\ w(0) &= \eta, \end{aligned}$$

then for all sequence $h_i > 0$ converging to zero there exists a sequence w_i converging to w in $W^{1,1}(0, T)$ such that $z + h_i w_i \in S_T(\xi + h_i \eta)$.

Consider a sequence $h_i > 0$ converging to zero. By Lemma 2.9 applied to $x_i = z$, $y_i = z'$, $u = w$, $v = w'$ there exists a sequence v_i converging to w' in $L^1(0, T)$ and such that for all i

$$(2.14) \quad z'(t) + h_i v_i(t) \in F(z(t) + h_i w(t)) \quad \text{a.e.}$$

For all $t \in [0, T]$ set

$$\pi_i(t) = \eta + \int_0^t v_i(\tau) d\tau,$$

and observe that

$$\begin{aligned} w - \pi_i &\text{ converges to zero in } C(0, T), \\ \pi'_i &\text{ converges to } w' \text{ in } L^1(0, T). \end{aligned}$$

By Lipschitzian character of F and (2.14) for a constant L and all large i we have

$$\text{dist}((z' + h_i \pi'_i)(t), F((z + h_i \pi_i)(t))) \leq L h_i \|w(t) - \pi_i(t)\|.$$

By the Filippov Theorem 2.0 for a constant M and for all large i there exists $x_i \in S_T(\xi + h_i \eta)$ satisfying

$$\|x'_i - z' - h_i \pi'_i\|_{L^1(0,T)} \leq M h_i \|w - \pi_i\|_{L^1(0,T)}.$$

Set

$$w_i = (x_i - z)/h_i.$$

Then $z + h_i w_i \in S_T(\xi + h_i \eta)$ and w_i converges to w in $W^{1,1}(0, T)$. \square

Remark. The set-valued map $dF(x, y)$ has a closed graph. Moreover if F has convex images on a neighborhood of x then $dF(x, y)$ has convex images. Indeed, if $v, v^1 \in dF(x, y)(u)$ then for all $h > 0$ there exist sequences v_h, v_h^1 converging to v, v^1 as $h \rightarrow 0+$, respectively, such that

$$y + h v_h \in F(x + h u), \quad y + h v_h^1 \in F(x + h u).$$

By convexity for all $\lambda \in [0, 1]$, $y + h(\lambda v_h + (1 - \lambda)v_h^1) \in F(x + h u)$ and therefore $\lambda v + (1 - \lambda)v^1 \in dF(x, y)(u)$. Since λ is arbitrary in $[0, 1]$ the proof follows.

3. The local controllability problem. Let $F: \mathbb{R}^n \rightrightarrows \mathbb{R}^n$ be a set-valued map, $\xi \in \mathbb{R}^n$ and $R(T, \xi)$ denote the reachable set from ξ at time T of the differential inclusion

$$(1) \quad x' \in F(x).$$

We seek to know whether $\xi \in \text{Int } R(T, \xi)$. If it holds true then we say that (1) is *locally controllable around ξ at time T* .

Our first result concerns a necessary condition for local controllability.

THEOREM 3.1. *Assume that a set-valued map F from \mathbb{R}^n into itself is bounded and upper semicontinuous at a point $\xi \in \mathbb{R}^n$. If the system (1) is locally controllable around ξ for all small time $T > 0$ then*

$$0 \in \overline{\text{co}} F(\xi)$$

(the closed convex hull of $F(\xi)$).

Proof. If $0 \notin \overline{\text{co}} F(\xi)$ then there exists $p \in S^{n-1}$ such that

$$\inf_{u \in F(\xi)} \langle p, u \rangle > 0.$$

Since F is upper semicontinuous at ξ there exists $\beta > 0$ such that

$$\inf \{ \langle p, u \rangle : u \in F(x), x \in \xi + \beta B \} \geq 0.$$

Let $M > 0$ be such that the image of F is contained in the ball MB . If $T \leq \beta/M$ then we have for all $x \in S_T(\xi)$

$$x([0, T]) \subset \xi + MTB \subset \xi + \beta B$$

and therefore

$$\langle p, x(T) \rangle = \langle p, \xi \rangle + \int_0^T \langle p, x'(t) \rangle dt \geq \langle p, \xi \rangle.$$

But this means that $\xi \notin \text{Int } R(T, \xi)$. \square

Remark. Consider the relaxed differential inclusion

$$x' \in \text{co } F(x)$$

and let $R^{\text{co}}(T, \xi)$ denote its reachable set from ξ at time T . When F has compact images and is Lipschitzian, the relaxation theorem (see, for example, Aubin and Cellina

[2, p. 124] or Clarke [10, p. 117]) implies that $\text{Int cl } R(T, \xi) = \text{Int } R^{\text{co}}(T, \xi)$. Hence the local controllability of the relaxed inclusion around ξ implies that the *closure* of the reachable set $R(T, \xi)$ contains ξ in its interior, $\xi \in \text{Int cl } R(T, \xi)$, i.e. that (1) is “almost” locally controllable around ξ .

We shall study next sufficient conditions for local controllability around the point of weak (or strong) equilibrium.

For this we shall use an Open Mapping Principle. We recall

DEFINITION 3.2 (of Clarke’s tangent cone). Let E be a Banach space, K be a subset of E and x be a point in the closure \bar{K} of K . We say that v belongs to $C_K(x)$ (Clarke’s tangent cone to K at x) if and only if for all sequence $x_i \in K$, $h_i > 0$ converging to x and zero, respectively, we can find a sequence $v_i \in E$ converging to v such that $x_i + h_i v_i \in K$.

The set $C_K(x)$ is a closed convex cone (see Clarke [10]).

For studying the local controllability problem we need several weaker topologies on E . This is why we shall adapt Definition 3.2 to our case.

DEFINITION 3.3. Let U, V, E be Banach spaces and $E \subset U$, $E \subset V$. We assume that the topology of E is stronger than the ones of U and V . Let K be a subset of E and let $x \in \bar{K}^U$ (the closure of K in U). We say that $w \in C_K^{U,V}(x)$ if and only if for all sequence $x_i \in K$ converging to x in the space U there exists a constant $m = m(w)$ such that for all sequence $h_i > 0$ converging to zero we can find a sequence $w_i \in E$ converging to w in the space V , verifying for all large i

$$x_i + h_i w_i \in K, \quad \|w_i\|_E \leq m.$$

The set $C_K^{U,V}(x)$ is a convex cone. Moreover,

$$C_K^{E,E}(x) = C_K(x) \subset C_K^{U,V}(x).$$

We denote by $\|\cdot\|_W, \|\cdot\|_C$ the usual norms of $W^{1,1}(0, T)$, $C(0, T)$, respectively.

THEOREM 3.4. Let $F: \mathbb{R}^n \rightrightarrows \mathbb{R}^n$ be a set-valued map with closed graph. If one of the following two assumptions holds true:

- (i) ξ is a weak equilibrium of F , i.e. $0 \in \overline{\text{co}} F(\xi)$, F is bounded, Lipschitzian on a neighborhood of ξ and $\{w(T): w \in C_{S_T(\xi)}^{C,C}(\xi)\} = \mathbb{R}^n$;
 - (ii) ξ is an equilibrium of F , i.e. $0 \in F(\xi)$, and $\{w(T): w \in C_{S_T(\xi)}^{W,C}(\xi)\} = \mathbb{R}^n$;
- then (1) is locally controllable around ξ at time T .

To prove the above theorem we shall use the following:

OPEN MAPPING PRINCIPLE 3.5. Let E, U, V be Banach spaces $E \subset U \subset V$, K be a closed subset of E and $x_0 \in \bar{K}^U$ (the closure of K in U). Let A be a continuously differentiable map from a neighborhood of K in V into \mathbb{R}^q . If

$$A'(x_0)C_K^{U,V}(x_0) = \mathbb{R}^q,$$

then

$$A(x_0) \in \text{Int } A(K).$$

The proof of this result is given at the end of the section.

Proof of Theorem 3.4. Since $\text{graph}(F) := \{(x, y): y \in F(x), x \in \mathbb{R}^n\}$ is a closed subset of \mathbb{R}^{2n} the set $K := S_T(\xi)$ is closed in $W^{1,1}(0, T)$. By Filippov–Ważewski’s relaxation theorem (see [10, p. 117]) the constant trajectory ξ belongs to the closure of K in the metric $\|\cdot\|_C$. If (i) holds use the open mapping principle with $E = W^{1,1}(0, T)$, $U = V = C(0, T)$ and $A: V \rightarrow \mathbb{R}^n$ defined by $Aw = w(T)$. If (ii) holds set $U = W^{1,1}(0, T)$ and E, V, A as in (i). \square

Remark. As in § 2 we can associate with Clarke's tangent cone $C_{\text{graph}(F)}(\xi, 0)$ the derivative $CF(\xi, 0)$ and we can prove that if ξ is an equilibrium then any solution of the "sublinearization"

$$(3.6) \quad w' \in CF(\xi, 0)(w), \quad w(0) = 0,$$

belongs to $C_{S_T(\xi)}(\xi) \subset C_{S_T(\xi)}^{w,C}(\xi)$. Thus the controllability of (3.6) around zero at time T implies the controllability of (1) around ξ at time T . In the next section we shall improve this result taking a larger sublinearization.

Proof of the open mapping principle. We assume for a moment that $A(x_0)$ does not belong to $\text{Int } A(K)$ and we shall derive a contradiction. Then for all $n \geq 1$ there exists $y_n \in R^q \setminus A(K)$ such that

$$\|A(x_0) - y_n\| \leq 1/2n^2.$$

Since $x_0 \in \bar{K}^U$ by continuity of A there exist $x_0^n \in K$ such that x_0^n converges to x_0 in U and $\|A(x_0^n) - A(x_0)\| \leq 1/n^2$. By Ekeland's variational principle applied to the function $x \rightarrow \|A(x) - y_n\|$ on the complete subset K of E (see Ekeland [11], Aubin and Ekeland [4, p. 255]) there exist $x_n \in K$ such that

$$(3.7) \quad \begin{aligned} \text{(i)} \quad & \|A(x_n) - y_n\| + \frac{1}{n} \|x_n - x_0^n\|_E \leq \|A(x_0^n) - y_n\| \leq \frac{1}{n^2}, \\ \text{(ii)} \quad & \text{for all } x \in K \quad \|A(x_n) - y_n\| \leq \|A(x) - y_n\| + \frac{1}{n} \|x_n - x\|_E. \end{aligned}$$

By (i) we know that x_n converges to x_0 . Introduce a function $f: V \rightarrow \mathbb{R}$ by

$$f(v) := \|A(v) - y_n\|.$$

By assumptions $A(x_n) - y_n \neq 0$. Let us set

$$p_n = \frac{A(x_n) - y_n}{\|A(x_n) - y_n\|} \in S^{q-1}.$$

Then

$$f'(x_n) = A'(x_n)^* p_n.$$

S^{q-1} being a compact, we can take a subsequence p_{n_i} converging to some $p \in S^{q-1}$. Let $w \in C_K^{U,V}(x_0)$ be such that $A'(x_0)w = -p$. From now on we set $x_i = x_{n_i}$, $p_i = p_{n_i}$, $y_i = y_{n_i}$. Since A is Fréchet differentiable on a neighborhood of x_i in V , for all $i > 0$ there exists $\eta_i > 0$ such that for all $v \in V$ of $\|v\|_V \leq \eta_i$

$$(3.8) \quad \|A(x_i + v) - y_i\| - \|A(x_i) - y_i\| \leq \langle p_i, A'(x_i)v \rangle + \frac{1}{n_i} \|v\|_V.$$

Let $m > 1$ be the constant associated with w and $\{x_i\}$ by Definition 3.3. Consider any sequence $h_i \in]0, \eta_i/m[$ converging to zero.

Then there exists a sequence $w_i \in E$ converging to w in V such that

$$(3.9) \quad \|w_i\|_E \leq m, \quad x_i + h_i w_i \in K.$$

Setting $x = x_i + h_i w_i$ in (3.7)(ii) and using (3.8), (3.9) we obtain

$$\begin{aligned} -\frac{m}{n_i} & \leq \frac{1}{h_i} (\|A(x_i + h_i w_i) - y_i\| - \|A x_i - y_i\|) \\ & \leq \langle p_i, A'(x_i) w_i \rangle + \frac{m}{n_i}. \end{aligned}$$

Since A' is continuous at x_0 and w_i converges to w in V , we obtain by passing to the limit in the last inequality when $i \rightarrow \infty$

$$\langle p, A'(x_0)w \rangle \geq 0$$

which contradicts the choice of w and achieves the proof. \square

4. Controllability through set-valued linearization. In this section we shall provide an application of the results of § 3. We recall first

DEFINITION 4.1. Let $F: \mathbb{R}^n \rightrightarrows \mathbb{R}^n$ be a set-valued map, Lipschitzian on a neighborhood of a point x_0 , $y_0 \in F(x_0)$. We denote by $CF(x_0, y_0)$ the set-valued map from \mathbb{R}^n into \mathbb{R}^n defined by

$$v \in CF(x_0, y_0)(u) \quad \text{if and only if} \quad \limsup_{\substack{(x,y) \rightarrow (x_0,y_0) \\ (x,y) \in \text{graph } F \\ h \rightarrow 0+}} \text{dist} \left(v, \frac{F(x+hu) - y}{h} \right) = 0.$$

If F is a single-valued function continuously differentiable at x_0 , then $y_0 = F(x_0)$ and $CF(x_0, y_0)$ is equal to the Fréchet derivative to F at x_0 .

The graph of the set-valued map $CF(x_0, y_0)$ is the tangent cone (of Clarke) to graph (F) at (x_0, y_0) . Thus it is a closed convex cone. We recall

DEFINITION 4.2. A set-valued map $\mathcal{A}: \mathbb{R}^n \rightrightarrows \mathbb{R}^n$ is called a closed convex process if its graph is a closed convex cone.

So $CF(x_0, y_0)$ is a closed convex process satisfying

$$\text{graph } CF(x_0, y_0) \subset \text{graph } dF(x_0, y_0).$$

Assume that ξ is an equilibrium and consider the first order approximation of (1) at $(\xi, 0)$

$$(4.3) \quad w' \in F(\xi) + CF(\xi, 0)w, \quad w(0) = 0.$$

From now on we write l.c. for locally controllable.

THEOREM 4.4. Assume that F has compact images and is Lipschitzian on a neighborhood of an equilibrium ξ . Then the inclusion (1) is l.c. around ξ at time $T > 0$ if the inclusion (4.3) is l.c. around zero at time T .

Observe that l.c. of the inclusion (4.3) implies implicitly that for all $x \in \mathbb{R}^n$, $CF(\xi, 0)(x) \neq \emptyset$, i.e. $\text{Dom } CF(\xi, 0) = \mathbb{R}^n$. By the Robinson-Ursescu theorem (see Robinson [23], Ursescu [29], Aubin and Ekeland [4, p. 132]) the set-valued map $CF(\xi, 0)$ is Lipschitzian. Using the proof of the Filippov-Ważewski relaxation theorem (see Clarke [10, p. 117]) one can verify that the set of solutions on $[0, T]$ to the inclusion (4.3) is dense with respect to the metric of uniform convergence on $[0, T]$ in the set of solutions on $[0, T]$ to the relaxed inclusion

$$(4.5) \quad w' \in \text{co } F(\xi) + CF(\xi, 0)w, \quad w(0) = 0.$$

On the other hand if (4.3) is l.c. around zero at time T , so is (4.5). For this reason instead of Theorem 4.4 we shall prove a stronger:

THEOREM 4.6. Assume that F has compact images and is Lipschitzian on a neighborhood of an equilibrium ξ . Then the inclusion (1) is l.c. around ξ at time $T > 0$ if the inclusion (4.5) is l.c. around zero at time T .

To prove Theorem 4.6 we need several lemmas. For the reader's convenience we first state all of them. Their proofs are provided in § 6.

The first lemma is a kind of bang-bang principle for the points around zero (for the bang-bang in the linear case, see, for example, Olech [22]).

LEMMA 4.7. Let $\mathcal{A}: \mathbb{R}^n \rightrightarrows \mathbb{R}^n$ be a closed convex process and $Y \subset \mathbb{R}^n$ be a nonempty convex compact set. Assume that for some $T > 0$ the inclusion

$$(4.8) \quad x' \in \mathcal{A}(x) + Y, \quad x(0) = 0,$$

is l.c. around zero at time T . Then there exists a finite subset $U \subset Y$ of extremal points of Y such that the inclusion

$$(4.9) \quad x' \in \mathcal{A}(x) + \text{co } U, \quad x(0) = 0,$$

is l.c. around zero at time T .

The second lemma is a technical result we need here.

LEMMA 4.10. Let U be a finite subset of \mathbb{R}^n , $0 \in U$, and $T > 0$ be given. Then there exists a constant C depending only on $\text{co } U$ such that for all measurable subset $A \subset [0, T]$ and $a \in \int_A \text{co } U$, we can find a selection $\alpha(t) \in U$ satisfying

$$a = \int_A \alpha(t) dt, \quad \int_A \|\alpha(t)\| dt \leq C \|a\|.$$

The third lemma is related to the sufficient conditions for the local controllability from § 3.

LEMMA 4.11. Assume that F satisfies all the assumptions of Theorem 4.6. Let $U \subset F(\xi)$ be a finite set, $0 \in U$. Then any solution $w \in W^{1,1}(0, T)$ of the inclusion

$$(4.12) \quad w' \in \text{co } U + CF(\xi, 0)w, \quad w(0) = 0,$$

belongs to $C_{S_T(\xi)}^{w,C}(\xi)$.

Proof of Theorem 4.6. The set-valued map $CF(\xi, 0)$ is a closed convex process and $\text{co } F(\xi)$ is a nonempty convex compact set. By Lemma 4.7 there exists a finite set $U \subset F(\xi)$ such that the inclusion (4.12) is l.c. around zero at time T . The set $U \cup \{0\}$ satisfies the assumptions of Lemma 4.11. Thus

$$0 \in \text{Int} \{w(T): w \in C_{S_T(\xi)}^{w,C}(\xi)\}.$$

But this is equivalent to the condition (ii) of Theorem 3.4. The proof is complete. \square

We state next a theorem which allows us to replace the right-hand side of (4.5) by a closed convex process.

Consider the closed convex cone generated by $\text{co } F(\xi)$

$$L = \text{cl} \{\lambda w: \lambda \geq 0; w \in \text{co } F(\xi)\}$$

(L is the tangent cone of convex analysis to $\text{co } F(\xi)$ at zero) and define a set-valued map $\mathcal{A}: \mathbb{R}^n \rightrightarrows \mathbb{R}^n$ by

$$\mathcal{A}(x) = \text{cl} (CF(\xi, 0)x + L).$$

PROPOSITION 4.13. If $\text{Dom } CF(\xi, 0) = \mathbb{R}^n$ then the map \mathcal{A} defined as above is a closed convex process.

Proof. By the Robinson–Ursescu theorem the map $CF(\xi, 0)$ is Lipschitzian. So is the map \mathcal{A} . Moreover \mathcal{A} has closed images. It implies that graph \mathcal{A} is closed. Since the graph of the set-valued map $x \rightarrow CF(\xi, 0)x + L$ is a convex cone so is graph \mathcal{A} .

THEOREM 4.14. Assume that F has compact images and is Lipschitzian on a neighborhood of an equilibrium ξ . Then inclusion (4.5) is l.c. around zero at time $T > 0$ if and only if the reachable set to the inclusion

$$(4.15) \quad x' \in \text{cl} (CF(\xi, 0)x + L), \quad x(0) = 0,$$

at time T is equal to \mathbb{R}^n .

This theorem follows from a more general lemma.

LEMMA 4.16. *Let $\mathcal{B}: \mathbb{R}^n \rightrightarrows \mathbb{R}^n$ be a closed convex process and Y be a convex compact subset of \mathbb{R}^n , $0 \in Y$. For all $x \in \mathbb{R}^n$ set*

$$\mathcal{A}(x) = \text{cl} \{ \mathcal{B}(x) + \lambda w : \lambda \geq 0, w \in Y \}.$$

Then the inclusion

$$(4.17) \quad x' \in \mathcal{B}(x) + Y, \quad x(0) = 0,$$

is l.c. around zero at time $T > 0$ if and only if the reachable set to the inclusion

$$(4.18) \quad x' \in \mathcal{A}(x), \quad x(0) = 0,$$

at time T is equal to \mathbb{R}^n .

Proof. Observe that for all $x \in \mathbb{R}^n$, $\mathcal{B}(x) + Y \subset \mathcal{A}(x)$. Thus the reachable set to (4.17) at time T is contained in the reachable set to (4.18) at time T . On the other hand, graph \mathcal{A} is a cone. Hence the reachable sets to (4.18) are also cones. Thus the l.c. of (4.17) around zero at time T implies that the reachable set to (4.18) at time T is equal to \mathbb{R}^n . Assume next that the reachable set to (4.18) at time T is the whole space. For all $x \in \mathbb{R}^n$ set

$$G(x) = \mathcal{B}(x) + Y.$$

Because $\text{Dom } \mathcal{A} = \mathbb{R}^n$ then $\text{Dom } \mathcal{B} = \mathbb{R}^n$ and by the Robinson-Ursescu theorem \mathcal{B} is Lipschitzian on \mathbb{R}^n . Therefore

$$(4.19) \quad \begin{aligned} &G \text{ is Lipschitzian on } \mathbb{R}^n, \quad 0 \in G(0), \\ &\text{graph } G \text{ is convex and closed.} \end{aligned}$$

This implies that the intermediate derivative $dG(0, 0)$ has a convex graph. Because $0 \in Y$, graph \mathcal{B} is a cone contained in graph G . This implies that

$$(4.20) \quad \text{graph } \mathcal{B} \subset \text{graph } dG(0, 0).$$

Since 0 belongs to $\mathcal{B}(0)$, we also have

$$(4.21) \quad \{ \lambda w : \lambda \geq 0, w \in Y \} \subset dG(0, 0)(0).$$

Since graph $dG(0, 0)$ is a closed convex cone we finally obtain by (4.20), (4.21) that for all $x \in \mathbb{R}^n$

$$(4.22) \quad \mathcal{A}(x) \subset dG(0, 0)x.$$

Let $S_T(0)$ denote the set of all solutions on $[0, T]$ to the inclusion

$$x' \in G(x), \quad x(0) = 0.$$

The set $S_T(0) \subset W^{1,1}(0, T)$ is convex. Therefore the Clarke tangent cone $C_{S_T(0)}(0)$ is equal to the tangent cone of convex analysis to $S_T(0)$ at zero (see, for example, Aubin and Ekeland [4, p. 407]). It implies that

$$(4.23) \quad dS_T(0, 0)(0) = C_{S_T(0)}(0).$$

By (4.19) G and $z \equiv 0$ satisfy all the assumptions of Theorem 2.7. This and (4.22) together yield

$$\begin{aligned} dS_T(0, 0)(0) &= \{ w \in W^{1,1}(0, T) : w'(t) \in dG(0, 0)w(t), w(0) = 0 \} \\ &\supset \{ x \in W^{1,1}(0, T) : x'(t) \in \mathcal{A}(x(t)), x(0) = 0 \}. \end{aligned}$$

Since the reachable set at time T to (4.18) is the whole space, using the last relation and (4.23) we obtain

$$\{w(T): w \in C_{S_T(0)}(0) = \mathbb{R}^n\}.$$

But $C_{S_T(0)}(0) \subset C_{S_T(0)}^{w,C}(0)$ and we complete the proof by Theorem 3.4(ii). \square

Remark. Necessary and sufficient conditions for the controllability of closed convex processes were recently studied by J. P. Aubin, C. Olech and the author in a forthcoming paper [5].

Remark. To simplify the matter we have assumed in Theorems 4.4, 4.6 and 4.14 that F has compact values. This assumption can be weakened. Namely it is sufficient for F to have closed (not necessarily bounded) images. The proofs will require a small modification based on density arguments and Filippov's Theorem 2.0.

Remark. One must pay attention to the high sensitivity that the derivative $CF(\xi, 0)$ inherits from the properties of the Clarke tangent cone. As an example, consider the closed unit ball B in \mathbb{R}^2 and the set $A = B \cup \{0\} \times [1, +\infty]$. Then, although the set A is larger than B we have

$$C_B(0, 1) = \mathbb{R} \times \mathbb{R}_-, \quad C_A(0, 1) = \{0\}.$$

When a similar thing happens to $CF(\xi, 0)$ it is often more appropriate (when it is possible) to consider a smaller differential inclusion

$$x' \in Q(x)$$

having the property

$$(\xi, 0) \in \text{graph } Q \subset \text{graph } F.$$

The local controllability of a "smaller" inclusion then will imply the local controllability of the inclusion (1).

The following example illustrates this remark.

Example. Consider the control system in \mathbb{R}^2

$$x' = u, \quad u \in [-1, 1],$$

$$y' = xv, \quad v \in [-1, 1].$$

Then $F(x, y) = [-1, 1] \times [-|x|, |x|]$. The direct computation of $CF(0)$ gives then $CF(0)(x) = \mathbb{R} \times \{0\}$ for all $x \in \mathbb{R}^n$ and $F(0) = \text{co } F(0) = [-1, 1] \times 0$.

The inclusion

$$w' \in CF(0)w + \text{co } F(0), \quad w(0) = 0,$$

is not l.c. at any time T . But for any $v_0 \in [-1, 1]$, $v_0 \neq 0$ the linear system

$$x' = u, \quad y' = xv_0,$$

is controllable.

5. Some applications.

5.1. The parametrized case. We show here how to derive from Theorem 4.6 a result on local controllability of parametrized systems, without assuming too much regularity. Let U be a compact metric space and let $f: \mathbb{R}^n \times U \rightarrow \mathbb{R}^n$ be a continuous

function. Assume that for some $(\xi, \bar{u}) \in \mathbb{R}^n \times U$, $f(\xi, \bar{u}) = 0$ and for some $L, \beta > 0$ and all $u \in U$, $f(\cdot, u)$ is L -Lipschitzian on $\xi + \beta B$ and

$$\frac{\partial f}{\partial x}(\cdot, \bar{u}) \text{ is continuous at } \xi.$$

Consider the control system

$$(5.1) \quad x' = f(x, u(t)), \quad u(t) \in U \text{ is measurable, } x(0) = \xi.$$

We wish to study the local controllability of system (5.1) at a given time T .

Set $F(x) = \{f(x, u) : u \in U\}$. By Aubin and Cellina [2, p. 91] the set of solutions of the control system (5.1) defined on the time interval $[0, T]$ coincides with $S_T(\xi)$. Moreover F has compact images and is Lipschitzian on $\xi + \beta B$, ξ being an equilibrium. Therefore, we can apply the results of § 4 to the control system (5.1).

THEOREM 5.2. *If the system*

$$x' = \frac{\partial f}{\partial x}(\xi, \bar{u})x + v, \quad v \in \text{co } f(\xi, U), \quad x(0) = 0,$$

is l.c. around zero at time T then the system (5.1) is l.c. around ξ at time T .

Proof. By Lemma 4.7 there exist a finite subset $U_1 \subset U$ such that for all $u, v \in U_1$, $u \neq v$, $f(\xi, u) \neq f(\xi, v)$ and the system

$$x' \in \frac{\partial f}{\partial x}(\xi, \bar{u})x + \text{co } f(\xi, U_1), \quad x(0) = 0,$$

is l.c. around zero at time T . It is not restrictive to assume that $\bar{u} \in U_1$. Indeed, if $f(x, \bar{u}) = f(x, u)$ for some $u \in U_1$ we can replace U_1 by $U_1 \cup \{\bar{u}\} \setminus \{u\}$.

Thus we may assume that $U = U_1$. We set then $F(x) = f(x, U)$. A simple computation then gives

$$CF(\xi, 0) = \frac{\partial f}{\partial x}(\xi, \bar{u})$$

and by Theorem 4.6 we complete the proof. \square

The above theorem generalizes a result of Yorke [32].

5.2. An example of controllability of a control system nondifferentiable in the parameter. Consider the following system in \mathbb{R}^2

$$(5.3) \quad x' = x - y + x\sqrt{u} + u, \quad y' = x + yu, \quad u \in [0, 1].$$

The point $(x, y) = 0$ is an equilibrium. The corresponding control is $\bar{u} = 0$. The linearized system from Theorem 5.2 can be written in the form

$$(5.4) \quad w' = Aw + Bu, \quad u \in U$$

where

$$A = \begin{pmatrix} 1 & -1 \\ 1 & 0 \end{pmatrix}, \quad B = \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \quad U = [0, 1].$$

The rank of A is equal to $\text{rg } A = 2$ and the transposed matrix A^* has no real eigenvalues. By a result of Saperstone and Yorke [27] the linear system (5.4) is l.c. around zero at some time $T > 0$.

Therefore

$$\liminf_j \hat{\Phi}_j|_D \cong \int_{D_{\varepsilon'}} r_K(\bar{z}(\tau), \bar{\chi}(\tau), \bar{\chi}'(\tau)) d\tau - \varepsilon''' M(\gamma_3 + \gamma_4) + \nu.$$

Let $\bar{\varepsilon} > 0$. Then for our fixed K and ε' such that

$$-(\gamma_2 - \gamma_1 M)\varepsilon' < \frac{\bar{\varepsilon}}{5}$$

we can choose, in order,

$$\varepsilon''' M(\gamma_3 + \gamma_4) < \frac{\bar{\varepsilon}}{5},$$

$$\varepsilon'' M < \frac{\bar{\varepsilon}}{5},$$

$$K\delta' \gamma_{K\varepsilon''} m(D_{\varepsilon'}) < \frac{\bar{\varepsilon}}{5},$$

$$-\delta'(\gamma_2 - \gamma_1 M)m(D_{\varepsilon'}) < \frac{\bar{\varepsilon}}{5}.$$

Hence,

$$\liminf_j \hat{\Phi}_j|_D \cong \int_{D_{\varepsilon'}} r_K(\bar{z}(\tau), \bar{\chi}(\tau), \bar{\chi}'(\tau)) d\tau - \bar{\varepsilon}.$$

Since K and ε' were arbitrary,

$$\begin{aligned} \liminf_j \hat{\Phi}(z_j, \chi_j, \alpha_j^{-1})|_D &\cong \int_D r(\bar{z}(\tau), \bar{\chi}(\tau), \bar{\chi}'(\tau)) d\tau - \bar{\varepsilon} \\ &= \hat{\Phi}(\bar{z}, \bar{\chi}, \bar{\alpha}^{-1})|_D - \bar{\varepsilon}. \end{aligned}$$

Therefore,

$$(3.6) \quad \liminf_j \hat{\Phi}(z_j, \chi_j, \alpha_j^{-1})|_D \cong \hat{\Phi}(\bar{z}, \bar{\chi}, \bar{\alpha}^{-1})|_D.$$

Let us define an x^* such that

$$x^*(\bar{z}(\tau)) = \bar{\chi}(\tau), \quad \tau \notin D.$$

Then x^* (or more precisely, its equivalence class) belongs to \mathcal{B} . What we wish to show is that x^* and \bar{x} are equivalent and that from $\bar{\chi}$ we can construct an $\tilde{x} \in \mathcal{A}_{M\bar{\theta}}(\bar{x})$ so that (3.6) implies a corresponding result for the $\Phi(x_j)$ and $\Phi(\bar{x})$. Firstly let us show the equivalence of x^* and \bar{x} . The function \bar{z} represents the original time variable t . We want to show that $x^*(\bar{z}(\tau)) = \bar{x}(\bar{z}(\tau))$ for almost all $\tau \notin D$ since the times $z(\tau)$, $\tau \in D$, correspond to the singular part of dx^* .

Fix $\varepsilon > 0$; then there exists an open set $A \subset T$ with $m(A) < \varepsilon$ such that $A \supset \bigcup_{j=1}^{\infty} z_j(D_j) \cup \bar{z}(D)$ where $D_j = \{\tau: z'_j(\tau) = 0\}$. Since $(z_j, \chi_j) \rightarrow (\bar{z}, \bar{\chi})$ uniformly, for every $\delta > 0 \exists k$ such that for $j \geq k$

$$|z_j(\tau) - \bar{z}(\tau)| < \delta \quad \text{and} \quad |\chi_j(\tau) - \bar{\chi}(\tau)| < \delta \quad \forall \tau.$$

But on $T \setminus A$, \bar{z}^{-1} and \bar{z}_j^{-1} , $j = 1, 2, \dots$ exist since we have $\bar{z}'(\tau) = 1$ when $\bar{z}(\tau) \in T \setminus A$ and $z'_j(\tau) = 1$ when $z_j(\tau) \in T \setminus A$. We will then have for $j \geq k$ and $t \in T \setminus A$

$$(3.7) \quad |z_j^{-1}(t) - \bar{z}^{-1}(t)| < \delta.$$

For a (Lebesgue) measurable subset $A \subset [0, T]$ we denote by $\mu(A)$ its Lebesgue measure. Then for all measurable $A \subset [0, T]$

$$(6.6) \quad \int_A U = \int_A \text{co } U = \mu(A) \text{co } U,$$

(see for example Olech [22]). The set $\text{ext co } U \subset U$ of extremal points of $\text{co } U$ is finite. Because $0 \in \text{co } U$ there exists a finite family $\{K_i\}_{i=1}^m$ of subsets $K_i \subset \text{ext co } U$ such that for all i , $0 \notin \text{co } K_i$ and $\bigcup_{i=1}^m \text{co } (K_i \cup \{0\}) = \text{co } U$. Set $U_i = K_i \cup \{0\}$. Then

$$(6.7) \quad \bigcup_{i=1}^m \text{co } U_i = \text{co } U.$$

For all i , zero is a vertex of the polyhedra $\text{co } U_i$. Thus there exist $C_1 > 0$ and $p_i \in S^{n-1}$, $i = 1, \dots, m$ such that for all i

$$(6.8) \quad \sup_{\substack{y, z \in \text{co } U_i \\ y \neq 0}} \left\{ \frac{\|z\|}{\|y\|} : \langle z, p_i \rangle \geq \langle y, p_i \rangle \right\} \leq C_1.$$

Set $C = (2n-1)C_1$. We claim that C is the constant we are looking for. Indeed fix a measurable set $A \subset [0, T]$ and $a \in \int_A \text{co } U$. Then $\int_A U_i = \mu(A) \text{co } U_i$ and by (6.6), (6.7) $\int_A \text{co } U = \bigcup_{i=1}^m \int_A U_i$. Thus for some i_0 , $a \in \int_A U_{i_0} = \mu(A) \text{co } U_{i_0}$. By (6.8)

$$\sup_{z \in \text{co } U_{i_0}} \{ \|z\| : \langle z, p_{i_0} \rangle \geq \langle a, p_{i_0} \rangle \} \leq C_1 \|a\|.$$

A lemma of Frankowska-Olech [14] implies the existence of $\bar{\alpha}(t) \in U_{i_0}$ such that

$$\int_A \bar{\alpha}(t) dt = a, \quad \int_A \|\bar{\alpha}(t)\| dt \leq (2n-1)C_1 \|a\| = C \|a\|.$$

Set

$$\alpha(t) = \begin{cases} \bar{\alpha}(t) & \text{if } t \in A, \\ 0 & \text{otherwise.} \end{cases}$$

Then $\alpha(t) \in U_{i_0} \subset U$ and α is the required function.

Proof of Lemma 4.12. Let $w \in W^{1,1}(0, T)$ be a solution to the inclusion 4.12 and let $u, v \in L^1(0, T)$ be such that for almost all $t \in [0, T]$

$$(6.9) \quad \begin{aligned} w'(t) &= u(t) + v(t), \\ u(t) &\in \text{co } U, \\ v(t) &\in CF(\xi, 0)(w(t)). \end{aligned}$$

Let $x_i \in S_T(\xi)$ be a sequence converging in $W^{1,1}(0, T)$ to the constant trajectory ξ and $h_i \in]0, 1]$ be a sequence converging to zero.

Because $0 \in \text{co } U$, for all i

$$(6.10) \quad h_i u(t) \in \text{co } U.$$

By Lemma 2.9 there exist $v_i \in L^1(0, T)$ such that for all large i

$$(6.11) \quad x'_i(t) + h_i v_i(t) \in F(x_i(t) + h_i w(t)) \quad \text{a.e. in } [0, T],$$

$$(6.12) \quad \lim_{i \rightarrow \infty} v_i = v \quad \text{in } L^1(0, T).$$

Assume for a moment that there exist $y_i \in W^{1,1}(0, T)$ such that:

$$(6.13) \quad y_i(0) = \xi,$$

$$(6.14) \quad \text{for a constant } M \text{ independent of } \{h_i\} \text{ and for all large } i, \|y'_i - x'_i\|_{L^1} \leq Mh_i,$$

$$(6.15) \quad (y_i - x_i)/h_i \rightarrow w \quad \text{uniformly on } [0, T],$$

$$(6.16) \quad \lim_{i \rightarrow \infty} \frac{1}{h_i} \int_0^T \text{dist}(y'_i(t), F(y_i(t))) dt = 0.$$

Then we can apply the Filippov Theorem 2.0 to deduce the existence of M_1 independent of h_i and $z_i \in S_T(\xi)$ such that $(z_i - x_i)/h_i \rightarrow w$ uniformly on $[0, T]$ and for all large i , $\|z'_i - x'_i\|_{L^1} \leq M_1 h_i$. But this yields $w \in C_{S_T(\xi)}^{w, C}(\xi)$.

Therefore it remains to construct functions y_i as described above. We proceed in two steps.

Step 1. We construct here functions $\alpha_i(t) \in U$, $i = 1, 2, \dots$ such that for a constant $C > 0$ and all $i \geq 1$

$$(6.17) \quad \mu\{t \in [0, T]: \alpha_i(t) \neq 0\} \leq Ch_i,$$

$$(6.18) \quad \lim_{i \rightarrow \infty} \sup_{t \in [0, T]} \frac{1}{h_i} \left\| \int_0^t (\alpha_i(\tau) - h_i u(\tau)) d\tau \right\| = 0.$$

Because $\|x'_i\|_{L^1} \rightarrow 0$ there exist $\varepsilon_i \rightarrow 0+$ such that for

$$A_i = \{t: \|x'_i(t)\| \leq \varepsilon_i\}.$$

we have

$$(6.19) \quad \lim_{i \rightarrow \infty} \mu(A_i) = T.$$

Fix i and let $q = q(i)$ be an integer satisfying

$$(6.20) \quad q \geq \frac{2T}{h_i^2} \sup_{e \in U} \|e\| + 1.$$

Denote by I_j the interval

$$\left[\frac{j-1}{q} T, \frac{j}{q} T \right], \quad j = 1, \dots, q.$$

By Lemma 4.10 and (6.10) there exist a constant \bar{C} independent of h_i and functions $\alpha^j(t) \in U$ such that for all j

$$(6.21) \quad \int_{I_j \cap A_i} \alpha^j(t) dt = \int_{I_j \cap A_i} h_i u(t) dt,$$

$$(6.22) \quad \int_{I_j \cap A_i} \|\alpha^j(t)\| dt \leq \bar{C} \left\| \int_{I_j \cap A_i} h_i u(t) dt \right\|.$$

Set

$$\alpha_i(t) = \begin{cases} \alpha^j(t) & \text{if } t \in I_j \cap A_i, \\ 0 & \text{if } t \in [0, T] \setminus A_i, \end{cases}$$

and let $\delta = \min \{\|e\|: e \in U \setminus \{0\}\} > 0$. By (6.22)

$$\begin{aligned} \delta \mu\{t: \alpha_i(t) \neq 0\} &\leq \int_0^T \|\alpha_i(t)\| dt = \sum_{j=1}^q \int_{I_j \cap A_i} \|\alpha^j(t)\| dt \\ &\leq \bar{C} \sum_{j=1}^q h_i \int_{I_j} \|u(t)\| dt = \bar{C} h_i \|u\|_{L^1}. \end{aligned}$$

Thus for a constant C independent of h_i , inequality (6.17) holds true. By (6.21) for all $t \in [0, T]$ and some $j = j(t)$

$$\begin{aligned} \left\| \int_0^t \alpha_i(\tau) d\tau - \int_0^t h_i u(\tau) d\tau \right\| &\leq \int_{I_j} (\|\alpha_i(\tau)\| + h_i \|u(\tau)\|) d\tau + h_i \int_{[0,t] \setminus A_i} \|u(\tau)\| d\tau \\ &\leq 2 \sup_{e \in U} \|e\| T/q + h_i \int_{[0,t] \setminus A_i} \|u(\tau)\| d\tau \end{aligned}$$

and (6.19), (6.20) imply (6.18).

Step 2. We define here the required functions y_i . Let L be the Lipschitz constant of F . Since $\alpha_i(t) \in U \subset F(\xi)$ for all large i there exist $\gamma_i \in L^1(0, T)$ satisfying

$$(6.23) \quad \gamma_i(t) \in F(x_i(t)),$$

$$(6.24) \quad \|\gamma_i(t) - \alpha_i(t)\| \leq L \|x_i(t) - \xi\|.$$

For all $t \in [0, T]$ set

$$(6.25) \quad y_i'(t) = \begin{cases} \gamma_i(t) & \text{if } \alpha_i(t) \neq 0, \\ x_i'(t) + h_i v_i(t) & \text{otherwise,} \end{cases}$$

and

$$y_i(t) = \xi + \int_0^t y_i'(\tau) d\tau.$$

By (6.17)

$$\begin{aligned} \|y_i' - x_i'\|_{L^1} &\leq h_i \|v_i\|_{L^1} + \int_{\{t: \alpha_i(t) \neq 0\}} \|\gamma_i(t) - x_i'(t)\| dt \\ &\leq h_i (\|v_i\|_{L^1} + 2C \sup \{\|e\|: e \in F(x_i(t))\}). \end{aligned}$$

Therefore $\{y_i\}$ satisfy (6.13), (6.14). Furthermore for all i, t

$$\begin{aligned} y_i(t) - x_i(t) &= \int_0^t (y_i' - x_i')(\tau) d\tau \\ &= \int_0^t h_i v_i(\tau) d\tau + \int_{[0,t] \cap \{\tau: \alpha_i(\tau) \neq 0\}} (\gamma_i - x_i' - h_i v_i)(\tau) d\tau \\ &= \int_0^t (h_i v + \alpha_i)(\tau) d\tau + h_i \int_0^t (v_i - v)(\tau) d\tau \\ &\quad + \int_{[0,t] \cap \{\tau: \alpha_i(\tau) \neq 0\}} [(\gamma_i - \alpha_i)(\tau) - (x_i' + h_i v_i)(\tau)] d\tau. \end{aligned}$$

Note that $\|x_i'(\tau)\| \leq x/\varepsilon_i$ a.e. on $\{\tau: \alpha_i(\tau) \neq 0\}$. Thus using (6.24), (6.12) and (6.17) we obtain

$$y_i(t) - x_i(t) = \int_0^t (h_i v + \alpha_i)(\tau) d\tau + o(t, h_i)$$

where $o(t, h_i)/h_i \rightarrow 0$ uniformly on $[0, T]$. Thus relations (6.18) and (6.9) imply (6.15).

By the Lipschitzianity of F for all large i

$$(6.26) \quad \text{dist}(y_i'(t), F(y_i(t))) \leq \text{dist}(y_i'(t), F(x_i(t) + h_i w(t))) + L \|y_i(t) - x_i(t) - h_i w(t)\|.$$

By (6.11), (6.25), (6.23) and Lipschitzian nature of F

$$\begin{aligned} & \int_0^T \text{dist}(y'_i(t), F(x_i(t) + h_i w(t))) dt \\ &= \int_{\{t: \alpha_i(t) \neq 0\}} \text{dist}(\gamma_i(t), F(x_i(t) + h_i w(t))) dt \\ &\leq Lh_i \int_{\{t: \alpha_i(t) \neq 0\}} \|w(t)\| dt. \end{aligned}$$

By (6.26), (6.17), (6.15) and the last inequality we obtain (6.16) and end the proof of the lemma. \square

Note added in proof. During the revision of this article, I became aware of a recent work by Soviet mathematicians Polovinkin and Smirnov who (in a book in preparation) use a similar approach to the variational equations of Theorem 2.7.

Acknowledgments. The author wishes to thank St. Łojasiewicz, Jr. for useful remarks and wishes to acknowledge the comments of the unknown referee which helped to correct and improve the presentation of this paper.

REFERENCES

- [1] H. A. ANTOSIEWICZ AND A. CELLINA (1975), *Continuous selections and differential relations*, J. Differential Equations, 19, pp. 386–398.
- [2] J. P. AUBIN AND A. CELLINA (1984), *Differential Inclusions*, Springer-Verlag, New York.
- [3] J. P. AUBIN AND F. H. CLARKE (1977), *Monotone invariant solutions to differential inclusions*, J. London Math. Soc., 16, pp. 357–366.
- [4] J. P. AUBIN AND I. EKELAND (1984), *Applied Nonlinear Analysis*, Wiley-Interscience, New York.
- [5] J. P. AUBIN, H. FRANKOWSKA AND C. OLECH (1986), *Controllability of convex processes*, this Journal, 24, pp. 1192–1211.
- [6] A. BRESSAN (1980), *On differential relations with lower semicontinuous right-hand side*, J. Differential Equations, 37, pp. 89–97.
- [7] T. BUSHAW (1963), *Dynamical systems and optimization*, Contrib. Differential Equations, (2), pp. 351–365.
- [8] C. CASTAING AND M. VALADIER (1977), *Convex Analysis and Measurable Multifunctions*, Lecture Notes in Mathematics 580, Springer-Verlag, Berlin.
- [9] F. H. CLARKE (1975), *Generalized gradients and applications*, Trans. Amer. Math. Soc., 205, pp. 247–262.
- [10] ——— (1983), *Optimization and Nonsmooth Analysis*, Wiley-Interscience, New York.
- [11] I. EKELAND (1974), *On the variational principle*, J. Math. Anal. Appl., 47, pp. 324–353.
- [12] A. F. FILIPPOV (1967), *Classical solutions of differential equations with multi-valued right-hand side*, this Journal, 5, pp. 609–621.
- [13] H. FRANKOWSKA, *The maximum principle for an optimal solution to a differential inclusion with end point constraints*, this Journal, to appear.
- [14] H. FRANKOWSKA AND C. OLECH (1981), *R-convexity of integral of set-valued function*, Contributions to Analysis and Geometry, John Hopkins Univ. Press, Baltimore, MD, pp. 117–129.
- [15] ——— (1982), *Boundary solutions to differential inclusions*, J. Differential Equations, 44, pp. 156–165.
- [16] G. HADDAD (1981), *Monotone trajectories of differential inclusions and functional differential inclusions with memory*, Israel J. Math., 4, pp. 149–169.
- [17] H. HERMES (1970), *The generalized differential equation $x' \in R(t, x)$* , Adv. Math., 4, pp. 149–169.
- [18] P. KLOEDEN (1975), *Asymptotic invariance and limit sets of general control systems*, J. Differential Equations, 19, pp. 91–105.
- [19] ——— (1979), *The fannell boundary of multivalued dynamical systems*, J. Austr. Math. Soc., A 27, pp. 108–124.
- [20] A. LE DONNE AND M. V. MARCHI (1980), *Representation of Lipschitzian compact convex valued mappings*, Rend. Acc. Naz. Lincei, 68, pp. 277–280.

- [21] C. OLECH (1975), *Existence of solutions of nonconvex orientor fields*, Boll. Un. Math. It. (4) 11, pp. 189–197.
- [22] ——— (1976), *Existence Theory in Optimal Control*, in Control Theory and Topics in Functional Analysis, I, Intern. AT. Energy Agency, Vienna, 1976, pp. 291–328.
- [23] S. ROBINSON (1972), *Normed convex processes*, Trans. Amer. Math. Soc., 174, pp. 127–140.
- [24] R. T. ROCKAFELLAR (1980), *Generalized directional derivatives and subgradients of nonconvex functions*, Canad. J. Math., 32, pp. 257–280.
- [25] E. ROXIN (1965), *Stability in general control systems*, J. Differential Equations, 1, pp. 115–150.
- [26] ——— (1966), *On stability of control systems*, this Journal, 3, pp. 357–372.
- [27] S. SAPERSTONE AND J. YORKE (1971), *Controllability of linear oscillatory systems using positive controls*, this Journal, pp. 253–262.
- [28] H. SUSSMAN (1978), *A sufficient condition for local controllability*, this Journal, 16, pp. 790–802.
- [29] C. URSESCU (1975), *Multifunctions with closed convex graph*, Czech. Math. J., 25, pp. 438–441.
- [30] D. M. WAGNER (1977), *Survey of measurable selection theorems*, this Journal, 15, pp. 859–903.
- [31] T. WAŻEWSKI (1964), *On an Optimal Control Problem*, Proc. Conference on Differential Equations and their Applications, Prague, pp. 229–242.
- [32] J. A. YORKE (1972), *The maximum principle and controllability of nonlinear equations*, this Journal, 10, pp. 334–338.

THE STRUCTURE OF TIME-OPTIMAL TRAJECTORIES FOR SINGLE-INPUT SYSTEMS IN THE PLANE: THE C^∞ NONSINGULAR CASE*

H. J. SUSSMANN†

Abstract. For single-input C^∞ systems in the plane, in which the control enters linearly, we prove, if the system is suitably nondegenerate, that the time-optimal trajectories are finite concatenations of “bang-bang” and singular arcs, with local bounds on the number of switchings.

Key words. time-optimal control, two-dimensional systems, regular synthesis

AMS(MOS) subject classifications. 93C10, 93B15, 93B20

1. Introduction. The purpose of this paper is to prove local bounds on the number of switchings for arbitrary optimal control problems in the plane, of the form

$$(1.1) \quad \dot{x} = F(x) + uG(x), \quad |u| \leq 1,$$

$$(1.2) \quad \text{minimize } \int L(x(t)) dt,$$

where the vector fields F , G and the scalar function L are of class C^∞ , $L(x) > 0$ for all x , and some additional nondegeneracy conditions are satisfied.

In a subsequent paper [Su 7], these results will be applied to obtain a complete description of the structure of optimal trajectories for real analytic problems *without any extra requirements*. Finally, in the third article of this series, the results on optimal trajectories will be used to prove existence of a regular synthesis for real analytic problems (cf. [Su 8]).

Our work is part of a general program of research, initiated by Boltyanskii in his paper [Bo], and later pursued by Brunovsky (cf. [Br 1], [Br 2]), this author [Su 1]–[Su 6], and Baytman [Ba].

The goal of this program is to prove the existence of a regular synthesis for large classes of optimal control problems. A precise definition of “regular synthesis” will be given in [Su 8], when we state and prove our existence theorem. But the intuitive idea is quite simple: a “regular synthesis” for an optimal control problem with target point \bar{p} is, roughly, a feedback control law that steers to \bar{p} every point from which \bar{p} can be reached, that satisfies certain “piecewise smoothness” conditions, and that is “extremal,” in the sense that every trajectory satisfies the Maximum Principle.

Boltyanskii proved in [Bo] that a “regular synthesis” in his sense is necessarily optimal, thereby generalizing the classical result that, when the whole space is smoothly covered by trajectories that satisfy the necessary conditions for optimality, then these trajectories are optimal. This leads naturally to the question: when does a regular synthesis exist? or, equivalently: when is there an optimal feedback which is “piecewise smooth” in some appropriate sense? In [Br 1], Brunovsky proved an existence theorem for a very special class of systems (time-optimal problems for linear systems with a normality condition and polyhedral control constraints), but his work introduced a fundamental idea: the use of the theory of subanalytic sets. It has now become clear that, modulo minor technical restrictions, the use of subanalytic sets makes it possible to reduce the problem of proving the existence of a regular synthesis to the problem

* Received by the editors March 12, 1984; accepted for publication (in revised form) March 6, 1986. This work was partially supported by National Science Foundation grant DMS83-01678-01.

† Mathematics Department, Rutgers University, New Brunswick, New Jersey 08903.

of *trajectory analysis*. Roughly speaking, a *finite-dimensional reduction* (FDR) of an optimal control problem is a family \mathcal{F} of trajectories which is parametrized by a finite-dimensional parameter, and which is *sufficient* (i.e. every optimal trajectory is in \mathcal{F}) or at least *weakly sufficient* (i.e. whenever there is an optimal trajectory from p to q , then there is an optimal trajectory $\gamma \in \mathcal{F}$ that goes from p and q). It turns out that, whenever an FDR exists and some other technical hypotheses hold, the existence of a regular synthesis follows. (This will be shown in full generality in [Su 6]. However, for the case that concerns us here, of problems of the form (1.1), (1.2) in the plane, a self-contained presentation will be given in [Su 8].)

The main problem then becomes that of carrying out the trajectory analysis for particular classes of systems, so as to prove the existence of an FDR. This had been done, so far, for several types of problems, all of whom shared the feature of not giving rise to singular controls. (Examples:

(a) various types of problems where all optimal controls are necessarily bang-bang, as in Brunovsky [Br 1]–[Br 3], Sussmann [Su 2], [Su 4],

(b) classical calculus of variations problems with a nondegenerate Lagrangian, as outlined in [Su 2],

(c) linear problems with quadratic cost, as in Brunovsky [Br 3], Sussmann [Su 5].)

The present work constitutes the first step in the direction of extending the theory to the case where singular controls appear. We prove that, for systems (1.1), (1.2), with the properties stated above, the required FDR's are possible. With the exception of certain "degenerate cases," we show that every point p in the state space has a neighborhood U such that every optimal trajectory in U is a finite concatenation of bang-bang and singular pieces, with a fixed finite bound on the number of pieces. This shows that, at least locally, there is a finite-dimensional sufficient family. In one of the "degenerate cases" we prove a weaker result, namely, that every p has a neighborhood U with the property that, whenever $q_1 \in U$ can be reached from $q_0 \in U$ by an optimal trajectory in U , then q_1 can be optimally reached in U from q_0 by a bang-bang trajectory, with a fixed bound on the number of switchings. In the other "degenerate cases" no such conclusion is true but, fortunately, when the system is real analytic, it is still possible to prove the existence of a regular synthesis. So we end up proving the existence of a regular synthesis for *all* systems (1.1), (1.2), with F , G , L analytic and $L > 0$, with no extra hypotheses whatsoever (except for the technical requirement that there be no trajectory that goes to ∞ with finite cost).

This paper, and its sequel [Su 7], deal exclusively with the time-optimal case. Only in [Su 8], where the existence of a regular synthesis is proved, do we return to the general situation, which turns out to be easily reduced to the time-optimal one (because we can reparametrize the trajectories using cost, rather than time, as the parameter, so that (1.1), (1.2) becomes a time-minimization problem with a modified F , G).

The paper is organized as follows: In § 2 we introduce some basic definitions and notations. In § 3 we review the Maximum Principle, we study properties of conjugate points, and we study the optimal trajectories near "ordinary points" p (i.e., points where F and G are independent, and where G and $[F, G]$ are independent). In § 4 and § 5 we prove some lemmas, and in § 6 we study the optimal trajectories near certain nonordinary points (called "good nonordinary points"), with the exception of one type of such points (called "antiturnpike points") which are studied in § 7. For suitably nondegenerate smooth problems, our results give a complete description of the optimal trajectories, except in the neighborhood of certain "branch points." In [Su 7] we will show that in the real analytic case the branch points can be handled as well.

The main results of this paper are Theorems 3.8, 3.9, 3.13, 6.1, 6.2, 6.3 and 6.4, which give a complete description of the optimal trajectories in regions where the system is sufficiently nondegenerate.

Our work partially overlaps with Baytman's book [Ba]. Baytman proved existence of a regular synthesis for a class of smooth systems, without using the theory of subanalytic sets. However, he has to make various nondegeneracy assumptions which, in particular, fail to be satisfied for arbitrary analytic systems.

2. Basic definitions. Throughout this paper, we use \mathbb{R} to denote the real line and \mathbb{R}^n to denote n -dimensional Euclidean space. Points in \mathbb{R}^n are always thought of as *column vectors*.

We use $]a, b[$, $[a, b]$ to denote, respectively, the open and the closed intervals with end points a, b . The self-explanatory notations $]a, b[$, $[a, b]$ are used for half-open, half-closed intervals.

A *curve* in \mathbb{R}^n is a continuous map $\gamma: I \rightarrow \mathbb{R}^n$, where I is some real interval. We use the symbol "Dom" for "domain" so that, for instance, if $\gamma: I \rightarrow \mathbb{R}^n$ is a curve, then $\text{Dom}(\gamma) = I$. The symbol " \upharpoonright " denotes "restriction." (Examples: $\gamma \upharpoonright J$ is the restriction of the curve γ to the subinterval J of $\text{Dom}(\gamma)$; if X is a vector field on an open set $U \subseteq \mathbb{R}^n$, and if $V \subseteq U$, V open, then $X \upharpoonright V$ is the restriction of X to V .)

In this paper, M always denotes an open subset of \mathbb{R}^2 . (More generally, M could also be taken to be an arbitrary two-dimensional smooth manifold.)

A C^∞ *chart* in M is a pair $(U, (\xi, \eta))$, where U is an open subset of M , and ξ, η are C^∞ real-valued functions on U that satisfy

(2.1.i) the mapping $Q: p \rightarrow (\xi(p), \eta(p))$ is one-to-one on U , and

(2.1.ii) the Jacobian matrix of the map Q is nonsingular at every point $p \in U$.

When $(U, (\xi, \eta))$ is a chart, then Q establishes a bijection between U and some open set $\tilde{U} \subseteq \mathbb{R}^2$. By means of this bijection, we can identify each point $p \in U$ with the pair (x, y) of its coordinates relative to the chart. We will often do so. The chart $(U, (\xi, \eta))$ will be called *analytic* if the functions (ξ, η) are real analytic.

If $\varepsilon > 0$, $\delta > 0$, we write

$$(2.1.a) \quad \mathcal{R}(\varepsilon, \delta) = \{(x, y): |x| < \varepsilon, |y| < \delta\}.$$

A chart $(U, (\xi, \eta))$ will be called *rectangular* if $\tilde{U} = \mathcal{R}(\varepsilon, \delta)$ for some ε, δ . If, in addition, $\varepsilon = \delta$, then $(U, (\xi, \eta))$ is a *square chart*. We say that $(U, (\xi, \eta))$ is *centered at* p if $\xi(p) = \eta(p) = 0$. We write

$$(2.1.b) \quad Sq(\varepsilon) = \mathcal{R}(\varepsilon, \varepsilon).$$

A C^∞ *vector field* on an open subset U of \mathbb{R}^2 is a C^∞ \mathbb{R}^2 -valued function on U . We use $\mathcal{V}(U)$ to denote the set of all C^∞ vector fields on U . Also, we use $C^\infty(U)$ to denote the class of all C^∞ real-valued functions on U . It is well known that each $X \in \mathcal{V}(U)$ gives rise to a *first order differential operator from $C^\infty(U)$ to $C^\infty(U)$* . As is customary, we will also use X to denote this operator so that, if $\alpha \in C^\infty(U)$, then $X\alpha$ is also a function, defined by

$$(2.2) \quad (X\alpha)(p) = \lim_{\varepsilon \rightarrow 0} \frac{\alpha(p + \varepsilon X(p)) - \alpha(p)}{\varepsilon}.$$

If $(U, (\xi, \eta))$ is a chart, and we identify U with \tilde{U} as above, then each vector field $X \in \mathcal{V}(U)$ can be written in a unique way as a linear combination

$$(2.3) \quad X = \alpha \partial_x + \beta \partial_y,$$

where ∂_x, ∂_y are the vector fields with components $(1, 0), (0, 1)$, respectively (i.e., the differential operators usually denoted by $\partial/\partial x, \partial/\partial y$), and α, β are C^∞ functions of the variables x, y , for $(x, y) \in \tilde{U}$. The functions α, β are the *components of X relative to the chart $(U, (\xi, \eta))$* . We always think of X as a column-valued function with components α, β . We then use DX to denote the matrix

$$(2.4) \quad DX = \begin{pmatrix} \partial_x \alpha & \partial_y \alpha \\ \partial_x \beta & \partial_y \beta \end{pmatrix}.$$

The *Lie-bracket* of two vector fields X, Y is the vector field $[X, Y]$ that corresponds to the differential operator

$$(2.5) \quad [X, Y] = XY - YX.$$

Equivalently, $[X, Y]$ can be computed relative to a coordinate chart by

$$(2.6) \quad [X, Y] = (DY) \cdot X - (DX) \cdot Y.$$

Throughout the paper, F and G are two fixed vector fields in $\mathcal{V}(M)$. We let Σ denote the control system

$$(2.7) \quad \dot{p} = F(p) + uG(p), \quad |u| \leq 1, \quad p \in M.$$

A *control* is a measurable function $u(\cdot): [a, b] \rightarrow [-1, 1]$, where $[a, b]$ is some bounded closed interval. We use \mathcal{U} to denote the class of all controls. Naturally, if $u(\cdot): [a, b] \rightarrow [-1, 1]$ is a control, then $\text{Dom}(u(\cdot)) = [a, b]$. A *trajectory of Σ* for a control $u(\cdot)$ is an absolutely continuous curve $\gamma: \text{Dom}(u(\cdot)) \rightarrow M$ which satisfies the equation

$$(2.8) \quad \dot{\gamma}(t) = F(\gamma(t)) + u(t)G(\gamma(t))$$

for almost every $t \in \text{Dom}(u(\cdot))$. The set of all trajectories of Σ will be denoted by $\text{Traj}(\Sigma)$.

Notice that, by definition, the domain $\text{Dom}(\gamma)$ of a $\gamma \in \text{Traj}(\Sigma)$ is necessarily a compact interval $[a, b]$. We use $\text{In}(\gamma)$ to denote $\gamma(a)$, i.e. the *initial point* of γ , and $\text{Term}(\gamma)$ to denote the *terminal point* $\gamma(b)$ of γ . Also, $T(\gamma)$ denotes $b - a$, i.e., the *time along γ* .

If $\gamma \in \text{Traj}(\Sigma)$, we say that γ is *time-optimal* if $T(\gamma) \leq T(\gamma')$ for every $\gamma' \in \text{Traj}(\Sigma)$ such that $\text{In}(\gamma) = \text{In}(\gamma')$, $\text{Term}(\gamma) = \text{Term}(\gamma')$. We use $\text{Opt}^1(\Sigma)$ to denote the set of all time-optimal $\gamma \in \text{Traj}(\Sigma)$. (The superscript 1 refers to the fact that time-optimal trajectories minimize the integral $\int L dt$, where the Lagrangian L is equal to 1. More generally, we will use $\text{Opt}^L(\Sigma)$ to denote the trajectories of Σ which are optimal for the cost functional $\int L$.)

If M' is an open subset of M , then the system Σ can be restricted to M' . This restriction is denoted by $\Sigma \upharpoonright M'$. Then $\text{Traj}(\Sigma \upharpoonright M')$ is the set of all trajectories of the restriction, i.e. the set of all $\gamma \in \text{Traj}(\Sigma)$ which are entirely contained in M' .

If $u_1(\cdot): [a, b] \rightarrow [-1, 1]$ and $u_2: [b, c] \rightarrow [-1, 1]$ are controls, we use $u_2 * u_1$ to denote the control defined on $[a, c]$ by

$$(2.9) \quad (u_2 * u_1)(t) = \begin{cases} u_1(t) & \text{for } t \in \text{Dom}(u_1(\cdot)), \\ u_2(t) & \text{for } t \in \text{Dom}(u_2(\cdot)). \end{cases}$$

(This control is called the *concatenation* of u_1 and u_2 .)

If $\gamma_1: [a, b] \rightarrow M$, $\gamma_2: [b, c] \rightarrow M$ are trajectories for $u_1(\cdot)$, $u_2(\cdot)$ such that $\gamma_1(b) = \gamma_2(b)$, then $\gamma_2 * \gamma_1$ is the trajectory

$$(2.10) \quad (\gamma_2 * \gamma_1)(t) = \begin{cases} \gamma_1(t) & \text{for } t \in \text{Dom}(\gamma_1), \\ \gamma_2(t) & \text{for } t \in \text{Dom}(\gamma_2). \end{cases}$$

Notice that $u_2 * u_1$ is only defined if $\max(\text{Dom}(u_1(\cdot))) = \min(\text{Dom}(u_2(\cdot)))$, and that, for $\gamma_2 * \gamma_1$ to be defined, it is necessary that, in addition, $\text{Term}(\gamma_1) = \text{In}(\gamma_2)$. Notice also the order: $\gamma_2 * \gamma_1$ is the trajectory obtained by following first γ_1 , and then γ_2 .

We write

$$(2.11.a) \quad X = F - G,$$

$$(2.11.b) \quad Y = F + G,$$

so that

$$(2.12.a) \quad F = \frac{1}{2}(Y + X),$$

$$(2.12.b) \quad G = \frac{1}{2}(Y - X).$$

If V is an arbitrary vector field, we use $t \rightarrow \Phi^V(t, p)$ to denote the integral curve of V which goes through p at time $t = 0$. We also write

$$(2.13) \quad \Phi_t^V(p) = \Phi^V(t, p).$$

The family of maps Φ_t^V will be referred to as the *flow* of V .

We use $\text{Traj}(X)$ to denote the set of all trajectories of Σ which correspond to the constant control $u(\cdot)$ whose value is equal to -1 . Therefore $\gamma \in \text{Traj}(X)$ iff γ is a trajectory of Σ which is an integral curve of X . (Notice that γ cannot possibly be a *maximal* integral curve of X , because $\text{Dom}(\gamma)$ must be a compact interval.) Equivalently, $\gamma \in \text{Traj}(X)$ iff $\gamma = \Phi^X(\cdot, p)|J$ for some $p \in M$ and some compact interval J . We define $\text{Traj}(Y)$ in a similar way, using the control $u = 1$ rather than $u = -1$. Elements of $\text{Traj}(X)$, $\text{Traj}(Y)$ will be called *X-trajectories* and *Y-trajectories*, respectively.

We need symbols to refer to more complicated trajectory types. If $\mathcal{C}_1, \dots, \mathcal{C}_N$ are collections of trajectories, we write $\mathcal{C}_1 * \dots * \mathcal{C}_N$ to denote the set of all concatenations $\gamma = \gamma_1 * \dots * \gamma_N$, where, for each $i \in \{1, \dots, N\}$, either $\gamma_i \in \mathcal{C}_i$ or γ_i is trivial. (A trivial trajectory is one whose domain is a single point. Thus, for instance, a concatenation $\gamma_1 * \gamma_2 * \gamma_3$ with γ_2 trivial is the same as the concatenation $\gamma_1 * \gamma_3$.) That is, $\mathcal{C}_1 * \dots * \mathcal{C}_N$ consists of all concatenations $\gamma_1 * \dots * \gamma_N$, with $\gamma_i \in \mathcal{C}_i$, with any number of γ_i 's allowed to be absent. If $\mathcal{C}_1 = \dots = \mathcal{C}_N = \mathcal{C}$, we will simply write \mathcal{C}^N for $\mathcal{C}_1 * \dots * \mathcal{C}_N$, so that \mathcal{C}^N is the set of all concatenations of at most N trajectories in \mathcal{C} .

If ξ_1, \dots, ξ_N are symbols for trajectory types, then we will write $\text{Traj}(\xi_1 * \dots * \xi_N)$ for $\text{Traj}(\xi_1) * \dots * \text{Traj}(\xi_N)$. (The elements of $\text{Traj}(\xi_1 * \dots * \xi_N)$ are called $\xi_1 * \dots * \xi_N$ -trajectories.) If $\xi_1 = \dots = \xi_N = \xi$, then we write ξ^N for $\xi_1 * \dots * \xi_N$. (Notice that, with these conventions, $\text{Traj}(\xi^N)$ and $[\text{Traj}(\xi)]^N$ denote the same class of trajectories.) If $\xi = \xi_1 * \dots * \xi_N$, then a *strict ξ -trajectory* is a ξ -trajectory that is not an η -trajectory for any symbol η obtained from $\xi_1 * \dots * \xi_N$ by deleting one or more ξ_i 's. We use $\text{Traj}_s(\xi)$ to denote the set of all strict ξ -trajectories. (Example: $\text{Traj}(\xi^4)$ is the disjoint union of the sets $\text{Traj}_s(\xi^i)$, $i = 0, 1, 2, 3, 4$.)

If ξ_1, \dots, ξ_N are trajectory type symbols, the symbol $\xi_1 \vee \dots \vee \xi_N$ will stand for a trajectory that is of one of the types ξ_1, \dots, ξ_N . That is

$$\text{Traj}(\xi_1 \vee \dots \vee \xi_N) = \text{Traj}(\xi_1) \cup \dots \cup \text{Traj}(\xi_N).$$

The following examples illustrate our notational conventions:

(a) An $X * Y * X * Y$ -trajectory is a concatenation of at most four pieces, of which the first and third are Y -trajectories and the second and fourth are X -trajectories;

however, any one, or any two, or any three of these pieces may be missing, and all four of them may even be missing, so that a trivial trajectory is in $\text{Traj}(X * Y * X * Y)$;

(b) An $(X \vee Y)^4$ -trajectory is a bang-bang trajectory with at most three switchings;

(c) A strict $(X \vee Y)^4$ -trajectory is a bang-bang trajectory with exactly three switchings;

(d) a strict $Z * (X \vee Y) * Z$ -trajectory is a concatenation of exactly three pieces, the first and third of which are in $\text{Traj}(Z)$, while the middle one is either an X -trajectory or a Y -trajectory (Z -trajectories will be defined later);

(e) There do not exist strict $X * X$ -trajectories, because $\text{Traj}(X * X) = \text{Traj}(X)$.

Now let \mathcal{C} be any set of trajectories. We write

$$(2.14) \quad \mathcal{C}^\infty = \bigcup_{N=1}^{\infty} \mathcal{C}^N.$$

A *bang-bang trajectory* is a trajectory which corresponds to a control $u(\cdot)$ such that $|u(t)| = 1$ for all t . A *regular bang-bang trajectory* is one that is a finite concatenation of X - and Y -trajectories. Therefore the set of regular bang-bang trajectories is precisely $[\text{Traj}(X \vee Y)]^\infty$.

If $\mathcal{C} \subseteq \text{Traj}(\Sigma)$, and U is a subset of M , we call \mathcal{C} *sufficient on U* if

$$(2.15) \quad \text{Opt}^1(\Sigma \upharpoonright U) \subseteq \mathcal{C}^\infty,$$

i.e., if every time-optimal trajectory in U is a finite concatenation of pieces that belong to \mathcal{C} . We call \mathcal{C} *boundedly sufficient on U* if there is an integer $N > 0$ such that

$$(2.16) \quad \text{Opt}^1(\Sigma \upharpoonright U) \subseteq \mathcal{C}^N,$$

i.e., if every time-optimal trajectory γ in U is a concatenation of a finite number $\nu(\gamma)$ of pieces which belong to \mathcal{C} , and if there is a finite upper bound, independent of γ , for the numbers $\nu(\gamma)$.

If $T > 0$, $\mathcal{C} \subseteq \text{Traj}(\Sigma)$, we use $\mathcal{C}[\leq T]$ to denote the set of all $\gamma \in \mathcal{C}$ such that $T(\gamma) \leq T$. We say that \mathcal{C} is *finite-time boundedly sufficient on U* if for every $T > 0$ there is an integer $N = N(T)$, $N > 0$, such that

$$(2.17) \quad \text{Opt}^1(\Sigma \upharpoonright U)[\leq T] \subseteq \mathcal{C}^N$$

(i.e., the integer $\nu(\gamma)$ considered above is allowed to depend on γ , but only through $T(\gamma)$).

Also, if \mathcal{C} , U are as before, we will call \mathcal{C} *weakly sufficient on U* if, whenever $q_0 \in U$, $q_1 \in U$ and $\gamma \in \text{Opt}^1(\Sigma \upharpoonright U)$ steers q_0 to q_1 , then there exists a $\gamma' \in \mathcal{C}$ that is also time-optimal, steers q_0 to q_1 , and is contained in U . We call \mathcal{C} *weakly boundedly sufficient on U* if there is an N such that \mathcal{C}^N is weakly sufficient on U .

3. The maximum principle, conjugate points, ordinary points. An *admissible pair* for the system Σ is a pair $(u(\cdot), \gamma)$, such that $u(\cdot) \in \mathcal{U}$ and γ is a trajectory corresponding to $u(\cdot)$. We use $\text{Adm}(\Sigma)$ to denote the set of admissible pairs.

Suppose $(u(\cdot), \gamma) \in \text{Adm}(\Sigma)$. A *variational vector field along $(u(\cdot), \gamma)$* is a vector-valued absolutely continuous function $v(\cdot): \text{Dom}(\gamma) \rightarrow \mathbb{R}^2$ that satisfies the equation

$$(3.1) \quad \dot{v}(t) = ((DF)(\gamma(t)) + u(t)(DG)(\gamma(t)))v(t).$$

Equation (3.1) is called the *variational equation along $(u(\cdot), \gamma)$* . It is a (time-varying) linear homogeneous system of two ordinary differential equations for the two components of the vector-valued function $v(\cdot)$, with coefficients that are continuous functions of t . Therefore (3.1) has one and only one solution $v(\cdot)$ for any initial

condition $v(t_0) = v_0$ ($t_0 \in \text{Dom}(\gamma)$, $v_0 \in \mathbb{R}^2$). We let $\text{VVF}(u(\cdot), \gamma)$ denote the set of all variational vector fields along $(u(\cdot), \gamma)$. So $\text{VVF}(u(\cdot), \gamma)$ is a two-dimensional linear space over \mathbb{R} .

The variational equation is stated in terms of a particular choice of coordinates. The following lemma shows that $\text{VVF}(u(\cdot), \gamma)$ is really a “coordinate free object” (i.e., that we can think of a variational vector field along γ as a map $v(\cdot)$ such that $v(t)$ is a tangent vector to M at $\gamma(t)$ for each t).

LEMMA 3.1. *Let $(u(\cdot), \gamma) \in \text{Adm}(\Sigma)$, and let $v(\cdot): \text{Dom}(\gamma) \rightarrow \mathbb{R}^2$ be absolutely continuous. Then $v(\cdot)$ is a variational vector field along $(u(\cdot), \gamma)$ if and only if there is an $\varepsilon > 0$, and a map*

$$\Gamma: [0, \varepsilon] \times \text{Dom}(\gamma) \rightarrow M$$

such that:

- (i) $s \rightarrow \Gamma(s, t)$ is of class C^1 for each fixed $t \in \text{Dom}(\gamma)$;
- (ii) $t \rightarrow \Gamma(s, t)$ is a trajectory of Σ corresponding to the control $u(\cdot)$ for each fixed $s \in [0, \varepsilon]$;

(iii) $\Gamma(0, t) = \gamma(t)$ for $t \in \text{Dom}(\gamma)$;

(iv) $v(t) = (d/ds)|_{s=0} \Gamma(s, t)$ for $t \in \text{Dom}(\gamma)$.

Proof. If Γ, ε are such that (i), \dots , (iv) hold, then

$$(3.2) \quad \frac{\partial}{\partial t} \Gamma(s, t) = F(\Gamma(s, t)) + u(t)G(\Gamma(s, t))$$

holds. If we differentiate with respect to s , set $s=0$, and formally interchange the t and s differentiations in the left side of the resulting equation, we get (3.1). The rigorous justification of the interchange of differentiation is carried out by a well-known argument, writing (3.2) as an integral equation, and differentiating under the integral sign. We omit the details.

Conversely, if $v(\cdot) \in \text{VVF}(u(\cdot), \gamma)$, pick an arbitrary C^1 curve $\delta: [0, \varepsilon] \rightarrow M$ such that $\dot{\delta}(a) = v(a)$, where we let $[a, b] = \text{Dom}(\gamma)$. If ε is sufficiently small, it follows from the existence, uniqueness and continuous dependence theorems of O.D.E. theory that there is a Γ for which (i), (ii) and (iii) hold and for which, in addition, $\Gamma(s, 0) = \delta(s)$ for $0 \leq s \leq \varepsilon$. Then, if we let

$$w(t) = \frac{d}{ds} \bigg|_{s=0} \Gamma(s, t),$$

it follows from the part already proved that $w(\cdot) \in \text{VVF}(u(\cdot), \gamma)$. Clearly, $w(a) = v(a)$. Since $v(\cdot) \in \text{VVF}(u(\cdot), \gamma)$ it follows from the uniqueness of solutions of (3.1) that $v(\cdot) = w(\cdot)$. So (iv) holds as well. \square

A *variational covector field* along a $(u(\cdot), \gamma) \in \text{Adm}(\Sigma)$ is an absolutely continuous $\lambda(\cdot): \text{Dom}(\gamma) \rightarrow \mathbb{R}_*^2$ that satisfies

$$(3.3) \quad \dot{\lambda}(t) = -\lambda(t)(DF(\gamma(t)) + u(t)DG(\gamma(t)))$$

for almost all $t \in \text{Dom}(\gamma)$. (Here \mathbb{R}_*^2 is the space of *row vectors*.) We let $\text{VVF}^*(u(\cdot), \gamma)$ denote the space of all variational covector fields along $(u(\cdot), \gamma)$. Equation (3.3), like (3.2), appears to depend on a particular choice of coordinates, but the following lemma shows that it is not so, and that the elements of $\text{VVF}^*(u(\cdot), \gamma)$ are covector-valued functions (i.e., $\lambda(t)$ is a cotangent vector at $\gamma(t)$ for each t). The proof is elementary, and we omit it.

LEMMA 3.2. Let $(u(\cdot), \gamma) \in \text{Adm}(\Sigma)$, and let $\lambda(\cdot): \text{Dom}(\gamma) \rightarrow \mathbb{R}_*^2$ be absolutely continuous. Then $\lambda(\cdot) \in \text{VVF}^*(u(\cdot), \gamma)$ if and only if the function $t \rightarrow \lambda(t) \cdot v(t)$ is constant for every $v(\cdot) \in \text{VVF}(u(\cdot), \gamma)$. \square

We define $\mathcal{H}: \mathbb{R}_*^2 \times M \times \mathbb{R} \rightarrow \mathbb{R}$ by

$$(3.4) \quad \mathcal{H}(\lambda, p, u) = \lambda \cdot (F(p) + uG(p)).$$

If $\lambda(\cdot) \in \text{VVF}^*(u(\cdot), \gamma)$, $(u(\cdot), \gamma) \in \text{Adm}(\Sigma)$, we say that $\lambda(\cdot)$ is *minimizing* if

$$(3.5) \quad \mathcal{H}(\lambda(t), \gamma(t), u(t)) = \min \{ \mathcal{H}(\lambda(t), \gamma(t), u) : -1 \leq u \leq 1 \}$$

for almost all $t \in \text{Dom}(\gamma)$.

The *Maximum Principle* says that, if $(u(\cdot), \gamma) \in \text{Adm}(\Sigma)$, and $\gamma \in \text{Opt}^1(\Sigma)$, then there exist: (a) a nontrivial, minimizing $\lambda(\cdot) \in \text{VVF}^*(u(\cdot), \gamma)$ and (b) a constant $\lambda_0 \geq 0$, such that

$$(3.6) \quad \mathcal{H}(\lambda(t), \gamma(t), u(t)) + \lambda_0 = 0$$

for almost all $t \in \text{Dom}(\gamma)$.

If $(u(\cdot), \gamma) \in \text{Adm}(\Sigma)$, and $\lambda(\cdot) \in \text{VVF}^*(u(\cdot), \gamma)$, $\lambda_0 \geq 0$, are such that $\lambda(\cdot)$ is nontrivial and minimizing, and (3.6) holds for almost all t , then the 4-tuple

$$(3.7) \quad \Lambda = (u(\cdot), \gamma, \lambda(\cdot), \lambda_0)$$

will be called an *extremal lift* of $(u(\cdot), \gamma)$. (So the Maximum Principle says that every time-optimal trajectory has an extremal lift.)

If Λ is an extremal lift of $(u(\cdot), \gamma)$, and (3.7) holds, we define the *switching function* ϕ_Λ along Λ to be the function $\phi_\Lambda: \text{Dom}(\gamma) \rightarrow \mathbb{R}$ given by

$$(3.8) \quad \phi_\Lambda(t) = \lambda(t) \cdot G(\gamma(t)).$$

Therefore, it is clear that the following holds.

LEMMA 3.3. Let $\Lambda = (u(\cdot), \gamma, \lambda(\cdot), \lambda_0)$ be an extremal lift of the admissible pair $(u(\cdot), \gamma)$. Then:

(a) The switching function ϕ_Λ is continuous on $\text{Dom}(\gamma)$;

(b) If $\phi_\Lambda(t) > 0$ for all t in some interval I , then $u(t) = -1$ for almost all $t \in I$, and therefore $\gamma \upharpoonright I \in \text{Traj}(X)$;

(c) If $\phi_\Lambda(t) < 0$ for $t \in I$, then $u(t) = 1$ for almost all $t \in I$, and therefore $\gamma \upharpoonright I \in \text{Traj}(Y)$.

A $t \in \text{Dom}(\gamma)$ will be said to be a *switching time* for γ if there does not exist an $\varepsilon > 0$ such that $\gamma \upharpoonright ([t - \varepsilon, t + \varepsilon] \cap \text{Dom}(\gamma)) \in \text{Traj}(X \vee Y)$. If t is a switching time for γ , and $p = \gamma(t)$, then we say that p is a *switching point* of γ , or that γ has a *switching* at p .

Lemma 3.3 implies the following.

LEMMA 3.4. If $(u(\cdot), \gamma) \in \text{Adm}(\Sigma)$, and $\Lambda = (u(\cdot), \gamma, \lambda(\cdot), \lambda_0)$ is an extremal lift of $(u(\cdot), \gamma)$, then $\phi_\Lambda(t) = 0$ for every switching time t of γ .

Suppose that $(u(\cdot), \gamma) \in \text{Adm}(\Sigma)$. Let t_0 belong to $\text{Dom}(\gamma)$, and let $v_0 \in \mathbb{R}^2$. Let us write $v(v_0, t_0, \cdot)$ to denote the unique variational vector field along $(u(\cdot), \gamma)$ whose value at t_0 is v_0 . If $t_0 \in \text{Dom}(\gamma)$ and $t_1 \in \text{Dom}(\gamma)$, we say that t_0 and t_1 are *conjugate along* $(u(\cdot), \gamma)$ if the vectors $v(G(\gamma(t_0)), t_0; t_1)$ and $G(\gamma(t_1))$ are linearly dependent. Notice that conjugacy is a symmetric relation. (If $v(G(\gamma(t_0)), t_0; t_1)$ and $G(\gamma(t_1))$ are linearly dependent, then $v(G(\gamma(t_0)), t_0; t)$ and $v(G(\gamma(t_1)), t_1; t)$ are dependent for all t , and in particular for $t = t_0$, so that $G(\gamma(t_0))$ and $v(G(\gamma(t_1)), t_1; t_0)$ are dependent.) However, it is possible for t_0 and t_2 to be conjugate to a time t_1 without being conjugate to each other. (This can happen if $G(\gamma(t_1)) = 0$.) If t_0 and t_1 are conjugate along $(u(\cdot), \gamma)$, we also say that the points $p_0 = \gamma(t_0)$, $p_1 = \gamma(t_1)$ are conjugate along $(u(\cdot), \gamma)$.

LEMMA 3.5. Let $(u(\cdot), \gamma) \in \text{Adm}(\Sigma)$, $\gamma \in \text{Opt}^1(\Sigma)$. Suppose that γ has switchings at p_0 and p_1 . Then p_0 and p_1 are conjugate along $(u(\cdot), \gamma)$.

Proof. Let $\Lambda = (u(\cdot), \gamma, \lambda(\cdot), \lambda_0)$ be an extremal lift of $(u(\cdot), \gamma)$. Let $\gamma(t_i) = p_i$, $i = 0, 1$. Since t_0 and t_1 are switching times for γ , we have

$$\phi_\Lambda(t_0) = \phi_\Lambda(t_1) = 0,$$

i.e.,

$$\lambda(t_0) \cdot G(\gamma(t_0)) = \lambda(t_1) \cdot G(\gamma(t_1)) = 0.$$

By Lemma 3.2, the function $t \rightarrow \lambda(t) \cdot v(G(\gamma(t_0)), t_0; t)$ is constant. So

$$\lambda(t_1) \cdot v(G(\gamma(t_0)), t_0; t_1) = \lambda(t_1) \cdot G(\gamma(t_1)) = 0.$$

Since $\lambda(t_1) \neq 0$, the vectors $v(G(\gamma(t_0)), t_0; t_1)$ and $G(\gamma(t_1))$ are linearly dependent. So t_0 and t_1 are conjugate. \square

The following results will be useful later.

LEMMA 3.6. Let U be the domain of a rectangular coordinate system (ξ, η) with respect to which X has components $(1, 0)$ (i.e. $X = \partial_x$). Let $\gamma_1: [a, b] \rightarrow U$ be a trajectory of Σ , of the form $t \rightarrow (x_1(t), y_1(t))$, such that the function $y_1(\cdot)$ is of class C^1 , and $\dot{y}_1(t) \neq 0$ for all $t \in \text{Dom}(\gamma_1)$. Suppose that $\tau: [a, b] \rightarrow]0, \infty[$ is a function of class C^1 such that, for each $t \in [a, b]$, the point $(x_1(t) + \tau(t), y_1(t))$ belongs to U and is conjugate to $(x_1(t), y_1(t))$ along the X -trajectory joining both points. Suppose that the curve $t \rightarrow (x_1(t) + \tau(t), y_1(t))$, $t \in \text{Dom}(\gamma_1)$ can be reparametrized by a parameter s such that $ds/dt > 0$, in such a way that the reparametrized curve is a trajectory γ_2 of Σ . Let δ_a, δ_b be the unique X -trajectories going from $(x_1(a), y_1(a))$ to $(x_1(a) + \tau(a), y_1(a))$ and from $(x_1(b), y_1(b))$ to $(x_1(b) + \tau(b), y_1(b))$, respectively. Then

$$(3.9) \quad T(\delta_b * \gamma_1) = T(\gamma_2 * \delta_a).$$

Proof. (cf. Fig. 1). Let Y have components (α, β) . Then

$$(3.10.a) \quad F = \frac{1}{2}(\alpha + 1)\partial_x + \frac{1}{2}\beta\partial_y$$

and

$$(3.10.b) \quad G = \frac{1}{2}(\alpha - 1)\partial_x + \frac{1}{2}\beta\partial_y.$$

Let $u(\cdot): [a, b] \rightarrow [-1, 1]$ be a control such that γ_1 is a trajectory for $u(\cdot)$. Then the functions $x_1(\cdot)$, $y_1(\cdot)$ satisfy the differential equations

$$(3.11.a) \quad \dot{x}_1(t) = \frac{1}{2}(\alpha(x_1(t), y_1(t)) + 1) + \frac{u(t)}{2}(\alpha(x_1(t), y_1(t)) - 1),$$

$$(3.11.b) \quad \dot{y}_1(t) = \frac{1 + u(t)}{2}\beta(x_1(t), y_1(t)).$$

Since $\dot{y}_1(t) \neq 0$ for all t , \dot{y}_1 is either > 0 for all t , or < 0 for all t . Let us assume that $\dot{y}_1(t) > 0$ for all $t \in [a, b]$. Then we can reparametrize γ_1 using the y coordinate as a parameter. Let $\tilde{a} = y_1(a)$, $\tilde{b} = y_1(b)$, let $y \rightarrow (\tilde{x}(y), y)$, $\tilde{a} \leq y \leq \tilde{b}$ be the reparametrized curve, and let $y \rightarrow \tilde{t}(y)$ be the time as a function of y . Then

$$\frac{d\tilde{x}}{dy} = \frac{\alpha}{\beta}(\tilde{x}(y), y) + \frac{1 - u(\tilde{t}(y))}{1 + u(\tilde{t}(y))} \cdot \frac{1}{\beta(\tilde{x}(y), y)}$$

and

$$\frac{d\tilde{t}}{dy} = \frac{2}{1 + u(\tilde{t}(y))} \cdot \frac{1}{\beta(\tilde{x}(y), y)}.$$

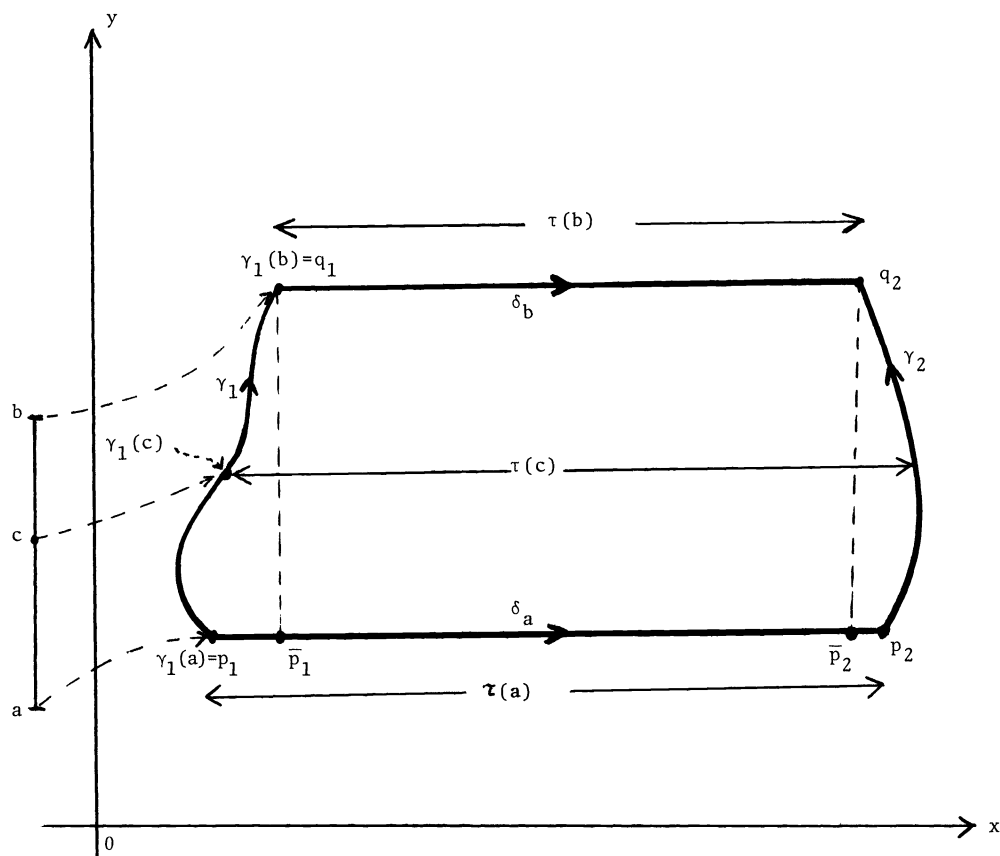


FIG. 1

Therefore

$$\frac{d}{dy}(\tilde{x}(y) - \tilde{t}(y)) = \frac{\alpha - 1}{\beta}(\tilde{x}(y), y).$$

So

$$(3.12.a) \quad x_1(b) - x_1(a) - T(\gamma_1) = \int_a^b \frac{\alpha - 1}{\beta}(\tilde{x}(y), y) dy.$$

Let $\gamma_2: [a', b'] \rightarrow U$ be of the form $t \rightarrow (x_2(t), y_2(t))$, and let $y \rightarrow (\tilde{x}^*(y), y)$ be γ_2 , reparametrized using y as the parameter. Then, a computation similar to the one leading to (3.12.a) gives

$$(3.12.b) \quad x_2(b') - x_2(a') - T(\gamma_2) = \int_a^b \frac{\alpha - 1}{\beta}(\tilde{x}^*(y), y) dy.$$

The hypothesis that $(x_1(t) + \tau(t), y_1(t))$ and $(x_1(t), y_1(t))$ are conjugate along the X -trajectory joining them implies that

$$\frac{\alpha - 1}{\beta}(\tilde{x}^*(y), y) = \frac{\alpha - 1}{\beta}(\tilde{x}(y), y)$$

for all y . (Because $(\alpha - 1)\beta^{-1}$ is the inverse of the slope of G .) So

$$(3.13) \quad x_1(b) - x_1(a) - T(\gamma_1) = x_2(b') - x_2(a') - T(\gamma_2).$$

Suppose that $x_1(b) \geq x_1(a)$, and that $x_2(b') \leq x_2(a')$. Let $\bar{p}_1 = (x_1(b), y_1(a))$, $\bar{p}_2 = (x_2(b'), y_2(b'))$.

Then

$$(3.14.a) \quad T(\delta_b * \gamma_1) = T(\gamma_1) + x_2(b') - x_1(b)$$

and

$$(3.14.b) \quad T(\gamma_2 * \delta_a) = x_1(b) - x_1(a) + x_2(b') - x_1(b) + x_2(a') - x_2(b') + T(\gamma_2).$$

Therefore (3.13) implies that $T(\delta_b * \gamma_1) = T(\gamma_2 * \delta_a)$. If $x_1(b) \geq x_1(a)$ but $x_2(b') > x_2(a')$, then the term $x_2(a') - x_2(b')$ does not occur in (3.14.b), but we must add $x_2(b') - x_2(a')$ to (3.14.a), so (3.9) still holds. The other cases are similar. So (3.9) is proved if $\dot{y}_1 > 0$ on $[a, b]$. The case when $\dot{y}_1 < 0$ is similar. \square

We now investigate the derivative of the switching function. After a straightforward computation, we get the following.

LEMMA 3.7. *Let $\Lambda = (u(\cdot), \gamma, \lambda(\cdot), \lambda_0)$ be an extremal lift of an admissible $(u(\cdot), \gamma)$. Then the switching function $\phi_\Lambda: \text{Dom}(\gamma) \rightarrow \mathbb{R}$ is of class C^1 , and its derivative is given by*

$$(3.15) \quad \dot{\phi}_\Lambda(t) = \lambda(t) \cdot [F, G](\gamma(t)).$$

The preceding result shows that the Lie bracket of F and G is very important, so we give it a special name. We let

$$(3.16) \quad H = [F, G].$$

Then H can also be expressed in terms of X and Y , using (2.12.a, b). We get

$$(3.17) \quad H = \frac{1}{2}[X, Y].$$

From Lemma 3.7, we get the following.

THEOREM 3.8. *Let $U \subseteq M$ be an open set such that $G(p)$ and $H(p)$ are linearly independent at every $p \in U$. Then*

$$\text{Opt}^1(\Sigma \upharpoonright U) \subseteq \text{Traj}(X \vee Y)^\infty.$$

Proof. Let $(u(\cdot), \gamma) \in \text{Adm}(\Sigma)$, $\gamma \in \text{Opt}^1(\Sigma \upharpoonright U)$. Let $\Lambda = (u(\cdot), \gamma, \lambda(\cdot), \lambda_0)$ be an extremal lift of $(u(\cdot), \gamma)$.

If $t_0 \in \text{Dom}(\gamma)$ is such that $\phi_\Lambda(t_0) = 0$, then (3.15) implies that

$$(3.18) \quad \dot{\phi}_\Lambda(t_0) = \lambda(t_0) \cdot H(\gamma(t_0)).$$

Since $G(\gamma(t_0))$ and $H(\gamma(t_0))$ are linearly independent, and $\lambda(t_0) \neq 0$, but $\lambda(t_0) \cdot G(\gamma(t_0)) = 0$, it follows that $\dot{\phi}_\Lambda(t_0) \neq 0$.

So all zeros of ϕ_Λ are isolated. Then ϕ_Λ has finitely many zeros, since $\text{Dom}(\gamma)$ is compact. Therefore γ is regular bang-bang. \square

If F and G are independent, then we can be much more precise. Let $\Omega(A)$ be the set of points $p \in M$ such that $F(p)$ and $G(p)$ are independent, and let $\Omega(B)$ be the set of p such that $G(p)$ and $H(p)$ are independent. Then it is clear that $\Omega(A)$ and $\Omega(B)$ are open. An *ordinary point* is a point p which belongs to $\Omega(A) \cap \Omega(B)$.

We can define smooth functions f, g on $\Omega(A)$ by

$$(3.19) \quad H = fF + gG.$$

THEOREM 3.9. *Let $U \subseteq M$ be open, and suppose that $U \subseteq \Omega(A) \cap \Omega(B)$. Then*

$$(3.20) \quad \text{Opt}^1(\Sigma \upharpoonright U) \subseteq \text{Traj}((X \vee Y)^2).$$

If $f > 0$ throughout U , then

$$(3.21) \quad \text{Opt}^1(\Sigma \upharpoonright U) \subseteq \text{Traj}(Y * X).$$

If $f < 0$ throughout U , then

$$(3.22) \quad \text{Opt}^1(\Sigma \upharpoonright U) \subseteq \text{Traj}(X * Y).$$

Proof. Any $\gamma \in \text{Traj}(\Sigma \upharpoonright U)$ is entirely contained in a connected component of U . If U is connected, then $f > 0$ throughout U , or $f < 0$ throughout U . So the last two assertions imply the first one.

We will prove that (3.21) holds if $f > 0$ throughout U . Let $(u(\cdot), \gamma) \in \text{Opt}^1(\Sigma \upharpoonright U)$, and let $\Lambda = (u(\cdot), \gamma, \lambda(\cdot), \lambda_0)$ be an extremal lift. Let t_0 be a time such that $\phi_\Lambda(t_0) = 0$. Then

$$\dot{\phi}_\Lambda(t_0) = \lambda(t_0) \cdot H(\gamma(t_0)) = f(\gamma(t_0))\lambda(t_0) \cdot F(\gamma(t_0))$$

(since $\lambda(t_0) \cdot G(\gamma(t_0)) = 0$).

Since $F(\gamma(t_0))$ and $G(\gamma(t_0))$ are linearly independent, and $\lambda(t_0) \cdot G(\gamma(t_0)) = 0$, we have

$$(3.23) \quad \lambda(t_0) \cdot F(\gamma(t_0)) \neq 0.$$

On the other hand,

$$(3.24) \quad \mathcal{H}(\lambda(t), \gamma(t), u(t)) \leq 0$$

for almost all t (because $\lambda_0 \geq 0$).

So

$$\lambda(t) \cdot F(\gamma(t)) \leq -u(t)\lambda(t) \cdot G(\gamma(t))$$

for almost all t . In particular, we can let $t \rightarrow t_0$, and get

$$(3.25) \quad \lambda(t_0) \cdot F(\gamma(t_0)) \leq 0.$$

Combining this with (3.23) we get

$$(3.26) \quad \lambda(t_0) \cdot F(\gamma(t_0)) < 0.$$

Since $f(\gamma(t_0)) > 0$, we get

$$(3.27) \quad \dot{\phi}_\Lambda(t_0) < 0.$$

This proves that, if t_0 is a switching time for γ , then γ is an X -trajectory for $t < t_0$, and a Y -trajectory for $t > t_0$. So γ cannot have more than one switching, and our conclusion is proved. \square

The conclusion of Theorem 3.9 can also be proved by a different method, which gives some other useful results. We define a 1-form ω in $\Omega(A)$ by

$$(3.28) \quad \langle \omega, F \rangle = 1, \quad \langle \omega, G \rangle = 0.$$

Then, if $\gamma \in \text{Traj}(\Sigma \upharpoonright \Omega(A))$, the time $T(\gamma)$ is the integral of ω along γ . Suppose that

(3.I.a) γ_1 and γ_2 are trajectories of Σ in $\Omega(A)$, going from a point p to a point q ,

(3.I.b) $\gamma_2^{-1} * \gamma_1$ is a simple closed curve whose interior is entirely contained in $\Omega(A)$ (Here γ_2^{-1} is γ_2 run backwards.),

(3.I.c) $\gamma_2^{-1} * \gamma_1$ is oriented counterclockwise, relative to some coordinate system.

Let \mathcal{R} be the region enclosed by $\gamma_2^{-1} * \gamma_1$. Then

$$(3.29) \quad T(\gamma_1) - T(\gamma_2) = \int_{\gamma_2^{-1} * \gamma_1} \omega.$$

By Stokes' Theorem, we get

$$(3.30) \quad T(\gamma_1) - T(\gamma_2) = \iint_{\mathcal{R}} d\omega.$$

On the other hand, we can compute $d\omega$ by the fomula

$$(3.31) \quad \langle d\omega, F \wedge G \rangle = F\langle \omega, G \rangle - G\langle \omega, F \rangle - \langle \omega, [F, G] \rangle$$

and we get (since $\langle \omega, G \rangle$ and $\langle \omega, F \rangle$ are constant functions, and so $F\langle \omega, G \rangle \equiv G\langle \omega, F \rangle \equiv 0$)

$$(3.32) \quad \langle d\omega, F \wedge G \rangle = -\langle \omega, [F, G] \rangle = -f.$$

In the preceding formulas, $F \wedge G$ is the exterior product of F and G . In coordinates, we can express $F \wedge G$ as

$$(3.33) \quad F \wedge G = \Delta_A \partial_x \wedge \partial_y,$$

where

$$(3.34) \quad \Delta_A = \det(F, G).$$

(Here “det” stands for “determinant.” Recall that F and G are column-vector-valued functions on M , and so we can form, for each $p \in M$, the two-by-two matrix with columns $F(p)$, $G(p)$.)

Then

$$(3.35) \quad d\omega = \rho \, dx \wedge dy$$

where ρ is computed by

$$(3.36) \quad \rho = \langle d\omega, \partial_x \wedge \partial_y \rangle = \frac{1}{\Delta_A} \langle d\omega, F \wedge G \rangle,$$

so that

$$(3.37) \quad d\omega = -\frac{f}{\Delta_A} \, dx \wedge dy.$$

Therefore, we have proved the following.

LEMMA 3.10. *Suppose (3.I.a, b, c) hold, and let \mathcal{R} be the region enclosed by $\gamma_2^{-1} * \gamma_1$. Then*

$$(3.38) \quad T(\gamma_1) - T(\gamma_2) = - \iint_{\mathcal{R}} \frac{f}{\Delta_A} \, dx \wedge dy.$$

As a consequence of Lemma 3.10, we get the following.

LEMMA 3.11. *Let U be an open subset of $\Omega(A)$ such that $f(p) > 0$ for p in some dense subset of U . Suppose that U is the domain of some coordinate chart with respect to which $\Delta_A(p) > 0$ for $p \in U$. Then, if γ_1 and γ_2 are trajectories of Σ in U such that $\text{In}(\gamma_1) = \text{In}(\gamma_2)$, $\text{Term}(\gamma_1) = \text{Term}(\gamma_2)$, and that $\gamma_2^{-1} * \gamma_1$ is a simple closed curve oriented counterclockwise whose interior is contained in U , it follows that*

$$(3.39) \quad T(\gamma_1) < T(\gamma_2).$$

Inequality (3.39) also holds if $f < 0$ on a dense set and $\Delta_A < 0$ on U . In the other two cases (i.e., when $f \cdot \Delta_A < 0$ on a dense set) the opposite inequality holds.

Proof. We only consider the first case. (The other ones are similar.) If \mathcal{R} is the region bounded by $\gamma_2^{-1} * \gamma_1$, then $f \geq 0$ on \mathcal{R} , but \mathcal{R} contains points where $f > 0$. So

$$\iint_{\mathcal{R}} \frac{f}{\Delta_A} dx dy > 0.$$

Formula (3.38) then implies (3.39). \square

In order to apply the preceding lemma to prove an extension of Theorem 3.9, we first establish the existence of a particular type of coordinate chart about any point $p \in \Omega(A)$.

LEMMA 3.12. *Let X and Y be smooth vector fields on an open set $M \subseteq \mathbb{R}^2$, and let $p \in M$ be such that $X(p)$ and $Y(p)$ are linearly independent. Then there is a coordinate chart $(U, (\xi, \eta))$, defined on a neighborhood U of p , relative to which X and Y have components $(\alpha, 0)$, $(0, \beta)$, respectively, where α and β are smooth functions on U such that $\alpha > 0$ and $\beta > 0$ throughout U .*

Proof. Let $Q_1(t, s) = \Phi_t^X \Phi_s^Y(p)$. Then Q_1 is a smooth map from some square $\text{Sq}(\varepsilon)$ onto an open set U_1 such that $p \in U$. Moreover, the differential of Q_1 at $(0, 0)$ is nonsingular. So, by taking ε small enough, we may assume that Q_1 is a diffeomorphism. Let $\eta: U_1 \rightarrow \mathbb{R}$ be Q_1^{-1} followed by the map $(t, s) \rightarrow s$. If $q \in U_1$ is of the form $\Phi_{\bar{t}}^X \Phi_{\bar{s}}^Y(p)$, then the integral curve of X through q is the curve $r \mapsto \Phi_{\bar{t}+r}^X \Phi_{\bar{s}}^Y(p)$, and η has the same value \bar{s} at all the points in this curve. Therefore $X\eta \equiv 0$. On the other hand, $\eta(p) = 0$ and $\eta(\Phi_t^Y(p)) = t$ for $|t| < \varepsilon$, so that $Y\eta(p) = 1$. By a similar reasoning, we get a function ξ on some open subset U_2 , such that $p \in U_2$, $\xi(p) = 0$, $X\xi(p) = 1$, and $Y\xi \equiv 0$. The formulas $X\xi(p) = Y\eta(p) = 1$, $X\eta(p) = Y\xi(p) = 0$, imply that the differentials of ξ and η at p are linearly independent. So, if $U \subseteq U_1 \cap U_2$, $p \in U$, and U is sufficiently small, the map $q \mapsto (\xi(q), \eta(q))$ is a diffeomorphism, so that $(U, (\xi, \eta))$ is a chart. The components of X in this chart are $X\xi, X\eta$. Since $X\eta \equiv 0$, we see that X has components $(\alpha, 0)$, for some smooth function α . Since $X\xi(p) = 1$, it follows that $\alpha(p) = 1$. So, by making U smaller, if necessary, we may assume that $\alpha > 0$ throughout U . A similar argument shows that Y has components $(0, \beta)$, and it may be assumed that $\beta > 0$ throughout U . \square

We are now ready to prove a slight generalization of Theorem 3.9.

THEOREM 3.13. *Let $U \subseteq M$ be open, and suppose that $U \subseteq \Omega(A)$, and that $f > 0$ on a dense subset U' of U . Then*

$$(3.40.a) \quad \text{Opt}^1(\Sigma \upharpoonright U) \subseteq \text{Traj}(Y * X).$$

If, instead, $f < 0$ throughout U' , then

$$(3.40.b) \quad \text{Opt}^1(\Sigma \upharpoonright U) \subseteq \text{Traj}(X * Y).$$

Proof. We prove the first assertion only. (The second one follows by a similar argument, or can be proved from the first one by interchanging X and Y .) So we assume that $f > 0$ on a dense subset of U .

The first step is to prove that no strict $X * Y$ -trajectory in U can be optimal. Suppose $\gamma: [a, b] \rightarrow U$ is an optimal strict $X * Y$ -trajectory. Let t_0 be the switching time, so that $a < t_0 < b$, $\gamma \upharpoonright [a, t_0]$ is a Y -trajectory, and $\gamma \upharpoonright [t_0, b]$ is an X -trajectory. Pick a coordinate chart $(V, (\xi, \eta))$, centered at $\gamma(t_0)$, and such that X, Y have components $(\alpha, 0)$, $(0, \beta)$, respectively, with $\alpha > 0$, $\beta > 0$ on V . We identify V with its image under (ξ, η) , and we assume that V is a square $\text{Sq}(\varepsilon)$. Then the X -trajectories are horizontal and go from left to right, and the Y -trajectories are vertical and go up. If $\delta > 0$ is small enough, then $\gamma_2 = \gamma \upharpoonright [t_0 - \delta, t_0 + \delta]$ is a strict $X * Y$ -trajectory in V .

Clearly, there is a unique strict $Y * X$ -trajectory γ_1 going from $\gamma(t_0 - \delta)$ to $\gamma(t_0 + \delta)$ in V . The curve $\gamma_2^{-1} * \gamma_1$ is simple, closed, and oriented counterclockwise. Then

$$(3.41) \quad T(\gamma_1) - T(\gamma_2) = - \iint_{\mathcal{R}} \frac{f}{\Delta_A} dx dy,$$

where \mathcal{R} is the rectangular region enclosed by $\gamma_2^{-1} * \gamma_1$. Since $\Delta_A > 0$ on U (because $\Delta_A = \alpha\beta$), and $f \geq 0$ on U , but \mathcal{R} contains at least one point where $f > 0$, we see that $T(\gamma_1) < T(\gamma_2)$. So γ_2 is not optimal. Therefore γ is not optimal, which is a contradiction. This completes the first step.

We now let $\gamma \in \text{Opt}^1(\Sigma \upharpoonright U)$ be arbitrary. Let t_0 be an arbitrary interior point of $\text{Dom}(\gamma)$, and pick a chart $(V, (\xi, \eta))$ as above. Also, pick $\delta > 0$ such that $\gamma \upharpoonright [t_0 - \delta, t_0 + \delta]$ is entirely contained in V . Then it may happen that

(i) $\gamma \upharpoonright [t_0 - \delta, t_0 + \delta] \in \text{Traj}(X \vee Y)$.

If (i) does not happen, then $\xi(\gamma(t_0 + \delta)) > \xi(\gamma(t_0 - \delta))$ and $\eta(\gamma(t_0 + \delta)) > \eta(\gamma(t_0 - \delta))$. Therefore the point $\gamma(t_0 + \delta)$ lies strictly above and strictly to the right of $\gamma(t_0 - \delta)$. Let $\gamma_2^* = \gamma \upharpoonright [t_0 - \delta, t_0 + \delta]$, and let γ_1^* be the unique strict $Y * X$ -trajectory in V going from $\gamma(t_0 - \delta)$ to $\gamma(t_0 + \delta)$. Let E be the set of points $\gamma_1^*(t)$, for $t \in \text{Dom}(\gamma_1^*)$. Then E is compact. If $\gamma_2^*(s) \in E$ for all $s \in \text{Dom}(\gamma_2^*)$, then it is easy to see that $\gamma_2^* = \gamma_1^*$, i.e., that

(ii) $\gamma \upharpoonright [t_0 - \delta, t_0 + \delta] \in \text{Traj}(Y * X)$.

If (ii) is not true, then there must be a $t \in [t_0 - \delta, t_0 + \delta]$ such that $\gamma_2^*(t) \notin E$. The set of such t is open in $[t_0 - \delta, t_0 + \delta]$, and does not contain the endpoints $t_0 - \delta, t_0 + \delta$, because $\gamma(t_0 - \delta)$ and $\gamma(t_0 + \delta)$ are in E . So there is a maximal interval $[t_1, t_2] \subseteq [t_0 - \delta, t_0 + \delta]$ such that $\gamma_2^*(t) \notin E$ for $t_1 < t < t_2$. Then $\gamma_2^*(t_1) \in E$ and $\gamma_2^*(t_2) \in E$. Let $\gamma_2 = \gamma_2^* \upharpoonright [t_1, t_2]$. If $\gamma_1^*(s_i) = \gamma_2^*(t_i)$ for $i = 1, 2$, let $\gamma_1 = \gamma_1^* \upharpoonright [s_1, s_2]$. Then γ_1 and γ_2 go from the same initial point to the same terminal point. Moreover, γ_1 and γ_2 have no points in common other than the endpoints, because $\gamma_2(t) \notin E$ for $t_1 < t < t_2$, but γ_1 is entirely contained in E . So $\gamma_2^{-1} * \gamma_1$ is a simple closed curve oriented counterclockwise. Exactly as in the first step, we conclude that $T(\gamma_1) < T(\gamma_2)$, so that γ_2 is not optimal, and then γ is not optimal. This contradiction arose from assuming that neither (i) nor (ii) hold.

So we have proved that, if t_0 is any interior point of $\text{Dom}(\gamma)$, then there is a $\delta > 0$ for which (i) or (ii) holds. Then it follows easily that, if J is a compact subinterval of the interior of $\text{Dom}(\gamma)$, then $\gamma \upharpoonright J$ is regular bang-bang. Since we know that $\gamma \upharpoonright J$ cannot have a strict $X * Y$ piece, it follows that $\gamma \upharpoonright J \in \text{Traj}(Y * X)$. From this it follows immediately that $\gamma \in \text{Traj}(Y * X)$. \square

We conclude this section with the definition of a new function which makes it possible to compute f as a quotient of determinants. We let

$$(3.42) \quad \Delta_B = \det(G, H).$$

Then we have

$$\Delta_B = \det(G, fF + gG)$$

so that

$$\Delta_B = f \det(G, F) = -f \Delta_A.$$

Therefore

$$(3.43) \quad f = -\frac{\Delta_B}{\Delta_A}.$$

The determinants Δ_A , Δ_B can also be expressed in terms of X and Y . The result is

$$(3.44) \quad \Delta_A = \frac{1}{2} \det(X, Y),$$

$$(3.45) \quad \Delta_B = \frac{1}{4} \det(Y - X, [X, Y]).$$

4. Barriers. Let $U \subseteq M$ be a connected open set. A *barrier* in U is a subset B of U such that

- (4.I.a) B is a smooth, one-dimensional, connected, relatively closed submanifold of U ,
- (4.I.b) $U - B$ has exactly two connected components,
- (4.I.c) one of the components of $U - B$, which we denote by U^+ , has the property that, at each $p \in B$, each of the vectors $X(p)$, $Y(p)$ is either tangent to B or points into U^+ .

LEMMA 4.1. *Let $U \subseteq M$ be open and connected. Let $B \subseteq U$ be a barrier in U , and let U^+ be the connected component of $U - B$ with the property described in (4.I.c). Let U^- be the other connected component of $U - B$. Then every trajectory $\gamma \in \text{Traj}(\Sigma \upharpoonright U)$ is a concatenation of at most three pieces $\gamma_1, \gamma_2, \gamma_3$, such that $\gamma_1(t) \in U^-$ for all $t \in \text{Dom}(\gamma_1)$, $t \neq \max \text{Dom}(\gamma_1)$, γ_2 is entirely contained in B , and $\gamma_3(t) \in U^+$ for all $t \in \text{Dom}(\gamma_3)$, $t \neq \min \text{Dom}(\gamma_3)$.*

Proof. Step 1: We prove that there is no integral trajectory γ of X or of Y that goes from a point $p \in B$ to a $q \in U^-$. Suppose γ is such a trajectory for, say, X . Let $\gamma(t_0) = p_0$, $\gamma(t_1) = p_1$. Let $\bar{t} \in [t_0, t_1]$ be the largest t such that $\gamma(t) \in B$. (Here we use the fact that B is closed.) Then $\gamma(t) \in U^-$ for $\bar{t} < t \leq t_1$. Pick a vector field X^* , on a neighborhood W of $\gamma(\bar{t})$, such that $X^*(q)$ points into U^+ for each $q \in B \cap W$. For each $\varepsilon > 0$, let γ_ε denote the integral curve of $\varepsilon X^* + X$ such that $\gamma_\varepsilon(\bar{t}) = \gamma(\bar{t})$. Pick a $\delta > 0$, $\delta < t_1 - \bar{t}$, such that $\gamma|[\bar{t}, \bar{t} + \delta]$ is entirely contained in W . Then by the continuous dependence on parameters for solutions of ordinary differential equations, the curve γ_ε is defined on $[\bar{t}, \bar{t} + \delta]$, and $\gamma_\varepsilon(t) \in W$ for $t \in [\bar{t}, \bar{t} + \delta]$, if $0 < \varepsilon < \varepsilon_0$, provided that ε_0 is small enough. Moreover, $\gamma_\varepsilon(t) \rightarrow \gamma(t)$ as $\varepsilon \rightarrow 0$. We claim that, if $0 < \varepsilon < \varepsilon_0$, then $\gamma_\varepsilon|[\bar{t}, \bar{t} + \delta]$ is entirely contained in U^+ . Indeed, if there is a $t \in [\bar{t}, \bar{t} + \delta]$ such that $\gamma_\varepsilon(t) \in U^-$, then this t is contained in a maximal open interval $]a, b[\subseteq [\bar{t}, \bar{t} + \delta]$ such that $\gamma_\varepsilon(]a, b[) \subseteq U^-$. Then $\gamma_\varepsilon(a) \in B$. Since $\dot{\gamma}_\varepsilon(a) = (\varepsilon X^* + X)(\gamma_\varepsilon(a))$, and $X(\gamma_\varepsilon(a))$ points to U^+ or is tangent to B , while $X^*(\gamma_\varepsilon(a))$ points to U^+ , we have

$$\gamma_\varepsilon(a + s) \in U^+ \quad \text{for } s > 0, s \text{ near } 0.$$

This contradicts the fact that $\gamma_\varepsilon(\tau) \in U^-$ for $\tau \in]a, b[$. So no t can exist for which $\gamma_\varepsilon(t) \in U^-$. On the other hand, if $\gamma_\varepsilon(t) \in B$, and if $t > \bar{t}$, then $\gamma_\varepsilon(t - s) \in U^-$ for small $s > 0$. Since we already know that no point of $\gamma_\varepsilon|[\bar{t}, \bar{t} + \delta]$ can be in U^- , it follows that $\gamma_\varepsilon(t) \in U^+$ for $t \in]\bar{t}, \bar{t} + \delta]$.

If we now let $\varepsilon \rightarrow 0$, we find that $\gamma(\bar{t} + \delta) \in B \cup U^+$. But this contradicts the fact that $\gamma(\bar{t} + \delta) \in U^-$.

Step 2: It follows easily from Step 1 that, if γ is an arbitrary regular bang-bang trajectory of Σ such that $\gamma(t_0) \in B$ for some t_0 , then $\gamma(t) \in B \cup U^+$ for all $t \geq t_0$, $t \in \text{Dom}(\gamma)$.

Step 3: Let $\gamma \in \text{Traj}(\Sigma \upharpoonright U)$, and let $t_0 \in \text{Dom}(\gamma)$ be such that $\gamma(t_0) \in B$. If γ corresponds to a control $u(\cdot): \text{Dom}(\gamma) \rightarrow [-1, 1]$, then $u(\cdot)$ is the weak limit of regular bang-bang controls $u_n(\cdot)$ with the same domain. Let γ_n be the corresponding trajectories, which satisfy $\gamma_n(t_0) = \gamma(t_0)$. Then $\gamma_n(t) \in B \cup U^+$ for $t \geq t_0$, by Step 2. Letting $n \rightarrow \infty$, we conclude that $\gamma(t) \in B \cup U^+$ for $t \geq t_0$.

Step 4: Let $\gamma \in \text{Traj}(\Sigma \upharpoonright U)$, and let $t_0 \in \text{Dom}(\gamma)$ be such that $\gamma(t_0) \in B$. Let γ^{-1} be γ run backwards, i.e., $\gamma^{-1}(t) = \gamma(-t)$, for $-t \in \text{Dom}(\gamma)$. Then γ^{-1} is a trajectory of Σ' , where Σ' is the system obtained from Σ by changing X, Y into $-X, -Y$. The set B is also a barrier in U for the new system, except that now U^- plays the role of U^+ . Since $\gamma^{-1}(-t_0) \in B$, it follows from Step 3 that $\gamma^{-1}(s) \in B \cup U^-$ for $s \geq t_0$, $s \in \text{Dom}(\gamma^{-1})$. Therefore $\gamma(t) \in B \cup U^-$ for $t \leq t_0$, $t \in \text{Dom}(\gamma)$.

Step 5: Let $\gamma \in \text{Traj}(\Sigma \upharpoonright U)$. If γ never goes through B , then γ is entirely contained in U^- or in U^+ , and the desired conclusion holds. If γ goes through B , let t_0 be the infimum, and let t_1 be the supremum, of the set of those $t \in \text{Dom}(\gamma)$ such that $\gamma(t) \in B$. Then $\gamma(t_0) \in B$ and $\gamma(t_1) \in B$. Since $\gamma(t_0) \in B$, it follows from Step 3 that $\gamma(t) \in B \cup U^+$ for $t > t_0$. Therefore $\gamma(t) \in U^+$ for $t > t_1$. Since $\gamma(t_1) \in B$, it follows that $\gamma(t) \in B \cup U^-$ for $t \leq t_1$. So $\gamma(t) \in U^-$ for $t < t_0$. If $t_0 \leq t \leq t_1$, then $\gamma(t) \in B \cup U^+$ and $\gamma(t) \in B \cup U^-$, so $\gamma(t) \in B$.

Summarizing: $\gamma(t) \in U^-$ for $t < t_0$, $\gamma(t) \in B$ for $t_0 \leq t \leq t_1$, and $\gamma(t) \in U^+$ for $t > t_1$. This is precisely the desired conclusion. \square

5. A lemma. We now prove a lemma that excludes certain types of switchings. We consider the following situation:

- (5.I.a) γ is a strict $X * Y$ - or $Y * X$ -trajectory, with a switching at a point q_0 ,
- (5.I.b) the vectors $X(q_0), Y(q_0)$ are linearly dependent, but point in opposite directions.

We want to conclude from these conditions that γ cannot be time-optimal. Let us first observe that such a conclusion cannot possibly be arrived at using solely the Maximum Principle. Indeed, a γ for which (5.I.a) and (5.I.b) hold may very well satisfy the Maximum Principle. For a simple example, suppose that X and Y satisfy (5.I.b), and that, in addition, the vectors $X(q_0)$ and $[X, Y](q_0)$ are linearly independent (i.e., that $G(q_0)$ and $H(q_0)$ are independent). Pick $\bar{\lambda} \in \mathbb{R}_*^2$ such that $\bar{\lambda} \neq 0$ but $\bar{\lambda} \cdot G(q_0) = 0$. Let γ be the concatenation of (a) the curve $t \rightarrow \Phi_t^Y(q_0)$, $-\delta \leq t \leq 0$, and (b) the curve $t \rightarrow \Phi_t^X(q_0)$, $0 \leq t \leq \delta$, where $\delta > 0$ is small. Let $\lambda(\cdot)$ be the unique variational covector along γ for which $\lambda(0) = \bar{\lambda}$. For $t > 0$, the derivative of $t \rightarrow \lambda(t) \cdot G(\gamma(t))$ is $t \rightarrow \lambda(t) \cdot [X, G](\gamma(t))$, i.e., $t \rightarrow \lambda(t) \cdot H(\gamma(t))$. In particular, the limit of this derivative at $t = 0$ is $\bar{\lambda} \cdot H(q_0)$. Similarly, the derivative for $t < 0$ also equals $\lambda(t) \cdot H(\gamma(t))$, which goes to $\bar{\lambda} \cdot H(q_0)$ as $t \rightarrow 0^-$. In view of our hypothesis that $H(q_0)$ and $G(q_0)$ are independent, we can conclude that $\bar{\lambda} \cdot H(q_0) \neq 0$. By changing $\bar{\lambda}$ to $-\bar{\lambda}$, if necessary, we may assume that $\bar{\lambda} \cdot H(q_0) > 0$. Then $\lambda(t) \cdot G(\gamma(t)) > 0$ for $t > 0$, and $\lambda(t) \cdot G(\gamma(t)) < 0$ for $t < 0$, as long as t is sufficiently close to 0. By making δ smaller, we may assume that $\lambda(t) \cdot G(\gamma(t)) > 0$ for all $t \in]0, \delta]$, $\lambda(t) \cdot G(\gamma(t)) < 0$ for all $t \in [-\delta, 0[$. Then the Hamiltonian $u \rightarrow \mathcal{H}(\lambda(t), \gamma(t), u)$ is minimized by $u = -1$ if $t > 0$, and by $u = 1$ if $t < 0$. So, if we let $u(t) = -1$ for $0 < t \leq \delta$, $u(t) = 1$ for $-\delta \leq t < 0$, we see that γ is a trajectory that corresponds to $u(\cdot)$, that $\lambda(\cdot)$ is a variational covector along $(u(\cdot), \gamma)$, and that $\lambda(\cdot)$ is nontrivial and minimizing. The value of the Hamiltonian along $(u(\cdot), \gamma, \lambda(\cdot))$ is $\mathcal{H}(\lambda(t), \gamma(t), u(t)) = \lambda(t) \cdot X(\gamma(t))$ for $t > 0$, and $\mathcal{H}(\lambda(t), \gamma(t), u(t)) = \lambda(t) \cdot Y(\gamma(t))$ for $t < 0$. Since $\gamma \upharpoonright [0, \delta] \in \text{Traj}(X)$, we see that $\mathcal{H}(\lambda(t), \gamma(t), u(t))$ is constant for $0 \leq t \leq \delta$. When $t = 0$, we get $\mathcal{H}(\lambda(0), \gamma(0), u(0)) = \bar{\lambda} \cdot F(q_0) = 0$ (since $F(q_0)$ and $G(q_0)$ are linearly dependent, and $\bar{\lambda} \cdot G(q_0) = 0$ but $G(q_0) \neq 0$). So \mathcal{H} vanishes along $\lambda(\cdot), \gamma, u(\cdot)$ for $t \geq 0$. A similar argument establishes that \mathcal{H} also vanishes for $t \leq 0$. So, if we let $\lambda_0 = 0$, all the conditions of the Maximum Principle hold.

The preceding remarks show that one can construct examples of trajectories γ for which (5.I.a) and (5.I.b) hold, but which satisfy the Maximum Principle. The fact that such a γ cannot be optimal is, therefore, a “high-order optimality condition.”

Actually, we will only prove our desired conclusion under an extra hypothesis, namely, that

- (5.I.c) the X - and Y -trajectories through q_0 have a tangency of finite order at q_0 .

We suspect that (5.I.c) is not really needed, but our proof does require it. In any case, notice that (5.I.c) is no restriction at all in the analytic case because, if (5.I.c) is violated, then the X - and Y -trajectories through q_0 must coincide (as sets) but be oriented in opposite directions, so that it is completely obvious that γ cannot be optimal.

LEMMA 5.1. *If (5.I.a, b, c) hold, then γ is not time-optimal.*

Proof. (cf. Fig. 2). We may assume that $\gamma \in \text{Traj}(X * Y)$. (The other case is identical.) Let $[a, b] = \text{Dom}(\gamma)$, and let $t_0 \in]a, b[$ be such that $\gamma(t_0) = q_0$.

Pick a square coordinate chart $(U, (\xi, \eta))$, of radius ε , and centered at q_0 , such that $X = \partial_x$ relative to these coordinates. Then $Y = \alpha \partial_x + \beta \partial_y$ on U , where α, β are smooth functions that satisfy

(5.1) $\alpha(0, 0) < 0, \quad \beta(0, 0) = 0.$

By making ε smaller, if necessary, we may assume that $\alpha < 0$ throughout U . Then, by making a larger, if necessary, we may assume that $\gamma|_{[a, t_0]}$ is entirely contained

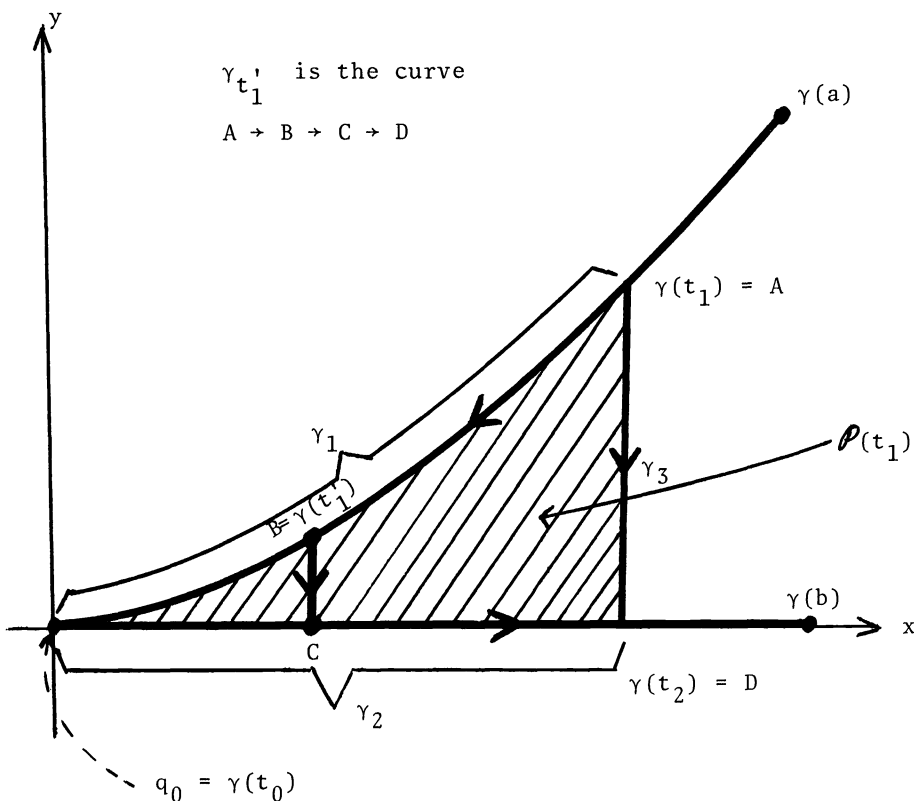


FIG. 2

in U . Let

$$(5.2) \quad t \rightarrow (x(t), y(t)), \quad a \leq t \leq t_0$$

be the expression of $\gamma| [a, t_0]$ relative to our coordinates. Then

$$(5.3) \quad \dot{x}(t) = \alpha(x(t), y(t)) < 0,$$

so that $\gamma| [a, t_0]$ is entirely contained in the half-square $\{(x, y) \in U: x > 0\}$. The hypothesis that the Y -trajectory through q_0 does not have an infinite-order tangency with the X -trajectory says that some derivative of $y(\cdot)$ is nonzero at t_0 . If we let k be the smallest integer such that

$$(5.4) \quad \frac{d^k y}{dt^k}(t_0) \neq 0,$$

then $y(\cdot)$ has the Taylor expansion

$$(5.5) \quad y(t) = \rho(t - t_0)^k + O(|t - t_0|^{k+1})$$

for $t \leq t_0$, where

$$(5.6) \quad \rho = \frac{1}{k!} \frac{d^k y}{dt^k}(t_0).$$

After making the change of coordinates $(x, y) \rightarrow (x, -y)$, if necessary, we may assume that the sign of ρ is $(-1)^k$, so that $y(t) > 0$ for $t < t_0$, t near t_0 . By making a even larger, if necessary, we may assume that

$$(5.7) \quad y(t) > 0 \quad \text{for } a \leq t < t_0.$$

On the other hand, we have

$$(5.8) \quad \dot{y}(t) = \beta(x(t), y(t))$$

and so, for $j \geq 1$:

$$(5.9) \quad \frac{d^j y}{dt^j}(t) = (Y^{j-1} \beta)(x(t), y(t)).$$

Therefore $k = \nu + 1$, where ν is the first integer j such that $(Y^j \beta)(q_0) \neq 0$. A simple computation shows that

$$(5.10) \quad Y^j \beta = \alpha^j (\partial_x^j \beta) + \sum_{0 \leq i < j} \psi_{ij} \cdot \partial_x^i \beta$$

for suitable functions ψ_{ij} . Since $\alpha(q_0) < 0$, this implies that ν is also the first integer j such that $(\partial_x^j \beta)(q_0) \neq 0$, and that

$$(5.11) \quad \rho = \frac{1}{k!} \alpha(q_0)^\nu (\partial_x^\nu \beta)(q_0).$$

Since ρ has sign $(-1)^{\nu+1}$, it follows that

$$(5.12) \quad \partial_x^\nu \beta(q_0) < 0.$$

We now pick a $t_1 \in [a, t_0[$, such that $x(t_1) < b - t_0$. Then the point $(x(t_1), 0)$ is in $\gamma| [t_0, b]$, and $(x(t_1), 0) = \gamma(t_2)$ for some t_2 . (Actually, $t_2 = t_0 + x(t_1)$.) We let $\gamma_1 = \gamma| [t_1, t_0]$, $\gamma_2 = \gamma| [t_0, t_2]$. Also, we let γ_3 denote the vertical segment from $(x(t_1), y(t_1))$ to $(x(t_1), 0)$. We let $\mathcal{P}(t_1)$ denote the closed region bounded by γ_1 , γ_2 and γ_3 .

The function β has the Taylor expansion

$$(5.13) \quad \beta(x, y) = \beta(x, 0) + O(|y|)$$

for (x, y) near 0. Therefore, we have

$$(5.14) \quad \beta(x, y) = \frac{x^\nu}{\nu!} (\partial_x^\nu \beta)(q_0) + O(|x|^{\nu+1}) + O(|y|).$$

Since the curve γ_1 satisfies $\dot{x}(t_0) = \alpha(0, 0) \neq 0$, and

$$(5.15) \quad y(t) = \rho(t - t_0)^{\nu+1} + O(|t - t_0|^{\nu+2}),$$

we see that, as $(x, y) \rightarrow 0$ along $\mathcal{P}(t_1)$, we have $y = O(x^{\nu+1})$. Hence

$$(5.16) \quad y = O(x^{\nu+1}) \quad \text{as } (x, y) \rightarrow 0 \text{ along } \mathcal{P}(t_1).$$

So, if we restrict ourselves to $\mathcal{P}(t_1)$, the Taylor expansion for β becomes

$$(5.17) \quad \beta(x, y) = \frac{x^\nu}{\nu!} (\partial_x^\nu \beta)(q_0) + O(x^{\nu+1}).$$

On the other hand, the derivative $\partial_x \beta$ satisfies

$$(5.18) \quad \partial_x \beta(x, y) = \partial_x \beta(x, 0) + O(|y|),$$

i.e.,

$$(5.19) \quad \partial_x \beta(x, y) = \frac{x^{\nu-1}}{(\nu-1)!} (\partial_x^\nu \beta)(q_0) + O(|x|^\nu) + O(|y|);$$

so that, if we restrict ourselves to $\mathcal{P}(t_1)$, we get

$$(5.20) \quad \partial_x \beta(x, y) = \frac{x^{\nu-1}}{(\nu-1)!} (\partial_x^\nu \beta)(q_0) + O(x^\nu).$$

Combining (5.17) and (5.20) we get

$$(5.21) \quad \beta(x, y) = \frac{x}{\nu} (\partial_x \beta)(x, y) [1 + O(x)]$$

as $(x, y) \rightarrow 0$, $(x, y) \in \mathcal{P}(t_1)$.

It is clear that

$$(5.22) \quad \Delta_A = \frac{1}{2} \beta,$$

$$(5.23) \quad \Delta_B = \frac{1}{4} ((\alpha - 1)(\partial_x \beta) + \beta(\partial_x \alpha)).$$

Using (5.21), we get

$$(5.24) \quad \Delta_B = \frac{1}{4} \left(\alpha - 1 + \frac{x}{\nu} \partial_x \alpha (1 + O(x)) \right) (\partial_x \beta),$$

i.e.,

$$(5.25) \quad \Delta_B = \frac{1}{4} (\alpha - 1 + O(x)) \partial_x \beta,$$

always for $(x, y) \rightarrow 0$, $(x, y) \in \mathcal{P}(t_1)$.

Now choose t_1 so small that the $O(x)$ of formula (5.25) is dominated by $\alpha - 1$ (recall that $\alpha < 0$ on U), and that the $O(x^{\nu+1})$, $O(x^\nu)$ that appear in the right sides

of (5.17), (5.20) are actually dominated by the lower order terms that precede them. Then (since $\partial_x^\nu \beta(q_0) < 0$), we see that, if $\mathcal{P}(t_1)^0 = \mathcal{P}(t_1) - \{q_0\}$:

$$(5.II) \quad \beta < 0 \text{ and } \partial_x \beta < 0 \text{ throughout } \mathcal{P}(t_1)^0,$$

and that

$$(5.III) \quad \Delta_A < 0 \text{ and } \Delta_B > 0 \text{ throughout } \mathcal{P}(t_1)^0.$$

Therefore $\mathcal{P}(t_1)^0 \subseteq \Omega(A)$, and f is well defined on $\mathcal{P}(t_1)^0$. Since $f = -\Delta_B/\Delta_A$, we conclude that

$$(5.IV) \quad f > 0 \text{ throughout } \mathcal{P}(t_1)^0.$$

At any point q of γ_3 , the vector $Y(q)$ points to the left (because $\alpha < 0$) and down (because $\beta < 0$ on $\mathcal{P}(t_1)^0$). Therefore the tangent vector $-\partial_y$ to γ_3 is a convex combination of $X(q)$ and $Y(q)$. So γ_3 is, after a suitable reparametrization, a trajectory of Σ . Moreover, the curve $\gamma_2 * \gamma_1$ is also a trajectory of Σ , and $\gamma_3^{-1} * \gamma_2 * \gamma_1$ is a simple closed curve, oriented counterclockwise.

Since $f > 0$ and $\Delta_A < 0$ in the region enclosed by $\gamma_3^{-1} * (\gamma_2 * \gamma_1)$, we are almost in a situation where Lemma 3.11 applies. If we could apply it, we would conclude that

$$(5.26) \quad T(\gamma_3) < T(\gamma_2 * \gamma_1).$$

(Naturally, the roles of the γ_1 , γ_2 of the statement of Lemma 3.11 are played by $\gamma_2 * \gamma_1$ and γ_3 , respectively.) The reason why Lemma 3.11 cannot actually be applied as it stands is that, in Lemma 3.11, it was required that the curves γ_1 , γ_2 , and the region enclosed by them, be entirely contained in $\Omega(A)$. Here this condition does not hold, because $q_0 \notin \Omega(A)$. However, it is a very simple matter to modify the argument so as to be able to apply Lemma 3.11 and conclude that (5.26) holds. Pick t'_1 between t_1 and t_0 , and let $\gamma_{t'_1}$ be the concatenation of: (a) $\gamma \upharpoonright [t_1, t'_1]$, (b) the vertical segment from $(x(t'_1), y(t'_1))$ to $(x(t'_1), 0)$, suitably reparametrized so that it becomes a trajectory of Σ , and (c) the X -trajectory from $(x(t'_1), 0)$ to $(x(t_1), 0)$. Now Lemma 3.11 can be applied rigorously, and we get

$$(5.27) \quad T(\gamma_3) < T(\gamma_{t'_1}).$$

Another application of Lemma 3.11 shows that, if $t'_1 < t'' < t_0$, then $T(\gamma_{t'_1}) < T(\gamma_{t''})$. If we let $t''_1 \rightarrow t_0$, we get $T(\gamma_{t'_1}) \leq T(\gamma_2 * \gamma_1)$. So (5.26) holds.

Since $\gamma_2 * \gamma_1$ is a piece of γ , we conclude from (5.26) that γ is not optimal. \square

Remark 5.1. Suppose that (5.I.a, b) hold, and that γ is actually a strict $X * Y$ -trajectory. Let \mathcal{R} denote the open region bounded by the curves $t \rightarrow \Phi_t^Y(q_0)$, $t < 0$, and $t \rightarrow \Phi_t^X(q_0)$, $t > 0$. Hypothesis (5.I.c) was only used in the proof of Lemma 5.1 to conclude that $f > 0$ on \mathcal{R} . Therefore the conclusion of Lemma 5.1 is still true if (5.I.c) does not hold, as long as $f > 0$ on \mathcal{R} .

6. Nonordinary arcs. We now study the time-optimal trajectories in the neighborhood of certain arcs which consist entirely of points p such that $\Delta_A(p) = 0$ or $\Delta_B(p) = 0$. Recall that a point $p \in M$ is called an *ordinary point* if both $\Delta_A(p)$ and $\Delta_B(p)$ are nonzero. A *nonordinary point* is therefore a point p where the function $\Delta_A \cdot \Delta_B$ vanishes. A *nonordinary arc* is a smooth one-dimensional connected embedded submanifold S of M , with the property that every $p \in S$ is nonordinary. An *isolated nonordinary arc* (henceforth abbreviated as INOA) is a nonordinary arc S with the property that there

is a U such that

- (6.I.a) U is an open, connected subset of M ,
- (6.I.b) $U - S \subseteq \Omega(A) \cap \Omega(B)$,
- (6.I.c) S is a relatively closed subset of U .

If U is such that (6.I.a, b) hold, then $U - S$ has at most two connected components. However, $U - S$ might be connected. (Example: let U be an annulus, and let S be a segment joining the inner and outer boundaries of U .) In any case, one can always make U smaller, and assume that, in addition to (6.I.a, b, c), a further condition holds, namely,

- (6.I.d) $U - S$ has exactly two connected components.

An INOA S will be called *regular* if it satisfies the following three conditions:

- (6.II.a) Each of the vector fields X, Y is either everywhere tangent to S or nowhere tangent to S ,
- (6.II.b) each of the functions Δ_A, Δ_B is either identically zero on S or nowhere zero on S ,
- (6.II.c) if X and Y are everywhere tangent to S , then each of the vector fields $X, Y, Y - X$ is either identically zero on S or never zero on S .

Suppose that S is a regular INOA. Let U be such that (6.I) hold. In view of (6.II.a) it is clear that, if either X or Y is everywhere tangent to S , then S is a barrier in U . Also, if both X and Y are nowhere tangent to S , then it may happen that

- (6.III) for each $p \in S$, the vectors $X(p), Y(p)$ point to opposite sides of S ,

or that $X(p)$ and $Y(p)$ point to the same side of S for all p . In the latter case, the submanifold S is a barrier in U . Hence a regular INOA is either a barrier in U or an arc for which (6.III) holds. (In the latter case, S will be called a *nonbarrier regular* INOA.)

THEOREM 6.1. *Let S be a regular INOA which is a barrier in U for some U for which (6.I) hold. Then every time-optimal trajectory in U is bang-bang with at most four switchings.*

Proof. Let U^+, U^- be the connected components of $U - S$, labelled in such a way that, for each $p \in S$, each of the vectors $X(p), Y(p)$ either is tangent to S or points into U^+ . Then Lemma 4.1 implies that every trajectory of $\Sigma \upharpoonright U$ is a concatenation $\gamma_3 * \gamma_2 * \gamma_1$ of at most three pieces such that, if $\text{Dom}(\gamma_i) = [a_i, b_i]$ (so that $b_1 = a_2, b_2 = a_3$), then $\gamma_1 \upharpoonright [a_1, b_1[$ is contained in U^- , $\gamma_3 \upharpoonright [a_3, b_3]$ is contained in U^+ , and γ_2 is contained in S .

Now suppose that γ is time-optimal. Since U^- is entirely contained in $\Omega(A) \cap \Omega(B)$, Theorem 3.9 implies that γ_1 is bang-bang with at most one switching. A similar conclusion holds for γ_3 . As for γ_2 , let us first observe that this piece can only occur if at least one of the vector fields X, Y is everywhere tangent to S . If only one of X, Y is everywhere tangent to S , then γ_2 is bang-bang with no switchings. If both X and Y are everywhere tangent to S , then (6.II.c) implies, first of all, that X is either always zero on S , or never zero on S . In the former case, γ_2 must be a Y -trajectory, and so γ_2 is bang-bang with no switchings. A similar conclusion follows if Y vanishes everywhere on S . So we need only consider the case when both X and Y are nonzero everywhere on S . In this case, if there is some $p \in S$ such that $X(p)$ and $Y(p)$ have

opposite directions, it follows that $X(p)$ and $Y(p)$ have opposite directions for all $p \in S$. But then it is clear that γ_2 must be an $X \vee Y$ -trajectory. Finally, we must consider the case when, for all $p \in S$, the vectors $X(p)$ and $Y(p)$ are nonzero, tangent to S , and positive multiples of each other. In this case, (6.II.c) says that either $X(p) = Y(p)$ for all $p \in S$, or $X(p) \neq Y(p)$ for all $p \in S$. If $X \equiv Y$ on S , then $\gamma_2 \in \text{Traj}(X)$. If $X(p) \neq Y(p)$ for all $p \in S$, then either $X(p)$ is shorter than $Y(p)$ for all $p \in S$, or it is longer for all $p \in S$. In the former case, γ_2 must be a Y -trajectory whereas, in the latter case, γ_2 has to be an X -trajectory. So we have shown that, in all possible cases, $\gamma_2 \in \text{Traj}(X \vee Y)$.

So γ is bang-bang with at most four switchings (namely: one switching for each of γ_1, γ_3 , and, possibly, the switchings at a_2 and b_2). \square

We now study what happens when (6.III) holds. Let U_X, U_Y be the connected components of $U - S$, labelled in such a way that, if V is either X or Y , and $p \in S$, then $V(p)$ points into U_V . The function f is well defined and nowhere vanishing on $U_X \cup U_Y$ (because $U_X \cup U_Y \subseteq \Omega(A) \cap \Omega(B)$).

Since U_X and U_Y are connected, we have three possibilities:

(6.IV.a) $f > 0$ throughout $U_X \cup U_Y$, or $f < 0$ throughout $U_X \cup U_Y$,

(6.IV.b) $f < 0$ on U_X and $f > 0$ on U_Y ,

(6.IV.c) $f > 0$ on U_X and $f < 0$ on U_Y .

The case when (6.IV.a) holds is easy to dispose of.

THEOREM 6.2. *Let S be a nonbarrier regular INOA, and let U be such that (6.I), (6.III) and (6.IV.a) hold. Then every time-optimal trajectory in U is bang-bang with at most one switching.*

Proof. Since S is a regular INOA, it satisfies (6.II.b). Therefore the function Δ_A either vanishes identically on S , or vanishes nowhere on S . Suppose first that Δ_A never vanishes on S . Then U is entirely contained in $\Omega(A)$. So the hypothesis of Theorem 3.13 holds, and we can conclude that

$$\text{Opt}^1(\Sigma \upharpoonright U) \subseteq \text{Traj}(X \vee Y)^2.$$

Now suppose that $\Delta_A \equiv 0$ on S . Let us consider the case when $f > 0$ on $U_X \cup U_Y$. (The case when $f < 0$ on $U_X \cup U_Y$ is similar.) Let $\gamma \in \text{Opt}^1(\Sigma \upharpoonright U)$. Let I_X, I_Y denote the subsets of $\text{Dom}(\gamma)$ that consist of all t such that $\gamma(t) \in U_X, \gamma(t) \in U_Y$, respectively. Then I_X, I_Y are relatively open subsets of $\text{Dom}(\gamma)$.

Let $\mathcal{F}_X, \mathcal{F}_Y$ denote the sets of connected components of I_X, I_Y , respectively. Suppose that $J \in \mathcal{F}_Y$. If $J =]a, b[$ for some a, b , then $\gamma(a) \in S$ and $\gamma(b) \in S$. Since $\gamma \upharpoonright J$ is entirely contained in U_Y , and $f > 0$ on U_Y , we can conclude from Thm. 3.9 that $\gamma \upharpoonright J$ is an X -trajectory, or a Y -trajectory, or an X -trajectory followed by a Y -trajectory. On the other hand, it is not possible for $\gamma \upharpoonright]a, a + \delta[$ to be an X -trajectory for any $\delta > 0$ (because X points into U_X , and $\gamma \upharpoonright J$ is contained in U_Y). Similarly, it is not possible for $\gamma \upharpoonright]b - \delta, b[$ to be a Y -trajectory. So we have reached a contradiction. Therefore no $J \in \mathcal{F}_Y$ can be an open interval. This shows that \mathcal{F}_Y consists of at most two intervals, and that each of these intervals contains one of the endpoints of $\text{Dom}(\gamma)$.

Now let $J \in \mathcal{F}_X$. Suppose that $J =]a, b[$. Exactly as above, we conclude that $\gamma(a) \in S$, that $\gamma(b) \in S$, and that $\gamma \upharpoonright J$ is an X -trajectory, or a Y -trajectory, or an X -trajectory followed by a Y -trajectory. If $\gamma \upharpoonright J$ were an X -trajectory, then γ would leave S at $\gamma(a)$ and enter U_X , but it would never be able to return to S (since $X(p)$ points into U_X for each $p \in S$). But this would contradict the fact that $\gamma(b) \in S$. Similarly, we also reach a contradiction if $\gamma \upharpoonright J$ is a Y -trajectory. Therefore $\gamma \upharpoonright J$ is an X -trajectory followed by a Y -trajectory. Choose $c \in]a, b[$ such that $\gamma \upharpoonright]a, c[$ is an X -trajectory and

that $\gamma|_{[c, b[}$ is a Y -trajectory. We now show that a and b cannot be switching times of γ . Suppose a were a switching time for γ . Then $\gamma(a)$ and $\gamma(c)$ would have to be conjugate along γ . Let $t \mapsto v(t)$ denote the variational vector field along γ such that $v(a) = G(\gamma(a))$. Since $G(\gamma(a)) = \frac{1}{2}(Y(\gamma(a)) - X(\gamma(a)))$, and $\Delta_A(\gamma(a)) = 0$, the vector $G(\gamma(a))$ must equal $\rho X(\gamma(a))$, for some ρ . Moreover, ρ must be nonzero, because $G(\gamma(a)) \neq 0$. If we let v' denote the variational vector field along γ such that $v'(a) = X(\gamma(a))$, then our preceding remarks show that $v(t) = \rho v'(t)$ for all $t \in \text{Dom}(\gamma)$. On the other hand, since $\gamma|_{[a, c]} \in \text{Traj}(X)$, we have $v'(t) = X(\gamma(t))$ for $t \in [a, c]$. In particular, $v'(c) = X(\gamma(c))$. Since $\gamma(a)$ and $\gamma(c)$ are conjugate along γ , the vectors $G(\gamma(c))$ and $v(c)$ are linearly dependent. Since $v(c) = \rho v'(c) \neq 0$, we have an equality $G(\gamma(c)) = \rho' X(\gamma(c))$ for some $\rho' \in \mathbb{R}$. Since $Y = X + 2G$, we conclude that $Y(\gamma(c))$ is a multiple of $X(\gamma(c))$, and therefore it follows that $\Delta_A(\gamma(c)) = 0$. But this contradicts the facts that $\gamma(c) \in U_X$, and $U_X \subseteq \Omega(A)$. So a cannot be a switching time for γ . A similar reasoning shows that b is not a switching time either.

Now suppose that $\min \text{Dom}(\gamma) < a$. Since γ does not have a switching at a , and $\gamma|_{[a, c]} \in \text{Traj}(X)$, it follows that $\gamma|_{[a - \delta, a]} \in \text{Traj}(X)$ for some $\delta > 0$. Since $X(\gamma(a))$ points into U_X , it follows that $\gamma(t) \in U_Y$ for $a - \delta \leq t < a$. Let J' be the connected component of I_Y such that $[a - \delta, a] \subseteq J'$. Then $J' \in \mathcal{F}_Y$ and so, as was shown before, J' must contain an endpoint of $\text{Dom}(\gamma)$. So $J' = \text{Dom}(\gamma) \cap] - \infty, a[$. The curve $\gamma|_{J'}$ is an X -trajectory, or a Y -trajectory, or an X -trajectory followed by a Y -trajectory. Since $\gamma|_{[a - \delta, a]} \in \text{Traj}(X)$, we conclude that $\gamma|_{(\text{Dom}(\gamma) \cap] - \infty, a])} \in \text{Traj}(X)$. Therefore $\gamma|_{(\text{Dom}(\gamma) \cap] - \infty, c])} \in \text{Traj}(X)$. This conclusion was arrived at under the assumption that $a > \min \text{Dom}(\gamma)$, but it is obviously true if $a = \min \text{Dom}(\gamma)$. Similarly, we can conclude that $\gamma|_{[c, \infty[\cap (\text{Dom}(\gamma))} \in \text{Traj}(Y)$.

We have therefore established that, if there is some interval $J \in \mathcal{F}_X$ which is open, then $\gamma \in \text{Traj}(Y * X)$.

Now suppose there is a $J \in \mathcal{F}_X$ which is not open. If $J = [a, b]$ then, necessarily, $J = \text{Dom}(\gamma)$ (because J is a connected component of I_X , and so J is relatively open in $\text{Dom}(\gamma)$). But then Theorem 3.9 implies that $\gamma \in \text{Traj}(Y * X)$. Now suppose that $J = [a, b[$. Then $a = \min \text{Dom}(\gamma)$, and $\gamma(b) \in S$. Theorem 3.9 implies that $\gamma|_{[a, b]} \in \text{Traj}(Y * X)$. If $\gamma|_{[a, b]}$ is a strict $Y * X$ -trajectory, with switching at a time $c \in]a, b[$, then the conjugate point argument used before shows that γ cannot have a switching at b , and that $\gamma|_{([c, \infty[\cap (\text{Dom}(\gamma))}]$ is a Y -trajectory. Therefore $\gamma \in \text{Traj}(Y * X)$, if $\gamma|_{[a, b]}$ is a strict $Y * X$ -trajectory.

If $\gamma|_{[a, b]}$ is not a strict $Y * X$ -trajectory, then $\gamma|_{[a, b]} \in \text{Traj}(X \vee Y)$. The possibility that $\gamma|_{[a, b]} \in \text{Traj}(X)$ is excluded, because $\gamma|_{[a, b]} \subseteq U_X$, and $\gamma(b) \in S$. So $\gamma|_{[a, b]} \in \text{Traj}(Y)$. Hence we have shown that, if $J \in \mathcal{F}_X$ and $J = [a, b[$, then $\gamma \in \text{Traj}(Y * X)$, except possibly if $\gamma|_J \in \text{Traj}(Y)$. Similarly, if $J \in \mathcal{F}_X$ and $J =]a, b[$, then $\gamma \in \text{Traj}(Y * X)$, except possibly in the case when $\gamma|_J \in \text{Traj}(X)$.

Now suppose that $\gamma \notin \text{Traj}(Y * X)$. Then one and only one of the following can happen:

- (6.V.i) \mathcal{F}_X is empty,
- (6.V.ii) \mathcal{F}_X consists of exactly one interval J , of the form $[\min \text{Dom}(\gamma), b[$, such that $\gamma|_J \in \text{Traj}(Y)$,
- (6.V.iii) \mathcal{F}_X consists of exactly one interval J , of the form $]a, \max \text{Dom}(\gamma)]$, such that $\gamma|_J \in \text{Traj}(X)$,
- (6.V.iv) \mathcal{F}_X consists of exactly two intervals of the form $J_1 = [\min \text{Dom}(\gamma), b[$, and $J_2 =]a, \max \text{Dom}(\gamma)]$ (with $b \leq a$), such that $\gamma|_{J_1} \in \text{Traj}(Y)$, and that $\gamma|_{J_2} \in \text{Traj}(X)$.

Suppose that \mathcal{F}_X is empty. Since \mathcal{F}_Y consists of at most two intervals, and each of these must contain an endpoint of γ , we see that γ is a concatenation of at most three pieces $\gamma_1, \gamma_2, \gamma_3$, such that $\text{Dom}(\gamma_i) = \text{Clos } J_i, J_i \in \mathcal{F}_Y$, for $i = 1, 3$. But then γ_2 must be entirely contained in S . Since $\dot{\gamma}_2(t)$ is a convex combination of $X(\gamma_2(t))$ and $Y(\gamma_2(t))$, and these vectors are multiples of each other and not tangent to S , it follows that $\dot{\gamma}_2(t) = 0$ for $t \in \text{Dom}(\gamma_2)$. So $\text{Dom}(\gamma_2)$ must be reduced to a single point. Therefore, if \mathcal{F}_Y consists of two intervals, then these intervals must be of the form $[\min \text{Dom}(\gamma), a[$ and $]a, \max \text{Dom}(\gamma)]$, for some interior point a of $\text{Dom}(\gamma)$, such that $\gamma(a) \in S$. By Theorem 3.9, $\gamma|((\text{Dom } \gamma) \cap]-\infty, a])$ is a $Y * X$ -trajectory. Since $\gamma|[\min \text{Dom}(\gamma), a[$ is contained in U_Y , there cannot be a piece $\gamma| [a - \delta, a]$ which is a Y -trajectory. So $\gamma|[\min \text{Dom}(\gamma), a]$ is an X -trajectory. Similarly, $\gamma|[a, \max \text{Dom}(\gamma)]$ is a Y -trajectory. Therefore $\gamma \in \text{Traj}(Y * X)$. If \mathcal{F}_Y consists of one interval only, then one shows as before that γ cannot contain a piece that is contained in S , and therefore one can also conclude that $\gamma \in \text{Traj}(Y * X)$. Finally, the case when $\mathcal{F}_Y = \emptyset$ is impossible, since it would imply that γ is entirely contained in S .

Summarizing, we have shown that, if \mathcal{F}_X is empty, then $\gamma \in \text{Traj}(Y * X)$. If $\gamma \notin \text{Traj}(Y * X)$ then, as indicated earlier, one of (6.V) hold. Moreover, we now know that (6.V.i.) cannot hold. Suppose (6.V.ii) holds. Let J, b be as in the statement of (6.V.ii). Then $\gamma| [b, \max \text{Dom}(\gamma)]$ is a time-optimal trajectory in U whose corresponding \mathcal{F}_X is empty. So we can apply what we have proved so far to this curve, and conclude that $\gamma| [b, \max \text{Dom}(\gamma)] \in \text{Traj}(Y * X)$. Therefore $\gamma \in \text{Traj}(Y * X * Y)$. We claim that γ cannot contain a switching from a Y - to an X -trajectory. Indeed, such a switching cannot occur at a point in $U_X \cup U_Y$, because $f > 0$ on $U_X \cup U_Y$ (cf. Theorem 3.9). Moreover, it cannot occur at a point $p \in S$, because $X(p)$ and $Y(p)$ are dependent and point in opposite directions, and so Lemma 5.1 applies. (The statement of Lemma 5.1 requires the extra hypothesis that, at p , the X - and Y -trajectories have a tangency of finite order. However, this hypothesis was only used to conclude that $f > 0$ on the region \mathcal{R} described in Remark 5.1. As explained in Remark 5.1, the conclusion of Lemma 5.1 still holds without the extra hypothesis, if we know that $f > 0$ on \mathcal{R} . Also, in our case, it is clear that $\mathcal{R} \subseteq U_Y$, and so $f > 0$ on \mathcal{R} .) So the switching from Y to X is excluded, and it follows that $\gamma \in \text{Traj}(Y * X)$.

If (6.V.iii) holds, then one proves in exactly the same way that $\gamma \in \text{Traj}(Y * X)$. Finally, if (6.V.iv) holds, then the restriction γ' of γ to $[b, \max \text{Dom}(\gamma)]$ is a time-optimal trajectory for which (6.V.ii) holds, and so $\gamma' \in \text{Traj}(Y * X)$. So $\gamma \in \text{Traj}(Y * X * Y)$. Exactly as before, the Y - to X -switching is excluded, and so $\gamma \in \text{Traj}(Y * X)$.

This completes the proof that $\text{Opt}^1(\Sigma \upharpoonright U) \subseteq \text{Traj}(Y * X)$, if $f > 0$ throughout $U_X \cup U_Y$. Similarly, if $f < 0$ on $U_X \cup U_Y$, it follows that $\text{Opt}^1(\Sigma \upharpoonright U) \subseteq \text{Traj}(X * Y)$. The proof of Theorem 6.2 is now complete. \square

We now consider the case when (6.IV.b) holds. A regular INOA S for which there is a U such that (6.I), (6.III) and (6.IV.b) hold will be said to be *of the turnpike type*. If, in addition, the function Δ_A never vanishes on S , then S will be called a *turnpike* (cf. Fig. 3).

THEOREM 6.3. *Let S be a regular INOA of the turnpike type. Let U be such that (6.I), (6.III) and (6.IV.b) hold. Then every time-optimal trajectory γ in U is a concatenation $\gamma_1 * \gamma_2 * \gamma_3$ of at most three pieces, such that γ_1 and γ_3 are in $\text{Traj}(X \vee Y)$, and that γ_2 is entirely contained in S . If S is not a turnpike, then every $\gamma \in \text{Opt}^1(\Sigma \upharpoonright U)$ is bang-bang with at most one switching (i.e., the γ_2 piece cannot occur).*

Proof. Exactly as in the proof of Theorem 6.2, let I_X be the set of those $t \in \text{Dom}(\gamma)$ such that $\gamma(t) \in U_X$, and let \mathcal{F}_X be the set of connected components of I_X . Define

The proof of Theorem 6.4 will be given in § 7.

We conclude this section with a corollary summarizing much of the information that we have just obtained. Let us say that p is a *near-ordinary point* if it is an ordinary point, or it belongs to a regular INOA S which, if it is of the antiturnpike type, is nondegenerate. We define a *Z-trajectory* to be a nontrivial trajectory γ of Σ entirely contained in $\text{Clos } S$ for some regular INOA S that is a turnpike.

COROLLARY 6.5. *Let p be a near ordinary point. Then p has a neighborhood U such that*

$$\text{Opt}^1(\Sigma \upharpoonright U) \subseteq [\text{Traj}(X \vee Y \vee Z)]^5.$$

Proof. If S is a barrier in U for some U for which (6.I) hold, then the conclusion follows from Theorem 6.1. If S is nonbarrier, then (6.III) holds. Let U be such that (6.I) hold. If (6.IV.a) holds, then we get the desired conclusion from Theorem 6.2. If S is of the turnpike type, then we can use Theorem 6.3. Finally, if S is of the antiturnpike type, then we use Theorem 6.4. \square

7. Nondegenerate antiturnpikes. In this section we prove Theorem 6.4. We assume that

- (7.I) S is a nondegenerate regular INOA of the turnpike type (i.e. a regular INOA for which (6.III) holds, and for which there exist a U such that (6.I) and (6.IV.c) hold, and a $k \geq 0$ such that (6.VI.a) or (6.VI.b) holds).

We pick a fixed U for which (6.I) and (6.IV.c) hold. Also, we fix a $k \geq 0$ for which either (6.VI.a) or (6.VI.b) holds, and we assume, without loss of generality, that it is actually (6.VI.a) that holds.

We first prove a local result. Pick a point $p \in S$. Since (6.III) holds, we know in particular that $X(p)$ is not tangent to S . So we can find a neighborhood $W(p)$ of p which is the domain of a square coordinate chart such that, if we use x, y to denote the coordinates, then

- (7.II.a) $W(p)$ is identified, via the coordinates, with the square,
 (7.1) $\text{Sq}(\varepsilon(p)) = \{(x, y) : |x| < \varepsilon(p), |y| < \varepsilon(p)\}$ for some $\varepsilon(p) > 0$,
 (7.II.b) $W(p) \subseteq U$,
 (7.II.c) p has coordinates $(0, 0)$,
 (7.II.d) $S \cap W(p) = \{(x, y) \in \text{Sq}(\varepsilon(p)) : x = 0\}$,
 (7.II.e) $X \upharpoonright W(p) = \partial_x$.

Let

$$(7.2) \quad Y = \alpha \partial_x + \beta \partial_y$$

be the expression of Y in our coordinates. Since $X(p)$ and $Y(p)$ point to opposite sides of S , we must have $\alpha(0, 0) < 0$ and therefore we may assume (after making $\varepsilon(p)$ smaller, if needed) that $\alpha < 0$ throughout $W(p)$. A simple computation shows that

$$(7.3) \quad \Delta_A = \frac{1}{2}\beta$$

and that

$$(7.4) \quad \Delta_B = \frac{1}{4}((\alpha - 1)(\partial_x \beta) - (\partial_x \alpha)\beta).$$

Let σ be the function

$$(7.5) \quad \sigma(q) = \frac{\beta(q)}{4(\alpha(q) - 1)},$$

so that $\sigma(q)$ is equal to one quarter of the slope of $G(q)$.

Let us write $q_1 \sim q_2$ if q_1, q_2 are points in $W(p)$ lying in the same X -trajectory (i.e., having the same y coordinate) and are conjugate along it. Then $(x_1, y_1) \sim (x_2, y_2)$ if and only if

$$(7.6.a) \quad y_1 = y_2$$

and

$$(7.6.b) \quad \sigma(x_1, y_1) = \sigma(x_2, y_2).$$

We now study the solutions of (7.6). Let

$$(7.7) \quad \zeta(x_1, x_2, y) = \sigma(x_1, y) - \sigma(x_2, y).$$

Then the study of the solutions of (7.6) is obviously equivalent to that of the solutions of

$$(7.8) \quad \zeta(x_1, x_2, y) = 0.$$

First observe that

$$(7.9) \quad \partial_x \sigma = \Delta_B.$$

It follows from (7.II.d), (7.II.b) and (6.I.b) that $\Delta_B \neq 0$ at all points $(x, y) \in W(p)$ such that $x \neq 0$. Therefore, if we fix any y , the function $x \rightarrow \sigma(x, y)$ is strictly monotonic for x in each of the intervals $]-\varepsilon(p), 0[,]0, \varepsilon(p)[$. So (7.7) has no solutions x_1, x_2, y for which $x_1 \neq x_2$ but they are both ≥ 0 , or both ≤ 0 . So all we have to do is study the solutions of (7.7) for which $x_1 < 0 < x_2$.

Since (7.9) holds, we have, for $i = 1, 2, \dots$

$$(7.10) \quad \partial_x^i \sigma = X^{i-1} \Delta_B.$$

Therefore, in view of (6.VI.a), we can write

$$(7.11) \quad \sigma(x, y) = \sigma(0, y) + \frac{x^{k+1}}{(k+1)!} (X^k \Delta_B)(0, y) + x^{k+2} \xi(x, y)$$

where ξ is a smooth function. This gives

$$(7.12) \quad \zeta(x_1, x_2, y) = \frac{x_1^{k+1} - x_2^{k+1}}{(k+1)!} (X^k \Delta_B)(0, y) + x_1^{k+2} \xi(x_1, y) - x_2^{k+2} \xi(x_2, y).$$

Since $\xi(x_2, y) - \xi(x_1, y)$ vanishes when $x_2 = x_1$, we can write

$$(7.13) \quad \xi(x_2, y) = \xi(x_1, y) + (x_2 - x_1) \eta(x_1, x_2, y),$$

where η is smooth. Therefore we have

$$(7.14) \quad \zeta(x_1, x_2, y) = (x_1 - x_2) \tilde{\zeta}(x_1, x_2, y)$$

where

$$(7.15) \quad \tilde{\zeta}(x_1, x_2, y) = \frac{1}{(k+1)!} \left(\sum_{i=0}^k x_1^i x_2^{k-i} \right) (X^k \Delta_B)(0, y) + O((|x_1| + |x_2|)^{k+1}).$$

Hence the solutions of $\zeta = 0$ for which $x_1 \neq x_2$ are exactly the solutions of $\tilde{\zeta} = 0$. We consider two cases:

(7.III.i) Δ_A has different signs on opposite sides of S , and

(7.III.ii) Δ_A has the same sign on both sides of S .

If (7.III.i) holds, then β has opposite signs on the two sides of S , and so $\sigma(x_1, y)\sigma(x_2, y) < 0$ whenever $x_1 < 0 < x_2$. (Recall that $\alpha - 1 < 0$.) Hence $\sigma(x_1, y) = \sigma(x_2, y)$ is impossible if $x_1 < 0 < x_2$, and we conclude that there are no solutions of $\zeta(x_1, x_2, y) = 0$ for which $x_1 < 0 < x_2$.

Now suppose that (7.III.ii) holds. Since $f = -\Delta_B/\Delta_A$, and f has opposite signs on the two sides of S , the function Δ_B must change sign on the y axis. This requires, in particular, that k be odd. But then $k > 0$, and so

$$(7.16) \quad \tilde{\zeta}(0, 0, y) = 0.$$

Define a function μ by

$$(7.17) \quad \mu(x, \rho, y) = \tilde{\zeta}(x, \rho x, y).$$

Then μ is well defined and smooth for $(x, y) \in \text{Sq } \frac{1}{2}\varepsilon(p)$, $|\rho| < 2$.

We have

$$(7.18) \quad \mu(x, \rho, y) = \frac{x^k}{(k+1)!} \nu(x, \rho, y),$$

where

$$(7.19) \quad \nu(x, \rho, y) = \left(\sum_{i=0}^k \rho^{k-i} \right) (X^k \Delta_B)(0, y) + O(|x|).$$

Since k is odd, we have

$$(7.20) \quad \nu(0, -1, y) = 0.$$

On the other hand, $\rho = -1$ is a simple zero of the polynomial

$$(7.21) \quad \sum_{i=0}^k \rho^{k-i} = \frac{\rho^{k+1} - 1}{\rho - 1}$$

and so

$$(7.22) \quad \frac{\partial \nu}{\partial \rho}(0, -1, y) \neq 0.$$

By the Implicit Function theorem, there exists a $\delta(p)$ such that $0 < \delta(p) < \frac{1}{2}\varepsilon(p)$, and that there is a smooth function $\rho: \text{Sq}(\delta(p)) \rightarrow \mathbb{R}$ such that

$$(7.23.a) \quad \rho(0, y) = -1 \quad \text{for } |y| < \delta(p),$$

that

$$(7.23.b) \quad |\rho(x, y)| < 2 \quad \text{for } (x, y) \in \text{Sq}(\delta(p))$$

and that

$$(7.24) \quad \nu(x, \rho(x, y), y) = 0 \quad \text{for } (x, y) \in \text{Sq}(\delta(p)).$$

Let us make $\delta(p)$ smaller, if necessary, and assume that

$$(7.25) \quad \rho(x, y) \leq -\frac{1}{2} \quad \text{for } (x, y) \in \text{Sq}(\delta(p)).$$

Let $\psi: \text{Sq}(\delta(p)) \rightarrow \mathbb{R}$ be given by

$$(7.26) \quad \psi(x, y) = x\rho(x, y).$$

Then

$$(7.27) \quad |\psi(x, y)| < \varepsilon(p)$$

and

$$(7.28) \quad x \cdot \psi(x, y) < 0 \quad \text{for } (x, y) \in \text{Sq}(\delta(p)), \quad x \neq 0.$$

It is clear that

$$(7.29) \quad \tilde{\zeta}(x, \psi(x, y), y) = 0.$$

So, for each $(x, y) \in \text{Sq}(\delta(p))$ such that $x \neq 0$, the point $(\psi(x, y), y)$ lies in the same X -trajectory as (x, y) , and is conjugate to (x, y) along this trajectory. We claim that, conversely, if $(x_1, y) \sim (x_2, y)$, and if $x_1 < 0 < x_2$, and $(x_i, y) \in \text{Sq}(\delta(p))$ for $i = 1, 2$, then $x_1 = \psi(x_2, y)$ and $x_2 = \psi(x_1, y)$. To prove this, simply observe that, since $\Delta_B(x, y) \neq 0$ for $0 < x < \varepsilon(p)$, the function $x \rightarrow \sigma(x, y)$, $0 < x < \varepsilon(p)$ is strictly monotonic on $]0, \varepsilon(p)[$ (in view of (7.9)), and so the equation $\sigma(x, y) = \sigma(x_1, y)$ has at most one solution $x \in]0, \varepsilon(p)[$. Since $0 < \psi(x_1, y) < \varepsilon(p)$, and $\sigma(\psi(x_1, y), y) = \sigma(x_1, y)$, the facts that $\sigma(x_2, y) = \sigma(x_1, y)$ and $0 < x_2 < \delta(p)$ imply that $x_2 = \psi(x_1, y)$. The proof that $x_1 = \psi(x_2, y)$ is similar.

We now summarize what we have proved. Let $\delta(p)$ be as above if (7.III.ii) holds. If (7.III.i) holds, let $\delta(p) = \varepsilon(p)$. We have shown that

- (7.IV.a) If (7.III.i) holds, then there are no solutions (x_1, x_2, y) of (7.8) such that $|x_i| < \delta(p)$, $|y| < \delta(p)$ and that $x_1 \neq x_2$,
- (7.IV.b) if (7.III.ii) holds, then there is a smooth function $\psi: \text{Sq}(\delta(p)) \rightarrow]-\varepsilon(p), \varepsilon(p)[$ such that (7.28) holds and that, if $|x_i| < \delta(p)$, $|y| < \delta(p)$, then x_1, x_2, y satisfy (7.8) if and only if $x_1 = x_2$ or $x_1 = \psi(x_2, y)$.

We now prove, using (7.IV), that no strict $Y * X * Y$ -trajectory in $\text{Sq}(\delta'(p))$ can be time-optimal, if $0 < \delta'(p) \leq \delta(p)$, and $\delta'(p)$ is small enough. To prove this, we let $\delta'(p)$ be a number that satisfies $0 < \delta'(p) \leq \delta(p)$, plus an extra condition that will be stated later, and we suppose that γ is a strict $Y * X * Y$ -trajectory in $\text{Sq}(\delta'(p))$, and that γ is time-optimal. We will reach a contradiction. If (7.III.i) holds, then this is quite easy. Indeed, if q_1 and q_2 are the points where the switchings of γ occur, then it is clear that $q_1 \neq q_2$ but $q_1 \sim q_2$. If $q_i = (x_i, \bar{y})$, $i = 1, 2$, then (x_1, x_2, \bar{y}) satisfies (7.8), and $x_1 \neq x_2$. By (7.IV.a), this is a contradiction. (Hence we may take, e.g., $\delta'(p) = \delta(p)$.)

We now consider the other case, namely, when (7.III.ii) holds. We let γ , the q_i , x_i , \bar{y} be as above, and we let t_i be such that $\gamma(t_i) = q_i$. Clearly, we may assume that $t_1 < t_2$. Then we have $x_1 < 0 < x_2$, and $x_2 = \psi(x_1, \bar{y})$. Let Q be the open set

$$(7.30) \quad Q = \{(x, y) \in \text{Sq}(\delta(p)) : |\psi(x, y)| < \delta(p)\}.$$

Then $(x_1, \bar{y}) \in Q$. Let $K: Q \rightarrow Q$ be the map

$$(7.31) \quad K(x, y) = (\psi(x, y), y).$$

Let $\gamma_1 = \gamma|_{[t_1 - \theta, t_1]}$, where $\theta > 0$ is so small that γ_1 is entirely contained in Q , and that γ_1 never intersects the y axis. Let

$$(7.32) \quad \tilde{\gamma}_2 = K \circ \gamma_1.$$

Then $\gamma_1(t) \sim \tilde{\gamma}_2(t)$ for all $t \in [t_1 - \theta, t_1]$. If we prove that, after a suitable reparametrization, $\tilde{\gamma}_2$ becomes a trajectory γ_2 of Σ , which is not bang-bang, then we will have reached a contradiction. (Indeed: let $\hat{\gamma}$ be the concatenation $\gamma_2 * \hat{\gamma}_1$, where $\hat{\gamma}_1$ is the X -trajectory from $\gamma_1(t_1 - \theta)$ to $K(\gamma_1(t_1 - \theta))$. Then Lemma 3.6 implies that

$$(7.33) \quad T(\gamma \upharpoonright [t_1 - \theta, t_2]) = T(\hat{\gamma}).$$

Since γ is time-optimal, we conclude that $\hat{\gamma}$ is time-optimal, which is a contradiction, because γ_2 is a piece of $\hat{\gamma}$, which is not bang-bang but is contained in $\Omega(A) \cap \Omega(B)$.)

So in order to reach a contradiction, we must show that $\tilde{\gamma}_2$ can be reparametrized so as to produce a trajectory of Σ which is not bang-bang. This will follow if we prove that, for each $t \in [t_1 - \theta, t_1]$, the vector $\dot{\tilde{\gamma}}_2(t)$ is a strict convex combination of $X(\tilde{\gamma}_2(t))$ and $Y(\tilde{\gamma}_2(t))$. Since we are assuming that (7.III.ii) holds, we may assume, without loss of generality, that $\beta(x, y) > 0$ for all $(x, y) \in \text{Sq}(\delta(p))$, $x \neq 0$. If K_* denotes the differential of the map K , we must show that $K_*(\gamma_1(t)) \cdot Y(\gamma_1(t))$ is a strict convex combination of $X(\tilde{\gamma}_2(t))$ and $Y(\tilde{\gamma}_2(t))$, for $t \in [t_1 - \theta, t_1]$. Since $\beta(q) > 0$ and $\alpha(q) < 0$ for all $q \in \text{Sq}(\delta(p))$, this conclusion will follow if we show that, for $q \in Q$, the vector $K_*(q) \cdot Y(q)$ has strictly positive components. Clearly:

$$(7.34) \quad K_*(q) \cdot Y(q) = \begin{bmatrix} (\partial_x \psi)(q) & (\partial_y \psi)(q) \\ 0 & 1 \end{bmatrix} \begin{bmatrix} \alpha(q) \\ \beta(q) \end{bmatrix}.$$

Since $\psi(0, y) = 0$, it is clear that $(\partial_x \psi)(0, 0) = 0$. On the other hand, it follows easily from the definition of ψ that $(\partial_x \psi)(0, 0) = -1$. Since $\alpha(p) < 0$, there is an r such that $0 < r \leq \delta(p)$ and that the function

$$(7.35) \quad q \rightarrow ((\partial_x \psi) \cdot \alpha + (\partial_y \psi) \cdot \beta)(q)$$

is strictly positive on $\text{Sq}(r)$. We now take this r to be $\delta'(p)$. Then $K_*(q) \cdot Y(q)$ is a strict convex combination of $X(K(q))$ and $Y(K(q))$ for all $q \in Q \cap \text{Sq}(\delta'(p))$. As explained above, this shows that γ is not time-optimal.

We now let $\delta'(p)$ be such that $0 < \delta'(p) \leq \delta(p)$, and that no strict $Y * X * Y$ -trajectory in $\text{Sq}(\delta'(p))$ is time-optimal. We want to analyze the structure of an arbitrary time-optimal trajectory γ in $\text{Sq}(\delta'(p))$. First, it is clear that γ cannot contain a strict $Y * X * Y$ piece. Therefore, if $J \subseteq \text{Dom}(\gamma)$ is a maximal interval such that $\gamma \upharpoonright J$ is an X -trajectory, it follows that either

$$(7.V.a) \quad J \text{ contains one of the endpoints of } \text{Dom}(\gamma), \text{ or}$$

$$(7.V.b) \quad J = [t_1, t_2], \text{ and one of the points } \gamma(t_1), \gamma(t_2) \text{ is in } S.$$

If (7.V.b) holds, then the points $\gamma(t_1), \gamma(t_2)$ satisfy $\gamma(t_1) \sim \gamma(t_2)$, $\gamma(t_1) \neq \gamma(t_2)$. If we let $\gamma(t_i) = (x_i, y)$, then either $x_1 = 0$ or $x_2 = 0$. Moreover, $x_1 \neq x_2$, and $\sigma(x_1, y) = \sigma(x_2, y)$. But then we have reached a contradiction, because (7.IV) clearly implies that no solution (x_1, x_2, y) of (7.8) can satisfy $x_1 \neq x_2$ but be such that $x_1 = 0$ or $x_2 = 0$.

So (7.V.b) cannot hold.

We conclude that, if γ is a time-optimal trajectory in $\text{Sq}(\delta'(p))$, then $\gamma = \gamma_3 * \gamma_2 * \gamma_1$, where γ_1 and γ_3 are X -trajectories, and γ_2 contains no X -trajectory. We now have to determine the structure of those time-optimal trajectories γ that contain no X -trajectory. Suppose γ has this property. Let J be a maximal subinterval of $\text{Dom}(\gamma)$ such that $\gamma \upharpoonright J$ is a Y -trajectory. Let $J = [t_1, t_2]$. Since Y always points to the same side of S , it is not possible for both $\gamma(t_1)$ and $\gamma(t_2)$ to be in S . If $\gamma(t_i) \notin S$, then t_i must be an endpoint of $\text{Dom}(\gamma)$. (Otherwise there would be a $\theta > 0$ such that $\gamma \upharpoonright [t_i - \theta, t_i + \theta]$ is entirely contained in $\Omega(A) \cap \Omega(B)$, and therefore belongs to

$\text{Traj}((X \vee Y)^2)$. Since γ contains no X -trajectory, it follows that $\gamma|_{[t_i - \theta, t_i + \theta]}$ is a Y -trajectory, contradicting the maximality of J .) So γ is a concatenation $\gamma_3 * \gamma_2 * \gamma_1$, where γ_1 and γ_3 are Y -trajectories, and γ_2 contains no X - or Y -trajectory. Then it follows that γ_2 is entirely contained in S . If γ_2 is nontrivial then it follows, in particular, that S is an arc where Δ_A never vanishes. (Otherwise, since S is a regular INOA, we would have $\Delta_A \equiv 0$ on S . So, if $q \in S$, the vectors $X(q)$ and $Y(q)$ are linearly dependent and have opposite directions. Therefore no convex combination of $X(q)$ and $Y(q)$ can be tangent to S at q , unless it equals zero. So no nontrivial trajectory of Σ can be contained in S .) So β never vanishes on $\{(x, y) \in \text{Sq}(\delta(p)) : x = 0\}$, and therefore β never vanishes on $\text{Sq}(\delta(p))$. So f is well defined on $\text{Sq}(\delta(p))$. Since S is of the antiturnpike type, we have

$$(7.36.a) \quad f(x, y) > 0 \quad \text{for } x > 0,$$

$$(7.36.b) \quad f(x, y) < 0 \quad \text{for } x < 0.$$

It is easy to see that, if $\tau_1 < \tau_2$, $\tau_i \in \text{Dom}(\gamma_2)$, and $\tau_2 - \tau_1$ is small enough, then there is a $Y * X$ -trajectory $\tilde{\gamma}_2$ that goes from $\gamma_2(\tau_1)$ to $\gamma_2(\tau_2)$ and is entirely contained in $\text{Sq}(\delta(p))$. Then $\tilde{\gamma}_2(t) \in \{(x, y) : x > 0\}$ for t in the interior of $\text{Dom}(\tilde{\gamma}_2)$. If $\beta > 0$ on $\text{Sq}(\delta(p))$, then the curve $\gamma_2^{-1} * \tilde{\gamma}_2$ is oriented counterclockwise, and $f > 0$ in the interior on the region \mathcal{R} it encloses. Therefore we can apply Lemma 3.10 and conclude that $T(\tilde{\gamma}_2) < T(\gamma_2)$. If $\beta < 0$ on $\text{Sq}(\delta(p))$, then $\tilde{\gamma}_2^{-1} * \gamma_2$ is oriented counterclockwise, and therefore Lemma 3.10 again implies that $T(\tilde{\gamma}_2) < T(\gamma_2)$. In either case, γ_2 is not optimal, and we have reached a contradiction. So γ_2 is actually trivial, and γ is a Y -trajectory.

So we have shown that, if γ is a time-optimal trajectory in $\text{Sq}(\delta'(p))$, then it follows that $\gamma \in \text{Traj}(X * Y * X)$. That is, we have shown the following.

LEMMA 7.1. *Let S be a nondegenerate regular INOA of the antiturnpike type. Suppose that there is a $k \geq 0$ for which (6.VI.a) holds. Then for every $p \in S$ there exists an open arc $S'(p) \subseteq S$, containing p , for which there is a $\delta'(p) > 0$ with the property that, if*

$$(7.37) \quad V(p) = \{\Phi_i^X(q) : q \in S'(p), |t| < \delta'(p)\},$$

*then every time-optimal trajectory in $V(p)$ is an $X * Y * X$ -trajectory.*

Using Lemma 7.1, it is now easy to conclude the proof of Theorem 6.4. Let S satisfy the hypotheses of Lemma 7.1. For each p , pick an $S'(p)$ and a $\delta'(p)$ with the property of Lemma 7.1, and define $V(p)$ by (7.37). Make $\delta'(p)$ smaller, if necessary, so that $V(p) \subseteq U$. Let

$$(7.38) \quad U_0 = \bigcup \{V(p) : p \in S\}.$$

Then U_0 is open, and $S \subseteq U_0 \subseteq U$. Let $\gamma \in \text{Traj}(\Sigma|_{U_0})$, and suppose that γ is time-optimal. Then, locally, γ is contained in the sets $V(p)$, and therefore γ is regular bang-bang. We now claim that γ cannot contain a strict $Y * X * Y$ piece. Indeed, suppose that γ_1 is a strict $Y * X * Y$ -trajectory that is contained in U_0 and is time-optimal. Let t_1, t_2 be the switching times, labelled so that $t_1 < t_2$. Let $\tilde{\gamma}_1 = \gamma_1|_{[t_1, t_2]}$. Then $\tilde{\gamma}_1$ is contained in η , where η is the maximal integral curve of $X|_U$ such that $\eta(t_1) = \gamma_1(t_1)$. It follows from Theorem 3.9 that $\tilde{\gamma}_1$ must intersect S at some point q . (Otherwise, since $U_0 \subseteq U$ and $U - S \subseteq \Omega(A) \cap \Omega(B)$, we would conclude that $\tilde{\gamma}_1$ is contained in $\Omega(A) \cap \Omega(B)$ and therefore that $\gamma_1|_{[t_1 - \theta, t_2 + \theta]}$ is also contained in $\Omega(A) \cap \Omega(B)$ for some $\theta > 0$. But then Theorem 3.9 would imply that $\gamma_1|_{[t_1 - \theta, t_2 + \theta]}$ is not time-optimal, and therefore γ_1 would not be time-optimal either.)

Since $U - S$ has two connected components, and $X(p)$ points towards the same component of $U - S$ for all $p \in S$, it is clear that there is at most one time $\bar{t} \in \text{Dom}(\eta)$ such that $\eta(\bar{t}) \in S$. Since $\eta(t) = q$ for some $t \in [t_1, t_2]$, we conclude that there is a unique $\bar{t} \in \text{Dom}(\eta)$ such that $\eta(\bar{t}) \in S$, and that this \bar{t} satisfies $t_1 \leq \bar{t} \leq t_2$, and $\eta(\bar{t}) = q$.

If q' is any point on η , it is clear that $q' \in U_0$ if and only if $|\tau| < \delta'(p)$ for some $p \in S$ such that $q \in S'(p)$, where $\tau \in \mathbb{R}$ is the unique time such that $q' = \Phi_\tau^X(q)$. From this it follows easily that there is a $p \in S$ such that $\tilde{\gamma}_1$ is contained in $V(p)$. Then $\gamma_1|_{[t_1 - \theta, t_2 + \theta]}$ is contained in $V(p)$ for sufficiently small $\theta > 0$. But then γ_1 is not time-optimal, since $V(p)$ satisfies the conclusion of Lemma 7.1.

So γ is regular bang-bang, and it does not contain any strict $Y * X * Y$ piece. Therefore γ is an $X * Y * X$ -trajectory. This completes the proof of Theorem 6.4. \square

REFERENCES

- [Ba] M. BAYTMAN, *The Optimal Synthesis of Trajectories In The Plane*, Zinatne, Riga (USSR), 1971. (In Russian.)
- [Bo] V. BOLTYANSKII, *Sufficient conditions for optimality and the justification of the dynamic programming method*, SIAM J. Control, 4 (1966), pp. 326–361.
- [Br 1] P. BRUNOVSKY, *Every normal linear system has a regular time-optimal synthesis*, Math. Slovaca, 28 (1978), pp. 81–100.
- [Br 2] ———, *On the structure of optimal feedback systems*, Proc. Int. Congr. of Mathematicians, Helsinki, 1978, pp. 841–846.
- [Br 3] ———, *Regular synthesis for the linear quadratic control problem with linear control constraint*, J. Differential Equations, 38 (1980), pp. 317–343.
- [Ha] R. M. HARDT, *Stratification of real analytic maps and images*, Invent. Math., 28 (1975), pp. 193–208.
- [Hi 1] H. HIRONAKA, *Subanalytic sets*, in Number Theory, Algebraic Geometry and Commutative Algebra, in Honor of Y. Akizuki, Kinokuniya, Tokyo, 1973.
- [Hi 2] ———, *Subanalytic sets*, in Lecture Notes of Istituto Matematico “Leonida Tonelli”, Pisa, 1965.
- [Loj] S. LOJASIEWICZ, *Ensembles semi-analytiques*, Lecture notes at I.H.E.S., Bures-sur-Yvette, 1965.
- [Su 1] H. J. SUSSMANN, *Subanalytic sets and feedback control*, J. Differential Equations, 31 (1979), pp. 31–52.
- [Su 2] ———, *Analytic stratifications and control theory*, Proc. Int. Congr. of Mathematicians, Helsinki, 1978, pp. 865–871.
- [Su 3] ———, *Les semigroupes sousanalytiques et la régularité des commandes en boucle fermée*, Astérisque (Soc. Math. de France), 75–76 (1980), pp. 219–226.
- [Su 4] ———, *A bang-bang theorem with bounds on the number of switchings*, this Journal, 17 (1979), pp. 629–651.
- [Su 5] ———, *Bounds on the number of switchings for trajectories of piecewise analytic vector fields*, J. Differential Equations, 43 (1982), pp. 399–418.
- [Su 6] ———, *Subanalytic sets and regular synthesis*, to appear.
- [Su 7] ———, *The structure of time-optimal trajectories for single-input systems in the plane: the general real analytic case*, this Journal, 25 (1987), to appear.
- [Su 8] ———, *Regular synthesis for time-optimal control of single-input real-analytic systems in the plane*, this Journal, 25 (1987), to appear.
- [SuJ] H. J. SUSSMANN AND V. JURDJEVIC, *Controllability of nonlinear systems*, J. Differential Equations, 12 (1972), pp. 95–116.
- [Ta] M. TAMM, *Subanalytic sets in the calculus of variation*, Acta Math., 146 (1981), pp. 167–199.

ASYMPTOTICALLY EFFICIENT SELF-TUNING REGULATORS*

T. L. LAI† AND C. Z. WEI‡

Abstract. This paper studies the problem of adaptive regulation of linear systems with white-noise disturbances. The apparent dilemma between the control objective and the need of information for parameter estimation is resolved by occasional use of white-noise probing inputs and by a reparametrization of the model. Insights into the question concerning how often and when such probing inputs should be introduced are provided by the concept of "asymptotic efficiency," which quantifies the asymptotically minimal cost due to parameter ignorance, or equivalently, due to the infeasibility of using the optimal regulator that assumes knowledge of the system parameters. Asymptotically efficient adaptive regulators are constructed by making use of certain basic properties of adaptive predictors involving recursive least squares for the reparametrized model.

Key words. adaptive control, linear systems, least squares identification, self-tuning, probing inputs, asymptotic efficiency

AMS(MOS) subject classifications. Primary 62L20, 93E20; secondary 60G42, 93E10

1. Introduction. We study herein the problem of efficient adaptive control for the linear system

$$(1.1) \quad y_n = \alpha_1 y_{n-1} + \cdots + \alpha_p y_{n-p} + \beta_1 u_{n-d} + \cdots + \beta_q u_{n-d-q+1} + \varepsilon_n,$$

where the y 's represent outputs and the u 's represent inputs at various times and the ε 's represent random disturbances. We assume that $\{\varepsilon_n\}$ is a martingale difference sequence with respect to an increasing sequence of σ -fields $\{\mathcal{F}_n\}$ (i.e., ε_n is \mathcal{F}_n -measurable and $E(\varepsilon_n | \mathcal{F}_{n-1}) = 0$ for all n) such that

$$(1.2) \quad \sup_n E(|\varepsilon_n|^\gamma | \mathcal{F}_{n-1}) < \infty \quad \text{a.s. (almost surely) for some } \gamma > 2.$$

Moreover, assume that $\beta_1 \neq 0$ and that the polynomials

$$(1.3) \quad A(s) = s^p - \alpha_1 s^{p-1} - \cdots - \alpha_p \quad \text{and} \quad B(s) = \beta_1 s^{q-1} + \beta_2 s^{q-2} + \cdots + \beta_q$$

have all zeros inside the unit circle.

An important problem in the literature is how to choose the inputs u_i , on the basis of current and past observations $y_i, y_{i-1}, u_{i-1}, \dots$, to regulate the outputs, say, such that $\sum_{i=1}^N y_{i+d}^2$ is minimized in some sense, at least in the long run as $N \rightarrow \infty$. Although one may in principle use a Bayesian approach, putting a prior distribution on the unknown parameters and applying dynamic programming when the disturbances ε_n are independent and identically distributed (i.i.d.) with a known common distribution (say for example, normal) and when a fixed horizon N is given, the dynamic programming equations are prohibitively difficult to handle, both computationally and analytically (cf. [1]).

A much more practical approach is that of the *self-tuning regulator*, proposed by Åström and Wittenmark [2]. This "self-tuning" idea is to start by considering the case where the system parameters are known, for which the optimal controller can be

* Received by the editors April 8, 1985; accepted for publication (in revised form) March 11, 1986. This work was supported by the National Science Foundation, and in the case of the first author, also by the Army Research Office and the John Simon Guggenheim Foundation.

† Department of Statistics, Columbia University, New York, New York 10027.

‡ Department of Mathematics, University of Maryland, College Park, Maryland 20742.

explicitly found by the separation principle, and then to substitute the system parameters in the optimal controller by their least squares estimates. To fix the ideas, consider first the case of unit delay (i.e., $d = 1$). Here the optimal controller assuming known parameters is given by

$$(1.4) \quad u_n = -(\alpha_1 y_n + \cdots + \alpha_p y_{n-p+1} + \beta_2 u_{n-1} + \cdots + \beta_q u_{n-q+1}) / \beta_1,$$

or equivalently, by the equation

$$(1.4') \quad \theta' \varphi_n = 0,$$

where $\theta = (\alpha_1, \dots, \alpha_p, \beta_1, \dots, \beta_q)'$ and $\varphi_n = (y_n, \dots, y_{n-p+1}, u_n, \dots, u_{n-q+1})'$. Letting $\theta_n = (\hat{\alpha}_{1,n}, \dots, \hat{\alpha}_{p,n}, \hat{\beta}_{1,n}, \dots, \hat{\beta}_{q,n})' = (\sum_1^n \varphi_i \varphi_i')^{-1} \sum_1^n \varphi_i y_{i+1}$ be the least squares estimate of θ based on $\varphi_1, y_2, \dots, \varphi_n, y_{n+1}$, the self-tuning regulator that assumes no prior knowledge of the parameters is given by

$$(1.5) \quad u_n = -(\hat{\alpha}_{1,n-1} y_n + \cdots + \hat{\alpha}_{p,n-1} y_{n-p+1} + \hat{\beta}_{2,n-1} u_{n-1} + \cdots + \hat{\beta}_{q,n-1} u_{n-q+1}) / \hat{\beta}_{1,n-1},$$

or implicitly by

$$(1.5') \quad \theta'_{n-1} \varphi_n = 0.$$

As is well known, θ_i can be determined for $i > \tau = \inf \{n: \sum_1^n \varphi_i \varphi_i' \text{ is nonsingular}\}$ by the recursive algorithm

$$(1.6a) \quad \theta_i = \theta_{i-1} + (y_{i+1} - \theta'_{i-1} \varphi_i) P_i \varphi_i,$$

$$(1.6b) \quad P_i = P_{i-1} - P_{i-1} \varphi_i \varphi_i' P_{i-1} / (1 + \varphi_i' P_{i-1} \varphi_i),$$

noting that $P_i^{-1} = P_{i-1}^{-1} + \varphi_i \varphi_i'$ and therefore $P_n = (\sum_1^n \varphi_i \varphi_i')^{-1}$ for $n \geq \tau$ (cf. [3]).

When the random disturbances are i.i.d. normal with mean 0 and variance $\sigma^2 > 0$, the least squares estimate θ_i is the same as the maximum likelihood estimator of θ . Thus, in this case, the self-tuning regulator (1.5) is simply the maximum likelihood estimator of the (unobservable) optimal controller (1.4) at every stage. While (1.5) may well represent one's best approximation to (1.4), how good the approximation is depends on how much information there is to estimate θ . A basic issue concerning this self-tuning approach is, therefore, whether there is enough information for adequately estimating θ . If there is not enough information, then perhaps one should use some inputs to probe the system for more information, instead of adhering to a poorly estimated version of (1.4'). While the use of white-noise perturbations to improve parameter estimation and the future controls is a well-known idea (cf. [2]), it remains an open problem concerning how often such probing inputs should be introduced. We provide an answer to this problem herein, making use of the concept of *asymptotic efficiency*, introduced in [4], of adaptive regulators.

If we use the (unrealizable) optimal regulator (1.4) that assumes knowledge of θ , then the outputs are $y_i = \varepsilon_i$. In view of this, we define the "regret," at stage n , of a sequence of inputs $\{u_i\}$ to be

$$(1.7) \quad R_n = \sum_1^n (y_i - \varepsilon_i)^2$$

(cf. [4]). An input sequence is called "globally convergent" if

$$(1.8) \quad \lim_{n \rightarrow \infty} R_n / n = 0 \quad \text{a.s.}$$

As shown in [4], considerable insight into how rapidly the convergence in (1.8) can occur, or equivalently, how small can the order of magnitude for the regret R_n be, is

provided by studying the following auxiliary Bayes problem in which β_1 is assumed known.

Suppose that in addition to (1.3), the polynomials

$$(1.9) \quad B(s) = \beta_1 s^{q-1} + \cdots + \beta_q \text{ and } \alpha_1 s^{p-1} + \cdots + \alpha_p \text{ are relatively prime,}$$

and that the random disturbances ε_n are i.i.d. normal with mean 0 and variance $\sigma^2 > 0$. Assume that β_1 is known, and let

$$(1.10) \quad \lambda = -\beta_1^{-1}(\alpha_1, \cdots, \alpha_p, \beta_2, \cdots, \beta_q)'$$

have a truncated normal prior distribution π , which is the restriction of a normal distribution with mean 0 and covariance matrix $\sigma_0^2 I_h$ to the stability region defined by (1.3), where $h = p + q - 1$ and I_h denotes the $h \times h$ identity matrix. Define the $(p + q - 1)$ -dimensional vector

$$(1.11) \quad \psi_n = (y_n, \cdots, y_{n-p+1}, u_{n-1}, \cdots, u_{n-q+1})'$$

and let $z_{n+1} = y_{n+1} - \beta_1 u_n$. Letting $\lambda_{n-1}^* = E_\pi[\lambda | \psi_1, z_2, \cdots, \psi_{n-1}, z_n]$ be the Bayes estimate of λ , it is noted in [4] that for any choice of the input sequence $\{u_i\}$

$$(1.12) \quad E_\pi \left\{ \sum_{i=1}^n (u_i - \lambda' \psi_i)^2 \right\} \geq E_\pi \sum_{i=1}^n \{(\lambda_{i-1}^* - \lambda)' \psi_i\}^2.$$

Restricting to input sequences $\{u_i\}$ that are globally convergent and that satisfy the growth condition

$$(1.13) \quad u_n^2 = O(n^\delta) \quad \text{for some } 0 < \delta < 1,$$

it is shown in [4] that for such input sequences

$$(1.14) \quad \beta_1^2 \sum_{i=1}^n \{(\lambda_{i-1}^* - \lambda)' \psi_i\}^2 \sim \sigma^2 h \log n \quad \text{a.s.,}$$

and therefore, by Fatou's Lemma and (1.12),

$$(1.15) \quad E_\pi \left\{ \sum_{i=1}^n (y_{i+1} - \varepsilon_{i+1})^2 \right\} = \beta_1^2 E_\pi \left\{ \sum_{i=1}^n (u_i - \lambda' \psi_i)^2 \right\} \geq (\sigma^2 h + o(1)) \log n.$$

In view of (1.12), (1.14) and (1.15), an input sequence $\{u_i\}$ is called "asymptotically efficient" in [4] if its regret is asymptotically no larger than the order $\sigma^2 h \log n$. More generally, without any distributional assumptions on the unobservable ε_n and without assuming any prior knowledge of β_1 and the other parameters, [4] provides the following definition of asymptotically efficient regulators.

DEFINITION. Consider the linear system (1.1) with $d = 1$ and such that $\beta_1 \neq 0$ and (1.3), (1.9) hold, and in which $\{\varepsilon_n\}$ is a martingale difference sequence satisfying (1.2). Let $v = \limsup_{n \rightarrow \infty} E(\varepsilon_n^2 | \mathcal{F}_{n-1})$. A sequence of inputs $\{u_n\}$ is said to be asymptotically efficient if

$$(1.16) \quad \limsup_{n \rightarrow \infty} R_n / \log n \leq v(p + q - 1) \quad \text{a.s.}$$

The construction of asymptotically efficient self-tuning regulators is given in § 3 below. This construction involves a refinement of the method developed in [5] for

constructing adaptive control schemes whose regrets satisfy the weaker conclusion

$$(1.17) \quad \limsup_{n \rightarrow \infty} R_n / \log n < \infty \quad \text{a.s.},$$

but in the more general setting of colored noise (i.e., where the ε_n in (1.1) is a finite moving average of martingale differences).

It is interesting to compare the definition (1.8) of global convergence with (1.17) and with the definition (1.16) of asymptotic efficiency, and to review in this connection some recent results on adaptive control of the linear system (1.1). By making use of stochastic approximation techniques instead of least squares to estimate the unknown θ in (1.4'), Goodwin, Ramadge and Caines [6] showed that the resultant rule is globally convergent, without any stability assumption on the polynomial $A(s)$ and even for the more general setting of colored noise. Their fundamental work opened an active area of research during the past few years. In particular, Sin and Goodwin [7] established the global convergence of an alternative control rule that involves a hybrid between stochastic approximation and recursive least squares. In another direction, Caines and Lafontaine [8], and Chen and Caines [9], proposed the introduction of white noise perturbations into the control algorithm of Goodwin et al. [6] to persistently excite the system. They showed that this approach of "continually disturbed control" leads to strongly consistent estimates by the method of least squares or by stochastic approximation, although the resultant control rule is no longer globally convergent. Subsequently, Becker, Kumar and Wei [10] showed that without introducing white noise perturbations, the parameter estimates in the original algorithm of Goodwin et al. [6] converge to a limit which is a random multiple of θ and which differs from θ with strictly positive probability. A survey of these and other results has been provided by Kumar [11].

Although these recent developments have established the global convergence of adaptive regulators that are generated by stochastic approximation recursions (i.e., having scalar gains), it remains an unsettled problem whether the classical least squares recursions also lead to globally convergent schemes which may even have better rates of convergence than those generated by stochastic approximation. After all, the least squares method coincides with the asymptotically efficient method of maximum likelihood for parameter estimation in the case where the ε_n are i.i.d. normal with mean 0 and variance σ^2 . It is therefore natural to expect that the self-tuning regulator (1.5) should be nearly optimal when there is adequate information to estimate θ . However, if the information content for estimating θ remains consistently low, then there is no guarantee that the self-tuning equation (1.5') would eventually be a good approximation to (1.4') and the self-tuning idea may not work (cf. [4]).

In [5], under the stability assumption (1.3) but in the more general setting of colored noise that is assumed to be a.s. bounded, we proposed a simple criterion for deciding at every stage, on the basis of the observed data, whether there is adequate information to estimate θ . When the data show inadequate information, instead of adhering rigidly to the self-tuning equation to determine the output u_n , we proposed to introduce white-noise perturbations to improve the information content of the design. We showed in [5] that the number of these perturbations up to stage n is kept within $O(\log n)$ and that the regret R_n of the regulator satisfies (1.17), which is much stronger than the global convergence property (1.8).

In the present setting of white noise, we show in this work that by a refinement of the algorithm used in [5] the property (1.17) can be further sharpened to (1.16), which gives the "best constant" for the left-hand side of (1.17). Central to this refinement

are the reparametrization of (1.1) as

$$(1.18) \quad y_{n+1} = \beta_1(u_n - \lambda' \psi_n) + \varepsilon_{n+1}$$

and the use of least-squares-type recursions estimating λ instead of the recursions (1.6) estimating θ , where λ and ψ_n are defined in (1.10) and (1.11).

Suppose that β_1 in (1.18) is known, and assume that the ε_n are i.i.d. normal with mean 0 and variance $v > 0$. Then the Bayes estimate λ_{i-1}^* of λ with respect to the truncated normal prior distribution π is asymptotically equivalent to the least squares estimate $\hat{\lambda}_{i-1}$ (cf. [4]), and in view of (1.12), (1.14) and (1.15), an asymptotically optimal order for the regret R_n is $v(p+q-1) \log n$, as given in (1.16). Without assuming β_1 to be known and without any distributional assumptions on ε_n , we show in § 3 how to construct a consistent estimate of β_1 and how to “self-tune” the reparametrized model (1.18) to attain the asymptotically optimal order $v(p+q-1) \log n$ for R_n .

Section 2 develops certain basic lemmas that will be needed in the sequel. Section 4 extends the methods and results of § 3 to the case of general delay.

2. Preliminary lemmas. A basic tool in §§ 3 and 4 is the following asymptotic property, proved in [4, Thm. 2], of adaptive predictors in the regression model (1.18).

LEMMA 1. *Suppose that in the regression model (1.18), $\{\varepsilon_n\}$ is a martingale difference sequence with respect to an increasing sequence of σ -fields $\{\mathcal{F}_n\}$ such that (1.2) holds, and that u_n and ψ_n are \mathcal{F}_n -measurable, ψ_n and λ being $h \times 1$ vectors. Assume that $\beta_1 \neq 0$ and that*

$$(2.1) \quad \lambda_{\min} \left(\sum_1^n \psi_i \psi_i' \right) \rightarrow \infty \quad \text{and} \quad \psi_n' \left(\sum_1^n \psi_i \psi_i' \right)^{-1} \psi_i \rightarrow 0 \quad \text{a.s.}$$

Let b_n be \mathcal{F}_n -measurable such that $b_n \rightarrow \beta_1$ a.s. For $i > \tau = \inf \{n: \sum_1^n \psi_i \psi_i' \text{ is nonsingular}\}$, define the recursions

$$(2.2a) \quad \lambda_i = \lambda_{i-1} + (u_i - b_i^{-1} y_{i+1} - \lambda_{i-1}' \psi_i) P_i \psi_i,$$

$$(2.2b) \quad P_i^{-1} = P_{i-1}^{-1} + \psi_i \psi_i'.$$

Let

$$(2.3) \quad e_{n+1} = y_{n+1} - b_n(u_n - \lambda'_{n-1} \psi_n)$$

be the prediction error of the \mathcal{F}_n -measurable predictor $\hat{y}_{n+1} = b_n(u_n - \lambda'_{n-1} \psi_n)$ of y_{n+1} . Then

$$(2.4) \quad \sum_{\tau+1}^n (e_{i+1} - \varepsilon_{i+1})^2 + \beta_1^2 (\lambda_n - \lambda)' P_n^{-1} (\lambda_n - \lambda) \\ \leq \{ \limsup_{i \rightarrow \infty} E(\varepsilon_i^2 | \mathcal{F}_{i-1}) + o(1) \} \log \det \left(\sum_1^n \psi_i \psi_i' \right) + o \left(\sum_1^n (u_i - \lambda' \psi_i)^2 \right) \quad \text{a.s.}$$

Remark. In (2.1) and the sequel, we use λ_{\min} and λ_{\max} to denote the minimum and maximum eigenvalues of a symmetric matrix. For the special case $b_i = \beta_1$ for all i , the recursions (2.2) coincide with the recursions defining the least squares estimates of λ in the model (1.18) when β_1 is known (cf. (1.6)). In general, without assuming β_1 to be known, we replace β_1 in these least squares recursions by consistent estimates b_i , and this leads to (2.2).

The next three lemmas are on the asymptotic behavior of the linear system (1.1) in which $\beta_1 \neq 0$ and (1.3) holds. Let

$$Y_n = (y_n, \dots, y_{n-p+1})', \quad U_n = (u_n, \dots, u_{n-q+2})', \\ A = \begin{pmatrix} \alpha_1 & \dots & \alpha_{p-1} & \alpha_p \\ & I_{p-1} & & 0 \end{pmatrix}, \quad B = \begin{pmatrix} -\beta_2/\beta_1 & \dots & -\beta_q/\beta_1 \\ & I_{q-2} & & 0 \end{pmatrix}.$$

While the above definition assumes that $q \geq 2$, define $U_n = u_n$ and $B = 0$ in the case $q = 1$. Likewise set $Y_n = y_n$ and $A = 0$ when $p = 0$. By (1.1),

$$(2.5) \quad \begin{aligned} Y_n &= A Y_{n-1} + \left(\sum_{j=1}^q \beta_j u_{n-d-j+1} + \varepsilon_n, 0, \dots, 0 \right)', \\ U_n &= B U_{n-1} - \beta_1^{-1} \left(\sum_{j=0}^p \alpha_j y_{n+d-j} + \varepsilon_{n+d}, 0, \dots, 0 \right)' \quad \text{where } \alpha_0 = -1. \end{aligned}$$

Regarding a $k \times k$ matrix H as a linear operator, define $\|H\| = \sup_{\|x\|=1} \|Hx\| = \lambda_{\max}^{1/2}(H'H)$. By (1.3), there exist $0 < \rho < 1$ and $C > 0$ such that $\|A^n\| \leq C\rho^n$ and $\|B^n\| \leq C\rho^n$ for all n , and therefore (2.5) implies the following (cf. [12, p. 361]).

LEMMA 2. For $n > m$,

$$(i) \quad \|Y_n\| \leq C\rho^{n-m} \|Y_m\| + C \sum_{i=0}^{n-m-1} \rho^i \left\{ |\varepsilon_{n-i}| + \sum_{j=1}^q |\beta_j| |u_{n-d-i-j+1}| \right\},$$

$$(ii) \quad \|U_n\| \leq C\rho^{n-m} \|U_m\| + C|\beta_1|^{-1} \sum_{i=0}^{n-m-1} \rho^i \left\{ |\varepsilon_{n+d-i}| + \sum_{j=0}^p |\alpha_j| |y_{n+d-i-j}| \right\}.$$

Consequently, there exists $D > 0$ such that for all $n > \nu > m$,

$$(iii) \quad \sum_{i=\nu}^n \|Y_i\|^2 \leq D \left\{ (\rho^{\nu-m} \|Y_m\|)^2 + \sum_{i=m+1}^n \varepsilon_i^2 + \sum_{i=m+2-d-q}^{n-d} u_i^2 \right\},$$

$$(iv) \quad \sum_{i=\nu}^n \|U_i\|^2 \leq D \left\{ (\rho^{\nu-m} \|U_m\|)^2 + \sum_{i=m+1}^n \varepsilon_{i+d}^2 + \sum_{i=m+1-d-p}^{n+d} y_i^2 \right\}.$$

As an immediate application of Lemma 2 (i), we have the following.

LEMMA 3. Let c_n be a nondecreasing sequence of positive constants such that $c_n \rightarrow \infty$. Suppose that $u_n = O(c_n)$ and $\varepsilon_n = O(c_n)$ a.s. Then $y_n = O(c_n)$ a.s.

LEMMA 4. Suppose that $\{\varepsilon_n\}$ is a martingale difference sequence with respect to $\{\mathcal{F}_n\}$ such that $\sup_n E(e^{\gamma|\varepsilon_n|} | \mathcal{F}_{n-1}) < \infty$ for some $\gamma > 0$. Then $\varepsilon_n = O(\log n)$ a.s.

Proof. Taking $b > 1/\gamma$, note that

$$\sum_{n=1}^{\infty} P\{|\varepsilon_n| \geq b \log n | \mathcal{F}_{n-1}\} \leq \left(\sup_n E(e^{\gamma|\varepsilon_n|} | \mathcal{F}_{n-1}) \right) \sum_1^{\infty} n^{-\gamma b} < \infty.$$

Hence, by the conditional Borel-Cantelli lemma (cf. [13, p. 55]), $P\{|\varepsilon_n| < b \log n \text{ for all large } n\} = 1$. \square

LEMMA 5. Let $\{c_n\}, \{c_n^*\}$ be nondecreasing sequences of positive numbers such that

$$(2.6) \quad c_n^* \rightarrow \infty, \quad c_{n+d}^* = O(c_n^*), \quad c_n / c_n^* \rightarrow \infty.$$

Suppose that

$$(2.7) \quad |u_n| \leq c_n \quad \text{and} \quad \varepsilon_n = O(c_n^*) \quad \text{a.s.}$$

Let $n_0 < n_0 + m_0 < n_1 < \dots < n_i < n_i + m_i < n_{i+1} < \dots$ be positive integer-valued random variables such that

$$(2.8) \quad m_i / \log c_{n_i} \rightarrow \infty \quad \text{a.s.}$$

Letting $I = \cup_i \{n_i + 1, \dots, n_i + m_i\}$, suppose that

$$(2.9) \quad \sup_{n \in I} |u_n| / c_n^* < \infty \quad \text{a.s.}$$

(i) If $b > 1/|\log \rho|$, then

$$(2.10) \quad \max \{ \|Y_n\| / c_n^* : n_i + b \log c_{n_i} \leq n \leq n_i + m_i + d \} = O(1) \quad \text{a.s.}$$

(ii) Let $\psi_n = (y_n, \dots, y_{n-r}, u_{n-1}, \dots, u_{n-s})'$. Suppose that for some $K > 0$, with probability 1

$$(2.11) \quad |u_n| \leq \min \{c_n, K\|\psi_n\|\} \quad \text{for all large } n \notin I,$$

$$(2.12) \quad |y_{n+d} - \varepsilon_{n+d}| \chi\{K\|\psi_n\| < c_n \text{ and } n \notin I\} = o(c_n),$$

where χ_A denotes the indicator function of an event A , i.e., $\chi_A(\omega) = 0$ if $\omega \notin A$ and $\chi_A(\omega) = 1$ if $\omega \in A$. Then $\psi_n = o(c_n)$ a.s., and consequently, with probability 1,

$$(2.13) \quad |u_n| \leq K\|\psi_n\| < c_n \quad \text{for all large } n \notin I.$$

Proof. (i) By Lemma 3, $\|Y_{n_i}\| = O(c_{n_i})$ a.s. Since $\rho^b < e^{-1}$,

$$(2.14) \quad \|Y_{n_i}\| \rho^{b \log c_{n_i}} \rightarrow 0 \quad \text{a.s.}$$

From (2.7), (2.9), (2.14) and Lemma 2(i) (where we set $m = n_i$), (2.10) follows.

(ii) By restricting to an event with probability 1, we can assume that (2.7) and (2.12) hold everywhere (instead of a.s.). Given $\delta > 0$ sufficiently small so that

$$(2.15) \quad \delta K(r+s+1) < 1 \quad \text{and} \quad C|\beta_1|^{-1}(1-\rho)^{-1} \left(1 + \sum_{j=0}^p |\alpha_j|\right) \delta < 1,$$

we can choose by (2.12) and (2.7) sufficiently large T_0 so that for all $T \geq T_0$,

$$(2.16) \quad n_T + m_T < n \leq n_{T+1} \quad \text{and} \quad K\|\psi_n\| < c_n \Rightarrow |y_{n+d}| \leq \delta^2 c_n.$$

We now prove by induction on n that

$$(2.17) \quad |u_n| \leq \delta c_n \quad \text{and} \quad |y_{n+d}| \leq \delta^2 c_n$$

for all n satisfying $n_T + \frac{1}{3}m_T \leq n \leq n_{T+1}$, provided that $T (\geq T_0)$ is sufficiently large, as specified below.

By (i), (2.6) and (2.9), (2.17) holds for $n_T + \frac{1}{3}m_T \leq n \leq n_T + m_T$, provided that $T \geq T_0$ is sufficiently large. Now take $\nu \geq n_T + m_T$ with $\nu < n_{T+1}$, and assume that (2.17) holds for all $(n_T + \frac{1}{3}m_T \leq) n \leq \nu$. Since

$$\begin{aligned} K\|\psi_{\nu+1}\| &\leq K\{|y_{\nu+1}| + \dots + |y_{\nu+1-r}| + |u_\nu| + \dots + |u_{\nu+1-s}|\} \\ &\leq K(r+s+1)\delta c_\nu \quad \text{by induction hypothesis,} \\ &< c_{\nu+1} \quad \text{by (2.15),} \end{aligned}$$

it then follows from (2.16) that

$$(2.18) \quad |y_{\nu+1+d}| \leq \delta^2 c_{\nu+1}.$$

By (2.6) and (2.7), when T is sufficiently large,

$$(2.19) \quad |\varepsilon_n| \leq \delta^2 c_n \quad \text{for all } n \geq n_T.$$

From Lemma 2(ii) (with $m = n_T + \lceil \frac{1}{2}m_T \rceil$ and T sufficiently large), (2.15), (2.18), and the induction hypothesis, it then follows that $|u_{\nu+1}| \leq \delta c_{\nu+1}$. This shows that (2.17) also holds for $n = \nu + 1$.

We have therefore shown that $u_n \chi_{\{n \notin I\}} = o(c_n)$. By (2.6) and (2.9), $u_n \chi_{\{n \in I\}} = o(c_n)$. Hence $u_n = o(c_n)$. Since $\varepsilon_n = o(c_n)$, it then follows from Lemma 2(i) that $y_n = o(c_n)$. Therefore $\psi_n = o(c_n)$. \square

Lemma 5 provides a useful technical device for the analysis of the control rules developed in the next two sections, where the inputs will be truncated by c_n and white-noise probing signals will be used to excite the system at stages $n \in I$.

The following lemma provides consistent estimation of the parameters of the linear system (1.1) in the presence of “occasional excitation” in the inputs. Define

$$(2.20) \quad \varphi_n = (y_{n+d-1}, \dots, y_{n+d-p}, u_n, \dots, u_{n-q+1})', \quad \theta = (\alpha_1, \dots, \alpha_p, \beta_1, \dots, \beta_q)',$$

and rewrite (1.1) as the regression model

$$(2.21) \quad y_{n+d} = \theta' \varphi_n + \varepsilon_{n+d}.$$

LEMMA 6. Consider the linear system (1.1) in which $\{\varepsilon_n\}$ is a martingale difference sequence satisfying (1.2). Let τ be a (finite) stopping time and let $n_1 < n_1 + m_1 < \dots < n_i < n_i + m_i < n_{i+1} < \dots$ be nonrandom positive integers such that

$$(2.22) \quad \lim_{i \rightarrow \infty} m_i = \infty \quad \text{and} \quad \limsup_{i \rightarrow \infty} m_i / m_{i-1} < \infty.$$

Let $I = \{1, \dots, \tau\} \cup (\cup_{i=1}^{\infty} \{n_i + 1, \dots, n_i + m_i\})$, and let $\#_n$ denote the number of elements in I that are $\leq n$. Defining φ_n and θ as in (2.20), suppose that for $n \in I$

$$(2.23) \quad u_n \text{ is a random variable independent of } y_n, \varphi_{n-d}, \dots, \varphi_1 \\ \text{such that } u_n \text{ has mean 0, variance } V > 0 \text{ and } |u_n| \leq M.$$

Assume that $\liminf_{n \rightarrow \infty} E(\varepsilon_n^2 | \mathcal{F}_{n-1}) > 0$ a.s. and that

$$(2.24) \quad \|\varphi_n\| = O(c_n) \quad \text{a.s.,}$$

where $\{c_n\}$ is a nondecreasing sequence of positive constants satisfying (2.8) and

$$(2.25) \quad c_n \rightarrow \infty, c_n \sim c_{n+1} \quad \text{and} \quad \log c_n = O(\log \#_n).$$

Then

$$(2.26) \quad \liminf_{n \rightarrow \infty} \lambda_{\min} \left(\sum_{t \in I, t \leq n} \varphi_t \varphi_t' \right) / \#_n > 0 \quad \text{a.s.}$$

Moreover, defining the least squares estimate

$$(2.27) \quad \theta_n^* = \left(\sum_{t \in I, t \leq n} \varphi_t \varphi_t' \right)^{-1} \left(\sum_{t \in I, t \leq n} \varphi_t y_{t+d} \right)$$

based only on (φ_t, y_{t+d}) with $t \in I$, we have

$$(2.28) \quad \theta_n^* - \theta = O((\log \#_n)^{1/2} / \#_n^{1/2}) \quad \text{a.s.,}$$

and therefore $\theta_n^* \rightarrow \theta$ a.s.

Proof. In view of (2.21), we can apply Theorem 1 of [14] to conclude that

$$(2.29) \quad \theta_n^* - \theta = O \left(\left\{ \log \lambda_{\max} \left(\sum_{t \in I, t \leq n} \varphi_t \varphi_t' \right) \right\}^{1/2} / \lambda_{\min}^{1/2} \left(\sum_{t \in I, t \leq n} \varphi_t \varphi_t' \right) \right) \quad \text{a.s.}$$

By (2.24),

$$\lambda_{\max} \left(\sum_{t \in I, t \leq n} \varphi_t \varphi_t' \right) \leq \sum_{t \in I, t \leq n} \|\varphi_t\|^2 = O(c_n^2 \#_n) \quad \text{a.s.,}$$

and therefore

$$\log \lambda_{\max} \left(\sum_{t \in I, t \leq n} \varphi_t \varphi_t' \right) = O(\log \#_n) \quad \text{a.s.}$$

by (2.25). Hence (2.28) follows from (2.29) if it can be shown that (2.26) holds.

To prove (2.26), first note by (1.2) that

$$(2.30) \quad \sum_{t \in I, t \leq n} \varepsilon_t^2 = \sum_{t \in I, t \leq n} E(\varepsilon_t^2 | \mathcal{F}_{t-1}) + o(\#_n) = O(\#_n) \quad \text{a.s.}$$

(cf. [14, p. 157]). Let $I^* = \cup_i \{n_i + [\frac{1}{2}m_i], \dots, n_i + m_i\}$. Since $\sup_{t \in I} |u_t| \leq M$, it then follows from Lemma 2(iii) (with $m = n_i + d + q$, $v = n_i + [\frac{1}{2}m_i]$) and (2.30), (2.24), (2.8) that

$$\sum_{t \in I^*, t \leq n} \|\varphi_t\|^2 = \sum_{i: n_i \leq n} \sum_{t=n_i + [\frac{1}{2}m_i]}^{\min\{n_i + m_i, n\}} \|\varphi_t\|^2 = O(\#_n) \quad \text{a.s.}$$

In view of this, (2.22), (2.23) and the assumption that $\liminf_{n \rightarrow \infty} E(\varepsilon_n^2 | \mathcal{F}_{n-1}) > 0$ a.s., we can apply Theorem 5 and Corollary 2 of [12] to conclude that

$$(2.31) \quad \liminf_{n \rightarrow \infty} \lambda_{\min} \left(\sum_{t \in I^*, t \leq n} \varphi_t \varphi_t' \right) / \#_n > 0 \quad \text{a.s.},$$

which implies (2.26). \square

Remark. Instead of using the stochastic projection theory of [12] as in the preceding proof, we can apply Corollary 2 of [15] on excitation properties to derive (2.31). However, the same proof as above using the stochastic projection theory of [12] can also be used to prove the following stronger result than (2.26): For $r \geq 0$ define

$$(2.32) \quad \varphi_{n,r} = (\varphi_n', u_{n+1}, \dots, u_{n+r})'.$$

Then

$$(2.33) \quad \liminf_{n \rightarrow \infty} \lambda_{\min} \left(\sum_{t \in I, t \leq n \text{ and } t \equiv j \pmod{d}} \varphi_{t,r} \varphi_{t,r}' \right) / \#_n > 0 \quad \text{a.s.}$$

for $j = 0, \dots, d-1$. We will make use of (2.33) in § 4.

3. An asymptotically efficient input sequence in the case of unit delay. In this section we consider the case $d = 1$ first for bounded ε_n and then extend the result to the unbounded case. We will assume that $\{\varepsilon_n\}$ is a martingale difference sequence satisfying (1.2) and

$$(3.1) \quad \liminf_{n \rightarrow \infty} E(\varepsilon_n^2 | \mathcal{F}_{n-1}) > 0 \quad \text{a.s.}$$

Suppose that the ε_n are bounded with probability 1. This implies that the optimal input sequence (assuming known parameters), defined by (1.4) and with outputs $y_n = \varepsilon_n$, is bounded a.s. (Lemma 2(ii)). We therefore restrict the inputs u_n within $[-c_n, c_n]$ such that

$$(3.2) \quad c_n \sim \log \log n.$$

This implies by Lemma 3 that (2.24) holds.

A basic idea in our construction of asymptotically efficient regulators is the introduction of white-noise probing inputs to obtain strongly consistent estimates, as in Lemma 6. Let $\varphi_n = (y_n, \dots, y_{n-p+1}, u_n, \dots, u_{n-q+1})'$, $V_n = \sum_1^n \varphi_i \varphi_i'$ and define

$$(3.3) \quad \tau = \inf \{n: V_{n-1} \text{ is nonsingular and } \hat{\beta}_{1,n-1} \neq 0\},$$

where $\hat{\beta}_{1,t}$ is the least squares estimate of β_1 based on $\varphi_1, y_2, \dots, \varphi_t, y_{t+1}$ as in (1.5). Let $\rho > 1$, $\delta > 0$ and take positive integers n_i and m_i such that

$$(3.4) \quad n_i = \exp \{i^\rho (1 + o(1))\},$$

$$(3.5) \quad m_i \sim (\log i)^\delta.$$

Let

$$(3.6) \quad I = \{1, \dots, \tau\} \cup \left(\bigcup_{i=1}^{\infty} \{n_i + 1, \dots, n_i + m_i\} \right)$$

be the set of stages when white-noise probing inputs are introduced. Thus, at stage $n \in I$, the input u_n satisfies (2.23) with $d = 1$. Let $\#_n$ be the number of probing inputs up to stage n (i.e., $\#_n$ = number of elements in I that are $\leq n$). Then from (3.4) and (3.5),

$$(3.7) \quad \#_n \sim (\log n)^{1/\rho} (\rho^{-1} \log \log n)^\delta \left(\sim \sum_{i: n_i \leq n} m_i \right).$$

By (3.2), (3.5) and (3.7), conditions (2.8), (2.22) and (2.25) also hold. Define the least squares estimate θ_n^* based only on probing inputs as in (2.27), and modify the estimate as follows to ensure a nonzero estimate of β_1 : Set $\tilde{\theta}_{\tau-1} = \theta_{\tau-1}^*$ and define for $n \geq \tau$

$$(3.8) \quad \begin{aligned} \tilde{\theta}_n &= \theta_n^* && \text{if } \beta_{1,n}^* \neq 0, \\ &= \tilde{\theta}_{n-1} && \text{if } \beta_{1,n}^* = 0. \end{aligned}$$

Since $\beta_1 \neq 0$, it follows from Lemma 6 that

$$(3.9) \quad \tilde{\theta}_n - \theta = O((\log \#_n)^{1/2} / \#_n^{1/2}) \quad \text{a.s.}$$

Another key idea in the construction of asymptotically efficient regulators is the reparametrization of (1.1) in the form $y_{t+1} = \beta_1(u_t - \lambda' \psi_t) + \varepsilon_{t+1}$, where λ and ψ_t are defined in (1.10) and (1.11). Making use of the consistent estimates $\tilde{\theta}_t = (\tilde{\alpha}_{1,t}, \dots, \tilde{\alpha}_{p,t}, \tilde{\beta}_{1,t}, \dots, \tilde{\beta}_{q,t})'$ and noting that $\tilde{\beta}_{1,t} \neq 0$, define

$$(3.10) \quad \tilde{\lambda}_t = -\tilde{\beta}_{1,t}^{-1}(\tilde{\alpha}_{1,t}, \dots, \tilde{\alpha}_{p,t}, \tilde{\beta}_{2,t}, \dots, \tilde{\beta}_{q,t})'$$

as an auxiliary estimate of $\lambda = -\beta_1^{-1}(\alpha_1, \dots, \alpha_p, \beta_2, \dots, \beta_q)'$ and conclude from (3.9) that

$$(3.11) \quad \tilde{\lambda}_t - \lambda = O((\log \#_t)^{1/2} / \#_t^{1/2}) \quad \text{a.s.}$$

Although the auxiliary estimates $\tilde{\lambda}_t$ are strongly consistent, they are not asymptotically efficient since they only use a small (albeit well designed or excited) subset of the observations. Nevertheless, because of their strong consistency, they can be used to provide diagnostic checks on other estimators. One such estimator, introduced in Lemma 1, is defined recursively for $t \geq \tau$ by

$$(3.12a) \quad \lambda_t = \lambda_{t-1} + (u_t - \tilde{\beta}_{1,t-1}^{-1} y_{t+1} - \lambda'_{t-1} \psi_t) P_t \psi_t,$$

$$(3.12b) \quad P_t^{-1} = P_{t-1}^{-1} + \psi_t \psi_t', \quad \left(P_{\tau-1} = \left(\sum_{i=1}^{\tau-1} \psi_i \psi_i' \right)^{-1}, \lambda_{\tau-1} = \tilde{\lambda}_{\tau-1} \right),$$

and in turn suggests the adaptive regulator $\lambda'_{n-1} \psi_n$ at stage $n \notin I$. The adaptive choice, given in (3.13) below, between $\lambda'_{n-1} \psi_n$ and $\tilde{\lambda}'_{n-1} \psi_n$ yields an asymptotically efficient (as defined by (1.16)) sequence of self-tuning regulators. This is the content of the next theorem.

THEOREM 1. Consider the linear system (1.1) with $d = 1$ and $\beta_1 \neq 0$ and such that the stability assumption (1.3) holds. Suppose that $\{\varepsilon_n\}$ is a martingale difference sequence such that (1.2), (3.1) hold and $\sup_n |\varepsilon_n| < \infty$ a.s. Defining I by (3.3)–(3.6), introduce probing inputs (2.23) at stages $n \in I$. Let $\{c_n\}$ be a nondecreasing sequence of positive numbers satisfying (3.2). Define the auxiliary estimates $\tilde{\lambda}_t$ of λ by (3.8)–(3.10) and the recursive estimates λ_t by (3.12), where λ and ψ_t are given in (1.10) and (1.11). Let $A > 0$, and let $\#_n$ be the number of probing inputs up to stage n . For $n \notin I$, define

$$(3.13) \quad u_n = \begin{cases} (-c_n) \vee (\lambda'_{n-1} \psi_n \Lambda c_n) & \text{if } |(\lambda_{n-1} - \tilde{\lambda}_{n-1})' \psi_n| \leq A \#_n^{-1/2} (\log \#_n) \|\psi_n\|, \\ (-c_n) \vee (\tilde{\lambda}'_{n-1} \psi_n \Lambda c_n) & \text{otherwise,} \end{cases}$$

where \vee and Λ denote maximum and minimum respectively. The regret R_n of this input sequence $\{u_n\}$ satisfies (1.16).

Proof. Since $|u_n| \leq c_n$ and $\sup_n |\varepsilon_n| < \infty$ a.s., it follows from Lemma 3 that $\|\varphi_n\| = O(c_n)$ a.s. and therefore

$$(3.14) \quad \|\psi_n\| = O(c_n) \quad \text{a.s.}$$

Since ψ_i is a subvector of φ_i , $\lambda_{\min}(\sum_1^n \psi_i \psi_i') \geq \lambda_{\min}(\sum_1^n \varphi_i \varphi_i')$ and therefore by Lemma 6 (see (2.26)),

$$(3.15) \quad \liminf_{n \rightarrow \infty} \lambda_{\min} \left(\sum_1^n \psi_i \psi_i' \right) / \#_n > 0 \quad \text{a.s.}$$

By (3.2), (3.7), (3.14) and (3.15),

$$(3.16) \quad \psi_n' \left(\sum_1^n \psi_i \psi_i' \right)^{-1} \psi_n \leq \|\psi_n\|^2 / \lambda_{\min} \left(\sum_1^n \psi_i \psi_i' \right) \rightarrow 0 \quad \text{a.s.}$$

and

$$(3.17) \quad \sum_{i \in I, i \leq n} (u_i - \lambda' \psi_i)^2 = O(c_n^2 \#_n) = o(\log n) \quad \text{a.s.}$$

Define

$$(3.18) \quad \begin{aligned} z_n &= \lambda'_{n-1} \psi_n \quad \text{if } |(\lambda_{n-1} - \tilde{\lambda}_{n-1})' \psi_n| \leq A \#_n^{-1/2} (\log \#_n) \|\psi_n\|, \\ &= \tilde{\lambda}'_{n-1} \psi_n \quad \text{otherwise.} \end{aligned}$$

From (3.7), (3.11) and (3.18), it follows that for $0 < \eta < 1/(2\rho)$,

$$(3.19) \quad z_n = \lambda' \psi_n + O((\log n)^{-\eta} \|\psi_n\|) \quad \text{a.s.}$$

By (3.19), with probability 1, $|z_n| \leq (\|\lambda\| + 1) \|\psi_n\|$ for all large n . Moreover, note that $u_n = (-c_n) \vee (z_n \wedge c_n)$ for $n \notin I$ and that $y_{n+1} - \varepsilon_{n+1} = \beta_1(u_n - \lambda' \psi_n)$. Hence, by (3.19),

$$|y_{n+1} - \varepsilon_{n+1}| \chi\{(\|\lambda\| + 1) \|\psi_n\| < c_n, n \notin I\} = O((\log n)^{-\eta} \|\psi_n\|) = o(c_n) \quad \text{a.s., by (3.14).}$$

Therefore, by Lemma 5(ii),

$$(3.20) \quad \begin{aligned} 1 &= P\{|u_n| < c_n \text{ for all large } n \notin I\}. \\ &= P\{u_n = z_n \text{ for all large } n \notin I\}. \end{aligned}$$

Let $e_{n+1} = y_{n+1} - \tilde{\beta}_{1,n-1}(u_n - \lambda'_{n-1} \psi_n) = \beta_1(u_n - \lambda' \psi_n) - \tilde{\beta}_{1,n-1}(u_n - \lambda'_{n-1} \psi_n) + \varepsilon_{n+1}$. We now show that with probability 1

$$(3.21) \quad \beta_1 |u_n - \lambda' \psi_n| \leq |e_{n+1} - \varepsilon_{n+1}| \quad \text{for all large } n \notin I.$$

By (3.20), $u_n = z_n$ for all large $n \notin I$, with probability 1. If $u_n = \lambda'_{n-1} \psi_n$, then $e_{n+1} - \varepsilon_{n+1} = \beta_1(u_n - \lambda' \psi_n)$. If $u_n = z_n \neq \lambda'_{n-1} \psi_n$, then by (3.18), $u_n = \tilde{\lambda}'_{n-1} \psi_n$ and

$$\begin{aligned} |u_n - \lambda'_{n-1} \psi_n| &> A \#_n^{-1/2} (\log \#_n) \|\psi_n\| \\ &> (\log \#_n)^{1/3} |(\tilde{\lambda}_{n-1} - \lambda)' \psi_n| \quad \text{for all large } n, \text{ by (3.11),} \end{aligned}$$

and therefore

$$|u_n - \lambda' \psi_n| = |(\tilde{\lambda}_{n-1} - \lambda)' \psi_n| < (\log \#_n)^{-1/3} |u_n - \lambda'_{n-1} \psi_n|,$$

implying (3.21) since $\tilde{\beta}_{1,n-1} \rightarrow \beta$ a.s.

Since $\log \det(\sum_1^n \psi_i \psi_i')$ is the sum of logarithms of the eigenvalues of $\sum_1^n \psi_i \psi_i'$, it follows from (3.14) that

$$(3.22) \quad \begin{aligned} \log \det \left(\sum_1^n \psi_i \psi_i' \right) &\leq (p+q-1) \log \left(\sum_1^n \|\psi_i\|^2 \right) \\ &\leq (p+q-1)(1+o(1)) \log n \quad \text{a.s.} \end{aligned}$$

In view of (3.16), Lemma 1 is applicable. From Lemma 1 and (3.21), (3.22), it follows that with probability 1, for all large n ,

$$(3.23) \quad \beta_1^2 \sum_{i \in I, i \leq n} (u_i - \lambda' \psi_i)^2 \leq (v + o(1))(p + q - 1) \log n + o\left(\sum_{i=1}^n (u_i - \lambda' \psi_i)^2\right).$$

From (3.17) and (3.23), the desired conclusion (1.16) follows since

$$R_n = \sum_{i=0}^{n-1} (y_{i+1} - \varepsilon_{i+1})^2 = \beta_1^2 \sum_{i=0}^{n-1} (u_i - \lambda' \psi_i)^2. \quad \square$$

We now extend Theorem 1 to the general case where the random disturbances ε_n need not be bounded a.s. To consider the unbounded case, we strengthen the assumption (1.2) into

$$(3.24) \quad \sup_n E(e^{\gamma|\varepsilon_n|} | \mathcal{F}_{n-1}) < \infty \quad \text{a.s. for some } \gamma > 0.$$

This includes the important example where the ε_n are i.i.d. with mean 0, variance $\sigma^2 > 0$ and a finite moment generating function. The truncation sequence $\{c_n\}$ given by (3.2) is too small for this setting. For example, when the ε_n are i.i.d. normal, it follows easily from the Borel-Cantelli lemma that

$$\limsup_{n \rightarrow \infty} |\varepsilon_n| / (2 \log n)^{1/2} = \sigma \quad \text{a.s.}$$

We therefore take a larger truncation sequence c_n such that

$$(3.25) \quad c_n \sim c_{n+1}, \quad c_n / \log n \rightarrow \infty \quad \text{but } c_n = o((\log n)^\alpha) \text{ for all } \alpha > 1.$$

With this choice of c_n , in order that condition (2.8) still holds, we replace the growth condition (3.5) on m_i by

$$(3.26) \quad m_i \sim i^\delta \quad \text{for some } 0 < \delta < \rho - 1.$$

Note that if $n_j \leq n < n_{j+1}$, where n_j is given in (3.4), then

$$(3.27) \quad \#_n \sim \sum_{i=1}^j m_i \sim (1 + \delta)^{-1} j^{1+\delta} \sim (1 + \delta)^{-1} (\log n)^{(1+\delta)/\rho} = o(\log n).$$

While it was easy to prove (3.16) in the case of bounded ε_n , we cannot use the same argument for the present unbounded setting since $\#_n = o(\log n)$ a.s. by (3.27) but $c_n / \log n \rightarrow \infty$. We shall show that, under the additional coprimality assumption (1.9), the ψ_n in fact have the persistent excitation property

$$(3.28) \quad \liminf_{n \rightarrow \infty} n^{-1} \lambda_{\min} \left(\sum_{i=1}^n \psi_i \psi_i' \right) > 0 \quad \text{a.s.}$$

THEOREM 2. *Suppose that we replace the a.s. boundedness assumption on $\{\varepsilon_n\}$ in Theorem 1 by (3.24), the condition (3.2) on $\{c_n\}$ by (3.25), and the condition (3.5) on m_i by (3.26). Then under the additional assumption (1.9), the conclusion of Theorem 1 still holds.*

Proof. Let $c_n^* = \log n (= o(c_n))$ by (3.25). By (3.24) and Lemma 4, $\varepsilon_n = O(c_n^*)$ a.s. Moreover, by (3.25)–(3.27), (2.8), (2.22) and (2.25) still hold. Define z_n as in (3.18) and note that (3.19) still holds with $0 < \eta < (1 + \delta)/(2\rho)$. The same argument as in Theorem 1 shows that (3.14) and (3.20) still hold.

From (1.18), (3.19) and (3.20), it follows that with probability 1

$$(3.29) \quad y_{n+1} = \varepsilon_{n+1} + O((\log n)^{-\eta} \|\psi_n\|) \quad \text{for all large } n \notin I.$$

By (3.14), (3.25) and Lemma 2, there exist $K > 0$ and $0 < a < 1$ such that with probability 1, for all large n ,

$$(3.30) \quad |u_n| \leq K \left\{ 1 + \sum_{0 \leq i \leq K \log \log n} a^i (|\varepsilon_{n+1-i}| + |y_{n+1-i}|) \right\},$$

$$(3.31) \quad |y_n| \leq K \left\{ 1 + \sum_{0 \leq i \leq K \log \log n} a^i (|\varepsilon_{n-i}| + |u_{n-1-i}|) \right\}.$$

We now show that

$$(3.32) \quad \sum_{i \in I, i \leq n} (u_i - \lambda' \psi_i)^2 = o(\log n) \quad \text{a.s.}$$

Since $|u_i| \leq M$ for $i \in I$ and since $\#_n = O((\log n)^{(1+\delta)/\rho})$ by (3.27), it suffices to show that

$$(3.33) \quad \sum_{i \in I, i \leq n} \|\psi_i\|^2 = o(\log n) \quad \text{a.s.}$$

We first note by (3.24) that

$$\sum_{i=1}^{\infty} \sum_{n_i - m_i < n \leq n_i + m_i} P\{|\varepsilon_n| \geq B \log i | \mathcal{F}_{n-1}\} \leq \left\{ \sup_n E(e^{\gamma|\varepsilon_n|} | \mathcal{F}_{n-1}) \right\} \sum_{i=1}^{\infty} 2m_i i^{-\gamma B} < \infty$$

by choosing B large enough so that $\gamma B > 1 + \delta$. Hence, by the conditional Borel–Cantelli lemma (cf. [13, p. 55]),

$$(3.34) \quad \max_{n_i - m_i < n \leq n_i + m_i} |\varepsilon_n| \leq B \log i \sim (B/\rho) \log \log n_i \quad \text{a.s.}$$

From (3.29) together with (3.14), (3.25) and (3.34), it follows that

$$(3.35) \quad \max_{n_i - m_i < n \leq n_i} |y_n| = O((\log n_i)^{1-\eta/2}) \quad \text{a.s.}$$

From (3.30), (3.34) and (3.35), we then obtain that

$$(3.36) \quad \max_{n_i - m_i/2 \leq n \leq n_i} \|\psi_n\| = O((\log n_i)^{1-\eta/2}) \quad \text{a.s.}$$

Putting (3.36) and (3.34) into (3.29) yields

$$(3.37) \quad \max_{n_i - m_i/2 < n \leq n_i} \|y_n\| = O((\log n_i)^{(1-\eta/2)-\eta}) \quad \text{a.s.}$$

Proceeding inductively in this way, we can show that if $(k-1)\eta > 1$ then

$$(3.38) \quad \max_{n_i - m_i/2^k \leq n \leq n_i} \|\psi_n\| = O(\log \log n_i) \quad \text{a.s.}$$

Since $|u_n| \leq M$ for $n_i < n \leq n_i + m_i$, it then follows from (3.31), (3.34) and (3.38) that

$$(3.39) \quad \max_{n_i < n \leq n_i + m_i} \|\psi_n\| = O(\log \log n_i) \quad \text{a.s.}$$

Therefore with probability 1

$$\sum_{i: n_i \leq n} \left\{ \sum_{r=n_i+1}^{n_i+m_i} \|\psi_r\|^2 \right\} = \sum_{i: n_i \leq n} O(m_i (\log \log n_i)^2) = o(\log n),$$

thus establishing (3.33) and therefore (3.32) also.

In view of (3.32), it remains to prove that

$$(3.40) \quad \beta_1^2 \sum_{i \notin I, i \leq n} (u_i - \lambda' \psi_i)^2 \leq \{v(p+q-1) + o(1)\} \log n \quad \text{a.s.}$$

We can prove this by applying Lemma 1 as in the last two paragraphs of the proof of Theorem 1. To show that Lemma 1 is applicable, we shall prove that (3.16) still holds by establishing the persistent excitation property (3.28) of $\{\psi_n\}$. First, in view of (3.29),

$$(3.41) \quad \sum_{i \notin I, i \leq n} y_{i+1}^2 \leq 2 \sum_{i \notin I, i \leq n} \varepsilon_{i+1}^2 + \sum_{i \notin I, i \leq n} o(\|\psi_i\|^2).$$

From (3.32) and (3.41), it follows that

$$(3.42) \quad \sum_{i=1}^n y_{i+1}^2 \leq 2 \sum_{i=1}^n \varepsilon_{i+1}^2 + o(\log n) + o\left(\sum_{i=1}^n \|\psi_i\|^2\right) = O(n) + o\left(\sum_{i=1}^n \|\psi_i\|^2\right) \quad \text{a.s.}$$

since $\sum_{i=1}^n \varepsilon_i^2 = \sum_{i=1}^n E(\varepsilon_i^2 | \mathcal{F}_{i-1}) + o(n)$ a.s. (cf. [14, p. 157]). By Lemma 2(iv) and (3.42), $\sum_{i=1}^n \|U_i\|^2 = O(n) + o(\sum_{i=1}^n \|\psi_i\|^2)$ a.s., so

$$\sum_{i=1}^n \|\psi_i\|^2 = O(n) + o\left(\sum_{i=1}^n \|\psi_i\|^2\right) \quad \text{a.s.}$$

This shows that $\sum_{i=1}^n \|\psi_i\|^2 = O(n)$ a.s., and therefore by (3.29),

$$(3.43) \quad \sum_{i \notin I, i \leq n} (y_{i+1} - \varepsilon_{i+1})^2 = \sum_{i \notin I, i \leq n} o(\|\psi_i\|^2) = o(n) \quad \text{a.s.}$$

From (3.32) and (3.43), it then follows that $\sum_{i=1}^n (y_{i+1} - \varepsilon_{i+1})^2 = o(n)$ a.s., i.e., $\{u_i\}$ is globally convergent. In view of the coprimality assumption (1.9), the global convergence of $\{u_i\}$ implies the desired persistent excitation property (3.28) for $\{\psi_n\}$, by Corollary 3 of [15]. \square

4. Extensions to the case of general delay. The approach in the preceding section can be readily extended to the case of general d . The key idea is to rewrite (1.1) in the regression form (4.4) below so that we can apply Lemma 1. Since $1 - \alpha_1 z - \cdots - \alpha_p z^p$ and z^d are relatively prime, we can find polynomials $F(z) = 1 + f_1 z + \cdots + f_{d-1} z^{d-1}$ and $G(z) = g_1 + g_2 z + \cdots + g_p z^{p-1}$ such that

$$(1 - \alpha_1 z - \cdots - \alpha_p z^p)F(z) + z^d G(z) = 1;$$

moreover, we can rewrite (1.1) as

$$(4.1) \quad y_{n+d} = g_1 y_n + \cdots + g_p y_{n-p+1} + \gamma_1 u_n + \cdots + \gamma_{q+d-1} u_{n-q-d+2} + \tilde{\varepsilon}_{n+d},$$

where $\gamma_j = \sum_{i+s=j} f_i \beta_s$ ($f_0 = 1$), so $\gamma_1 = \beta_1 \neq 0$, and

$$(4.2) \quad \tilde{\varepsilon}_{n+d} = \varepsilon_{n+d} + f_1 \varepsilon_{n+d-1} + \cdots + f_{d-1} \varepsilon_{n+1} = y_{n+d} - E(y_{n+d} | \mathcal{F}_n)$$

(cf. [6]). Defining

$$(4.3) \quad \begin{aligned} \lambda &= -\beta_1^{-1}(g_1, \cdots, g_p, \gamma_2, \cdots, \gamma_{q+d-1})', \\ \psi_n &= (y_n, \cdots, y_{n-p+1}, u_{n-1}, \cdots, u_{n-q-d+2})', \end{aligned}$$

we can write (4.1) in the form of d regression models: For $j = 0, \cdots, d-1$,

$$(4.4) \quad y_{j+d(t+1)} = \beta_1(u_{j+dt} - \lambda' \psi_{j+dt}) + \tilde{\varepsilon}_{j+d(t+1)}, \quad t = 1, 2, \cdots$$

Example. Consider the linear system

$$(4.5) \quad y_n = \alpha y_{n-1} + \beta u_{n-2} + \varepsilon_n.$$

Here $d = 2$ and we can write (4.5) as

$$(4.6) \quad \begin{aligned} y_{n+2} &= \alpha y_{n+1} + \beta u_n + \varepsilon_{n+2} = \alpha(\alpha y_n + \beta u_{n-1} + \varepsilon_{n+1}) + \beta u_n + \varepsilon_{n+2} \\ &= \alpha^2 y_n + \beta u_n + \alpha \beta u_{n-1} + \tilde{\varepsilon}_{n+2}, \end{aligned}$$

where $\tilde{\varepsilon}_i = \varepsilon_i + \alpha \varepsilon_{i-1}$.

When the system parameters are known, the optimal controller chooses the input u_n at stage n such that $E(y_{n+d}|\mathcal{F}_n) = 0$, and the outputs are $y_i = \tilde{\varepsilon}_i$. We therefore define the “regret,” at stage n , of a sequence of inputs $\{u_i\}$ to be

$$(4.7) \quad R_n(d) = \sum_1^n (y_i - \tilde{\varepsilon}_i)^2.$$

Without assuming any prior knowledge of the system parameters, we now construct an input sequence such that

$$(4.8) \quad \limsup_{n \rightarrow \infty} R_n(d)/\log n \leq \tilde{v}d(p+q+d-2) \quad \text{a.s., where}$$

$$\tilde{v} = \limsup_{n \rightarrow \infty} E(\tilde{\varepsilon}_{n+d}^2|\mathcal{F}_n).$$

Note that (4.8) reduces to (1.16) in the case $d = 1$.

Suppose that the martingale difference sequence $\{\varepsilon_n\}$ is bounded a.s. and satisfies (1.2) and (3.1). As in § 3, we introduce white-noise probing inputs to obtain strongly consistent estimates. For the given d , define φ_n and θ by (2.20). Let $V_n = \sum_1^n \varphi_i \varphi_i'$,

$$(4.9) \quad \tau = \inf \{n: V_{n-d} \text{ is nonsingular and } \hat{\beta}_{1,n-d} \neq 0\},$$

where $\hat{\beta}_{1,t}$ is the least squares estimate of β_1 based on $\{(\varphi_i, y_{i+d}): i \leq t\}$ in the regression model (2.21). With (4.9) replacing (3.3), we can define n_i , m_i and I as in (3.4)–(3.6) and introduce white-noise probing input (2.23) at stage $n \in I$. Define the auxiliary estimates $\tilde{\theta}_n = (\tilde{\alpha}_{1,n}, \dots, \tilde{\alpha}_{p,n}, \tilde{\beta}_{1,n}, \dots, \tilde{\beta}_{q,n})'$ as in (3.8). Since (3.9) still holds by Lemma 6 and since the parameters g_1, \dots, g_p , $\gamma_1 (= \beta_1), \dots, \gamma_{q+d-1}$ of the reparametrized model (4.1) are polynomials in the α 's and β 's (see the preceding example), we can obtain from the auxiliary estimates $\tilde{\theta}_n$ the corresponding strongly consistent estimates $\tilde{\beta}_{1,n} \neq 0$ and $\tilde{\lambda}_n$ such that

$$(4.10) \quad \tilde{\lambda}_n - \lambda = O((\log \#_n)^{1/2} / \#_n^{1/2}) \quad \text{a.s.,}$$

where $\#_n$ is the number of probing inputs up to stage n .

In view of (4.4), the general-delay version of the recursive estimates λ_t introduced in (3.12) is of the following form. For $n = j + dt$ (with $t = 1, 2, \dots$ and $j = 0, \dots, d-1$), let

$$(4.11) \quad P_n^{-1} = \sum_{s=1}^t \psi_{j+ds} \psi_{j+ds}' = \sum_{i \leq n, i \equiv j \pmod{d}} \psi_i \psi_i',$$

and define the estimates of λ recursively by

$$(4.12) \quad \lambda_n = \lambda_{n-d} + (u_n - \tilde{\beta}_{1,n-d}^{-1} y_{n+d} - \lambda'_{n-d} \psi_n) P_n \psi_n.$$

Let $A > 0$. Analogous to (3.13), we define at stage $n \notin I$ the input

$$(4.13) \quad u_n = \begin{cases} (-c_n) \vee (\lambda'_{n-d} \psi_n \Lambda c_n) & \text{if } |(\lambda_{n-d} - \tilde{\lambda}_{n-d})' \psi_n| \leq A \#_n^{-1/2} (\log \#_n) \|\psi_n\|, \\ (-c_n) \vee (\tilde{\lambda}'_{n-d} \psi_n \Lambda c_n) & \text{otherwise.} \end{cases}$$

THEOREM 3. *Consider the linear system (1.1) with $\beta_1 \neq 0$ and such that (1.3) holds. Suppose that $\{\varepsilon_n\}$ is a martingale difference sequence such that (1.2), (3.1) hold and $\sup_n |\varepsilon_n| < \infty$ a.s. Let $\{c_n\}$ be a nondecreasing sequence of positive numbers satisfying (3.2). Then the regret of the input sequence $\{u_n\}$ defined above (by (4.13) for $n \notin I$, and by (2.23) for $n \in I$) satisfies (4.8).*

Proof. Note by (4.2) that for fixed j , $\{\tilde{\varepsilon}_{j+dt}: t \geq 1\}$ is a martingale difference sequence such that $\sup_t E(|\tilde{\varepsilon}_{j+dt}|^\gamma | \mathcal{F}_{j+d(t-1)}) < \infty$ and $\sup_t |\tilde{\varepsilon}_{j+dt}| < \infty$ a.s. By Lemma 3, (3.14) still holds. Moreover, it follows from (4.4) and an argument similar to the proof of Theorem 1 that

$$|y_{n+d} - \tilde{\varepsilon}_{n+d}| \chi\{(\|\lambda\| + 1)\|\psi_n\| < c_n, n \notin I\} = o(c_n) \quad \text{a.s.}$$

Since ψ_n is a subvector of $\varphi_{n-d+1,d}$ defined in (2.32), it follows from (2.33) that

$$\liminf_{n \rightarrow \infty} \lambda_{\min} \left(\sum_{t \leq n, t \equiv j \pmod{d}} \psi_t \psi_t' \right) / \#_n > 0 \quad \text{a.s.}$$

for $j = 0, \dots, d-1$, and therefore $\psi_n' P_n \psi_n \rightarrow 0$ a.s. Hence Lemma 1 can again be applied to each of the d regression models in (4.4), and an argument similar to the proof of Theorem 1 shows that for $j = 0, \dots, d-1$,

$$(4.14) \quad \beta_1^2 \sum_{j+dt \notin I, j+dt \leq n} (u_{j+dt} - \lambda' \psi_{j+dt})^2 \\ \leq (\tilde{v} + o(1))(p + q + d - 2) \log n + o \left(\sum_{i=1}^n (u_i - \lambda' \psi_i)^2 \right),$$

noting that $p + q + d - 2$ is the dimensionality of the vector ψ_n . Clearly (3.17) still holds as before. From (4.14) and (3.17), the desired conclusion (4.8) follows. \square

REFERENCES

- [1] K. J. ÅSTRÖM, *Theory and applications of adaptive control—A survey*, Automatica—J. IFAC, 19 (1983), pp. 471–486.
- [2] K. J. ÅSTRÖM AND B. WITTENMARK, *On self-tuning regulators*, Automatica—J. IFAC, 9 (1973), pp. 195–199.
- [3] G. C. GOODWIN AND R. L. PAYNE, *Dynamic System Identification*, Academic Press, New York, 1977.
- [4] T. L. LAI, *Asymptotically efficient adaptive control in stochastic regression models*, Adv. in Appl. Math., 7 (1986), pp. 23–45.
- [5] T. L. LAI AND C. Z. WEI, *Extended least squares and their applications to adaptive control and prediction in linear systems*, IEEE Trans. Automat. Control, AC-31 (1986), to appear.
- [6] G. C. GOODWIN, P. J. RAMADGE AND P. E. CAINES, *Discrete time stochastic adaptive control*, this Journal, 19 (1981), pp. 829–853.
- [7] K. S. SIN AND G. C. GOODWIN, *Stochastic adaptive control using a modified least squares algorithm*, Automatica—J. IFAC, 18 (1982), pp. 315–321.
- [8] P. E. CAINES AND S. LAFORTUNE, *Adaptive control with recursive identification for stochastic linear systems*, IEEE Trans. Automat. Control, AC-29 (1984), pp. 312–321.
- [9] H. F. CHEN AND P. E. CAINES, *The strong consistency of the stochastic gradient algorithm of adaptive control*, IEEE Trans. Automat. Control, AC-30 (1985), pp. 189–192.
- [10] A. H. BECKER, P. R. KUMAR AND C. Z. WEI, *Adaptive control with stochastic approximation algorithm: Geometry and convergence*, IEEE Trans. Automat. Control, AC-30 (1985), pp. 330–338.
- [11] P. R. KUMAR, *A survey of some results in stochastic adaptive control*, this Journal, 23 (1985), pp. 329–380.
- [12] T. L. LAI AND C. Z. WEI, *Asymptotic properties of projections with applications to stochastic regression problems*, J. Multivariate Anal., 12 (1982), pp. 346–370.
- [13] W. F. STOUT, *Almost Sure Convergence*, Academic Press, New York, 1974.
- [14] T. L. LAI AND C. Z. WEI, *Least squares estimates in stochastic regression models with applications to identification and control of dynamic systems*, Ann. Statist., 10 (1982), pp. 154–166.
- [15] ———, *On the concept of excitation in least squares identification and adaptive control*, Stochastics, 16 (1986), pp. 227–254.

OPTIMAL CONTROL FOR PARABOLIC VARIATIONAL INEQUALITIES*

AVNER FRIEDMAN†

Abstract. Consider the problem of maximizing a functional that depends on a control function $k(x, t)$ and on the solution $u(x, t)$ of a parabolic variational inequality with k appearing in the data. Necessary conditions are obtained for the maximizers $k_0(x, t)$, and the structure of k_0 is then analyzed. An application to the Stefan problem is given.

Key words. optimal control, parabolic variational inequality, elliptic variational inequality, bang-bang principle, Stefan problem

AMS(MOS) subject classifications. 49A29, 49B22, 35K55

Introduction. Consider a parabolic variational inequality with control $k(x, t)$ appearing either in the inhomogeneous term or in the boundary data. Denote the corresponding solution by $u(x, t)$ and introduce a functional

$$J(k) = \iint F(x, t, u) \, dx \, dt + \iint \Phi(k) \, dx \, dt.$$

Let k_0 be a solution of the problem

$$(0.1) \quad J(k_0) = \max_{k \in \alpha} J(k), \quad k_0 \in \alpha$$

where α is a given class of control functions. We are interested in studying properties of k_0 .

In most control problems for partial differential equations (see [9], [10]) the functional $J(k)$ is differentiable. This is not the case here. For this reason there are difficulties in deriving effective necessary conditions on the maximizers k_0 ; see [1], [11], [12] and the reference given there.

In a recent paper [4] we have treated the case of an elliptic variational inequality under the assumption that

$$(0.2) \quad F_u > 0.$$

Using a comparison argument, we were able to obtain a general simple necessary condition for the maximizing control. This condition yields a bang-bang principle for some classes α . As pointed out in [4], the method does not carry over to parabolic variational inequalities.

In the present paper we use a different method which allows us to treat both elliptic and parabolic variational inequalities; furthermore, the restriction (0.2) is removed.

The method consists in approximating the maximizer (k_0, u_0) with maximizers $(k_\varepsilon, u_\varepsilon)$ of ε -penalized problems which are "smooth." We first obtain necessary conditions of the maximizers k_ε (in § 1) and then use them (in subsequent sections) to find the structure of k_0 .

* Received by the editors April 22, 1985; accepted for publication (in revised form) March 11, 1986. This work was partially supported by National Science Foundation grant MCS-8300293.

† Department of Mathematics, Purdue University, West Lafayette, Indiana 47907.

In order to avoid duplication we shall treat in detail only parabolic variational inequalities; elliptic variational inequalities can be treated in precisely the same way.

In case

$$\alpha = \left\{ 0 \leq k \leq M, \int \int k \, dx \, dt = H \right\}$$

where k appears as the nonhomogeneous term, it is shown (in § 2) that in the noncoincidence set of u_0

$$(0.3) \quad k_0 = \begin{cases} M & \text{if } Q < \lambda, \\ 0 & \text{if } Q > \lambda \end{cases}$$

for a suitable solution Q of a parabolic equation.

Other functionals and control sets α are considered in § 3. In § 4 we consider the case where the control appears in the boundary data. An application to the Stefan problem in N -dimension is given in § 5.

We finally remark that the necessary conditions derived in this paper are not sufficient in general.

1. The ε -problem. Let D be a bounded domain in \mathbb{R}^N with C^2 boundary S and set

$$\begin{aligned} D_T &= D \times (0, T), & S_T &= S \times (0, T), \\ \partial_p D_T &= S_T \cup D \times \{0\}, & \partial_{-p} D_T &= S_T \cup D \times \{T\} \end{aligned}$$

for any $T > 0$. We denote points in D_T by (x, t) .

Let $U^0(x, t)$ be a given function satisfying

$$(1.1) \quad \begin{aligned} D_t U^0, D_x^2 U^0 &\in L^\infty(D_T), \\ U^0 &> 0 \quad \text{on } S_T, \quad U^0 \geq 0 \quad \text{on } D \times \{0\}. \end{aligned}$$

Let f be a given function satisfying

$$(1.2) \quad f \in L^p(D_T)$$

where

$$(1.3) \quad p > \frac{N+2}{2}, \quad p \geq 2.$$

We introduce a class α of control functions $k(x, t)$ with the following properties:

$$(1.4) \quad \alpha \text{ is a closed, bounded and convex set in } L^p(D_T).$$

For any $k \in \alpha$ consider the parabolic variational inequality

$$(1.5) \quad \int \int_{D_T} u_t(\psi - u) + \int \int_{D_T} \nabla u \cdot \nabla(\psi - u) \geq - \int \int_{D_T} (f + k)(\psi - u) \quad \forall \psi \in K, \quad u \in K$$

where

$$(1.6) \quad \begin{aligned} K &= \{ \psi; \psi - U^0 \in L^2(0, T; H_0^1(D)), \psi_t \in L^2(0, T; H^{-1}(D)), \\ &\quad \psi(\cdot, 0) = U_0(\cdot, 0), \psi \geq 0 \text{ a.e. in } D_T \}; \end{aligned}$$

here $H^{-1}(D)$ is the dual of $H_0^1(D)$. Notice that $\psi \in C^0([0, T]; H^{-1}(D))$ so that the condition $\psi(\cdot, 0) = U_0(\cdot, 0)$ is meaningful.

We introduce the functional

$$(1.7) \quad J(k) = \iint_{D_T} F(x, t, u) \, dx \, dt + \iint_{D_T} \Phi(k) \, dx \, dt$$

and assume that

$$(1.8) \quad \begin{aligned} &k \rightarrow \iint_{D_T} \Phi(k) \text{ is upper semicontinuous under weak convergence in } L^p(D_T), \\ &\text{i.e., if } k_m \rightarrow k \text{ weakly in } L^p(D_T) \text{ where } k_m, k \in \mathfrak{a} \text{ then} \\ &\iint_{D_T} \Phi(k) \geq \limsup_{k \rightarrow \infty} \iint_{D_T} \Phi(k_m). \end{aligned}$$

We also assume that

$$(1.9) \quad F(x, t, u) \text{ is continuous in } \overline{D_T} \times \mathbb{R}^1.$$

Consider the problem: find k_0 such that

$$(1.10) \quad J(k_0) = \max_{k \in \mathfrak{a}} J(k), \quad k_0 \in \mathfrak{a}.$$

By the general theory of variational inequalities [2], [3], [7], for any $k \in \mathfrak{a}$ there exists a unique solution u of (1.5) and $u, D_x^2 u$ belong to $L^p(D_T)$; hence, by the Sobolev imbedding and (1.3) (cf. [8])

$$(1.11) \quad |u|_{C^\alpha(D_T)} \leq C$$

for some α positive; C and α are independent of k . Take a maximizing sequence k_m and their corresponding solutions u_m such that

$$\begin{aligned} k_m &\rightarrow k_0 \quad \text{weakly in } L^p(D_T), \\ u_m &\rightarrow u_0 \quad \text{in } C^\beta(D_T), \quad 0 < \beta < \alpha. \end{aligned}$$

Then one can easily verify that (1.5) is satisfied for $k = k_0$, $u = u_0$ and that k_0 is a solution of problem (1.10).

Our main purpose is to find necessary conditions for maximizers k_0 . For this we need the additional assumptions:

$$(1.12) \quad F_u(x, t, u) \text{ is continuous in } \overline{D_T} \times \mathbb{R}^1,$$

and $\Phi(k)$ is differentiable, i.e.

$$(1.13) \quad \begin{aligned} &\text{for any } k \in \mathfrak{a} \text{ there exists a function in } L^q(D_T) (1/p + 1/q = 1) \text{ denoted} \\ &\text{by } \Phi'(k) \text{ such that for any } l \in L^p(D_T) \text{ with } k + \delta l \in \mathfrak{a} \text{ if } 0 < \delta < 1 \\ &\text{there holds: } \lim_{\delta \rightarrow 0} 1/\delta \iint_{D_T} [\Phi(k + \delta l) - \Phi(k)] = \iint_{D_T} \Phi'(k) l. \end{aligned}$$

The difficulty in deriving effective necessary conditions on k_0 stems from the fact that the functional $J(k)$ does not have a Frechet derivative (unless the free boundary happens to be smooth). To overcome this difficulty we consider, for any $\varepsilon \in (0, 1)$, the penalized problem

$$(1.14) \quad \begin{aligned} u_t - \Delta u + \beta_\varepsilon(u) &= -(f + k) \quad \text{in } D_T, \\ u &= U^0 \quad \text{on } \partial_p D_T \end{aligned}$$

where $\beta_\varepsilon(s) \in C^\infty(\mathbb{R}^1)$ and

$$(1.15) \quad \begin{aligned} \beta_\varepsilon(s) &\rightarrow 0 \quad \text{if } s > 0, \quad \varepsilon \rightarrow 0, \\ \beta_\varepsilon(s) &\rightarrow -\infty \quad \text{if } s < 0, \quad \varepsilon \rightarrow 0, \\ \beta'_\varepsilon(s) &\geq 0. \end{aligned}$$

Denote by u_ε the unique solution of (1.14) and (following [1]) introduce the functional

$$(1.16) \quad J_\varepsilon(k) = \iint_{D_T} F(x, t, u_\varepsilon) + \iint_{D_T} \Phi(k) - \frac{1}{2} \iint_{D_T} (k - k_0)^2.$$

DEFINITION 1.1. The ε -problem consists in finding a function k_ε satisfying

$$(1.17) \quad J_\varepsilon(k_\varepsilon) = \max_{k \in \mathfrak{a}} J_\varepsilon(k), \quad k_\varepsilon \in \mathfrak{a}.$$

Suppose u_ε is the solution of (1.14) corresponding to a maximizer k_ε . Denote by Q_ε the solution of the parabolic problem

$$(1.18) \quad \begin{aligned} Q_t + \Delta Q - \beta'_\varepsilon(u_\varepsilon)Q &= -F_u(x, t, u_\varepsilon) \quad \text{in } D_T, \\ Q &= 0 \quad \text{on } \partial_{-p}D_T. \end{aligned}$$

LEMMA 1.1. If $k_\varepsilon + \delta l \in \mathfrak{a}$ for any $0 < \delta < \delta_0$ and some $\delta_0 > 0$, then

$$(1.19) \quad \iint_{D_T} (-Q_\varepsilon - \Phi'(k_\varepsilon) + k_\varepsilon - k_0)l \geq 0.$$

Proof. Denote by $u_{\varepsilon,\delta}$ the solution of (1.14) corresponding to $k_\varepsilon + \delta l$. Then

$$(1.20) \quad \begin{aligned} 0 &\geq \overline{\lim}_{\delta \rightarrow 0} \frac{1}{\delta} (J(k_\varepsilon + \delta l) - J(k_\varepsilon)) \\ &= \overline{\lim}_{\delta \rightarrow 0} \iint_{D_T} \frac{1}{\delta} (F(x, t, u_{\varepsilon,\delta}) - F(x, t, u_\varepsilon)) + \iint_{D_T} \Phi'(k_\varepsilon)l - \iint_{D_T} (k_\varepsilon - k_0)l \\ &= I_1 + I_2 + I_3; \end{aligned}$$

the integral in I_1 is equal to

$$\iint_{D_T} F_u(x, t, \tilde{u}_{\varepsilon,\delta}(x, t)) \frac{u_{\varepsilon,\delta} - u_\varepsilon}{\delta}$$

where $\tilde{u}_{\varepsilon,\delta}(x, t)$ lies in the interval with end-points $u_\varepsilon(x, t)$, $u_{\varepsilon,\delta}(x, t)$. Denote by $Q_{\varepsilon,\delta}$ the solution of (1.18) corresponds to $F_u(x, t, \tilde{u}_{\varepsilon,\delta})$. Then $\lim_{\delta \rightarrow 0} Q_{\varepsilon,\delta} = Q_\varepsilon$ uniformly in D_T . Also

$$\lim_{\delta \rightarrow 0} \frac{u_{\varepsilon,\delta} - u_\varepsilon}{\delta} = z \quad \text{uniformly in } D_T$$

where

$$(1.21) \quad \begin{aligned} z_t - \Delta z + \beta'_\varepsilon(u_\varepsilon)z &= -l \quad \text{in } D_T, \\ z &= 0 \quad \text{on } \partial_p D_T. \end{aligned}$$

It follows that

$$\begin{aligned} I_1 &= - \iint_{D_T} [Q_{\varepsilon,t} + \Delta Q_\varepsilon - \beta'_\varepsilon(u_\varepsilon)Q_\varepsilon]z \\ &= \iint_{D_T} [z_t - \Delta z + \beta'_\varepsilon(u_\varepsilon)z]Q = \iint_{D_T} Ql, \quad \text{by (1.21)}. \end{aligned}$$

When we substitute this into (1.20), the assertion (1.19) follows.

LEMMA 1.2. As $\varepsilon \rightarrow 0$

$$(1.22) \quad \begin{aligned} k_\varepsilon &\rightarrow k_0 \quad \text{strongly in } L^p(D_T), \\ u_\varepsilon &\rightarrow u_0 \quad \text{uniformly in } D_T. \end{aligned}$$

This result is due to Barbu [1]. For convenience we recall the proof.

Proof. Denote by u_ε^* the solution of (1.14) corresponding to k_0 . By a standard argument [2], [3]

$$\|\beta_\varepsilon(u_\varepsilon)\|_{L^p(D_T)} \leq C, \quad \|\beta_\varepsilon(u_\varepsilon^*)\|_{L^p(D_T)} \leq C.$$

Hence (1.11) holds for u_ε and u_ε^* . It follows that for any sequence of ε 's there is a subsequence such that

$$\begin{aligned} k_\varepsilon &\rightarrow \tilde{k} \quad \text{weakly in } L^p(D_T), \\ u_\varepsilon &\rightarrow \tilde{u}, \quad u_\varepsilon^* \rightarrow \hat{u} \quad \text{uniformly in } D_T \end{aligned}$$

and

$$\begin{aligned} J_\varepsilon(k_0) &\rightarrow J(k_0), \\ \lim J_\varepsilon(k_\varepsilon) &= J(\tilde{k}) - \frac{1}{2} \lim \iint_{D_T} |k_\varepsilon - k_0|^2. \end{aligned}$$

Since $J_\varepsilon(k_0) \leq J_\varepsilon(k_\varepsilon)$ and $J(k_0) \geq J(\tilde{k})$, it follows that

$$\lim \iint_{D_T} |k_\varepsilon - k_0|^2 = 0.$$

Thus $\tilde{k} = k_0$, $\tilde{u} = u_0$, and (1.22) follows.

2. The bang-bang principle. In this section we take

$$(2.1) \quad \alpha = \left\{ k \in L^\infty(D_T), 0 \leq k \leq M, \iint_{D_T} k = H \right\}$$

where M is such that $M \operatorname{meas}(D_T) > H$, so that α is nontrivial. For simplicity we take

$$(2.2) \quad J(k) = \iint_{D_T} F(x, t, u) \, dx \, dt$$

and assume that

$$(2.3) \quad f \in L^\infty(D_T).$$

We shall use the notation of § 1, and set

$$\Omega_\varepsilon = \{u_\varepsilon > 0\}, \quad \Omega_0 = \{u_0 > 0\}.$$

Let q be the solution of

$$(2.4) \quad -\Delta q = C_0 \quad \text{in } D, \quad q = 0 \quad \text{on } \partial D.$$

Then $q > 0$ in D and, by comparison,

$$-q < Q_\varepsilon < q \quad \text{in } D_T$$

provided

$$C_0 > \sup_{D_T} |F_u(x, t, u_\varepsilon)| \quad \forall \varepsilon \in (0, 1)$$

(recall that $\beta'_\varepsilon(u_\varepsilon) \geq 0$). It follows that

$$(2.5) \quad -C \leq Q_\varepsilon \leq C \quad \text{in } D_T;$$

further

$$(2.6) \quad \text{if } F_u(x, t, u) \geq 0 \quad \text{for } u \geq 0 \quad \text{then } 0 \leq Q_\varepsilon \leq C.$$

Let G be any compact subset of Ω_0 . Then $u_\varepsilon \geq c > 0$ in G if ε is small enough and, consequently, $\beta_\varepsilon(u_\varepsilon) = 0$ in G . It follows that, for a subsequence, $Q_\varepsilon \rightarrow Q$ uniformly in G and $Q_t + \Delta Q = -F_u(x, t, u_0)$ in G . Since G is arbitrary we conclude that, for a subsequence,

$$(2.7) \quad Q_\varepsilon \rightarrow Q \quad \text{uniformly in compact subsets of } \Omega_0,$$

$$(2.8) \quad Q_t + \Delta Q = -F_u(x, t, u_0) \quad \text{in } \Omega_0$$

and

$$(2.9) \quad -C \leq Q \leq C \quad \text{in } \Omega_0.$$

Also, by (2.6) and the strong maximum principle,

$$(2.10) \quad \text{if } F_u(x, t, u) > 0 \text{ for } u > 0 \text{ then } 0 < Q \leq C \text{ in } \Omega_0.$$

THEOREM 2.1. *There exists a constant λ such that*

$$(2.11) \quad \begin{aligned} k_0 &= M \quad \text{a.e. in } \{Q < \lambda\} \cap \Omega_0, \\ k_0 &= 0 \quad \text{a.e. in } \{Q > \lambda\} \cap \Omega_0. \end{aligned}$$

Proof. From Lemma 1.2 it follows that for any $\eta > 0$ there exists a set G_η of measure $< \eta$ such that

$$|k_\varepsilon - k_0| < \eta \quad \text{in } \Omega' \equiv D_T \setminus G_\eta.$$

If we take in Lemma 1.1 l with support in Ω' , then we obtain the inequality

$$\iint_{\Omega'} (Q_\varepsilon + \theta) l \geq 0 \quad \text{where } |\theta| \leq \eta;$$

here $\theta = k_\varepsilon - k_0$ is independent of l .

We can now argue as in Theorem 2.1 of [4] and deduce that there exists a λ_ε such that in Ω'

$$k_\varepsilon = \begin{cases} M & \text{a.e. in } \{Q_\varepsilon < \lambda_\varepsilon - \eta\}, \\ 0 & \text{a.e. in } \{Q_\varepsilon > \lambda_\varepsilon + \eta\}. \end{cases}$$

Since $|\lambda_\varepsilon| \leq C$ (by (2.9)), we may assume that $\lambda_\varepsilon \rightarrow \lambda$ as $\varepsilon \rightarrow 0$.

Let G be any compact subset of Ω_0 and let $G' = \Omega' \cap G$. Then, by (2.7),

$$\{Q_\varepsilon < \lambda_\varepsilon - \eta\} \cap G' \supset \{Q < \lambda - 2\eta\} \cap G'$$

if ε is sufficiently small. It follows that $k_\varepsilon = M$ in $\{Q < \lambda - 2\eta\} \cap G'$, and the same is then true of the limit k_0 . Since η and G are arbitrary we conclude that $k_0 = M$ a.e. in $\{Q < \lambda\} \cap \Omega_0$. Similarly $k_0 = 0$ a.e. in $\{Q > \lambda\} \cap \Omega_0$.

Remark 2.1. If $F_u(x, t, u) \neq 0$ for all $u > 0$ then the set $\{Q = \lambda\}$ has measure zero. Indeed, a.e. on this set $Q_t = 0$ and $\Delta Q = 0$, so that also $F_u(x, t, u(x, t)) = 0$ (by (2.8)).

Notation. By V_δ ($\delta > 0$) we denote δ D -neighborhood of S .

THEOREM 2.2. *Suppose $F_u(x, t, u) > 0$ if $(x, t) \in \bar{D}_T$, $u > 0$. Then (i) $\lambda \geq 0$; (ii) if $\lambda > 0$ then $k_0 = M$ in some set $V_{\delta_0} \times (0, T)$; (iii) if $\lambda > 0$ then $k_0 = M$ a.e. in $D_T \setminus \Omega_0$; (iv) if $f \leq 0$ then $\lambda > 0$.*

Proof. The assertion (i) follows from (2.10). Next, since $U^0 > 0$ on S we have

$$u_\varepsilon \geq c > 0 \quad \text{in } V_{\delta_1} \times (0, T)$$

for some $\delta_1 > 0$, and

$$Q_{\varepsilon,t} + \Delta Q_{\varepsilon} = -F_u(x, t, u_{\varepsilon}) \quad \text{in } V_{\delta_1} \times (0, T).$$

Recalling that $Q_{\varepsilon} = 0$ on S_T , $Q_{\varepsilon} \geq 0$, we can deduce (using the strong maximum principle and parabolic regularity) that in $V_{\delta_2} \times (0, T)$ ($\delta_2 < \delta_1$),

$$(2.12) \quad c_0 d(x, t) \leq Q_{\varepsilon}(x, t) \leq c_1 d(x, t) \quad (0 < c_0 < c_1)$$

where $d(x, t)$ is the distance to the parabolic boundary $\partial V_{\delta_1} \times (0, T) \cup V_{\delta_1} \times \{t = T\}$. The second inequality in (2.12) and Theorem 2.1 yield the assertion (ii).

Suppose next that $\lambda > 0$. If the assertion (iii) is not true, then there exists a $\delta > 0$ and a subset $G \subset (D_T \setminus \Omega_0)$ with positive measure η such that $k_0 < M - \delta$ in G . Let $k_1 = k_0$ in $D_T \setminus G$, $k_1 = k_0 + \delta$ in G . Then the solution u_1 of the variational inequality (1.5) corresponding to k_1 is the same solution u_0 as for k_0 . Notice that η can be chosen arbitrarily small.

Since $\lambda > 0$, by (ii) there is a set G' in Ω_0 such that $k_0 = M$ on G' , and $\text{meas}(G') = \eta$ (if η was chosen small enough). Let $k_2 = k_1$ in $D_T \setminus G'$, $k_2 = M - \delta$ in G' . Then by comparison, the solution u' corresponding to k_2 satisfies: $u' \geq u_0$ and $u' \neq u_0$. It follows that $J(k_2) > J(k_0)$, which is a contradiction since $k_2 \in \alpha$.

To prove (iv) observe that if $\lambda = 0$ then $k_0 = 0$ in Ω_0 and therefore

$$\frac{\partial u_0}{\partial t} - \Delta u_0 = -f \geq 0 \quad \text{in } \Omega_0.$$

We can construct a ball

$$E: |x - \bar{x}|^2 + (t - \bar{t})^2 \leq \alpha$$

in Ω_0 such that ∂E intersect the free boundary at a point (x_0, t_0) which is not the top or bottom points of E . In fact, to construct E we move a plane which has a fixed nonhorizontal direction until it touches $\{u_0 = 0\}$ at some point (x_0, t_0) . Then, by the strong maximum principle, $\nabla_x u_0(x_0, t_0) \neq 0$, contradicting the fact that $\nabla_x u$ (is continuous and) vanishes on the free boundary [3], [7].

Remark 2.2 If $F_u(x, t, u) < 0$ for $(x, t) \in \bar{D}_T$, $u > 0$, then $Q < 0$ in Ω_0 and $\lambda \leq 0$. Further, if $\lambda < 0$ then $k_0 = 0$ in $V_{\delta_0} \times (0, T)$ and $k_0 = 0$ a.e. in $D_T \setminus \Omega_0$. Finally, if $f \leq -M$ then $\lambda < 0$. The proof of these statements is analogous to the proof of Theorem 2.2.

Remark 2.3. Theorems 2.1, 2.2 and Remark 2.2 exhibit the bang-bang nature of the maximizer k_0 .

THEOREM 2.3. *If $f \geq -M$, $F_u(x, t, u) > 0$ for all $(x, t) \in \bar{D}_T$, $u > 0$, then $\text{meas}[\Omega_0 \cap \{k_0 = 0\}] > 0$.*

Proof. If the assertion is not true, then

$$\frac{\partial u_0}{\partial t} - \Delta u_0 = -f - M \quad \text{in } \Omega_0$$

and

$$\frac{\partial u_0}{\partial t} - \Delta u_0 = 0 \geq -f - M \quad \text{a.e. in } D_T \setminus \Omega_0.$$

Thus u_0 is a solution of the parabolic variational inequality with the right-hand side $-f - M$, which is $\leq (-f - k)$ for any $k \in \alpha$. By comparison we conclude that $u_0 \leq u$ for any solution u of (1.5), and clearly $u_0 \neq u$ if $k_0 \neq k$ in Ω_0 . It follows that $J(k_0) < J(k)$ if $k_0 \neq k$ in Ω_0 , a contradiction.

Remark 2.4. Similarly, if $F_u < 0$ and $f \geq 0$ then $\text{meas}[\Omega_0 \cap \{k_0 = M\}] > 0$.

Remark 2.5. The results of this section extend to variational inequalities with any obstacle $\phi(x, t)$; one simply has to work with $\tilde{u} = u - \phi$ and replace f by $\tilde{f} = f - \Delta\phi$. This remark applies as well to the results of the subsequent sections.

Remark 2.6. All the results of this section extend to the case of parabolic operators

$$\frac{\partial}{\partial t} - \sum_{i,j=1}^N a_{ij}(x, t) \frac{\partial^2}{\partial x_i \partial x_j} + \sum_{i=1}^N b_i(x, t) \frac{\partial}{\partial x_i}$$

with $a_{ij} \in C^2$, $b_i \in C^1$; the same is true of the results of the subsequent sections.

Remark 2.7. The results of this section can be extended to the case where $U^0 \geq 0$ on S_T .

3. Generalizations. We begin by extending the results of § 2 to the functional

$$(3.1) \quad J(k) = \sum_{i=1}^m \mu_i \int_D F(x, u(x, t_i)) dx + \iint_{D_T} \Phi(k)$$

where μ_i are given constants and $0 < t_1 < \dots < t_m \leq T$ are given.

Let (k_0, u_0) be a maximizer and let $(k_\varepsilon, u_\varepsilon)$ be a maximizer for the ε -problem. Define Q_ε^i to be the solution of

$$(3.2) \quad \begin{aligned} Q_t + \Delta Q - \beta'_\varepsilon(u_\varepsilon)Q &= 0 \quad \text{in } D_{t_i}, \\ Q(x, t_i) &= F_u(x, u_\varepsilon(x, t_i)) \quad \text{if } x \in D, \\ Q &= 0 \quad \text{on } S_{t_i} \end{aligned}$$

and set

$$(3.3) \quad Q_\varepsilon(x, t) = \sum_{i=1}^m \mu_i Q_\varepsilon^i(x, t) \quad \text{if } (x, t) \in D_{t_i}.$$

Then

$$\begin{aligned} \int_D F_u(x, u_\varepsilon(x, t_i)) z(x, t_i) &= \int_D Q_\varepsilon^i(x, t_i) z(x, t_i) = \iint_{D_{t_i}} (Q_\varepsilon^i z)_t \\ &= \iint_{D_{t_i}} (Q_{\varepsilon,t}^i z + Q_\varepsilon^i z_t) = - \iint_{D_{t_i}} Q_\varepsilon^i L. \end{aligned}$$

Proceeding analogously to Lemma 1.1, we get:

LEMMA 3.1. *If $(k_\varepsilon, u_\varepsilon)$ is a maximizer for the ε -problem corresponding to (3.1) then for any l such that $k_\varepsilon + \delta l \in \mathfrak{a} \forall 0 < \delta < \delta_0$,*

$$\iint_{D_T} (Q^\varepsilon(x, t) - \Phi'(k_\varepsilon) + (k_\varepsilon - k_0))l \geq 0.$$

We can now proceed as in § 2 and prove:

THEOREM 3.2. *Let \mathfrak{a} be given by (2.1). Then for any maximizer (k_0, u_0) for (3.1) there exists a constant λ such that (2.11) holds, where $Q = \lim Q_\varepsilon$ and the Q_ε are given by (3.2), (3.3); the convergence of Q_ε to Q is uniform in compact subsets of Ω_0 .*

Similarly one can extend Theorem 2.2.

Consider next an admissible set

$$(3.4) \quad \mathfrak{a} = \left\{ k \in L^p(D_T), k \geq \delta, \iint_{D_T} k^p(x, t) \leq H \right\}$$

where $\delta > 0$, with functional

$$(3.5) \quad J(k) = \iint_{D_T} F(x, t, u) + \iint_{D_T} k;$$

for simplicity we assume that

$$(3.6) \quad F_u(x, t, u) > 0 \quad \text{if } u > 0.$$

THEOREM 3.3. *For any maximizer k_0 there exists a positive constant μ such that*

$$(3.7) \quad Q - 1 = \mu k_0^{p-1} \quad \text{a.e. in } \Omega_0 \cap \{Q < 1\}.$$

Proof. For the ε -problem we have (cf. [4])

$$(3.8) \quad \begin{aligned} Q_\varepsilon - 1 + k_\varepsilon - k_0 &= -\mu_\varepsilon k_\varepsilon^{p-1} \quad \text{a.e. in } \Omega_\varepsilon \cap \{k_\varepsilon > \delta\}, \\ &\geq 0 \quad \text{a.e. in } \Omega_\varepsilon \cap \{k_\varepsilon = \delta\}, \end{aligned}$$

and $Q_\varepsilon \geq 0$ since $F_u \geq 0$. As in § 2, $Q_\varepsilon \rightarrow Q$ uniformly in compact subsets of Ω_0 , and

$$(3.9) \quad -1 \leq Q_\varepsilon - 1 < -\frac{1}{2} \quad \text{in } (V_{\delta_0} \times (\eta, T)) \setminus G_\varepsilon, \quad \text{meas } G_\varepsilon \rightarrow 0 \quad \text{if } \varepsilon \rightarrow 0,$$

where V_{δ_0} is some δ_0 D -neighborhood of S and η is any positive number; ε is sufficiently small depending on η .

From (3.8), (3.9) and (1.22) it follows that $k_\varepsilon > \delta$ in $(V_{\delta_0} \times (\eta, T)) \setminus G_\varepsilon$ where $\text{meas}(G_\varepsilon) \rightarrow 0$ if $\varepsilon \rightarrow 0$, and $0 < \mu_\varepsilon \leq C$. Integrating the first relation in (3.8) over $(V_{\delta_0} \times (\eta, T)) \setminus G_\varepsilon$, we find that

$$\mu_\varepsilon \iint k_\varepsilon^{p-1} \geq c > 0.$$

Since also $\iint k_\varepsilon^{p-1} \leq C$ for all $k_\varepsilon \in \mathfrak{a}$, we find that $\mu_\varepsilon \geq c > 0$. Hence, for a subsequence, $\mu_\varepsilon \rightarrow \mu$, $0 < \mu < \infty$.

Taking $\varepsilon \rightarrow 0$ in (3.8), (3.7) follows.

Remark 3.1. Theorem 3.1 can be extended to more general $\Phi(k)$ and to other control sets \mathfrak{a} .

Remark 3.2. The results of §§ 2 and 3 can trivially be extended to elliptic variational inequalities. Thus, all the results obtained in [4] are valid also when the condition $F_u > 0$ is dropped.

4. Control on the boundary. In this section we extend the results of the previous sections to problems in which the control occurs on the boundary. For definiteness we take zero initial conditions and Neumann boundary conditions.

Consider the parabolic variational inequality

$$(4.1) \quad \begin{aligned} &\iint_{D_T} u_t(\psi - u) + \iint_{D_T} \nabla u \cdot \nabla(\psi - u) \\ &\geq \iint_{S_T} k(\psi - u) - \iint_{D_T} f(\psi - u) \quad \forall \psi \in K, \quad u \in K \end{aligned}$$

where

$$(4.2) \quad K = \{\psi; \psi \in L^2(0, T; H^1(D)), \psi_t \in L^2(0, T; H^{-1}(D)), \psi(\cdot, 0) = 0, \psi \geq 0 \text{ a.e. in } D_T\}$$

and k is a control variable in the class

$$(4.3) \quad \mathfrak{a} = \left\{ k \in L^\infty(S_T), 0 \leq k \leq M, \iint_{S_T} k = H \right\}.$$

We also introduce the functional

$$(4.4) \quad J(k) = \iint_{D_T} F(x, t, u) + \iint_{S_T} \Phi(k)$$

and consider the problem: Find k_0 such that

$$(4.5) \quad J(k_0) = \max_{k \in \mathfrak{a}} J(k), \quad k_0 \in \mathfrak{a}.$$

We assume that $F(x, t, u)$ satisfies (1.9), (1.12) and that Φ satisfies (1.8), (1.13) with D_T replaced by S_T . We further assume that, for some $\delta_0 > 0$,

$$(4.6) \quad f < 0 \quad \text{in } V_{\delta_0} \times (0, T)$$

where V_δ is a δ D -neighborhood of S .

Let (k_0, u_0) be a solution of (4.5). We introduce the ε -penalized problem

$$(4.7) \quad \begin{aligned} u_t - \Delta u + \beta_\varepsilon(u) &= -f \quad \text{in } D_T, \\ \frac{\partial u}{\partial \nu} &= k \quad \text{on } S_T, \\ u(x, 0) &= 0 \quad \text{if } x \in D. \end{aligned}$$

Denoting by \tilde{u}_ε the solution of (4.7), the corresponding functional for the ε -problem is

$$(4.8) \quad J_\varepsilon(k) = \iint_{D_T} F(x, t, \tilde{u}_\varepsilon) + \iint_{S_T} \Phi(k) - \frac{1}{2} \iint_{S_T} (k_\varepsilon - k_0)^2.$$

Let $(k_\varepsilon, u_\varepsilon)$ be a maximizer for the ε -problem; as in Lemma 1.2, $k_\varepsilon \rightarrow k_0$ strongly in $L^2(S_T)$ and $u_\varepsilon \rightarrow u_0$ uniformly in D_T .

LEMMA 4.1. *If ε is small enough then*

$$(4.9) \quad u_\varepsilon \geq c > 0 \quad \text{in } V_{\delta_0} \times (0, T)$$

for some $c > 0$, $\delta_0 > 0$.

Proof. In the coincidence set of u_0 there holds: $0 \geq -f$, that is $f \geq 0$. In view of (4.6) we conclude that $u_0 > 0$ in $V_{\delta_0} \times (0, T)$. Since further $\partial u_0 / \partial \nu \geq 0$ on S_T , the strong maximum principle shows that $u_0 > 0$ on S_T and, thus, $u_0 \geq 2c > 0$ in $V_{\delta_0} \times (0, T)$. The assertion (4.9) now follows by the uniform convergence of u_ε to u_0 .

Let Q_ε be the solution of

$$(4.10) \quad \begin{aligned} Q_{\varepsilon,t} + \Delta Q_\varepsilon - \beta'_\varepsilon(u_\varepsilon)Q &= -F_u(x, t, u_0) \quad \text{in } D_T, \\ \frac{\partial Q_\varepsilon}{\partial \nu} &= 0 \quad \text{on } S_T, \\ Q_\varepsilon(x, T) &= 0 \quad \text{if } x \in D. \end{aligned}$$

Suppose $k_\varepsilon + \delta l$ is an admissible control for any small $\delta > 0$. Proceeding as in § 2, we deduce that

$$\begin{aligned} 0 &\geq \overline{\lim}_{\delta \rightarrow 0} \frac{1}{\delta} (J_\varepsilon(k_\varepsilon + \delta l) - J_\varepsilon(k_\varepsilon)) \\ &= \overline{\lim}_{\delta \rightarrow 0} \iint_{D_T} F_u(x, t, \tilde{u}_{\varepsilon, \delta}) \frac{u_{\varepsilon, \delta} - u_\varepsilon}{\delta} + \iint_{S_T} \Phi'(k_\varepsilon)l - \iint_{S_T} (k_\varepsilon - k_0)l \\ &= \iint_{S_T} Q_\varepsilon l + \iint_{S_T} \Phi'(k_\varepsilon)l - \iint_{S_T} (k_\varepsilon - k)l. \end{aligned}$$

Thus

$$(4.11) \quad \iint_{S_T} (Q_\varepsilon + \Phi'(k_\varepsilon) - (k_\varepsilon - k))l \leq 0 \quad \text{if } k_\varepsilon + \delta l \in \alpha \quad \forall 0 < \delta < \delta_1.$$

From this we deduce that there is a constant λ_ε and a subset G_ε of S_T of measure $\leq \eta(\varepsilon)$ such that in $S_T \setminus G_\varepsilon$

$$(4.12) \quad k_\varepsilon = \begin{cases} 0 & \text{a.e. on } \{Q_\varepsilon + \Phi'(k_\varepsilon) < \lambda_\varepsilon - \eta(\varepsilon)\}, \\ M & \text{a.e. on } \{Q_\varepsilon + \Phi'(k_\varepsilon) > \lambda_\varepsilon + \eta(\varepsilon)\}, \end{cases}$$

and $\eta(\varepsilon) \rightarrow 0$ if $\varepsilon \rightarrow 0$.

By comparing u_ε with a solution of

$$-\Delta w = C_1 \quad \text{in } D, \quad w_\nu = M \quad \text{on } S$$

where $C_1 > \|f\|_\infty$, we find that

$$(4.13) \quad 0 \leq u_\varepsilon \leq C.$$

Similarly, by comparing Q_ε with a solution of

$$\frac{\partial q}{\partial t} - \Delta q = C_2 \quad \text{in } D_T,$$

$$\frac{\partial q}{\partial \nu} = 0 \quad \text{on } S_T,$$

$$q(x, 0) = q_0,$$

where $C_2 > \sup |F_u|$, we deduce that

$$(4.14) \quad -C \leq Q_\varepsilon \leq C.$$

Making use of Lemma 4.1 and (4.14), we can now go to the limit in (4.12) and deduce:

THEOREM 4.2. *If (k_0, u_0) is a maximizer for (4.5), then there exists a constant λ such that*

$$k_0 = M \quad \text{a.e. in } Q + \alpha_M > \lambda,$$

$$k_0 = 0 \quad \text{a.e. in } Q + \beta_M < \lambda,$$

where $\alpha_M \leq \Phi'(S) \leq \beta_M$ for $0 \leq S \leq M$; in particular, $\alpha_M = \beta_M = 0$ if $\Phi \equiv 0$.

5. An application to the Stefan problem. In this section we consider the Stefan problem in N dimensions, i.e., the problem of melting ice. Given a certain amount of heating energy, we ask: how should we spread it in space and time so as to achieve the largest volume of melted ice? A more precise formulation will be given below. This problem was studied in [5] in case $N = 1$.

Let G (the metal core) be a bounded $C^{2,\alpha}$ domain in \mathbb{R}^N and let E_1 be a domain containing \bar{G} ; set $E = E_1 \setminus \bar{G}$. Let B_R denote a ball of large radius R in \mathbb{R}^N and set $D = B_R \setminus G$.

Physically, E is initially occupied by water and $B_R \setminus E_1$ is occupied by ice at zero temperature. Denote the temperature by θ . Then

$$\theta(x, 0) = h(x) \quad \text{if } x \in E \quad (h > 0)$$

and

$$\theta_t - \Delta \theta = 0 \quad \text{in the water.}$$

On the free boundary $\partial\{\theta > 0\} \cap D_T$ we have $\theta = 0$ as well as the conservation of energy. On the boundary of the core $S \equiv \partial G$

$$(5.1) \quad \frac{\partial \theta}{\partial \nu} = k$$

where ν is the inner normal to ∂G . The left-hand side in (5.1) represents the flux, and k is a control function in the class

$$(5.2) \quad \mathfrak{a} = \left\{ k; k \in L^\infty(S_T), 0 \leq k \leq M, \int \int_{S_T} k = H \right\}.$$

Using the Duvaut transformation

$$u(x, t) = \begin{cases} \int_0^t \theta(x, \tau) d\tau & \text{if } x \in E, \\ \int_{s(x)}^t \theta(x, \tau) d\tau & \text{if } t > s(x), \\ 0 & \text{if } t < s(x) \end{cases}$$

where $t = s(x)$ is the free boundary, one can formally transform the problem into a parabolic variational inequality (for details see [3], [6], [7]):

$$(5.3) \quad \begin{aligned} & \int \int_{D_T} u_t(v - u) + \int \int_{D_T} \nabla u \cdot \nabla(v - u) \\ & \geq \int \int_{S_T} \tilde{k}(v - u) + \int \int_{D_T} f(v - u) \quad \forall v \in K, \quad u \in K \end{aligned}$$

where

$$(5.4) \quad f = \begin{cases} h(x) & \text{if } x \in E, \\ -1 & \text{if } x \in B_R \setminus E, \end{cases}$$

$$(5.5) \quad \tilde{k}(x, t) = \int_0^t k(x, \tau) d\tau$$

and

$$(5.6) \quad K = \{v \in H^1(D_T), v \geq 0 \text{ a.e., } v = 0 \text{ on } \partial B_R \times (0, T), v(\cdot, 0) = 0 \text{ on } D \times \{0\}\}.$$

Using comparison as in [6], [3, p. 87] one can show that if R is chosen sufficiently large (depending on h, M, H, T) then the solution of (5.3) must vanish in a neighborhood of $\partial B_R \times (0, T)$ and, therefore, is independent of the truncating parameter R .

We also have, as in [3], [6], [7], that $u_t \geq 0$ in D_T and $u > 0$ in $\bar{E} \times (0, T)$, and u_t is continuous [3].

The volume of the melted ice at time T is given by

$$\int_D \chi_{\{\theta(x, T) > 0\}} dx = \int_D \chi_{\{u(x, T) > 0\}} dx,$$

which is easily seen (by integrating the variational inequality over $D \times \{T\}$) to be equal to

$$-\int_D u_t(x, T) dx + \text{const.}$$

This functional is not sufficiently regular for our method. We shall therefore average it “slightly,” replacing it by

$$-\frac{1}{h} \int_{T-h}^T \int_D u_t(x, t) dx dt + \text{const.}$$

for any small $h > 0$. Thus the control problem is reduced to minimizing the functional

$$(5.7) \quad J(k) = \int_D u(x, T) dx - \int_D u(x, T_0) dx$$

where $T_0 = T - h$. We shall accordingly replace \mathfrak{a} by

$$(5.8) \quad \mathfrak{a}_{T_0} = \{k \in \mathfrak{a}, k(x, t) = 0 \text{ if } t > T_0\}.$$

Thus the problem is to find k_0 satisfying

$$(5.9) \quad J(k_0) = \min_{k \in \mathfrak{a}_{T_0}} J(k), \quad k_0 \in \mathfrak{a}_{T_0}.$$

Let (k_0, u_0) be a solution of problem (5.9).

Introduce the ε -penalized problem

$$(5.10) \quad \begin{aligned} u_t - \Delta u + \beta_\varepsilon(u) &= f_\varepsilon, \\ u(x, 0) &= u_{0,\varepsilon}(x) \quad \text{if } x \in D, \\ \frac{\partial u}{\partial \nu} &= k \quad \text{on } S_T, \\ u &= 0 \quad \text{on } \partial B_R \times (0, T); \end{aligned}$$

here, as in [3], [6], [7], f_ε is a mollification of f and $u_{0,\varepsilon}$ is a suitably chosen positive function which converges to 0 if $\varepsilon \rightarrow 0$; the functions $\beta_\varepsilon(s)$ are chosen to satisfy

$$(5.11) \quad \begin{aligned} \beta_\varepsilon(s) &= 0 \quad \text{if } s > \varepsilon, \\ \beta'_\varepsilon(s) &= \frac{1}{\varepsilon} \quad \text{if } s < \varepsilon; \end{aligned}$$

notice that $\beta_\varepsilon(0) = -1$. Recall [3], [6] that $\partial u_\varepsilon / \partial t \geq 0$ and thus

$$(5.12) \quad -1 \leq \beta_\varepsilon(u) \leq 0 \quad \text{in } D_T.$$

For the ε -problem the functional is

$$(5.13) \quad J_\varepsilon(k) = J(k) + \frac{1}{2} \iint_{S_T} (k - k_0)^2$$

where the u corresponding to k is the solution of (5.10).

Let $(k_\varepsilon, u_\varepsilon)$ be a minimizer for the ε -problem associated with (5.8), (5.10) and (5.13), (5.7). Denote by Q_ε^T the solution of

$$(5.14) \quad \begin{aligned} Q_t + \Delta Q - \beta'_\varepsilon(u_\varepsilon)Q &= 0 \quad \text{in } D_T, \\ Q_\nu &= 0 \quad \text{on } S_T, \\ Q &= 0 \quad \text{on } \partial B_R \times (0, T), \\ Q(x, T) &= 1 \quad \text{if } x \in D. \end{aligned}$$

Similarly we define $Q_\varepsilon^{T_0}$ and set

$$(5.15) \quad \begin{aligned} \tilde{Q}_\varepsilon &= Q_\varepsilon^T - Q_\varepsilon^{T_0}, \\ W_\varepsilon(x, t) &= \int_t^{T_0} \tilde{Q}_\varepsilon(x, \tau) d\tau. \end{aligned}$$

If $k_\varepsilon + \delta l \in \mathfrak{a}$ for any $0 < \delta < \delta_1$, then proceeding similarly to § 4 we get

$$\iint_{S_{T_0}} \tilde{Q}_\varepsilon \tilde{l} + \iint_{S_{T_0}} (k_\varepsilon - k) l \geq 0$$

where $\tilde{l}(x, t) = \int_0^t l(x, \tau) d\tau$, or

$$(5.16) \quad \iint_{S_{T_0}} (W_\varepsilon + (k_\varepsilon - k)) l \geq 0.$$

Since $u_\varepsilon \geq c > 0$ in $V_{\delta_0} \times (\eta, T)$ where V_{δ_0} is a δ_0 D -neighborhood of S and $\eta > 0$, provided ε is small enough, we deduce, as in § 4, that for a subsequence

$$(5.17) \quad Q_\varepsilon^T \rightarrow Q^T, \quad Q_\varepsilon^{T_0} \rightarrow Q^{T_0} \quad \text{uniformly in } V_{\delta_0} \times (\eta, T_0).$$

LEMMA 5.1. *For any small $\eta > 0$, $\delta_0 > 0$ there exists an $\alpha > 0$ such that*

$$(5.18) \quad Q^T - Q^{T_0} \leq -\alpha \quad \text{in } V_{\delta_0} \times (\eta, T_0).$$

Proof. By the maximum principle $Q_\varepsilon^T(x, T_0) < 1$; hence, by comparison,

$$(5.19) \quad \tilde{Q}_\varepsilon = Q_\varepsilon^T - Q_\varepsilon^{T_0} \leq 0 \quad \text{in } D_{T_0}.$$

Since

$$\frac{\partial}{\partial t} Q_\varepsilon^T + \Delta Q_\varepsilon^T = \beta'_\varepsilon(u_\varepsilon) Q_\varepsilon^T \geq 0,$$

we can compare Q_ε^T with the solution of

$$\begin{aligned} Z_t + \Delta Z &= 0 \quad \text{in } D_T, \\ Z_\nu &= 0 \quad \text{on } S_T, \\ Z &= 0 \quad \text{on } \partial B_R \times (0, T), \\ Z(x, T) &= 0 \quad \text{if } x \in D \end{aligned}$$

and deduce that

$$Q_\varepsilon^T(x, T_0) \leq Z(x, T_0) \leq 1 - \alpha_0 \quad \text{if } x \in D$$

where $\alpha_0 > 0$ is independent of ε . Hence

$$(5.20) \quad \tilde{Q}_\varepsilon(x, T_0) \leq -\alpha_0 \quad \text{if } x \in D.$$

Also

$$\frac{\partial}{\partial t} \tilde{Q}_\varepsilon + \Delta \tilde{Q}_\varepsilon = 0 \quad \text{in } V_{2\delta_0} \times (\eta, T)$$

if ε is small enough (so that $u_\varepsilon \geq c > 0$ in $V_{2\delta_0} \times (\eta, T)$). We compare \tilde{Q}_ε with the solution \tilde{Z} of

$$\frac{\partial}{\partial t} \tilde{Z} + \Delta \tilde{Z} = 0 \quad \text{in } V_{2\delta_0} \times (\eta, T),$$

$$\begin{aligned}\tilde{Z}_\nu &= 0 \quad \text{on } S_{T_0}, \\ \tilde{Z} &= 0 \quad \text{on } (\partial V_{2\delta_0} \cap D) \times (\eta, T), \\ \tilde{Z}(x, T_0) &= -\alpha_0.\end{aligned}$$

When we recall (5.19) and (5.20) it is clear that $\tilde{Q}_\varepsilon \leq \tilde{Z}$. On the other hand, by the maximum principle, $\tilde{Z} \leq -\alpha < 0$ in $V_{\delta_0} \times (\eta, T)$ for some α depending on δ_0 and α_0 . Consequently,

$$(5.21) \quad Q_\varepsilon^T - Q_\varepsilon^{T_0} \leq \tilde{Z} \leq -\alpha \quad \text{in } V_{\delta_0} \times (\eta, T_0)$$

and, taking $\varepsilon \rightarrow 0$, the assertion (5.18) follows.

Set

$$(5.22) \quad W(x, t) = \int_t^{T_0} (Q^T(x, \tau) - Q^{T_0}(x, \tau)) d\tau.$$

From Lemma 5.1 we have that

$$(5.23) \quad W_t(x, t) \geq \alpha \quad \text{in } V_{\delta_0} \times (\eta, T_0).$$

Now, from (5.16) we deduce that there is a subset G_ε of S_T of measure $\leq \delta(\varepsilon)$ such that in $S_T \setminus G_\varepsilon$

$$\begin{aligned}k_\varepsilon &= 0 & \text{if } W_\varepsilon > \lambda_\varepsilon + \eta(\varepsilon), \\ k_\varepsilon &= M & \text{if } W_\varepsilon < \lambda_\varepsilon - \eta(\varepsilon),\end{aligned}$$

and $\delta(\varepsilon) \rightarrow 0$ if $\varepsilon \rightarrow 0$. Going to the limit with $\varepsilon \rightarrow 0$ and proceeding as in § 4 we find that the strong L^2 limit k_0 satisfies:

$$(5.24) \quad \begin{aligned}k_0 &= 0 & \text{if } W > \lambda, \\ k_0 &= M & \text{if } W < \lambda\end{aligned}$$

for some λ . In view of (5.23) we also conclude that

$$(5.25) \quad \begin{aligned}k_0(x, t) &= M & \text{if } 0 < t < \phi(x), \quad x \in S, \\ k_0(x, t) &= 0 & \text{if } \phi(x) < t < T_0, \quad x \in S\end{aligned}$$

for some function $\phi(x)$; further, since W is continuously differentiable on S_{T_0} and W_t is strictly negative, it easily follows that $\phi(x)$ is continuously differentiable about any point x where $\phi(x) > 0$. We have thus proved

THEOREM 5.2. *If (k_0, u_0) is a solution of problem (5.9) then there exists a function $W(x, t)$ defined by (5.22), (5.17) and a continuous function $\phi(x)$ defined for $x \in S$ such that k_0 satisfies (5.24), (5.25); $\phi(x)$ is continuously differentiable on the set $\{\phi > 0\}$.*

We conclude this section by considering the maximum problem for the functional

$$(5.26) \quad J(k) = \iint_{D_T} \theta(x, t) dx dt$$

in the class \mathfrak{a} defined in (5.2); this functional represents the thermal energy of the fluid. Since $\theta = u_t$, we have, up to an additive constant, that

$$(5.27) \quad J(k) = \int_D u(x, T) dx.$$

The maximization problem leads to

$$(5.28) \quad \iint_{S_T} \left[\left(\int_{t_0}^T Q_\varepsilon(x, \tau) d\tau \right) - (k_\varepsilon - k)(x, t) \right] l(x, t) \leq 0$$

if $k_\varepsilon + \delta l \in \alpha$ for all small $\delta > 0$, where Q_ε is the solution of

$$Q_t + \Delta Q - \beta'_\varepsilon(u_\varepsilon)Q = 0 \quad \text{in } D_T,$$

$$Q_\nu = 0 \quad \text{on } S_T,$$

$$Q = 0 \quad \text{on } \partial B_R \times (0, T),$$

$$Q(x, T) = 1 \quad \text{if } x \in D.$$

From this we can deduce, as before, that any maximizer k_0 satisfies (5.24), (5.25), with replaced by the limit (say \tilde{W}) of the functions $\int_{t_0}^T Q_\varepsilon(x, \tau) d\tau$, and $\phi(x)$ is replaced by another continuous function (say $\tilde{\phi}(x)$).

REFERENCES

- [1] V. BARBU, *Optimal Control of Variational Inequalities*, Pitman, London, 1984.
- [2] A. FRIEDMAN, *Stochastic Differential Equations and Applications*, Vol. 2, Academic Press, New York, 1976.
- [3] ———, *Variational Principles and Free Boundary Problems*, John Wiley, New York, 1982.
- [4] ———, *Optimal control for variational inequalities*, this Journal, 24 (1986), pp. 439–451.
- [5] A. FRIEDMAN AND L. JIANG, *Nonlinear optimal control problems in heat conduction*, this Journal, 21 (1983), pp. 940–952.
- [6] A. FRIEDMAN AND D. KINDERLEHRER, *A one phase Stefan problem*, Indiana Univ. Math. J., 24 (1975), pp. 1005–1035.
- [7] D. KINDERLEHRER AND G. STAMPACCHIA, *An Introduction to Variational Inequalities and Their Applications*, Academic Press, New York, 1980.
- [8] O. A. LADYZENSKAJA, V. A. SOLONNIKOV AND N. N. URAL'CEVA, *Linear and Quasilinear Equations of Parabolic Type*, Amer. Math. Soc. Transl., Providence, RI, 1968.
- [9] J. L. LIONS, *Optimal Control of Systems Governed by Partial Differential Equations*, Springer-Verlag, Berlin, 1971.
- [10] ———, *Some aspects of optimal control of distributed parameter systems*, Regional Conference Series in Applied Mathematics, Soc. Indust. Appl. Math., Philadelphia, 1972.
- [11] F. MIGNOT AND J. P. PUEL, *Optimal control in some variational inequalities*, this Journal, 22 (1984), pp. 466–476.
- [12] ———, *Contrôle optimal d'un système gouverné par une inéquation variationnelle parabolique*, C.R. Acad. Sci. Paris Sér. I. Math., 298 (1984), pp. 277–280.

ACAUSAL REALIZATION THEORY, PART I; LINEAR DETERMINISTIC SYSTEMS*

A. J. KRENER†

Abstract. We study acausal linear systems, their controllability and observability properties and the weighting patterns that they realize. A complete classification is given of all minimal real analytic realizations of a given weighting pattern and of all minimal autonomous realizations of a stationary weighting pattern.

Key words. acausal linear systems, controllable and observable on and off, minimal realization

AMS(MOS) subject classifications. 93B20, 34B05, 47A50

1. Introduction. Acausal linear systems theory is concerned with mathematical entities of the form

$$(1.1a) \quad \dot{x} = Ax + Bu,$$

$$(1.1b) \quad V^0 x(t_0) + V^1 x(t_1) = v,$$

$$(1.1c) \quad y = Cx + Du,$$

$$(1.1d) \quad w = W^0 x(t_0) + W^1 x(t_1)$$

where $x(t)$, v , $w \in \mathbb{R}^n$, $u(t) \in \mathbb{R}^m$ and $y(t) \in \mathbb{R}^p$. The matrices A , B , C , D , V^0 , V^1 , W^0 and W^1 are dimensioned accordingly and A , B , C , D may be bounded measurable functions of t . We refer to $x(t)$ as the *state*, $u(t)$ as the *input* and $y(t)$ the *output* at time t , although t may actually represent a spatial parameter. The vector v is called the *boundary input* and the vector w the *boundary output*. We refer to (1.1) as the acausal system Σ .

We always assume that (1.1a), (1.1b) is a *well-posed problem*, i.e., for every boundary input v and square integrable $u(t)$ there exists a unique solution $x(t)$. In this case (1.1) defines a linear mapping also denoted by Σ .

$$(1.2a) \quad \Sigma: \mathbb{R}^{n \times 1} \times L_2^{m \times 1}[t_0, t_1] \rightarrow \mathbb{R}^{n \times 1} \times L_2^{p \times 1}[t_0, t_1],$$

$$(1.2b) \quad \Sigma: (v, u(t)) \mapsto (w, y(t)).$$

We say that such a mapping is the *input output map* of the acausal system (1.1) or equivalently that the acausal system (1.1) is a *realization* of the input output mapping.

In § 2 we discuss situations where such models naturally arise. Of course (1.1) is a generalization of the usual linear system where $V^0 = W^1 = I$ and $V^1 = W^0 = 0$. Such a system is *causal* because future inputs do not affect past states or outputs. Systems of the form (1.1) do not necessarily have this property, hence the term *acausal*. There are many possible generalizations of (1.1) which are of interest. We shall mention some of these in § 2, but we shall not discuss them in any great depth.

Section 3 is essentially a review of [1] where systems of the form (1.1) were first introduced under the name of *boundary value linear systems*.

* Received by the editors November 5, 1984; accepted for publication (in revised form) February 24, 1986. This research was supported in part by a Senior Postdoctoral Research Fellowship from the Science Research Council of Great Britain, by the National Science Foundation under grant MCS 8300884, and by the National Aeronautics and Space Administration under grant NAG 2-268.

† Department of Mathematics, University of California, Davis, California 95616.

In § 4, we exhibit two other processes closely related to $x(t)$ but which are causal in some generalized sense. These processes supply the foundations for the definitions of controllability and observability given in § 5. Also in this section we begin to relate controllability and observability to minimality.

The key results of this paper are found in §§ 6 and 7. The first is a complete classification of all the minimal real analytic realizations of the mapping

$$(v=0, u(\cdot)) \mapsto y(\cdot)$$

induced by (1.1). The second is complete classification of all the minimal autonomous realizations of such maps which are stationary.

Recently Gohberg and Kaashoek [9], [10] have made an excellent study of such systems. Their work is based on completely different concepts of controllability and observability. They do not discuss the question of minimality but treat a different question of irreducibility. At the end of § 5 we give an example that illustrates the differences between their work and ours.

2. Examples and extensions. Acausal systems naturally arise when the independent variable t is spatial rather than temporal. For example, consider a static, approximately horizontal beam, clamped at both ends, which supports a continuously distributed load. We can view this from a system theoretic point of view where the input $u(t)$ is the load density and the output $y(t)$ is the deflection of the beam. ($y(t) > 0$ indicates downward deflection.) The variable t measures length along the beam. The relationship between input and output is given by

$$(2.1) \quad E(t)I(t) \frac{d^4 y}{dt^4} = u(t)$$

where $E(t)$ is the modulus of elasticity and $I(t)$ is the moment of inertia of the cross section. Clamping at both ends imposes boundary conditions on $y(t_0)$, $y(t_1)$, $\dot{y}(t_0)$ and $\dot{y}(t_1)$.

This can be put in state space form (1.1) by letting $x = (x_1, x_2, x_3, x_4) = (y, \dot{y}, \ddot{y}, \dddot{y})$; then

$$\begin{aligned} A &= \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \end{bmatrix}, & B &= \begin{bmatrix} 0 \\ 0 \\ 0 \\ 1/EI \end{bmatrix}, \\ C &= [1 \ 0 \ 0 \ 0], & D &= [0], \\ V^0 &= \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}, & V^1 &= \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix}. \end{aligned}$$

A similar example was considered in [2] but with the bending moment as input, which resulted in a second order system.

The acausal nature of this system is apparent, the load at point s affects the deflection at every point t along the beam. One goal of the loading might be to force the beam to assume a desired shape. Such a problem can be cast as a linear quadratic optimal control problem [2].

The boundary condition (1.1) can be used to force $x(t)$ to be *cyclic*, $x(t_0) = x(t_1)$ by letting $V^0 = -V^1$ and $v = 0$. Similarly if $V^0 = V^1$ and $v = 0$ we obtain an *anticyclic* state process $x(t_0) = -x(t_1)$. Such phenomena cannot be modeled by causal systems.

If we drive (1.1) by white Gaussian noise process $u(t)$ and an independent Gaussian boundary value v then we obtain two Gaussian processes $x(t)$ and $y(t)$. If a causal system is so driven, then the state process $x(t)$ is Markov but for most acausal systems the state process is not Markov.

These stochastic systems are a convenient way of representing stochastic processes. We study the associated realization theory in the sequel [4] to this paper. Using this class of models one can formulate and solve various estimation problems for spatially distributed processes [3], [5], [6].

Even processes where the independent variable t is temporal can have an acausal character. These are systems which are anticipatory. There is an intelligent controller who modifies the evolution of the system in order to achieve a desired goal at some future time. This may be on a fixed time interval or over a moving time interval. Most messages are of this type. Before composing the message, the author usually has a fairly clear idea of what the contents should include, and particularly how it should begin and end. Another example is the tracking of an object whose ultimate destination is already known.

Causality is a property of the mapping from inputs $u(\cdot)$ to outputs $y(\cdot)$. Suppose one is studying a system where the inputs and outputs are not known a priori as is frequently the case in network theory. If it is impossible to decide a priori whether the process one wishes to model is causal or not, why restrict a priori to causal models? Besides (1.1), Luenberger's descriptor systems can be used to model acausality [12].

There are numerous extensions of the acausal linear system (1.1) which we will not go into in any depth. A straightforward one is to discretize t or we could let $t_0 = -\infty$ and/or $t_1 = \infty$. A more substantial generalization is to allow t to be a multidimensional variable. Such systems arise in distributed parameter control, image processing, and seismic data processing. It is somewhat surprising considering all the effort that has gone into these areas that one-dimensional acausal systems have not received more study.

Throughout this paper we consider only well-posed systems. This rules out many interesting problems. For example in the stochastic setting, we rule out a pinned Wiener process (Brownian Bridge). Generally we restrict our attention to two point boundary value processes where the solution $x(t)$ of (1.1a) is partially constrained at only two times t_0, t_1 as in (1.1b). But multipoint constrained problems will arise even in this paper. They have wide applicability in many other contexts.

3. Basic facts. Let $\Phi(t, s)$ be $n \times n$ matrix valued function satisfying

$$(3.1a) \quad \frac{\partial}{\partial t} \Phi(t, s) = A(t) \Phi(t, s),$$

$$(3.1b) \quad \Phi(t, t) = I.$$

Since $A(t)$ is assumed to be bounded and measurable, the existence, uniqueness and absolute continuity of $\Phi(t, s)$ follows from standard theorems on ODE's.

The boundary value problem (1.1a), (1.1c) is *well posed* iff the matrix

$$(3.2) \quad F = V^0 + V^1 \Phi(t_1, t_0)$$

is invertible. If this is satisfied then the solution to (1.1a), (1.1b) is given by

$$(3.3) \quad x(t) = \Phi(t, t_0) F^{-1} v + \int_{t_0}^{t_1} G(t, s) B(s) u(s) ds.$$

Green's matrix $G(t, s)$ is given by splicing together two matrix valued functions

$$(3.4a) \quad G(t, s) = \begin{cases} G^0(t, s) & \text{if } t > s, \\ G^1(t, s) & \text{if } s > t \end{cases}$$

along the line $t = s$, where

$$(3.4b) \quad G^0(t, s) = \Phi(t, t_0)F^{-1}V^0\Phi(t_0, s),$$

$$(3.4c) \quad G^1(t, s) = -\Phi(t, t_0)F^{-1}V^1\Phi(t_1, s) = \Phi(t, t_0)F^{-1}(V^0 - I)\Phi(t_0, s).$$

The output $y(t)$ is given by

$$(3.5) \quad y(t) = C(t)\Phi(t, t_0)F^{-1}v + \int_{t_0}^{t_1} W(t, s)u(s) ds$$

where the *weighting pattern* $W(t, s)$ is given by

$$(3.6) \quad W(t, s) = C(t)G(t, s)B(s) + D(t)\delta(t - s).$$

The system (1.1) defines a linear mapping Σ

$$\begin{aligned} \Sigma: \mathbb{R}^{n \times 1} \times L_2^{m \times 1}[t_0, t_1] &\rightarrow \mathbb{R}^{n \times 1} \times L_2^{p \times 1}[t_0, t_1], \\ \Sigma: \begin{pmatrix} v \\ u(\cdot) \end{pmatrix} &\mapsto \begin{pmatrix} w \\ y(\cdot) \end{pmatrix} = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix} \begin{pmatrix} v \\ u(\cdot) \end{pmatrix}. \end{aligned}$$

The maps Σ_{11} , Σ_{12} and Σ_{21} are of finite rank since they map from and/or to finite-dimensional vector spaces. The most important part of Σ is the infinite rank part Σ_{22} which maps between the function spaces. The usefulness of the model (1.1) depends ultimately on the fact that it describes an infinite rank mapping in a very concise and tractable fashion.

Suppose we have an integral operator Σ_{22}

$$\begin{aligned} \Sigma_{22}: L_2^{m \times 1}[t_0, t_1] &\rightarrow L_2^{p \times 1}[t_0, t_1], \\ \Sigma_{22}: u(\cdot) &\mapsto y(\cdot) \end{aligned}$$

where

$$y(t) = \int_{t_0}^{t_1} W(t, s)u(s) ds.$$

The system (1.1) is said to be a *realization* of Σ_{22} (or equivalently the kernel $W(t, s)$) if $W(t, s)$ is the *weighting pattern* of (1.1) as given by (3.6). Realization theory (in the deterministic sense) is concerned with the existence and classification of the realizations of $W(t, s)$ and related questions. As an example of such a question consider the adjoint map Σ^* of Σ

$$(3.7a) \quad \Sigma^*: \mathbb{R}^{1 \times n} \times L_2^{1 \times p}[t_0, t_1] \rightarrow \mathbb{R}^{1 \times n} \times L_2^{1 \times m}[t_0, t_1],$$

$$(3.7b) \quad \Sigma^*: (\zeta, \mu(t)) \mapsto (\xi, \nu(t))$$

defined by the equation

$$(3.8) \quad \xi v + \int_{t_0}^{t_1} \nu(t)u(t) dt = \zeta w + \int_0^t \mu(t)y(t) dt$$

for all $(v, u(t))$ and $(\zeta, \mu(t))$. An obvious question is whether Σ^* can be realized by an acausal linear system. As was shown in [1] the following system (given in adjoint

form) does the job. The functions $\lambda(t)$, $\mu(t)$ and $\nu(t)$ are the state, input and output, respectively, ζ denotes the boundary input and ξ the boundary output.

$$(3.9a) \quad \dot{\lambda} = -\lambda A - \mu C,$$

$$(3.9b) \quad \lambda(t_0)M^0 + \lambda(t_1)M^1 = \zeta,$$

$$(3.9c) \quad \nu = \lambda B + \mu D,$$

$$(3.9d) \quad \xi = \lambda(t_0)N^0 + \lambda(t_1)N^1.$$

The boundary matrices are fixed by

$$(3.10) \quad \begin{bmatrix} V^0 & V^1 \\ W^0 & W^1 \end{bmatrix} \begin{bmatrix} -M^0 & -N^0 \\ M^1 & N^T \end{bmatrix} = \begin{bmatrix} 0 & -I \\ I & 0 \end{bmatrix}.$$

It is convenient to make a change of coordinates in the space of boundary input values v so that

$$(3.11) \quad F = V^0 + V^1\Phi(t_1, t_0) = I.$$

If V^0 and V^1 satisfy this, then the boundary conditions are in *standard form* and Σ is a *standard realization* of $W(t, s)$. Causal linear systems where $V^0 = I$ and $V^1 = 0$ have boundary conditions in standard form.

Frequently W^0 and W^1 are not explicitly given, but the dual systems can still be determined if one assumes that the dual boundary conditions (3.9c) satisfy the dual of (3.11), i.e.,

$$(3.12) \quad \Phi(t_1, t_0)M^0 + M^1 = I.$$

By equating the upper left blocks of (3.10) and (3.12) we see that

$$(3.13a) \quad M^0 = V^1,$$

$$(3.13b) \quad M^1 = \Phi(t_1, t_0)V^0\Phi(t_0, t_1).$$

These conditions (3.13) are called the *standard dual boundary conditions*.

Notice in the causal case when $V^0 = I$ and $V^1 = 0$, the standard dual boundary conditions are $M^0 = 0$ and $M^1 = I$. Systems with such boundary conditions are said to be *anticausal*, because under time reversal they become causal.

Having computed the standard dual boundary conditions we can in the same way compute the *standard* boundary output equation (1.1d). We assume that

$$(3.14) \quad W^0\Phi(t_0, t_1) + W^1 = I;$$

then this and the lower left block of (3.10) imply that

$$(3.15a) \quad W^0 = \Phi(t_1, t_0)(V^0 - I),$$

$$(3.15b) \quad W^1 = I + \Phi(t_1, t_0)V^1.$$

This normalization has been chosen so that for causal systems $W^0 = 0$ and $W^1 = I$, and the boundary output is $w = x(t_1)$.

It is a simple exercise to verify that for the standard boundary equations (1.1b), (1.1d) the matrix

$$(3.16) \quad \begin{bmatrix} V^0 & V^1 \\ W^0 & W^1 \end{bmatrix}$$

is nonsingular. This means that the boundary inputs and outputs are linearly independent variables. A *standard acausal system* (1.1) is one where the coefficients of the boundary equations (1.1b), (1.1d) satisfy the normalizations (3.11) and (3.14).

As we have seen, (1.1) defines a linear mapping

$$(3.17) \quad \Sigma_{22}: u(t) \mapsto y(t) = \int_{t_0}^{t_1} W(t, s) u(s) ds$$

where

$$(3.18) \quad W(t, s) = C(t)G(t, s)B(s) + D(t)\delta(t - s).$$

A kernel $W(t, s)$ is called *proper* if it is the sum of a bounded measurable term and a Dirac delta term as is (3.18). It is called *strictly proper* if the second term is missing, e.g., $D(t) = 0$. The next theorem is trivial but occasionally useful.

THEOREM 3.1. *A strictly proper $p \times m$ kernel $W(t, s)$ can be realized by an acausal linear system (1.1) iff there exists bounded measurable $p \times n$ and $n \times m$ matrices $C(t)$ and $B(s)$ and $n \times n$ matrices V^0 and V^1 such that*

$$(3.19) \quad V^0 + V^1 = I,$$

$$(3.20) \quad W(t, s) = \begin{cases} C(t)V^0B(s), & t > s, \\ -C(t)V^1B(s), & t < s. \end{cases}$$

Proof. If $C(t)$, $B(s)$, V^0 and V^1 exist then $W(t, s)$ is realized by (1.1) with $A = 0$. On the other hand, if $W(t, s)$ is realized by any acausal system (1.1) then the time dependent change of state coordinates given by $\tilde{x} = \Phi(t_0, t)x$ transforms (1.1) into an acausal system $\tilde{\Sigma}$ where $\tilde{A} = 0$, $\tilde{C}(t) = C(t)\Phi(t, t_0)$ and $\tilde{B}(s) = \Phi(t_0, s)B(s)$. The transformed boundary conditions $\tilde{V}^0 = V^0$ and $\tilde{V}^1 = V^1\Phi(t_1, t_0)$ are in standard form. It is straightforward to verify that $\tilde{\Sigma}$ also realizes $W(t, s)$. Q.E.D.

4. The inward and outward boundary value processes. In general a system of the form (1.1) defines an acausal mapping from $u(\cdot)$ to $y(\cdot)$. The output $y(t)$ at time t depends on the input $u(s)$ at all times $s \in [t_0, t_1]$, not just those $s \in [t_0, t]$ as for causal systems. A natural question to ask is whether there is any causal way of looking at (1.1).

It turns out that a related question is how to view the boundary input condition (1.1c) and the boundary output equation (1.1d). For causal systems the boundary input condition is just $x(t_0) = v$ and the boundary output equation is $w = x(t_1)$. The initial value v can be thought of as the medium which transmits the effects of past controls $u(s)$, $s \in (-\infty, t_0)$ to the state $x(t)$ and output $y(t)$ for $t \in [t_0, t_1]$. The state $x(t)$ and output $y(t)$ can also be determined if w and $u(s)$, $s \in [t, t_1]$ are known. But this is not really an anticausal representation because w depends on v and $u(s)$ for all $s \in [t_0, t_1]$.

We now present a similar interpretation of v and w in the acausal case. We need to introduce moving boundary conditions. Suppose $t_0 \leq \tau_0 \leq \tau_1 \leq t_1$, then define four matrices

$$(4.1a) \quad K^0 = \Phi(\tau_0, t_0)V^0\Phi(t_0, \tau_0),$$

$$(4.1b) \quad K^1 = \Phi(\tau_0, t_0)V^1\Phi(t_1, \tau_1),$$

$$(4.1c) \quad L^0 = \Phi(\tau_1, t_1)W^0\Phi(t_0, \tau_0),$$

$$(4.1d) \quad L^1 = \Phi(\tau_1, t_1)W^1\Phi(t_1, \tau_1).$$

These matrices are functions of τ_0 and τ_1 but for notational simplicity we suppress the arguments. We shall assume that V^0 , V^1 , W^0 and W^1 are in standard form on the interval $[t_0, t_1]$, then it is easy to see that K^0 , K^1 , L^0 and L^1 are in standard form on the interval $[\tau_0, \tau_1]$.

Let $x(t)$ be the solution of (1.1) for some $u(t)$ and v . We define the inward boundary process $k(\tau_0, \tau_1)$ by

$$(4.2) \quad k(\tau_0, \tau_1) = K^0 x(\tau_0) + K^1 x(\tau_1).$$

There are two important points to be made about this process. A simple calculation shows that

$$(4.3) \quad k(\tau_0, \tau_1) = \Phi(\tau_0, t_0)v + \left(\int_{t_0}^{\tau_0} + \int_{\tau_1}^{t_1} \right) G(\tau_0, s)B(s)u(s) ds.$$

We can interpret this as causality in some generalized sense for the mapping $u(\cdot) \mapsto k(\cdot, \cdot)$. Think of the pair $\{\tau_0, \tau_1\}$ as being the present; the past is $[t_0, t_1] \setminus [\tau_0, \tau_1]$ and the future is (τ_0, τ_1) . Then (4.3) says that the present value of $k(\tau_0, \tau_1)$ does not depend on future values of the input $u(\tau)$ for $\tau \in (\tau_0, \tau_1)$.

The second point is that given the present value $k(\tau_0, \tau_1)$ and future values $u(\tau)$, $\tau \in (\tau_0, \tau_1)$ we can compute future values $x(\tau)$ and $y(\tau)$ for $\tau \in (\tau_0, \tau_1)$. This is because the boundary value problem

$$(4.4a) \quad \dot{x} = Ax + Bu, \quad t \in [\tau_0, \tau_1],$$

$$(4.4b) \quad K^0 x(\tau_0) + K^1 x(\tau_1) = k,$$

is well posed, and its solution for given u and v agrees with that of (1.1) on $[\tau_0, \tau_1]$ provided $k = k(\tau_0, \tau_1)$ given by (4.3).

Note that $k(t_0, t_1) = v$ and so v can be thought of as the medium that transmits the effects of $u(s)$ for $s \notin [t_0, t_1]$ to $x(t)$ and $y(t)$ for $t \in [t_0, t_1]$.

If we consider the process $l(\tau_0, \tau_1)$ defined by

$$(4.5) \quad l(\tau_0, \tau_1) = L^0 x(\tau_0) + L^1 x(\tau_1),$$

then

$$(4.6) \quad l(\tau_0, \tau_1) = \Phi(\tau_1, \tau_0)k(\tau_0, \tau_1) + \int_{\tau_0}^{\tau_1} \Phi(\tau_1, s)B(s)u(s) ds.$$

From knowledge of $l(\tau_0, \tau_1)$ and $u(t)$ for $t \in [t_0, t_1] \setminus (\tau_0, \tau_1)$ we can reconstruct $x(t)$ for $t \in [t_0, t_1] \setminus (\tau_0, \tau_1)$ as the solution of the well-posed four point boundary value problem

$$(4.7a) \quad \dot{x} = Ax + Bu, \quad t \in [t_0, t_1] \setminus (\tau_0, \tau_1),$$

$$(4.7b) \quad V^0 x(t_0) + V^1 x(t_1) = v,$$

$$(4.7c) \quad L^0 x(\tau_0) + L^1 x(\tau_1) = l(\tau_0, \tau_1).$$

However the map $u(\cdot) \mapsto l(\cdot, \cdot)$ is not causal in any generalized sense because $l(\tau_0, \tau_1)$ depends on $u(s)$ for all $s \in [t_0, t_1]$.

We have just seen that the inward boundary value process of an acausal system plays a role similar to that of the forward moving state of a causal system. For acausal systems there is also an outward boundary value process which plays the same role as the future jump $x(t_1) - \Phi(t_1, \tau_0)x(\tau_0)$ of the state of a causal process caused by $u(\tau)$ differing from 0 on $[\tau_0, t_1]$.

Given $u(\tau)$ for $\tau \in [\tau_0, \tau_1]$, define $z(\tau)$ and $j(\tau_0, \tau_1)$ by

$$(4.8a) \quad \dot{z} = Az + Bu,$$

$$(4.8b) \quad z(\tau_0) = 0,$$

$$(4.8c) \quad j(\tau_0, \tau_1) = z(\tau_1).$$

The *outward boundary value process* is $j(\tau_0, \tau_1)$ and has two important properties similar to those of $k(\tau_0, \tau_1)$. The first is that

$$j(\tau_0, \tau_1) = \int_{\tau_0}^{\tau_1} \Phi(\tau_1, s) B(s) u(s) ds,$$

so if we think of $\{\tau_0, \tau_1\}$ as the present, $[t_0, t_1] \setminus [\tau_0, \tau_1]$ as the past and (τ_0, τ_1) as the future, then the mapping $u(\cdot) \mapsto j(\tau_0, \tau_1)$ is anticausal. In other words $j(\tau_0, \tau_1)$ does not depend on past values of $u(\tau)$.

The second point is that given the present value $j(\tau_0, \tau_1)$ and past values $u(t)$, $t \in [t_0, t_1] \setminus [\tau_0, \tau_1]$ we can compute past values $x(t)$ and $y(t)$ for $t \in [t_0, t_1] \setminus [\tau_0, \tau_1]$. This is because the solution $x(t)$ of the well-posed four point boundary value problem (4.7a), (4.7b), and

$$(4.7d) \quad -\Phi(\tau_1, \tau_0)x(\tau_0) + x(\tau_1) = j$$

agrees with the solution of (1.1) if $j = j(\tau_0, \tau_1)$ given by (4.8).

We have chosen the letter j for the outward boundary value process because it represents the jump that the state experiences between times τ_0 and τ_1 because of the control $u(t) \neq 0$, $t \in [\tau_0, \tau_1]$.

This suggests another viewpoint on the boundary value v . It is possible that the process $x(t)$ lies on some compactified version of the real line and v is the jump that the state experiences from time t_1 through infinity to time t_0 because of the effects of the control $u(t)$ for $t \notin [t_0, t_1]$.

5. Controllability and observability. Suppose the map $\Sigma: (v, u(\cdot)) \mapsto (w, y(\cdot))$ arises from the state space model (1.1); then it factors into a mapping $(v, u(\cdot)) \mapsto x(\cdot)$ followed by a mapping $x(\cdot) \mapsto (w, y(\cdot))$. It is natural that realization theory be concerned with these factor mappings. The critical issues for the minimality of a realization Σ are whether the first factor is onto and the second is one to one in some sense. In the systems literature this first property is called controllability and the second is called observability. For linear time invariant causal systems any two reasonable definitions of controllability are equivalent. The same holds for any two reasonable definitions of observability. But for time varying and/or nonlinear causal systems there are several nonequivalent definitions whose utility varies with the problem of the moment. Therefore it should come as no surprise that there are at least two useful definitions of both controllability and observability for acausal systems.

The first two definitions relate to the inward boundary value process $k(\tau_0, \tau_1)$.

DEFINITION. The system (1.1) is *controllable off* $[\tau_0, \tau_1]$ if the map

$$(5.1) \quad \{u(t): t \in [t_0, t_1] \setminus [\tau_0, \tau_1]\} \mapsto k(\tau_0, \tau_1) = \left(\int_{t_0}^{\tau_0} + \int_{\tau_1}^{t_1} \right) G(\tau_0, s) B(s) u(s) ds,$$

defined by (4.3) (or equivalently (4.2)) where $v = 0$, is onto.

DEFINITION. The system (1.1) is *observable on* $[\tau_0, \tau_1]$ if the map

$$(5.2) \quad k \mapsto \{y(\tau) = C\Phi(\tau, \tau_0)k: \tau \in [\tau_0, \tau_1]\},$$

defined for $\tau \in [\tau_0, \tau_1]$ by (4.4a), (4.4b) and (1.1c), is 1-1. The control $u(t)$ is assumed to be zero on $[\tau_0, \tau_1]$.

As mentioned before, in a general sense controllability and observability are the two halves of minimality. The next result shows that we are on the right track.

THEOREM 5.1. Let the system (1.1) be a realization of the weighting pattern $W(t, s)$. If there exists τ_0 and τ_1 where $t_0 < \tau_0 < \tau_1 < t_1$ such that (1.1) is controllable off $[\tau_0, \tau_1]$

and observable on $[\tau_0, \tau_1]$, then (1.1) is minimal, i.e., of minimal state dimension among all realizations of $W(t, s)$. Moreover, if $\tilde{\Sigma}$ is any other minimal realization of $W(t, s)$ then $\tilde{\Sigma}$ is also controllable off and observable on $[\tau_0, \tau_1]$.

We defer the proof of this theorem for the moment. To check controllability off $[\tau_0, \tau_1]$ and observability on $[\tau_0, \tau_1]$ we need to compute the Gramians

$$(5.3) \quad \mathcal{C}[\tau_0, \tau_1] = \left(\int_{\tau_0}^{\tau_1} + \int_{\tau_1}^{t_1} \right) G(\tau_0, s) B(s) B^*(s) G^*(\tau_0, s) ds,$$

$$(5.4) \quad \mathcal{O}[\tau_0, \tau_1] = \int_{\tau_0}^{\tau_1} \Phi^*(t, \tau_0) C^*(t) C(t) \Phi(t, \tau_0) dt$$

where $*$ denotes transpose. The following is a standard exercise in linear systems theory; see, for example, Desoer [11] or Brockett [17].

PROPOSITION 5.2. *The system (1.1) is controllable off $[\tau_0, \tau_1]$ iff $\mathcal{C}[\tau_0, \tau_1]$ is positive definite. The system (1.1) is observable on $[\tau_0, \tau_1]$ iff $\mathcal{O}[\tau_0, \tau_1]$ is positive definite.*

Remark. $\mathcal{O}[\tau_0, \tau_1]$ is the observability Gramian of the causal system with the same A and C matrices. Therefore observability on is the same as causal observability.

Proof of Theorem 5.1. Let Σ and $\tilde{\Sigma}$ be realizations of $W(t, s)$. (Σ is given by (1.1) and $\tilde{\Sigma}$ by a similar acausal system with tildes, i.e., \tilde{x} , \tilde{A} , etc.) Since Σ is controllable off $[\tau_0, \tau_1]$, $\mathcal{C}[\tau_0, \tau_1]$ is invertible. Given $k \in \mathbb{R}^n$ define a control $u(t; k)$ with support off (τ_0, τ_1) (i.e., with support in $[t_0, t_1] \setminus (\tau_0, \tau_1)$) by

$$u(t; k) = B^*(t) G^*(\tau_0, t) (\mathcal{C}[\tau_0, \tau_1])^{-1} k.$$

Under (5.1)

$$u(t; k) \mapsto k(\tau^0, \tau^1) = k.$$

If we drive Σ and $\tilde{\Sigma}$ with $u(t; k)$ ($= \tilde{u}(t; k)$) and $v = 0$, $\tilde{v} = 0$, then we obtain $k(\tau_0, \tau_1)$ and $\tilde{k}(\tau_0, \tau_1)$. We have the commuting diagram (Fig. 5.1), which defines a linear mapping

$$T: k(\tau_0, \tau_1) \mapsto \tilde{k}(\tau_0, \tau_1).$$

The outputs $y(t)$ and $\tilde{y}(t)$ of Σ and $\tilde{\Sigma}$ corresponding to $u(t; k)$ must agree. Since the support of $u(t; k)$ is off (τ_0, τ_1) , the output $y(\tau)$ on $[\tau_0, \tau_1]$ is given by (5.2) as a function of $k(\tau_0, \tau_1)$. A similar expression holds for $\tilde{y}(\tau)$ for $\tau \in [\tau_0, \tau_1]$. Therefore we have another commuting diagram (Fig. 5.2). From this we see that the kernel of T is contained in the kernel of the upper right mapping of Fig. 5.2 which is given by (5.2). But this latter kernel is 0 since Σ is observable on $[\tau_0, \tau_1]$.

This shows that the state dimension of $\tilde{\Sigma}$ must be greater than or equal to that of Σ . Hence Σ is minimal. If the dimensions are equal, then T is a linear isomorphism. From this it follows that $\tilde{\Sigma}$ is controllable off and observable on $[\tau_0, \tau_1]$. Q.E.D.

We note for future reference that if Σ is controllable off $[\tau_0, \tau_1]$ but not observable on $[\tau_0, \tau_1]$, the map T is well defined but not necessarily 1-1. Its kernel is contained in the kernel of $\mathcal{O}[\tau_0, \tau_1]$. In particular, for all $t \in [\tau_0, \tau_1]$

$$(5.5) \quad \tilde{C}(t) \tilde{\Phi}(t, \tau_0) T = C(t) \Phi(t, \tau_0).$$

There is an analogous development based on the outward boundary value process $j(\tau_0, \tau_1)$.

DEFINITION. The system (1.1) is *controllable on* $[\tau_0, \tau_1]$ if the map

$$(5.6) \quad \{u(\tau): \tau \in [\tau_0, \tau_1]\} \mapsto j(\tau_0, \tau_1) = \int_{\tau_0}^{\tau_1} \Phi(\tau_1, s) B(s) \dot{u}(s) ds$$

as defined by the system (4.8) is onto.

DEFINITION. The system (1.2) is *observable off* $[\tau_0, \tau_1]$ if the map defined by (4.7a), (4.7b), (4.7d)

$$(5.7) \quad j(\tau_0, \tau_1) \mapsto \{y(t) = C(t)G(t, \tau_1)j(\tau_0, \tau_1) : t \in [t_0, t_1] \setminus (\tau_0, \tau_1)\}$$

is one to one. The control is restricted to be zero off (τ_0, τ_1) .

The associated Gramians are

$$(5.8a) \quad \mathcal{C}[\tau_0, \tau_1] = \int_{\tau_0}^{\tau_1} \Phi(\tau_1, s)B(s)B^*(s)\Phi^*(\tau_1, s) ds,$$

$$(5.8b) \quad \mathcal{O}_{\tau_0, \tau_1} = \left(\int_{t_0}^{\tau_0} + \int_{\tau_1}^{t_1} \right) G^*(t, \tau_1)C^*(t)C(t)G(t, \tau_1) dt.$$

PROPOSITION 5.3. *The system (1.1) is controllable on $[\tau_0, \tau_1]$ iff $\mathcal{C}[\tau_0, \tau_1]$ is positive definite. The system (1.1) is observable off $[\tau_0, \tau_1]$ iff $\mathcal{O}_{\tau_0, \tau_1}$ is positive definite.*

Remark. $\mathcal{C}[\tau_0, \tau_1]$ is the controllability Gramian of the causal system with the same A and B matrices. Therefore controllability on is just causal controllability.

THEOREM 5.4. *Let the system (1.1) be a realization of the weighing pattern $W(t, s)$. If there exists a τ_0 and τ_1 where $t_0 < \tau_0 < \tau_1 < t_1$ such that the system is controllable on $[\tau_0, \tau_1]$ and observable off $[\tau_0, \tau_1]$ then (1.1) is minimal. Moreover if $\tilde{\Sigma}$ is any other minimal realization of $W(t, s)$ then $\tilde{\Sigma}$ is also controllable on and observable off $[\tau_0, \tau_1]$.*

Proof. It is essentially the same as the proof of Theorem 5.1. Let Σ and $\tilde{\Sigma}$ be two realizations of $W(t, s)$. Since Σ is controllable on $[\tau_0, \tau_1]$, $\mathcal{C}[\tau_0, \tau_1]$ is invertible. Given $j \in \mathbb{R}^n$, define a control $u(t; j)$ with support in $[\tau_0, \tau_1]$ by

$$u(t; j) = B^*(t)\Phi^*(\tau_1, t)(\mathcal{C}[\tau_0, \tau_1])^{-1}j;$$

then under (5.6)

$$j \mapsto u(t; j) \rightarrow j(\tau_0, \tau_1) = j.$$

Let $\tilde{j}(\tau_0, \tau_1)$ be the value of the outward boundary value process of $\tilde{\Sigma}$ corresponding to $u(t; j)$. Then we have the commuting diagram (Fig. 5.3), which defines the linear mapping

$$S: j(\tau_0, \tau_1) \mapsto \tilde{j}(\tau_0, \tau_1).$$

The outputs $y(t)$ and $\tilde{y}(t)$ of Σ and $\tilde{\Sigma}$ must agree. Since the support of $u(\tau; j)$ is on $[\tau_0, \tau_1]$, the outputs off (τ_0, τ_1) are functions (5.7) of $j(\tau_0, \tau_1)$, $\tilde{j}(\tau_0, \tau_1)$. Therefore, we have a second commuting diagram (Fig. 5.4). The kernel of S is contained in the kernel of the upper right mapping of Fig. 5.4, given by (5.6). Since Σ is observable off $[\tau_0, \tau_1]$, the latter is zero. Hence S is one to one. The state dimension of $\tilde{\Sigma}$ must be greater than or equal to that of Σ . Hence Σ is minimal. If the dimensions are equal, then S is an isomorphism. From this it follows that $\tilde{\Sigma}$ is controllable on and observable off $[\tau_0, \tau_1]$. Q.E.D.

We note for future reference that if Σ is not controllable on $[\tau_0, \tau_1]$ then the map S can be defined on the range of the map (5.6) which is the range of $\mathcal{C}[\tau_0, \tau_1]$. If Σ is observable off $[\tau_0, \tau_1]$ then on this domain S is 1-1 by the above argument. In particular, for all $s \in [\tau_0, \tau_1]$

$$(5.9) \quad S\Phi(\tau_1, s)B(s) = \tilde{\Phi}(\tau_1, s)\tilde{B}(s).$$

In general the concepts of controllability on and controllability off are independent, i.e., we can construct an example that is one but not the other. A system is *real analytic*

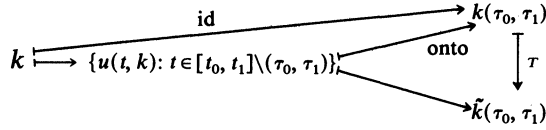


FIG. 5.1

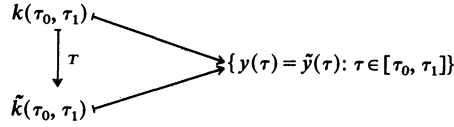


FIG. 5.2

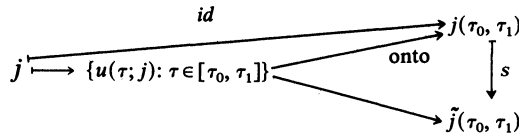


FIG. 5.3

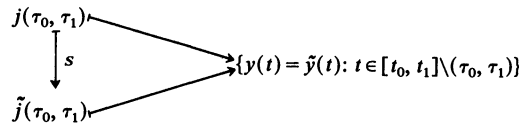


FIG. 5.4

if the matrices A , B , C and D are real analytic functions of t . Of course this includes *autonomous* systems where A , B , C , and D are constant. For real analytic systems, controllability (observability) on implies controllability (observability) off but they are not equivalent.

PROPOSITION 5.5. Suppose (1.1) is a real analytic system and $t_0 < \tau_0 < \tau_1 < t_1$.

(a) If (1.1) is controllable (observable) off some $[\tau_0, \tau_1]$ then it is controllable (observable) off every $[\tau_0, \tau_1]$.

(b) If (1.1) is controllable (observable) on some $[\tau_0, \tau_1]$ then it is controllable (observable) on every $[\tau_0, \tau_1]$.

(c) If (1.1) is controllable (observable) on some $[\tau_0, \tau_1]$ then it is controllable (observable) off every $[\tau_0, \tau_1]$.

Proof. (a) By definition, the Gramian $\mathcal{G}[\tau_0, \tau_1]$ is a real analytic, nonnegative definite matrix valued function of τ_0 and τ_1 . Also it is monotone nonincreasing in $[\tau_0, \tau_1]$, i.e., if $[\sigma_0, \sigma_1] \supseteq [\tau_0, \tau_1]$ then $\mathcal{G}[\sigma_0, \sigma_1] \leq \mathcal{G}[\tau_0, \tau_1]$. Suppose for some τ_0 and τ_1 , $t_0 < \tau_0 < \tau_1 < t_1$, the Gramian $\mathcal{G}[\tau_0, \tau_1]$ is not positive definite then it is also not positive definite for all σ_0 and σ_1 such that $[\sigma_0, \sigma_1] \supseteq [\tau_0, \tau_1]$. For some open set of σ_0 and σ_1 , the determinant of $\mathcal{G}[\sigma_0, \sigma_1]$ is zero. Real analyticity implies it is zero everywhere.

The other parts of (a) and (b) are proved in a similar fashion. (c) Suppose (1.1) is controllable on some $[\tau_0, \tau_1]$; then by (b) it is controllable on every $[\tau_0, \tau_1]$. Suppose it is not controllable off some $[\tau_0, \tau_1]$. Then we can find $0 \neq \lambda \in \mathbb{R}^{1 \times n}$ such that

$$\lambda(\mathcal{G}[\tau^0, \tau^1])\lambda^* = 0.$$

$\mathcal{C}[\tau^0, \tau^1]$ is the sum of two nonnegative definite matrices so λ must annihilate each:

$$\begin{aligned} 0 &= \lambda \int_{t_0}^{\tau_0} G(\tau_0, s) B(s) B^*(s) G^*(\tau_0, s) ds \lambda^* \\ &= \lambda \Phi(\tau_0, t_0) V^0 \Phi(t_0, t_1) \mathcal{C}[t_0, t_1] \Phi^*(t_0, t_1) V^{0*} \Phi^*(\tau_0, t_0) \lambda^*. \end{aligned}$$

Since $\mathcal{C}[t_0, \tau_0]$ is positive definite and $\Phi(t_0, t_1)$ is invertible we observe that

$$0 = \lambda \Phi(\tau_0, t_0) V^0.$$

From the other integral of $\mathcal{C}[\tau^0, \tau^1]$ we derive in a similar fashion that

$$0 = \lambda \Phi(\tau_0, t_0) V^1 \Phi(t_1, t_0).$$

But this implies $\lambda = 0$ for $V^0 + V^1 \Phi(t_1, t_0) = I$.

The other part of (c) is proved similarly. Q.E.D.

While Theorems 5.1 and 5.4 give sufficient conditions for minimality, these conditions are not necessary, as is shown by the following example, similar to one found in [10].

Example 5.6. Let $[t_0, t_1] = [0, 1]$. Consider the acausal system

$$\begin{aligned} \dot{x}_1 &= 0, & x_2(0) + x_1(1) - x_2(1) &= v_1, \\ \dot{x}_2 &= u, & x_2(0) &= v_2, \\ y &= x_1. \end{aligned}$$

It is a simple exercise to verify that this system is controllable and observable off every $[\tau_0, \tau_1]$ but it is not controllable nor observable on any $[\tau_0, \tau_1]$. The Gramians are

$$\begin{aligned} \mathcal{C}[\tau^0, \tau^1] &= \begin{vmatrix} 1 - \tau_1 + \tau_0 & \tau_0 \\ \tau_0 & \tau_0 \end{vmatrix}, & \mathcal{C}[\tau^0, \tau^1] &= \begin{vmatrix} 0 & 0 \\ 0 & \tau_1 - \tau_0 \end{vmatrix}, \\ \mathcal{O}[\tau^0, \tau^1] &= \begin{vmatrix} 1 - \tau_1 & \tau_1 - 1 \\ \tau_1 - 1 & 1 - \tau_1 + \tau_0 \end{vmatrix}, & \mathcal{O}[\tau^0, \tau^1] &= \begin{vmatrix} \tau_1 - \tau_0 & 0 \\ 0 & 0 \end{vmatrix}. \end{aligned}$$

By studying this system we get an understanding of what controllability and observability off $[\tau_0, \tau_1]$ really mean. The boundary conditions allow us to control x_1 in an indirect fashion. For example, to achieve a desired $x(\tau)$, $0 < \tau < 1$, we use the control on $[0, \tau]$ to fix $x_2(\tau)$. We use the control on $[\tau_1, 1]$ to fix $x_2(1)$. The first boundary condition and dynamics imply that $x_1(\tau) = x_2(1)$.

This indirect controllability through the boundary conditions is possible in acausal systems that are controllable off every $[\tau_0, \tau_1]$. Given any $\tau \in (t_0, t_1)$ and any $x^0 \in \mathbb{R}$ there exists a $u(\cdot)$ such that $x(\tau) = x^0$. The controllability off hypothesis implies that the map $u(\cdot) \rightarrow k(\tau, \tau)$ is onto, and it follows immediately from (4.1) and (4.2) that $x(\tau) = k(\tau, \tau)$.

The boundary conditions also allow us to detect jumps or breaks in both state coordinate trajectories even though we can only observe x_1 . For example, if for some unknown reason x_2 jumps at $\tau \in (0, 1)$ ($x_2(\tau^+) - x_2(\tau^-) = j_2$), then this affects $x_1(t)$ through the first boundary condition and we can detect it through $y(t)$.

This indirect observability through the boundary conditions is possible in any system that is observable off every $[\tau_0, \tau_1]$. Given any $\tau \in (t_0, t_1)$ and jump $j = x(\tau^+) - x(\tau^-)$, the mapping $j \rightarrow y(\cdot)$ is one to one. Therefore if we know the time τ of the jump, we can detect it.

The weighting pattern of this system is $W(t, s) = 1$ for all $t, s \in [t_0, t_1]$. From Theorem 3.1 it can be seen that this is a minimal realization of $W(t, s) = 1$. For if Σ is a one-dimensional realization then there exists 1×1 matrices $C(t)$, $B(s)$, V^0 and V^1 such that

$$V^0 + V^1 = 1$$

and

$$1 = C(t)V^0B(s), \quad t > s,$$

$$1 = -C(t)V^1B(s), \quad t < s.$$

The latter equations imply $C(t)$ and $B(s)$ are constant. If one is subtracted from the other we have

$$0 = C(V^0 + V^1)B = CB,$$

so either C or B is zero, a contradiction.

6. Minimal real analytic realizations. In the last section we gave sufficient conditions for minimality. In this one we give necessary and sufficient conditions for a real analytic realization to be minimal within the class of real analytic realizations. We describe how a real analytic realization can be reduced to a minimal real analytic realization, and how minimal real analytic realizations of the same weighting pattern can possibly differ. It may come as a bit of a surprise to readers familiar with causal systems theory that two minimal real analytic realizations can differ by more than a change of coordinates in the state space.

THEOREM 6.1. *Suppose Σ , given by (1.1), is a standard real analytic realization of $W(t, s)$. Σ is a minimal real analytic realization if and only if Σ is controllable and observable off every $[\tau_0, \tau_1]$ and*

$$(6.1) \quad \text{Kernel } \mathcal{O}[t_0, t_1] \subseteq \text{Range } \Phi(t_0, t_1)C[t_0, t_1]\Phi^*(t_0, t_1).$$

Any real analytic realization can be reduced to a minimal real analytic realization. If Σ and $\tilde{\Sigma}$ are two standard minimal real analytic realizations of the same $W(t, s)$, then there exists a real analytic invertible $n \times n$ matrix valued function $R(t)$ such that

$$(6.2a) \quad \tilde{A}(t) = R(t)A(t)R^{-1}(t) + \dot{R}(t)R^{-1}(t),$$

$$(6.2b) \quad \tilde{B}(t) = R(t)B(t),$$

$$(6.2c) \quad \tilde{C}(t) = C(t)R^{-1}(t),$$

$$(6.2d) \quad \tilde{D}(t) = D(t),$$

$$(6.2e) \quad R(t)\Phi(t, s) = \tilde{\Phi}(t, s)R(s),$$

and

$$(6.3a) \quad \mathcal{O}[t_0, t_1](V^0 - R^{-1}(t_0)\tilde{V}^0R(t_0))\Phi[t_0, t_1]\mathcal{G}[t_0, t_1] = 0,$$

$$(6.3b) \quad \mathcal{O}[t_0, t_1](V^1 - R^{-1}(t_0)\tilde{V}^1R(t_1))\mathcal{G}[t_0, t_1] = 0.$$

On the other hand, if Σ is a minimal real analytic realization of $W(t, s)$ and $\tilde{\Sigma}$ satisfies (6.2) and (6.3) for some real analytic invertible $R(t)$, then $\tilde{\Sigma}$ is also a minimal real analytic realization of $W(t, s)$.

Remarks. Condition (6.1) means that any state which is unobservable on $[t_0, t_1]$ must be controllable on $[t_0, t_1]$. Equations (6.2) are the same as those that arise from a time varying change of coordinates $\tilde{x} = R(t)x$, but this is not the whole story because of (6.3). If Σ is both controllable and observable on $[t_0, t_1]$ then (6.3) becomes

$$\tilde{V}^0 = R(t_0)V^0R^{-1}(t_0), \quad \tilde{V}^1 = R(t_0)V^1R^{-1}(t_1).$$

Then (6.2) and (6.3) represent a time varying change of state coordinates $\tilde{x} = R(t)x$ and a corresponding change of coordinates in the space of boundary inputs, $\tilde{v} = R(t_0)v$, so that $\tilde{\Sigma}$ is standard, $\tilde{V}^0 + \tilde{V}^1\tilde{\Phi}(t_1, t_0) = I$. Since $\tilde{\Sigma}$ is standard, (6.3a) and (6.3b) are equivalent.

Proof. Suppose Σ and $\tilde{\Sigma}$ are standard real and analytic realizations of $W(t, s)$. Clearly $D(t) = \tilde{D}(t)$. For simplicity henceforth we assume that $D(t) = 0$. Suppose for some τ_0 and τ_1 , the system Σ is controllable and observable off $[\tau_0, \tau_1]$ and (6.1) is satisfied.

It is convenient to make a time varying change of state coordinate $x_{\text{new}}(t) = \Phi(t_0, t)x_{\text{old}}(t)$ so that in these new coordinates $A(t) = 0$. We make a similar change of coordinates on $\tilde{\Sigma}$ so that $\tilde{A}(t) = 0$. Then $\Phi(t, s) = I$, $\tilde{\Phi}(t, s) = I$ and

$$(6.4a) \quad W(t, s) = C(t)V^0B(s) = \tilde{C}(t)\tilde{V}^0\tilde{B}(s) \quad \text{if } t > s,$$

$$(6.4b) \quad W(t, s) = C(t)(V^0 - I)B(s) = \tilde{C}(t)(\tilde{V}^0 - I)\tilde{B}(s) \quad \text{if } t < s.$$

By real analyticity we conclude that these formulas must hold for all t and s so

$$(6.5) \quad C(t)B(s) = \tilde{C}(t)\tilde{B}(s).$$

Relative to the controllability on and observability on Gramians $\mathcal{C}[t_0, t_1]$ and $\mathcal{O}[t_0, t_1]$, there is a nested family of subspaces of the state space

$$\begin{aligned} \mathbb{R}^n &\supseteq \text{Range}(\mathcal{C}[t_0, t_1]) + \text{Kernel}(\mathcal{O}[t_0, t_1]) \\ &\supseteq \text{Range}(\mathcal{C}[t_0, t_1]) \\ &\supseteq \text{Range}(\mathcal{C}[t_0, t_1]) \cap \text{Kernel}(\mathcal{O}[t_0, t_1]) \supseteq 0. \end{aligned}$$

We can choose coordinates $x = (x_1, x_2, x_3, x_4)$, which respect this flag, i.e.,

$$\text{Range } \mathcal{C}[t_0, t_1] \cap \text{Kernel } \mathcal{O}[t_0, t_1] = \{x: x_i = 0, i = 1, 2, 3\},$$

$$\text{Range } \mathcal{C}[t_0, t_1] = \{x: x_i = 0, i = 1, 2\},$$

$$\text{Range } \mathcal{C}[t_0, t_1] + \text{Kernel } \mathcal{O}[t_0, t_1] = \{x: x_1 = 0\}.$$

This is essentially the Kalman 4 part decomposition of the state space; see Kalman [16] or Desoer [11, p. 187] for more details.

The x_1 and x_3 coordinates are observable and the x_1 and x_2 coordinates are uncontrollable on $[t_0, t_1]$. We make the same change of coordinates in the space of boundary inputs v to ensure that $V^0 + V^1 = I$.

Relative to this partition of x we have that

$$(6.6a) \quad B^*(s) = \begin{pmatrix} 0 & 0 & B_3^*(s) & B_4^*(s) \end{pmatrix},$$

$$(6.6b) \quad C(t) = \begin{pmatrix} C_1(t) & 0 & C_3(t) & 0 \end{pmatrix}.$$

Let d_i be the dimension of x_i ; then $d_1 + d_2 + d_3 + d_4 = n$. Condition (6.1) ensures that $d_2 = 0$.

We make a similar decomposition of the state space of $\tilde{\Sigma}$, $\tilde{x} = (\tilde{x}_1, \tilde{x}_2, \tilde{x}_3, \tilde{x}_4)$ of dimensions $\tilde{d}_1 + \tilde{d}_2 + \tilde{d}_3 + \tilde{d}_4 = \tilde{n}$. Of course \tilde{d}_2 need not be zero.

Relative to this partition of \tilde{x} we have

$$(6.7a) \quad \tilde{B}^*(s) = \begin{pmatrix} 0 & 0 & \tilde{B}_3^*(s) & \tilde{B}_4^*(s) \end{pmatrix},$$

$$(6.7b) \quad \tilde{C}(t) = \begin{pmatrix} \tilde{C}_1(t) & 0 & \tilde{C}_3(t) & 0 \end{pmatrix}.$$

From (6.5), (6.6) and (6.7) we see that

$$(6.8) \quad C_3(t)B_3(s) = \tilde{C}_3(t)\tilde{B}_3(s).$$

We can consider (6.8) as the weighting pattern of a causal system which can be realized on either x_3 or \tilde{x}_3 space. Both these realizations are minimal because the x_3 and \tilde{x}_3 coordinates are both controllable and observable on $[t_0, t_1]$. Hence $d_3 = \tilde{d}_3$.

Now consider the map $T: k(\tau_0, \tau_1) \mapsto \tilde{k}(\tau_0, \tau_1)$ constructed in the proof of Theorem 5.1. Since Σ is controllable off $[\tau_0, \tau_1]$, T is well defined but it need not be one to one. Restricted to the range of $\mathcal{O}[\tau_0, \tau_1]$, it is one to one. The ranks of $\mathcal{O}[\tau_0, \tau_1]$ and $\tilde{\mathcal{O}}[\tau_0, \tau_1]$ are $d_1 + d_3$ and $\tilde{d}_1 + \tilde{d}_3$, respectively. The counting diagram (Fig. 5.2) implies that $d_1 + d_3 \leq \tilde{d}_1 + \tilde{d}_3$. Since $d_3 = \tilde{d}_3$ this implies that $d_1 \leq \tilde{d}_1$.

Next consider the map $S: j(\tau_0, \tau_1) \mapsto \tilde{j}(\tau_0, \tau_1)$ of Theorem 5.4. This is not defined for all $j(\tau_0, \tau_1)$ but only those in the range of $\mathcal{C}[\tau_0, \tau_1]$. On this domain it is one to one. The ranks of $\mathcal{C}[\tau_0, \tau_1]$ and $\tilde{\mathcal{C}}[\tau_0, \tau_1]$ are $d_3 + d_4$ and $\tilde{d}_3 + \tilde{d}_4$, respectively. The commuting diagram (Fig. 5.4) implies that $d_3 + d_4 \leq \tilde{d}_3 + \tilde{d}_4$ and hence $d_4 \leq \tilde{d}_4$.

We have shown that $d_i \leq \tilde{d}_i$ for $i = 1, 2, 3, 4$ so $n \leq \tilde{n}$ and hence Σ is minimal.

Suppose $\Sigma(t, s)$ is any standard real analytic realization of $W(t, s)$; we now show that Σ can be reduced to obtain a lower-dimensional realization of $W(t, s)$ which is controllable and observable off every $[\tau_0, \tau_1]$ and such that (6.1) is satisfied. In this manner we see that if a realization is minimal it is controllable and observable off every $[\tau_0, \tau_1]$, and every state unobservable on $[\tau_0, \tau_1]$ is controllable on $[\tau_0, \tau_1]$.

We assume that we have made the preliminary change of state coordinates so that $A(t) = 0$. Suppose Σ is controllable and observable off every $[\tau_0, \tau_1]$. We decompose the state space relative to the Gramians $\mathcal{C}[t_0, t_1]$ and $\mathcal{O}[t_0, t_1]$ as above.

The x_2 coordinate is irrelevant to the weighting pattern of the system. From (6.6) we have that

$$W(t, s) = \begin{cases} \sum_{i=1,3} \sum_{j=3,4} C_i(t) V_{ij}^0 B_j(s) & \text{if } t > s, \\ -\sum_{i=1,3} \sum_{j=3,4} C_i(t) V_{ij}^1 B_j(s) & \text{if } t < s. \end{cases}$$

Therefore we can delete x_2 and the second boundary condition. The new system satisfies (6.1) so we obtain a new realization of $W(t, s)$, of lower dimension, which is minimal among real analytic realizations.

Suppose Σ is not controllable or not observable off every $[\tau_0, \tau_1]$; then we can reduce it to one that is. We decompose the state space into 4 parts relative to Gramians $\mathcal{C}[\tau_0, \tau_0]$ and $\mathcal{O}[\tau_0, \tau_0]$ for any $t_0 < \tau_0 < \tau_1 < t_1$. We obtain a flag of subspaces

$$\begin{aligned} \mathbb{R}^n &\supseteq \text{Range}(\mathcal{C})_{\tau_0, \tau_1}] + \text{Kernel}(\mathcal{O})_{\tau_0, \tau_1}] \\ &\supseteq \text{Range}(\mathcal{C})_{\tau_0, \tau_1}] \\ &\supseteq \text{Range}(\mathcal{C})_{\tau_0, \tau_1}] \cap \text{Kernel}(\mathcal{O})_{\tau_0, \tau_1}] \supseteq 0. \end{aligned}$$

As before we choose coordinates $x = (x_1, x_2, x_3, x_4)$ which respect this flag and we make the same change of coordinates in the space of boundary values v to keep the system in standard form.

From the choice of coordinates

$$G(t, s)B(s) = V^i B(s) = \begin{bmatrix} 0 \\ 0 \\ * \\ * \end{bmatrix} \quad \text{where } i = 0 \text{ if } t > s \text{ and } i = 1 \text{ if } t < s,$$

and

$$C(t)G(t, s) = C(t)V^i = [* \ 0 \ * \ 0] \quad \text{where } i = 0 \text{ if } t > s \text{ and } i = 1 \text{ if } t < s.$$

Since $V^0 + V^1 = I$ we obtain

$$(6.9a) \quad B(s) = \begin{bmatrix} B_1(s) \\ B_2(s) \\ B_3(s) \\ B_4(s) \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ * \\ * \end{bmatrix},$$

$$(6.9b) \quad C(t) = [C_1(t)C_2(t)C_3(t)C_4(t)] = [* \ 0 \ * \ 0],$$

$$(6.9c) \quad V^i B(s) = \begin{bmatrix} V_{13}^i B_3(s) + V_{14}^i B_4(s) \\ V_{23}^i B_3(s) + V_{24}^i B_4(s) \\ V_{33}^i B_3(s) + V_{34}^i B_4(s) \\ V_{44}^i B_3(s) + V_{44}^i B_4(s) \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ * \\ * \end{bmatrix},$$

$$(6.9d) \quad \begin{aligned} C(t)V^i &= [C_1(t)V_{11}^i + C_3(t)V_{31}^i; C_1(t)V_{12}^i + C_3(t)V_{32}^i; \\ &\quad C_1(t)V_{13}^i + C_3(t)V_{33}^i; C_1(t)V_{14}^i + C_3(t)V_{34}^i] \\ &= [* \ 0 \ * \ 0]. \end{aligned}$$

We can calculate $W(t, s)$ from (6.9a), (6.9d) as

$$(6.10a) \quad W(t, s) = \begin{cases} C_1(t)V_{13}^0 B_3(s) + C_3(t)V_{33}^0 B_3(s), & t > s, \\ -C_1(t)V_{13}^1 B_3(s) - C_3(t)V_{33}^1 B_3(s), & t < s \end{cases}$$

or from (6.9b), (6.9c) as

$$(6.10b) \quad W(t, s) = \begin{cases} C_3(t)V_{33}^0 B_3(s) + C_3(t)V_{34}^0 B_4(s), & t > s, \\ -C_3(t)V_{33}^1 B_3(s) - C_3(t)V_{34}^1 B_4(s), & t < s. \end{cases}$$

Equation (6.10a) shows that the x_2 and x_4 coordinates are unnecessary to realize $W(t, s)$, so we delete them and the corresponding boundary conditions to obtain a new system which is observable off every $[\tau_0, \tau_1]$. From the first component of (6.9c) we see that $V_{13}^i B_3(s) = 0$ so we can delete the x_1 coordinate and realize $W(t, s)$ by

$$(6.11a) \quad \dot{x}_3 = B_3 u,$$

$$(6.11b) \quad V_{33}^0 x_3(t_0) + V_{33}^1 x_3(t_1) = v_3,$$

$$(6.11c) \quad y = C_3 x_3.$$

This realization is controllable and observable off every $[\tau_0, \tau_1]$. As previously seen, such realizations can be further reduced so that (6.1) holds, thereby obtaining a minimal real analytic realization.

Next we study the relationship between minimal real analytic realizations. Let Σ and $\tilde{\Sigma}$ be two standard minimal real analytic realizations of $W(t, s)$. Hence they are controllable and observable off every $[\tau_0, \tau_1]$ and (6.1) is satisfied. We transform

state coordinates on Σ and $\tilde{\Sigma}$ as in the first part of the proof so that $A(t) = \tilde{A}(t) = 0$ and $x = (x_1, x_3, x_4)$, $\tilde{x} = (\tilde{x}_1, \tilde{x}_3, \tilde{x}_4)$. Because they are minimal $d_i = \tilde{d}_i$, $i = 1, 3, 4$ and $d_2 = \tilde{d}_2 = 0$.

The decomposition of the state space induces a similar decomposition of the boundary value processes, $k = (k_1, k_3, k_4)$, $\tilde{k} = (\tilde{k}_1, \tilde{k}_3, \tilde{k}_4)$, $j = (j_1, j_3, j_4)$ and $\tilde{j} = (\tilde{j}_1, \tilde{j}_3, \tilde{j}_4)$. We consider the maps T and S in more detail. If $k \in \text{kernel } \mathcal{O}[\tau_0, \tau_1]$ then $T(k) = \tilde{k} \in \text{kernel } \tilde{\mathcal{O}}[\tau_0, \tau_1]$, so $\tilde{k} = Tk$ is given by

$$(6.12) \quad \begin{pmatrix} \tilde{k}_1 \\ \tilde{k}_3 \\ \tilde{k}_4 \end{pmatrix} = \begin{pmatrix} T_{11} & T_{13} & 0 \\ T_{31} & T_{33} & 0 \\ T_{41} & T_{43} & T_{44} \end{pmatrix} \begin{pmatrix} k_1 \\ k_3 \\ k_4 \end{pmatrix}.$$

By the commuting diagram (Fig. 5.2) the upper left 2×2 block of T is invertible.

On the other hand S is only defined on the range of $C[\tau_0, \tau_1]$ (i.e. $j = (0, j_3, j_4)$), but it is one to one. So $\tilde{j} = S(0, j_3, j_4)$ is given by

$$(6.13) \quad \begin{pmatrix} \tilde{j}_1 \\ \tilde{j}_3 \\ \tilde{j}_4 \end{pmatrix} = \begin{pmatrix} 0 & S_{13} & S_{14} \\ 0 & S_{33} & S_{34} \\ 0 & S_{43} & S_{44} \end{pmatrix} \begin{pmatrix} 0 \\ j_3 \\ j_4 \end{pmatrix}$$

where the matrix is of rank $d_3 + d_4$.

From (5.5) and (5.9) we have

$$(6.14a) \quad C(t) = \tilde{C}(t)T,$$

$$(6.14b) \quad \tilde{B}(s) = SB(s),$$

so (6.5) becomes

$$(6.15) \quad \tilde{C}(t)TB(s) = \tilde{C}(t)SB(s).$$

If we multiply by $\tilde{C}^*(t)$ and $B^*(s)$ and integrate we obtain

$$(6.16) \quad \tilde{\mathcal{O}}[t_0, t_1](T - S)\mathcal{C}[t_0, t_1] = 0.$$

This implies that $T_{ij} = S_{ij}$ for $i = 1, 3$ and $j = 3, 4$. We define an $n \times n$ matrix R by

$$(6.17) \quad R = \begin{bmatrix} T_{11} & T_{13} & 0 \\ T_{31} & T_{33} & 0 \\ T_{41} & S_{43} & S_{44} \end{bmatrix}.$$

Since $S_{14} = T_{14} = 0$ and $S_{34} = T_{34} = 0$, S_{44} must be invertible for S to be of rank $d_3 + d_4$. The upper left 2×2 block of R is from T , hence is invertible. This shows that R is invertible. Since R agrees with T in rows indexed by 1 and 3, from (6.7b) and (6.14a) we see that

$$(6.18a) \quad C(t) = \tilde{C}(t)R.$$

Since R agrees with S in columns indexed by 3 and 4, from (6.7a) and (6.14b) we see that

$$(6.18b) \quad \tilde{B}(t) = RB(t).$$

From (6.4a) and (6.18), we obtain

$$C(t)V^0B(s) = C(t)R^{-1}\tilde{V}^0RB(s).$$

We multiply on both sides and integrate to obtain

$$(6.19) \quad \mathcal{O}[t_0, t_1](V^0 - R^{-1}\tilde{V}^0R)C[t_0, t_1] = 0.$$

Let $\Phi(t, s)$ and $\tilde{\Phi}(t, s)$ denote the fundamental solutions for Σ and $\tilde{\Sigma}$ in their original state coordinates. Define

$$(6.20) \quad R(t) = \tilde{\Phi}(t, t_0) R \Phi(t_0, t);$$

then it is straightforward but tedious to verify (6.2) and (6.3) from (6.18), (6.19) and (6.20).

On the other hand suppose Σ and $\tilde{\Sigma}$ are two systems related by (6.2) and (6.3). Then if $t > s$ the weighting pattern of $\tilde{\Sigma}$ is given by

$$\begin{aligned} \tilde{W}(t, s) &= \tilde{C}(t) \tilde{\Phi}(t, t_0) \tilde{V}^0 \tilde{\Phi}(t_0, s) \tilde{B}(s) \\ &= C(t) \Phi(t, t_0) V^0 \Phi(t_0, s) B(s) + C(t) \Phi(t, t_0) (V^0 - R^{-1}(t_0) \tilde{V}^0 R(t_0)) \Phi(t_0, s) B(s). \end{aligned}$$

But (6.2a) implies that this second term is zero, so $\tilde{W}(t, s) = W(t, s)$ for $t > s$. A similar calculation holds for $t < s$. Q.E.D.

Suppose Σ is a minimal realization of $W(t, s)$ and we choose state coordinates as before so that $A(t) = 0$ and $x = (x_1, x_3, x_4)$ respects the flag of subspaces associated to $\mathcal{C}[t_0, t_1]$ and $\mathcal{O}[t_0, t_1]$. Since $V^0 + V^1 = I$

$$(6.21a) \quad W(t, s) = \sum_{i=1,3} \sum_{j=3,4} C_i(t) V_{ij}^0 B_j(s) \quad \text{if } t > s,$$

$$(6.21b) \quad W(t, s) = \sum_{i=1,3} \sum_{j=3,4} C_i(t) (V^0 - I)_{ij} B_j(s) \quad \text{if } t < s.$$

Then

$$(6.22a) \quad W(t, s) = W_1(t, s) + W_2(t, s),$$

$$(6.22b) \quad W_1(t, s) = \begin{cases} C_3(t) V_{33}^0 B_3(s), & t > s, \\ C_3(t) (V_{33}^0 - I) B_3(s), & t < s \end{cases}$$

and for all t, s

$$(6.23) \quad W_2(t, s) = C_1(t) V_{13}^0 B_3(s) + C_1(t) V_{14}^0 B_4(s) + C_3(t) V_{34}^0 B_4(s).$$

Each kernel $W_i(t, s)$ defines a mapping

$$u(t) \mapsto y_i(t) = \int_{t_0}^{t_1} W_i(t, s) u(s) ds.$$

The first kernel $W_1(t, s)$ defines a mapping of infinite rank which can be realized on x_3 space. The second kernel $W_2(t, s)$ defines a mapping of finite rank

$$y_2(t) = C_1(t) V_{13}^0 \int_{t_0}^{t_1} B_3(s) u(s) ds + (C_1(t) V_{14}^0 + C_3(t) V_{34}^0) \int_{t_0}^{t_1} B_4(s) u(s) ds.$$

There is an alternate decomposition of $W(t, s)$:

$$(6.24a) \quad W(t, s) = \tilde{W}_1(t, s) + \tilde{W}_2(t, s)$$

where

$$(6.24b) \quad \tilde{W}_1(t, s) = \begin{cases} C_3(t) B_3(s), & t > s, \\ 0, & t < s, \end{cases}$$

$$(6.24c)$$

$$\tilde{W}_2(t, s) = C_1(t) V_{13}^0 B_3(s) + C_1(t) V_{14}^0 B_4(s) + C_3(t) (V_{33}^0 - I) B_3(s) + C_3(t) V_{34}^0 B_4(s).$$

$\tilde{W}_1(t, s)$ is a causal weighing pattern which can be realized on \tilde{x}_3 space. As before, \tilde{W}_1 and \tilde{W}_2 are maps of infinite and finite rank, respectively.

In their excellent study of acausal systems, Gohberg and Kaashoek [9], [10] have introduced the concepts of multicontrollability and multiobservability, which are different generalizations of causal controllability and observability from those discussed above. The following example is a slight modification of one found in [10, § II.1 and illustrates some of the differences between their work and ours.

Example 6.2. Let $[t_0, t_1] = [0, 1]$.

$$\dot{x}_1 = 0, \quad x_2(0) + x_1(1) - x_2(1) = v_1,$$

$$\dot{x}_2 = 0, \quad x_2(0) + x_3(0) - x_3(1) = v_2,$$

$$\dot{x}_3 = u, \quad x_3(0) = v_3,$$

$$y = x_1.$$

Gohberg and Kaashoek associate with an acausal system a sequence of weighting patterns of which $W(t, s)$ given by (3.6) is the first. Relevant to these weighting patterns are the concepts of multicontrollability and multiobservability. This example is a 3 controllable and 3 observable system. They show that such systems are irreducible under similarity and reduction and are characterized up to similarity by their sequence of weighting patterns. They do not discuss minimality.

By our definition this system is not controllable nor observable off any $[\tau_0, \tau_1]$ and (6.1) is not satisfied. Therefore by Theorem 6.1 it is not minimal among real analytic realizations. If we reduce it as described above we arrive at the trivial system of state dimension 0. The weighting pattern (3.6) is $W(t, s) = 0$, but the other weighting patterns of Gohberg and Kaashoek are not all zero.

7. Autonomous and stationary systems. An acausal linear system is *autonomous* if A, B, C, D are constant with respect to t . The transition matrix and Green's matrix are given by

$$(7.1) \quad \Phi(t, s) = e^{(t-s)A},$$

$$(7.2a) \quad G(t, s) = e^{(t-t_0)A} V^0 e^{(t_0-s)A} \quad \text{if } t > s,$$

$$(7.2b) \quad G(t, s) = -e^{(t-t_0)A} V^1 e^{(t_1-s)A} \quad \text{if } t < s.$$

From this it is easy to give alternate tests for controllability and observability.

PROPOSITION 7.1. Let Σ be an autonomous acausal system and $t_0 < \tau_0 < \tau_1 < t_1$:

(a) Σ is controllable on $[\tau_0, \tau_1]$ iff the matrix

$$(7.3a) \quad \mathcal{C} = [B, \dots, A^{n-1}B]$$

is of rank n .

(b) Σ is controllable off $[\tau_0, \tau_1]$ iff the matrix

$$(7.3b) \quad \mathcal{C}^b = [V^0 B, \quad V^1 B, \quad \dots, \quad V^0 A^{n-1} B, \quad V^1 A^{n-1} B]$$

is of rank n .

(c) Σ is observable on $[\tau_0, \tau_1]$ iff the matrix

$$(7.3c) \quad \mathcal{O} = \begin{bmatrix} C \\ \vdots \\ CA^{n-1} \end{bmatrix}$$

is of rank n .

(d) Σ is observable off $[\tau_0, \tau_1]$ iff the matrix

$$(7.3d) \quad \mathcal{O}^b = \begin{bmatrix} CV^0 \\ CV^1 \\ \vdots \\ CA^{n-1}V^0 \\ CA^{n-1}V^1 \end{bmatrix}$$

is of rank n .

(e) Condition (6.1) is satisfied iff

$$(7.3e) \quad \text{Kernel } \mathcal{O} \text{ Range } \mathcal{C}.$$

Proof. Assertion (a) and (c) are well known from the causal theory, and (e) is straightforward. We only prove (b) since the proof of (d) is essentially the same. Suppose (7.3b) fails; then there exists an $1 \times n$ vector $\lambda \neq 0$ such that

$$(7.4) \quad \lambda V^0 A^k B = \lambda V^1 A^k B = 0$$

for any $k \geq 0$ (by Cayley-Hamilton). Hence for any t

$$(7.5) \quad \lambda V^0 e^{At} B = \lambda V^1 e^{At} B = 0.$$

Now for any $t_0 < \tau_0 < \tau_1 < t_1$,

$$(7.6) \quad \begin{aligned} 0 &= \lambda \Phi(t_0, \tau_0)(\mathcal{C})\tau_0, \tau_1[\Phi^*(t_0, \tau_0)\lambda^* \\ &= \int_{t_0}^{\tau_0} \lambda V^0 e^{A(t_0-s)} BB^* e^{A^*(t_0-s)} V^{0*} \lambda^* ds \\ &\quad + \int_{\tau_1}^{t_1} \lambda V^1 e^{A(t_1-s)} BB^* e^{A^*(t_1-s)} V^{1*} \lambda^* ds. \end{aligned}$$

Therefore $\mathcal{C}[\tau_0, \tau_1]$ is not positive definite and the system is not controllable off $[\tau_0, \tau_1]$.

On the other hand, if there exists λ such that (7.6) holds for all $t_0 < \tau_0 < \tau_1 < t_1$, then (7.5) must hold. Differentiating (7.5) yields (7.4). Q.E.D.

Gohberg and Kaashoek [13] have demonstrated that a weighting pattern can have a minimal autonomous realization and a minimal real analytic realization of lower dimension. They showed that the weighting pattern $W(t, s) = 1 - s$ has this property. Their realizations are as follows.

Example 7.2. Let $[t_0, t_1] = [0, 1]$.

$$\begin{aligned} \dot{x}_1 &= 0, & x_2(0) + x_3(0) + x_1(1) - x_2(1) &= v_1, \\ \dot{x}_2 &= x_3, & x_2(0) &= v_2, \\ \dot{x}_3 &= u, & x_3(0) &= v_3, \\ y &= x_1. \end{aligned}$$

The system is controllable off any $[\tau_0, \tau_1]$ but not observable off any $[\tau_0, \tau_1]$, where $0 < \tau_0 < \tau_1 < 1$. Therefore by Theorem 6.1 it is not a minimal realization within the class of real analytic realizations. In fact $W(t, s) = 1 - s$ can also be realized by the following.

Example 7.3. Let $[t_0, t_1] = [0, 1]$.

$$\begin{aligned} \dot{x}_1 &= (1-t)u, & x_1(0) &= v_1, \\ \dot{x}_2 &= (1-t)u, & x_2(1) &= v_2, \\ y &= x_1 - x_2. \end{aligned}$$

This is controllable and observable off every $[\tau_0, \tau_1]$ and (6.1) is satisfied. Hence by Theorem 6.1 this is a minimal real analytic realization of $W(t, s) = 1 - s$. Moreover the other minimal real analytic realizations of $W(t, s)$ are described by this theorem. If one of them were autonomous there would exist an invertible 2×2 matrix $R(t)$ such that for some constant matrices $\tilde{A}, \tilde{B}, \tilde{C}$

$$(7.7a) \quad \tilde{A} = \dot{R}(t)R^{-1}(t),$$

$$(7.7b) \quad \tilde{B} = R(t) \begin{pmatrix} 1-t \\ 1-t \end{pmatrix},$$

$$(7.7c) \quad \tilde{C} = (1 - 1)R^{-1}(t).$$

Without loss of generality we can assume that $R(0) = I$ so $\tilde{B} = (1 \ 1)^*$ and $\tilde{C} = (1 \ -1)$. Equation (7.7b) implies that \tilde{B} is an eigenvector of $R(t)$ with eigenvalue $1/(1-t)$. On the other hand, equation (7.7a) implies that $R(t) = e^{\tilde{A}t}$ and hence $1/(1-t)$ cannot be an eigenvalue of $R(t)$. We conclude that there are no two-dimensional autonomous realizations of $W(t, s) = 1 - s$ but there are two-dimensional real analytic realizations.

We were a bit surprised by these examples for we had conjectured the opposite, namely that the class of autonomous models admitted a self-contained minimal realization theory. Being autonomous is a property of the system and not of the weighting pattern. Therefore we should have expected that we need a "nice" class of weighting patterns to obtain a self-contained minimal realization theory. The stationary weighting patterns are such a class.

A weighting pattern $W(t, s)$ is *stationary* if it is only a function of $t - s$, in abuse of notation $W(t, s) = W(t - s)$. An acausal linear system is *stationary* if it is autonomous and its weighting pattern is stationary. Every autonomous causal system is stationary, and hence the stationary acausal systems generalize the autonomous causal systems. The corollary to the following theorem was first stated in [3]; see also [9], [10].

PROPOSITION 7.4. *A standard autonomous acausal linear system is stationary iff for all $k, l = 0, \dots, n-1$.*

$$(7.8a) \quad CA^k[A, V^0]A^lB = 0,$$

$$(7.8b) \quad CA^k[A, V^1]A^lB = 0$$

where $[A, V^i] = AV^i - V^iA, i = 1, 2$.

Proof. Suppose the system is stationary; then

$$(7.9) \quad Ce^{A(t-t_0)}V^0e^{A(t_0-s)}B = Ce^{A(t+r-t_0)}V^0e^{A(t_0-r-s)}B.$$

If we differentiate this with respect to r at $r = 0$ we obtain

$$(7.10) \quad 0 = Ce^{A(t-t_0)}[A, V^0]e^{A(t_0-s)}B.$$

Differentiation of this with respect to t and s one or more times yields (7.8). The steps are reversible, so the converse holds. Q.E.D.

COROLLARY 7.5. *Suppose Σ is a standard autonomous acausal system which is controllable and observable on $[t_0, t_1]$. Σ is stationary iff*

$$(7.11a) \quad [A, V^0] = 0,$$

$$(7.11b) \quad [A, V^1] = 0.$$

The Gohberg-Kaashoek phenomenon cannot happen for stationary weighting patterns; in other words, a minimal autonomous realization is also a minimal real analytic realization.

THEOREM 7.6. (i) Suppose Σ (1.1) is a standard stationary realization of a stationary weighting pattern $W(t-s)$. Σ is a minimal stationary (equivalently, autonomous) realization iff

$$(7.12a) \quad \text{rank } \mathcal{C}^b = n,$$

$$(7.12b) \quad \text{rank } \mathcal{O}^b = n,$$

$$(7.12c) \quad \text{Kernel } \mathcal{O} \subseteq \text{Range } \mathcal{C}.$$

(ii) A minimal stationary realization is also a minimal real analytic realization.

(iii) Any stationary realization can be modified and reduced to a minimal stationary realization.

(iv) Suppose Σ and $\tilde{\Sigma}$ are stationary minimal realizations of $W(t-s)$ then there exists an invertible constant matrix R such that

$$(7.13a) \quad (A - R^{-1}\tilde{A}R)\mathcal{C} = 0,$$

$$(7.13b) \quad \mathcal{O}(A - R^{-1}\tilde{A}R) = 0,$$

$$(7.13c) \quad \tilde{B} = RB,$$

$$(7.13d) \quad \tilde{C} = CR^{-1},$$

$$(7.13e) \quad \tilde{D} = D$$

and

$$(7.14a) \quad \mathcal{O}(V^0 - R^{-1}\tilde{V}^0R)\mathcal{C} = 0,$$

$$(7.14b) \quad \mathcal{O}(V^1 - R^{-1}\tilde{V}^1R)\mathcal{C} = 0.$$

On the other hand, if Σ is a minimal stationary realization of $W(t-s)$ and $\tilde{\Sigma}$ is an autonomous system related to Σ by (7.13) and (7.14) for some invertible R then $\tilde{\Sigma}$ is also a minimal stationary realization of $W(t-s)$.

Proof. (i) Suppose Σ is an autonomous realization of a stationary weighting pattern which satisfies (7.12); then by Proposition 7.1, Σ is controllable and observable off every $[\tau_0, \tau_1]$ and (6.1) is satisfied. By Theorem 6.1, Σ is a minimal real analytic realization and hence a minimal stationary realization.

Suppose Σ is an autonomous realization of a stationary weighting pattern $W(t-s)$ which does not satisfy (7.12). To show that Σ is not minimal we shall construct a new autonomous realization $\tilde{\Sigma}$ of smaller state dimension which does satisfy (7.12) and hence is minimal. As the reader has seen, the way one obtains a lower-dimensional realization of a weighting pattern is to find an appropriate subspace of the state space which is left invariant by the dynamics. One either restricts to this subspace or quotients by this subspace to reduce the dimension of the state space. In the context of real-analytic systems we have the luxury of making a time varying change of coordinates so that the invariant subspaces are time invariant. In the context of autonomous systems we do not have this option.

The natural subspaces associated with an autonomous system (1.1) are formed from the matrices found in Proposition 7.1, e.g.,

$$(7.15a) \quad \text{Range } \mathcal{C},$$

$$(7.15b) \quad \text{Kernel } \mathcal{O},$$

$$(7.15c) \quad \text{Range } \mathcal{C}^b,$$

$$(7.15d) \quad \text{Kernel } \mathcal{O}^b.$$

The first and second are clearly invariant by definition.

$$(7.16a) \quad A(\text{Range } \mathcal{C}) \subseteq \text{Range } \mathcal{C},$$

$$(7.16b) \quad A(\text{Kernel } \mathcal{O}) \subseteq \text{Kernel } \mathcal{O}.$$

The third and fourth are generally not. It is this latter fact which seems to cause the Gohberg–Kaashoek phenomenon.

Even if the system (1.1) is stationary it is not always true that (7.15b), (7.15d) are A invariant. However, they are nearly so, for (7.8) implies that

$$(7.17a) \quad A(\text{Range } \mathcal{C}^b) \subseteq \text{Range } \mathcal{C}^b + \text{Kernel } \mathcal{O},$$

$$(7.17b) \quad A(\text{Kernel } \mathcal{O}^b \cap \text{Range } \mathcal{C}) \subseteq \text{Kernel } \mathcal{O}^b.$$

The reader with a background in geometric linear control theory recognizes (7.17a) as a form of (A, B) invariance (or controlled invariance) and (7.17b) as a form of (C, A) (or conditioned invariance). For details see [14] and [15].

Let D be a matrix such that

$$(7.18) \quad \text{Range } D = \text{Kernel } \mathcal{O};$$

then by a standard lemma there exists a matrix F such that

$$(7.19) \quad (A + DF) \text{Range } \mathcal{C}^b \subseteq \text{Range } \mathcal{C}^b.$$

Moreover, since (7.16a) holds we can choose F so that

$$(7.20) \quad \text{Kernel } F \supset \text{Range } \mathcal{C}.$$

Let $\tilde{A} = A + DF$; then (7.18) and (7.20) imply that

$$(7.21a) \quad C\tilde{A}^k = CA^k,$$

$$(7.21b) \quad \tilde{A}^k B = A^k B$$

for all k , so for all t, s

$$(7.22a) \quad C e^{\tilde{A}t} = C e^{At},$$

$$(7.22b) \quad e^{\tilde{A}s} B = e^{As} B.$$

Therefore we can modify Σ by replacing A by \tilde{A} and not change the weighting pattern $W(t-s)$.

In this way we obtain another autonomous realization of $W(t-s)$ such that (7.15c) is \tilde{A} invariant. By restricting this system to the subspace of the state space given by (7.15c) we obtain a smaller autonomous realization of $W(t-s)$ which satisfies (7.12a).

The property described by equation (7.17b) is called (C, A) invariance (or conditioned invariance). It is the dual of the property described by (7.17a). If we choose a matrix E such that

$$(7.23) \quad \text{Range } \mathcal{C} = \text{Kernel } E,$$

then it is a standard exercise to show that there exists a matrix G such that

$$(7.24) \quad (A + GE) \text{Kernel } \mathcal{O}^b \subseteq \text{Kernel } \mathcal{O}^b.$$

Moreover, because of (7.16b) we can choose G such that

$$(7.25) \quad \text{Range } G \subset \text{Kernel } \mathcal{O}.$$

If we define $\tilde{A} = A + GE$ then because of (7.23) and (7.25), (7.21) and (7.22) hold. We can replace A by \tilde{A} without changing the weighting pattern. For this new realization (7.15d) is \tilde{A} invariant and we can project it out. The resulting realization satisfies (7.12b).

In this way we obtain a realization satisfying (7.12a), (7.12b); in other words, one that is controllable and observable off every $[\tau_0, \tau_1]$. To reduce this to a realization satisfying (7.12c) we choose coordinates that respect the flag of subspaces

$$\begin{aligned}\mathbb{R}^n &\supseteq \text{Range } \mathcal{C} + \text{Kernel } \mathcal{O} \\ &\supseteq \text{Range } \mathcal{C} \\ &\supseteq \text{Range } \mathcal{C} \cap \text{Kernel } \mathcal{O} \supseteq 0.\end{aligned}$$

In other words, $x = (x_1, x_2, x_3, x_4)^*$ and

$$\begin{aligned}\text{Range } \mathcal{C} + \text{Kernel } \mathcal{O} &= \{x: x_1 = 0\}, \\ \text{Range } \mathcal{C} &= \{x: x_1 = 0, x_2 = 0\}, \\ \text{Range } \mathcal{C} \cap \text{Kernel } \mathcal{O} &= \{x: x_1 = 0, x_2 = 0, x_3 = 0\}.\end{aligned}$$

If we define $B(s) = e^{As}B$ and $C(t) = C e^{At}$ then in these coordinates

$$(7.26a) \quad e^{As}B = B(s) = \begin{bmatrix} 0 \\ 0 \\ B_3(s) \\ B_4(s) \end{bmatrix},$$

$$(7.26b) \quad C e^{At} = C(t) = [C_1(t) \ 0 \ C_3(t) \ 0],$$

and

$$(7.26c) \quad W(t-s) = \begin{cases} \sum_{i=1,3} \sum_{j=3,4} C_i(t-t_0) V_{ij}^0 B_j(t_0-s), & t > s, \\ -\sum_{i=1,3} \sum_{j=3,4} C_i(t-t_0) V_{ij}^1 B_j(t_1-s), & t < s. \end{cases}$$

Hence we can delete the x_2 coordinate and the corresponding boundary input condition without changing $W(t-s)$. This completes the proof of statement (i).

(ii) By (i) a minimal stationary realization must satisfy (7.12). Hence by Theorem 6.1 and Proposition 7.1 it must also be a minimal real analytic realization.

(iii) In the proof of (i) we showed how a stationary realization can be modified and reduced to a realization satisfying (7.12). By Theorem 6.1 such a system is a minimal real analytic realization, hence a minimal stationary realization.

(iv) If Σ is a minimal stationary realization of $W(t-s)$ and $\tilde{\Sigma}$ is an autonomous realization related to Σ by (7.13) and (7.14), then it is easy to verify that $\tilde{\Sigma}$ realizes $W(t-s)$, hence is a minimal stationary realization.

If Σ and $\tilde{\Sigma}$ are two minimal stationary realizations of $W(t-s)$ then by Theorem 6.1 there exists a real analytic matrix valued function $R(t)$ satisfying (6.2) and (6.3). In particular (6.2b) implies that

$$\tilde{B} = R(t)B,$$

so $R(t)$ is constant on $\text{Range } B$. If we differentiate this expression using (6.2a) we obtain

$$\tilde{A}\tilde{B} = R(t)AB.$$

Further differentiations yield

$$(7.27a) \quad \tilde{C} = R(t)\mathcal{C},$$

$$(7.27b) \quad 0 = \dot{R}(t)\mathcal{C}.$$

In a similar fashion repeated differentiations of (6.2c) yield

$$(7.28a) \quad \tilde{\mathcal{O}} = \mathcal{O}R^{-1}(t),$$

$$(7.28b) \quad 0 = \mathcal{O}R^{-1}(t) = -\mathcal{O}\dot{R}^{-1}(t)R(t)R^{-1}(t).$$

Rewrite (6.2a) as

$$(7.29) \quad A - R^{-1}(t)\tilde{A}R(t)R(t)\dot{R}(t)$$

and multiply by \mathcal{C} on the right using (7.27b) to obtain

$$(7.30a) \quad (A - R^{-1}(t)\tilde{A}R(t))\mathcal{C} = 0.$$

We multiply (7.29) by \mathcal{O} on the left and use (7.28b) to obtain

$$(7.30b) \quad \mathcal{O}(A - R^{-1}(t)\tilde{A}R(t)) = 0.$$

If we let $R = R(t_0)$, a constant matrix, then (7.13) follows immediately. Moreover, (6.3a) implies (7.14a) and (7.27a) and (6.3b) imply (7.14b). Q.E.D.

Remark. While minimal stationary realizations are related by (7.13) and (7.14) for some constant matrix R , they can also be related by nonconstant matrices.

Example 7.7. Consider the time varying change of state coordinates $\tilde{x} = R(t)x$ for Example 5.6, where

$$R(t) = \begin{bmatrix} 1 & 0 \\ t & 1 \end{bmatrix}.$$

The new system is given by

$$\dot{\tilde{x}}_1 = 0, \quad \tilde{x}_2(0) + 2\tilde{x}_1(1) - \tilde{x}_2(1) = v_1,$$

$$\dot{\tilde{x}}_2 = \tilde{x}_1 + u, \quad \tilde{x}_2(0) = v_2,$$

$$y = \tilde{x}_1.$$

These two systems are also related by (7.13) and (7.14), where $R = I$.

Recall that an acausal system is *causal* if $V^0 = I$ and $V^1 = 0$ and *anticausal* if $V^0 = 0$ and $V^1 = I$. It is *strictly acausal* if both V^0 and V^1 are invertible.

PROPOSITION 7.8. *Suppose Σ is a standard stationary acausal system which is controllable and observable on $[t_0, t_1]$. The state and boundary space of Σ can be decomposed $x = (x_1, x_2, x_3)$, $v = (v_1, v_2, v_3)$ into a causal part x_1 , anticausal part x_2 , and an acausal part x_3 .*

$$(7.31a) \quad \dot{x}_1 = A_{11}x_1 + B_1u,$$

$$(7.31b) \quad x_1(t_0) = v_1,$$

$$(7.32a) \quad \dot{x}_2 = A_{22}x_2 + B_2u,$$

$$(7.32b) \quad x_2(t_1) = e^{A_{22}(t_1-t_0)}v_2,$$

$$(7.33a) \quad \dot{x}_3 = A_{31}x_1 + A_{32}x_2 + A_{33}x_3 + B_3u,$$

$$(7.33b) \quad V_{33}^0x_3(t_0) + V_{33}^1x_3(t_1) = v_3 - V_{31}^0x_1(t_0) - V_{32}^0x_2(t_0) - V_{31}^1x_1(t_1) - V_{32}^1x_2(t_1).$$

Proof. Consider the flag of subspaces

$$\mathbb{R}^m \supseteq \text{Range } V^1 \supseteq \text{Range } V^0 \cap \text{Range } V^0 \subseteq 0$$

and choose state coordinates that respect this flag $x = (x_1, x_2, x_3)$. In other words,

$$\text{Range } V^1 = \{x: x_1 = 0\},$$

$$\text{Range } V^0 \cap \text{Range } V^1 = \{x: x_1 = 0, x_2 = 0\},$$

$$\text{Range } V^0 = \{x: x_2 = 0\}.$$

We choose the same coordinates on the space of boundary input values v . Since A commutes with V^0 and V^1 it leaves their ranges invariant. Because $V^0 + V^1 e^{A(t_1-t_0)} = I$, V^0 and V^1 also commute hence leave their ranges invariant. This implies that

$$A = \begin{bmatrix} A_{11} & 0 & 0 \\ 0 & A_{22} & 0 \\ A_{31} & A_{32} & A_{33} \end{bmatrix},$$

$$V^0 = \begin{bmatrix} I & 0 & 0 \\ 0 & 0 & 0 \\ V_{31}^0 & V_{32}^0 & V_{33}^0 \end{bmatrix},$$

$$V^1 = \begin{bmatrix} 0 & 0 & 0 \\ 0 & e^{A_{22}(t_0-t_1)} & 0 \\ V_{31}^1 & V_{32}^1 & V_{33}^1 \end{bmatrix}$$

where $V_{33}^0 + V_{33}^1 e^{A_{33}(t_1-t_0)} = I$. Q.E.D.

Consider the acausal part of the above system, assuming that the causal state coordinates $x_1(t)$ and anticausal state coordinates $x_2(t)$ are identically zero. This yields an acausal system

$$(7.34a) \quad \dot{x}_3 = A_{33}x_3 + B_3u,$$

$$(7.34b) \quad V_{33}^0 x_3(t_0) + V_{33}^1 x_3(t_1) = v_3,$$

which can also be decomposed into causal, anticausal and acausal parts. The decomposition process can be repeated until the acausal part is strictly acausal. In this way the original system can be decomposed into causal and anticausal parts which feed into causal and anticausal parts through the state differential equations and boundary conditions. The pattern may be repeated several times until it terminates in a strictly acausal system.

The boundary condition (1.1b) of a stationary system is said to be *separable* if $\text{Range } V^0 \cap \text{Range } V^1 = 0$.

COROLLARY 7.9. *Suppose Σ is a standard stationary acausal system which is controllable and observable on $[t_0, t_1[$. If the boundary condition is separable then Σ separates into independent causal and anticausal systems.*

REFERENCES

- [1] A. J. KRENER, *Boundary value linear systems*, Astérisque, 75-76 (1980), pp. 149-165.
- [2] ———, *Acausal linear systems*, Proc. 18th IEEE CDC, Fort Lauderdale, 1979.
- [3] ———, *Smoothing of stationary cyclic processes*, Proc. MTNS, Santa Monica, 1981.
- [4] ———, *Acausal realization theory, Part II; Linear stochastic systems*, in preparation.
- [5] J. E. WALL, A. S. WILLSKY AND N. R. SANDELL, *On the fixed-interval smoothing problem*, Stochastics, 5 (1981), pp. 1-41.
- [6] M. B. ADAMS, A. S. WILLSKY AND B. C. LEVY, *Linear estimation of boundary value stochastic processes, Parts I and II*, IEEE Trans. Automat. Control, AC-29 (1984), pp. 803-821.
- [7] J. B. EDWARDS, *Wider application of multipass systems theory, Parts I and II*, Proc. IEEE, 125 (1978), pp. 447-459.
- [8] D. H. OWENS, *Stability of multipass processes*, Proc. IEEE, 124 (1977), pp. 1079-1082.
- [9] I. GOHBERG AND M. A. KAASHOEK, *Time varying linear systems with boundary conditions and integral operators, I. The transfer function and its properties*, Integral Equations Operator Theory, 7 (1984), pp. 325-391.

- [10] I. GOHBERG AND M. A. KAASHOEK, *Time varying linear systems with boundary conditions and integral operators*, II. *Similarity and reduction*, Report NR. 261, Dept. of Mathematics and Computer Science, Vrije Universiteit, Amsterdam, 1984.
- [11] C. A. DESOER, *Notes For a Second Course on Linear Systems*, Van Nostrand-Reinhold, New York, 1970.
- [12] D. G. LUENBERGER, *Dynamic equations in descriptor form*, IEEE Trans. Automat. Control, AC-22 (1977), pp. 312-321.
- [13] I. GOHBERG AND M. A. KAASHOEK, private communication, 1984.
- [14] W. M. WONHAM, *Linear Multivariable Control, A Geometric Approach*, 2nd edition, Springer, New York, 1979.
- [15] G. BASILE AND G. MARRO, *Controlled and conditioned invariant subspaces in linear systems theory*, J. Optim. Theory Appl., 3 (1969), pp. 306-315.
- [16] R. KALMAN, *Mathematical description of linear dynamical systems*, this Journal, 1 (1963), pp. 152-192.
- [17] R. W. BROCKETT, *Finite Dimensional Linear Systems*, John Wiley, New York, 1970.

MODELING, STABILIZATION AND CONTROL OF SERIALLY CONNECTED BEAMS*

G. CHEN†¶, M. C. DELFOUR‡, A. M. KRALL† AND G. PAYRE§

Abstract. Many flexible structures consist of a large number of components coupled end to end in the form of a chain. In this paper, we consider the simplest type of such structures which is formed by N serially connected Euler-Bernoulli beams, with N actuators and sensors co-located at nodal points. When these N beams are strongly connected at all intermediate nodes and their material coefficients satisfy certain properties, uniform exponential stabilization can be achieved by stabilizing at one end point of the composite beam.

We use finite elements to discretize the partial differential equation and compute the spectra of these boundary damped operators. Numerical results are also illustrated.

Key words. serially connected beams, actuators and sensors, exponential stabilization

AMS(MOS) subject classifications. 93D15, 73K12, 73K25

1. Introduction. In this paper, we study the stabilization and control of composite beams or serially connected beams as shown in Fig. 1 where beam i , $1 \leq i \leq N$, is represented by the line segment $J_i = [x_{i-1}, x_i]$, $x_{i-1} < x_i$. We assume that each beam is uniform, with constant mass density m_i and flexural rigidity $E_i I_i$, $i = 1, 2, \dots, N$. Each beam satisfies the partial differential equation

$$(1.1) \quad m_i \frac{\partial^2 y}{\partial t^2} + E_i I_i \frac{\partial^4 y}{\partial x^4} = 0 \quad \text{on }]x_{i-1}, x_i[, \quad t > 0,$$

with initial conditions

$$(1.2) \quad y(x, 0) = y_0(x), \quad \frac{\partial y}{\partial t}(x, 0) = y_1(x), \quad 0 \leq x \leq L,$$

and boundary conditions determined by the coupling with neighboring beams.

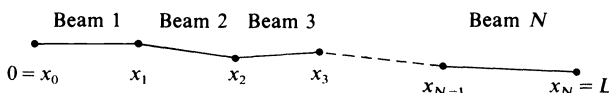


FIG. 1. Serially connected beams.

Our study is motivated by problems in structural dynamics. Many civil engineering structures such as bridges, and flexural members in large space structures are modeled by beam equations like (1.1). Due to operation, safety and performance considerations, engineers use many “in-span indeterminate conditions” to design those structures. Four basic types of such intermediate conditions have been mentioned in [8, p. II-42]:


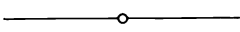
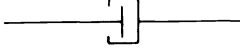
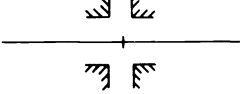
* Received by the editors October 16, 1984; accepted for publication (in revised form) February 18, 1986.

† Department of Mathematics, Pennsylvania State University, University Park, Pennsylvania 16802.

‡ Centre de recherche de mathématiques appliquées, Université de Montréal, C.P. 6128, Succ. A., Montréal, Québec, Canada H3C 3J7. The research of this author was supported in part by the Natural Sciences and Engineering Research Council of Canada, operating grant A-8730, and in part by project 24-ST.36001-3-1898, Department of Communications, Ottawa, Canada.

§ Département de génie chimique, Université de Sherbrooke, Sherbrooke, Québec, Canada J1K 2R1.

¶ The research of this author was supported by National Science Foundation grant DMS 84-01297 and Air Force Office of Scientific Research grant 85-0253.

- (1.3) Rigid support  (also called interior support),
- (1.4) Moment release  (also called interior hinge),
- (1.5) Shear release 
- (1.6) Angle guide 

Each of the above “joints” is known to have certain advantages in the design of structures.

The main objective of this paper is to study the suppression of vibration in structures. Our study is motivated by stabilization problems in large flexible space structures. These structures are dynamic in nature and require good designs of active control and passive damping mechanisms. We have observed several proposed designs of a future space station. They all contain flexible truss beam structures, some of which are linked by rotatable joints with damping devices. Those joints serve the important dual purposes of maneuvering and stabilization. Most or perhaps all of them are nonlinear in nature and hard to model mathematically. Linear joint conditions are only simplifications and approximations for those complicated joints. It will take some time before we can actually understand those structures.

For the Euler–Bernoulli beam equation (1.1), the physical meanings of various mathematical quantities are shown below:

y : transverse deflection,

$\frac{\partial y}{\partial t}$: transverse velocity,

$\theta = \frac{\partial y}{\partial x}$: rotation,

$\frac{\partial \theta}{\partial t} = \frac{\partial^2 y}{\partial t \partial x}$: angular velocity,

$\mathbb{M} = -E_i I_i \frac{\partial^2 y}{\partial x^2}$: bending moment,

$\frac{\partial \mathbb{M}}{\partial t} = -E_i I_i \frac{\partial^3 y}{\partial t \partial x^2}$: rate of change of moment,

$\mathbb{V} = -E_i I_i \frac{\partial^3 y}{\partial x^3}$: shear,

$\frac{\partial \mathbb{V}}{\partial t} = -E_i I_i \frac{\partial^4 y}{\partial t \partial x^3}$: rate of change of shear.

At a linear joint x_i , depending on the mechanical design, part or all of the physical variables from x_i^- and x_i^+ achieve equilibrium at x_i , yielding 4 mathematical relations.

The four basic joints (1.3)–(1.6) are the simplest special cases of such mathematical relations. They are known [8] to cause discontinuities (across the joint) in one of the state variables y , θ , \mathbb{M} and \mathbb{V} , and leave the other 3 state variables continuous, particularly, the “dual” (or “complementary”) state variable, as indicated in Table 1.

In the mathematical modeling of controller action at joints, it is natural and reasonable to designate the discontinuity across a joint as the control variable at x_i , because its values can be adjusted—under our command and with different mechanical

TABLE 1

Joint	Discontinuous state variables	Continuous state variables
Rigid support	\mathbb{V}	$y; \theta, \mathbb{M}$
Moment release	θ	$\mathbb{M}; y, \mathbb{V}$
Shear release	y	$\mathbb{V}; \theta, \mathbb{M}$
Angle guide	\mathbb{M}	$\theta; y, \mathbb{V}$

The first state variable in each entry above is the dual variable.

designs—to influence the dynamic response of the structure. For example, let x_i be a rigid support joint with control u_i . Then the intermediate boundary conditions at x_i are given as follows:

$$\begin{aligned} y(x_i^-, t) &= y(x_i^+, t), \\ \frac{\partial y}{\partial x}(x_i^-, t) &= \frac{\partial y}{\partial x}(x_i^+, t) \quad (\text{i.e., } \theta^- = \theta^+), \\ -E_i I_i \frac{\partial^2 y}{\partial x^2}(x_i^-, t) &= -E_{i+1} I_{i+1} \frac{\partial^2 y}{\partial x^2}(x_i^+, t) \quad (\text{i.e., } \mathbb{M}^- = \mathbb{M}^+), \\ E_i I_i \frac{\partial^3 y}{\partial x^3}(x_i^-, t) - E_{i+1} I_{i+1} \frac{\partial^3 y}{\partial x^3}(x_i^+, t) &= u_i(t). \end{aligned}$$

This is just one of the cases to be studied here by us.

Due to the limited scope of this paper, here we will only consider in-span controllers at joints that are rigid supports or angle guides, or combinations. Moment and shear release or other types of joints seem to have different advantages and disadvantages which are currently being investigated. We hope that our paper can motivate further study in this direction.

Our problem is to study the control and stabilization of serially connected beams as described above. We wish to establish a strong form of stabilization—uniform exponential stabilization—with one stabilizer only. This stabilizer, formed by velocity feedback, is considered to be located at one end of the span. Another situation, more practical perhaps, is to consider the stabilizer located at an interior node of the span. This corresponds to a much harder problem and is briefly discussed in § 4.3.

The organization of this paper is as follows.

In § 2 we first recall a fundamental theorem of existence for a second order (in time) hyperbolic equation. A control-theoretic formulation for a vibrating system consisting of a composite clamped beam with N pairs of co-located sensors and actuators is then given.

In § 3, we present the main result of the paper in Theorem 3.1. A special case of this theorem says that N serially connected beams such that

$$(1.7) \qquad m_i \leq m_{i+1}, \quad E_i I_i \geq E_{i+1} I_{i+1}, \quad 1 \leq i < N-1$$

can be uniformly exponentially stabilized with only one sensor and one actuator at the end ($x = L$) of the beam:

$$(1.8) \qquad E_N I_N \frac{\partial^3 y}{\partial x^3}(L, t) - k_{0N} \frac{\partial y}{\partial t}(L, t) = 0, \qquad k_{0N} > 0.$$

Previously, results have been established for exponential stabilization of all of the (finitely many) modes of the finite element model, but none for the continuum model.

In § 4, we discuss various cases, including different conservative left end boundary conditions, a beam with varying mass density and flexural rigidity, and stabilization at an intermediate node.

In § 5, we use the finite element numerical method to compute the spectrum. Some interesting information on the distribution of eigenvalues is illustrated.

Notation. \mathbb{R} will be the field of all real numbers. For $n \geq 1$, \mathbb{R}^n will denote the n -dimensional Euclidean space with norm $|x|$ and inner product $x \cdot y$ for x and y in \mathbb{R}^n . Given a Banach space E , its topological dual will be denoted E' and the duality pairing between an element x^* of E' and x of E by $\langle x^*, x \rangle_{E' \times E}$. Given a continuous linear map L between two Banach spaces E and F , L^* will be the corresponding dual map between F' and E' . Let a and b , $a < b$, be two real numbers and E be a Banach space. $L^p(a, b; E)$ will be the Banach space of all equivalence classes of Lebesgue measurable functions from $[a, b]$ into E which are p -integrable ($1 \leq p < \infty$) or essentially bounded ($p = \infty$). Given a function $y: [a, b] \rightarrow E$, \dot{y} and \ddot{y} will denote the first and second vectorial distribution derivatives of y . $C^m(a, b; E)$ will be the space of m -times boundedly continuously differentiable functions from $[a, b]$ into E . For $m \geq 1$ and an open interval I , $H^m(I)$ will be the Sobolev space of all functions in $L^2(I; \mathbb{R})$ whose first m distributional derivatives belong to $L^2(I; \mathbb{R})$.

2. Control-theoretic formulation of a clamped serially connected beam with N pairs of co-located actuators and sensors. Let V and H be two real Hilbert spaces with norms $\|\cdot\|$ and $|\cdot|$, respectively. Identify the elements of the dual H' of H with those of H and assume that the chains of injection are continuous with dense image

$$(2.1) \quad V \rightarrow H \equiv H' \rightarrow V'.$$

Let $a(t; v, w)$ be a family of bilinear forms on V such that

$$(2.1) \quad V \rightarrow H \equiv H' \rightarrow V'.$$

Let $a(t; v, w)$ be a family of bilinear forms on V such that

$$(2.2) \quad \text{for all } v, w, \quad t \rightarrow a(t; v, w) \text{ is } C^1([0, T]; \mathbb{R}) \quad \text{for all } T > 0,$$

$$(2.3) \quad \text{for all } v, w, \quad a(t; v, w) = a(t; w, v) \quad \text{for all } t \geq 0,$$

$$(2.4) \quad \text{there exists a } \lambda \in \mathbb{R}, \alpha > 0 \quad \text{for all } v, \quad a(t; v, v) + \lambda |v|^2 \geq \alpha \|v\|^2.$$

We let $A(t) \in \mathcal{L}(V, V')$ denote the operator defined by $a(t; v, w)$.

Let

$$(2.5) \quad C: V \rightarrow \mathbb{R}^m \quad (m \geq 1 \text{ an integer})$$

be an arbitrary linear map and $u \in L^2(0, T; \mathbb{R}^m)$, $T > 0$, be a control function. We now consider the controlled system

$$(2.6) \quad A(t)y(t) + \ddot{y}(t) + C^*u(t) = 0, \quad y_0 \in V, \quad y_1 \in H'.$$

This problem is well posed [5] and has a unique solution under appropriate hypotheses on A .

Let the observation be given by

$$(2.7) \quad z(t) = C\dot{y}(t)$$

in a sense to be made clear in (2.14). Let K be an $m \times m$ matrix such that

$$(2.8) \quad \text{there exists a } \beta \geq 0, \quad \text{such that for all } u \in \mathbb{R}^m, \quad Ku \cdot u \geq \beta |u|^2.$$

Define the feedback law

$$(2.9) \quad u(t) = K\dot{z}(t).$$

The resulting closed loop system will be

$$(2.10) \quad A(t)y(t) + C^*KC\dot{y}(t) + \ddot{y}(t) = 0, \quad y(0) = y_0, \quad \dot{y}(0) = y_1.$$

The perturbing term is

$$(2.11) \quad G\dot{y}(t), \quad G = C^*KC \in \mathcal{L}(V, V').$$

The next theorem shows that system (2.10) is well posed.

THEOREM 2.1. (i) *Under the hypotheses on the matrix K , equation (2.10) has a unique solution*

$$(2.12) \quad y \in C([0, T]; V), \quad \dot{y} \in C([0, T]; H), \quad C\dot{y} \in L^2(0, T; \mathbb{R}^m), \quad \ddot{y} \in L^2(0, T; V').$$

(ii) *Let $A(t) \equiv A$, $0 \leq t < \infty$, be autonomous. Then the densely defined closed operator*

$$\mathcal{A} \equiv \begin{bmatrix} 0 & I \\ A & 0 \end{bmatrix}$$

is dissipative with domain $D(\mathcal{A}) = D(A) \oplus V$ and generates a C_0 -semigroup of contractions $\{S(t) | t \geq 0\}$ on $V \oplus H$. The solution y of (2.10) forms a C_0 -semigroup in $V \times H$ defined by

$$(2.13) \quad S(t)(y_0, y_1) = (y(t), \dot{y}(t)), \quad t \geq 0.$$

(iii) *For (y_0, y_1) in $D(\mathcal{A})$,*

$$(2.14)$$

$$y \in C^1([0, T]; V), \quad \dot{y} \in C^1([0, T]; H), \quad C\dot{y} \in C([0, T]; \mathbb{R}^m), \quad \ddot{y} \in C([0, T]; V').$$

Proof. (ii) is a simple consequence of the Lumer-Phillips theorem. (i) is a simple extension of a theorem in [5]. \square

With the preceding formalism in mind, we now consider the control-theoretic formulation of a composite clamped beam with N pairs of co-located sensors and actuators.

Let a beam of length L be made up of N connected beams $[x_{i-1}, x_i]$, $1 \leq i \leq N$. For each i th segment beam the equation is

$$(2.15) \quad m_i \frac{\partial^2 y}{\partial t^2} + E_i I_i \frac{\partial^4 y}{\partial x^4} = 0, \quad x_{i-1} < x < x_i, \quad 1 \leq i \leq N.$$

The beam is clamped at $x = 0$:

$$(2.16) \quad y(0, t) = 0, \quad \frac{\partial y}{\partial x}(0, t) = 0, \quad t > 0.$$

At each interior node x_i , the beam is strongly connected:

$$(2.17) \quad y(x_i^-, t) = y(x_i^+, t), \quad \frac{\partial y}{\partial x}(x_i^-, t) = \frac{\partial y}{\partial x}(x_i^+, t), \quad t > 0,$$

with a point force $u_{0i}(t)$ and a point bending moment $u_{1i}(t)$ applied at x_i :

$$(2.18) \quad \begin{aligned} E_i I_i \frac{\partial^3 y}{\partial x^3}(x_i^-, t) - E_{i+1} I_{i+1} \frac{\partial^3 y}{\partial x^3}(x_i^+, t) &= u_{0i}(t), \\ - \left[E_i I_i \frac{\partial^2 y}{\partial x^2}(x_i^-, t) - E_{i+1} I_{i+1} \frac{\partial^2 y}{\partial x^2}(x_i^+, t) \right] &= u_{1i}(t), \end{aligned} \quad 1 \leq i \leq N-1.$$

At the right end $x_N = L$, a point force $u_{0N}(t)$ and a point bending moment $u_{1N}(t)$ are applied:

$$(2.19) \quad \begin{aligned} E_N I_N \frac{\partial^3 y}{\partial x^3}(L, t) &= u_{0N}(t), \\ -E_N I_N \frac{\partial^2 y}{\partial x^2}(L, t) &= u_{1N}(t), \end{aligned} \quad i = N.$$

Remark 2.2. The intermediate conditions (2.17) and (2.18) signify that the joint at x_i is:

- (i) A rigid support, if $u_{1i}(t) \equiv 0$, but $u_{0i}(t) \neq 0$;
- (ii) An angle guide, if $u_{0i}(t) \equiv 0$, but $u_{1i}(t) \neq 0$;
- (iii) A combination of (i) and (ii), if $u_{0i}(t) \neq 0$ and $u_{1i}(t) \neq 0$.

The above feedback laws (2.18) and (2.19) are realizable as all variables $E_i I_i (\partial^3 y / \partial x^3)$ (=shear) and $E_i I_i (\partial^2 y / \partial x^2)$ (=bending moment) on the left of (2.18) and (2.19) are determinable.

Remark 2.3. The boundary conditions (2.17)–(2.18) are the natural conditions arising from the presence of point forces $u_{0i}(t)$ and point bending moments $u_{1i}(t)$ at each point x_i . In the static case the calculus of variations can be used to show that the solution $y \in H^2(0, L)$ of the boundary value problem

$$\begin{aligned} \frac{\partial^4 y}{\partial x^4}(x) &= 0 \quad \text{on }]x_{i-1}, x_i[, \quad 1 \leq i \leq N, \\ E_i I_i \frac{\partial^3 y}{\partial x^3}(x_i^-) - E_{i+1} I_{i+1} \frac{\partial^3 y}{\partial x^3}(x_i^+) &= u_{0i}, \\ -\left[E_i I_i \frac{\partial^2 y}{\partial x^2}(x_i^-) - E_{i+1} I_{i+1} \frac{\partial^2 y}{\partial x^2}(x_i^+) \right] &= u_{1i}, \quad 1 \leq i \leq N-1, \\ E_N I_N \frac{\partial^3 y}{\partial x^3}(L) &= u_{0N}, \quad \frac{\partial y}{\partial x}(0) = 0, \\ -E_N I_N \frac{\partial^2 y}{\partial x^2}(L) &= u_{1N}, \quad y(0) = 0, \end{aligned}$$

is precisely the function that minimizes the strain energy plus the sum of the potential energy of the external forces u_{0i} and bending moments u_{1i} :

$$U(v) = \sum_{i=1}^N \frac{1}{2} E_i I_i \int_{x_{i-1}}^{x_i} |D^2 v(x)|^2 dx + u_{0i} v(x_i) + u_{1i} Dv(x_i)$$

over all deflection distributions v in $H^2(0, L)$ such that

$$v(0) = 0 \quad \text{and} \quad Dv(0) = 0.$$

The spaces H , V and the operators A and C are

$$(2.20) \quad \begin{aligned} H &= L^2(0, L), \quad V = \{v \in H^2(0, L) | v(0) = 0, Dv(0) = 0\}, \\ \langle Av, w \rangle &= \sum_{i=1}^N E_i I_i \int_{x_{i-1}}^{x_i} D^2 v(x) \cdot D^2 w(x) dx, \\ v \rightarrow Cv &= (C_1 v, C_2 v, \dots, C_N v), \quad C_i v = (v(x_i), Dv(x_i)). \end{aligned}$$

Denote by

$$(2.21) \quad u(t) = (u_1(t), u_2(t), \dots, u_N(t)), \quad u_i(t) = (u_{0i}(t), u_{1i}(t)),$$

the vector of control functions at time t in $(\mathbb{R}^2)^N$. Then the system of equations (2.15) to (2.19) can be rewritten in the following way:

$$(2.22) \quad A(t)y(t) + M\ddot{y}(t) + C^*u(t) = 0,$$

where

$$C^*u(t) = \sum_{i=1}^N C_i^* u_i(t)$$

and $M = M(x)$ is defined by

$$(2.23) \quad M(x) = \sum_{i=1}^N m_i \chi_{J_i}(x),$$

$\chi_{J_i}(x)$ is the characteristic function of the interval

$$J_i = [x_{i-1}, x_i].$$

At each point x_i , $1 \leq i \leq N$, we observe the vector

$$(2.24) \quad z_i(t) = \left(\frac{\partial y}{\partial t}(x_i, t), \frac{\partial^2 y}{\partial x \partial t}(x_i, t) \right), \quad t > 0.$$

The global observation is given by

$$(2.25) \quad z(t) = (z_1(t), z_2(t), \dots, z_N(t)) \in (\mathbb{R}^2)^N$$

and the feedback law is chosen as

$$(2.26) \quad u(t) = Kz(t), \quad K = \text{a } 2N \times 2N \text{ matrix.}$$

By construction

$$(2.27) \quad z(t) = C\dot{y}(t) \quad \text{and} \quad u(t) = KC\dot{y}(t).$$

So finally

$$(2.28) \quad A(t)y(t) + M\ddot{y}(t) + C^*KC\dot{y}(t) = 0.$$

Note that (2.28) differs slightly from (2.10) in Theorem 2.1 due to the presence of M . Nevertheless, the conclusions of Theorem 2.1 remain valid and applicable to our problem, as a straightforward modification will enable the proof to go through without difficulty.

Also note that for (2.28), we use the equivalent norm

$$\|(w, z)\|_{V \times H}^2 = \|w\|_V^2 + \int_0^L M(x)z^2(x) dx.$$

We will restrict our attention to special feedback matrices which are made up of N 2×2 matrices K_i , $1 \leq i \leq N$, along the diagonal. All other entries are zero.

$$(2.29) \quad K = \begin{bmatrix} K_1 & & & & & \\ & \ddots & & & & \\ & & K_2 & & & \\ & & & \ddots & & \\ & & & & K_3 & \\ & & & & & \ddots \\ & & 0 & & & & K_N \end{bmatrix}.$$

The above hypothesis on K amounts to assuming that sensors and actuators are co-located with local feedback.

3. Uniform exponential stabilization of serially connected beams. For the model proposed in the last section, we now wish to find certain feedback matrices K to obtain uniform exponential stabilization.

We assume that the matrices K_i are of the form

$$(3.1) \quad K_i = \begin{bmatrix} k_{0i} & -c_i \\ c_i & k_{1i} \end{bmatrix}, \quad k_{0i} \geq 0, \quad k_{1i} \geq 0, \quad c_i \in \mathbb{R}, \quad 1 \leq i \leq N.$$

It is readily seen that property (2.8) is satisfied, so Theorem 2.1 is applicable.

The next theorem is the main theorem which gives a set of sufficient conditions on the parameters to ensure uniform exponential stabilizability.

THEOREM 3.1. Assume that the $2N \times 2N$ matrix K satisfies (2.29) where the submatrices K_i satisfy (3.1). Also assume that

$$(3.2) \quad m_i \leq m_{i+1}, \quad E_i I_i \geq E_{i+1} I_{i+1} \quad \text{for } 1 \leq i \leq N-1.$$

Let the initial conditions y_0 and y_1 in (1.2) be given in $V \times H$. Furthermore, assume that either of the following occurs:

(i) Case 1.

$$(3.3) \quad \begin{aligned} k_{0N} > 0, \quad k_{1N} \geq 0, \quad c_N = 0, \\ k_{0i} \geq 0, \quad k_{1i} = 0, \quad c_i = 0, \quad 1 \leq i \leq N-1; \end{aligned}$$

(ii) Case 2.

$$(3.4) \quad \begin{aligned} k_{0N} > 0, \quad k_{1N} > 0, \quad c_N \in \mathbb{R} \quad (\text{arbitrary}), \\ k_{0i} \geq 0, \quad k_{1i} = 0, \quad c_i = 0, \quad 1 \leq i \leq N-1. \end{aligned}$$

Then there exist $\mu > 0$, $\omega > 0$ such that for all $(y_0, y_1) \in V \times H$, the energy of the system (2.28) decays uniformly exponentially:

$$\begin{aligned} \frac{1}{2} \sum_{i=1}^N \int_{x_{i-1}}^{x_i} [m_i |\dot{y}(t)|^2 + E_i I_i |D^2 y(t)|^2] dx &\leq \mu e^{-\omega t} \\ &\cdot \frac{1}{2} \sum_{i=1}^N \int_{x_{i-1}}^{x_i} [m_i |y_1|^2 + E_i I_i |D^2 y_0|^2] dx. \end{aligned}$$

Proof. The proof necessitates the construction of a Lyapunov functional and uses energy multipliers (see [1], [4] for similar techniques applied to the wave equation).

Let the Lyapunov functional be

$$v(t) = v(t, y(t), \dot{y}(t)),$$

where (y, \dot{y}) is the solution in § 2 with initial condition $(y_0, y_1) \in D(\mathcal{A})$, satisfying

$$(y, \dot{y}) \in C([0, T]; D(\mathcal{A})) \cap C^1([0, T]; V \times H) \quad \text{for any } T > 0.$$

We wish to show that

$$(3.5) \quad \text{there exists } T' > 0 \text{ such that for } t \geq T', (dv(t)/dt) \leq 0.$$

We proceed stepwise as follows.

(i) *Defining and utilizing the functional $v(t)$.*

On each interval $J_i = [x_{i-1}, x_i]$, $1 \leq i \leq N$, we let

$$(3.6) \quad \begin{aligned} v_i(t) = (1-\varepsilon)t \int_{J_i} [m_i |\dot{y}(x, t)|^2 + E_i I_i |D^2 y(x, t)|^2] dx \\ + \int_{J_i} 2m_i x \dot{y}(x, t) Dy(x, t) dx \quad \text{with } 0 < \varepsilon < 1, \end{aligned}$$

and define

$$(3.7) \quad v(t) = \sum_{i=1}^N v_i(t).$$

If (3.5) can be justified, then there exists some constant $c(T') > 0$, such that

$$(3.8) \quad v(t) \leq v(T') \leq c(T')E(y(0), \dot{y}(0)),$$

where E is the elastic energy norm

$$E(y(t), \dot{y}(t)) = \frac{1}{2} \int_0^L [M(x)|\dot{y}(x, t)|^2 + (EI)(x)|D^2y(x, t)|^2] dx$$

and

$$(EI)(x) = \sum_{i=1}^N E_i I_i \chi_{J_i}(x).$$

It is easy to show that there exists $c > 0$ such that (for T' sufficiently large and $t \geq T'$)

$$(3.9) \quad \left| \int_0^L 2x\dot{y}(x, t)Dy(x, t) dx \right| \leq cE(y(t), \dot{y}(t)).$$

Thus from (3.6)–(3.9) we deduce that

$$(3.10) \quad [(1-\varepsilon)t - c]E(y(t), \dot{y}(t)) \leq c(T')E(y(0), \dot{y}(0)), \quad t \geq T',$$

and consequently, for some $\bar{c}(\bar{T}) > 0$,

$$(3.11) \quad E(y(t), \dot{y}(t)) \leq \bar{c}(\bar{T})E(y(0), \dot{y}(0)), \quad t \geq T = \max \left\{ T', \frac{2c}{1-\varepsilon} \right\}.$$

The energy E is bounded on $[0, \bar{T}]$ and the inequality extends to initial conditions $(y_0, y_1) \in V \times H$, yielding

$$\int_{\bar{T}}^{\infty} E(y(t), \dot{y}(t))^2 dt \leq \bar{c}(T)^2 E(y_0, y_1)^2 \int_{\bar{T}}^{\infty} \left[\frac{1}{(1-\varepsilon)t - c} \right]^2 dt < \infty.$$

A theorem in [7, p. 121] applies, and we conclude that there exist $\mu \geq 1$, $\omega > 0$, such that

$$(3.12) \quad E(y(t), \dot{y}(t)) \leq \mu e^{-\omega t} E(y_0, y_1).$$

This proves uniform exponential stabilizability of the system.

(ii) *Computation of $(d/dt)v_i(t)$.* We want to show that

$$(3.13) \quad \frac{d}{dt} v_i(t) + \int_{J_i} DQ_i(x, t) dx + \int_{J_i} R_i(x, t) dx = 0,$$

with

$$(3.14) \quad \begin{aligned} Q_i(x, t) = & 2E_i I_i (1-\varepsilon)t(\dot{y}D^3y - D\dot{y}D^2y) - m_i x|\dot{y}|^2 \\ & + E_i I_i [2xDyD^3y - 2DyD^2y - x|D^2y|^2] \end{aligned}$$

and

$$(3.15) \quad R_i(x, t) = \varepsilon|\dot{y}|^2 + E_i I_i (2+\varepsilon)|D^2y|^2.$$

Differentiating v_i in (3.6), we get

$$\begin{aligned}\frac{d}{dt} v_i(t) &= \int_{J_i} \{(1-\varepsilon)[m_i|\dot{y}|^2 + E_i I_i |D^2 y|^2] \\ &\quad + 2(1-\varepsilon)t[m_i \dot{y}\ddot{y} + E_i I_i D^2 y D^2 \dot{y}] + 2m_i x[\ddot{y} D y + \dot{y} D \dot{y}]\} dx \\ &= (1-\varepsilon) \int_{J_i} [m_i|\dot{y}|^2 + E_i I_i |D^2 y|^2] dx + 2 \int_{J_i} m_i \ddot{y}[(1-\varepsilon)t\dot{y} + x D y] dx \\ &\quad + 2(1-\varepsilon)t \int_{J_i} E_i I_i D^2 y D^2 \dot{y} dx + \int_{J_i} m_i x D(\dot{y}^2) dx.\end{aligned}$$

On J_i ,

$$m_i \ddot{y} + E_i I_i D^4 y = 0$$

holds, so it can be used to eliminate \ddot{y} . Furthermore,

$$\begin{aligned}D(D^3 y \dot{y}) &= D^3 y D \dot{y} + D^4 y \dot{y}, \\ D(D^2 y D \dot{y}) &= D^3 y D \dot{y} + D^2 y D^2 \dot{y}, \\ D^2 y D^2 \dot{y} - D^4 y \dot{y} &= D(D^2 y D \dot{y} - D^3 y \dot{y});\end{aligned}$$

thus

$$\begin{aligned}(3.16) \quad \frac{d}{dt} v_i(t) &= (1-\varepsilon) \int_{J_i} [m_i|\dot{y}|^2 + E_i I_i |D^2 y|^2] dx - \int_{J_i} 2E_i I_i x D^4 y D y dx \\ &\quad + 2(1-\varepsilon)t E_i I_i \int_{J_i} D[D^2 y D \dot{y} - D^3 y \dot{y}] dx + \int_{J_i} m_i x D(\dot{y})^2 dx.\end{aligned}$$

Also

$$(3.17) \quad x D(\dot{y})^2 = D(x|\dot{y}|^2) - |\dot{y}|^2$$

and

$$(3.18) \quad D[2x(D^3 y D y - \tfrac{1}{2}|D^2 y|^2) - 2D^2 y D y] = 2x D^4 y D y - 3|D^2 y|^2.$$

The substitution of (3.17) and (3.18) in (3.16) yields (3.13)–(3.15).

(iii) *Verification of inequality (3.5).* In view of (3.13), the critical property (3.5) will be verified if we can show that

$$\begin{aligned}(3.19) \quad &\sum_{i=1}^N \left\{ [Q_i(x_i, t) - Q_i(x_{i-1}, t)] + \int_{J_i} R_i(x, t) dx \right\} \\ &= -\frac{d}{dt} \sum_{i=1}^N v_i(t) = -\frac{d}{dt} v(t) \geq 0, \quad t \geq T' .\end{aligned}$$

Since all feedback laws are local we need only verify that

$$\begin{aligned}(3.20) \quad &S_i(t) + \int_{J_i} R_i(x, t) dx \geq 0, \quad 1 \leq i \leq N, \quad t \geq T', \\ &S_0(t) \geq 0,\end{aligned}$$

where

$$\begin{aligned}(3.21) \quad &S_0(t) \equiv -Q_1(0, t), \\ &S_i(t) \equiv Q_i(x_i, t) - Q_{i+1}(x_i, t), \quad 1 \leq i \leq N-1, \\ &S_N(t) = Q_N(x_N, t),\end{aligned}$$

We first deal with the term R_i :

$$R_i(x, t) \cong \varepsilon [m_i |\dot{y}|^2 + E_i I_i |D^2 y|^2].$$

To deal with the Q_i 's we use the decomposition

$$\begin{aligned} Q_{i1}(x, t) &\equiv 2(1 - \varepsilon)t E_i I_i [\dot{y} D^3 y - D\dot{y} D^2 y], \\ Q_{i2}(x, t) &\equiv -m_i x |\dot{y}|^2, \\ Q_{i3}(x, t) &\equiv 2E_i I_i x D\dot{y} D^3 y, \\ Q_{i4}(x, t) &\equiv -2E_i I_i D\dot{y} D^2 y, \\ Q_{i5}(x, t) &\equiv -E_i I_i x |D^2 y|^2; \end{aligned} \quad (3.22)$$

thus

$$Q_i(x, t) = \sum_{j=1}^5 Q_{ij}(x, t). \quad (3.23)$$

Next, for $0 \leq i \leq N$ and $1 \leq j \leq 5$ we define

$$\begin{aligned} S_{ij}(t) &\equiv Q_{ij}(x_i, t) - Q_{i+1,j}(x_i, t), \quad 1 \leq i \leq N-1, \\ S_{0j}(t) &\equiv -Q_{1j}(0, t), \quad S_{Nj}(t) \equiv Q_{Nj}(x_N, t), \quad i = 0 \text{ or } N, \end{aligned} \quad (3.24)$$

and let

$$S_i(t) \equiv \sum_{j=1}^5 S_{ij}(t), \quad 0 \leq i \leq N. \quad (3.25)$$

For each i we compute $S_i(t)$. When $i = 0$, at the left end,

$$S_0(t) = 0. \quad (3.26)$$

For $i = N$, at the right end,

$$\begin{aligned} S_N(t) &= E_N I_N [2(1 - \varepsilon)t \dot{y} + 2x_N D\dot{y}] D^3 y - E_N I_N [2(1 - \varepsilon)t D\dot{y} + 2D\dot{y}] D^2 y \\ &\quad - m_N x_N |\dot{y}|^2 - E_N I_N x_N |D^2 y|^2 \\ &= [2(1 - \varepsilon)t \dot{y} + 2x_N D\dot{y}] [k_{0N} \dot{y} - c_N D\dot{y}] + [2(1 - \varepsilon)t D\dot{y} + 2D\dot{y}] [c_N \dot{y} + k_{1N} D\dot{y}] \\ &\quad - \frac{x_N}{E_N I_N} [c_N \dot{y} + k_{1N} D\dot{y}]^2 - m_N x_N |\dot{y}|^2 \\ &= 2(1 - \varepsilon)t [k_{0N} |\dot{y}|^2 + k_{1N} |D\dot{y}|^2] + D\dot{y} [2x_N (k_{0N} \dot{y} - c_N D\dot{y}) + 2(c_N \dot{y} + k_{1N} D\dot{y})] \\ &\quad - \frac{x_N}{E_N I_N} [c_N \dot{y} + k_{1N} D\dot{y}]^2 - m_N x_N |\dot{y}|^2. \end{aligned}$$

For any $\delta > 0$,

$$|D\dot{y}| \cdot |\dot{y}| \leq \delta^2 |D\dot{y}|^2 + \frac{1}{\delta^2} |\dot{y}|^2.$$

Also, for any $\alpha, \beta \in \mathbb{R}$,

$$(\alpha + \beta)^2 \leq 2(\alpha^2 + \beta^2).$$

Therefore

$$\begin{aligned}
 S_N(t) &\geq 2(1-\varepsilon)t[k_{0N}|\dot{y}|^2 + k_{1N}|D\dot{y}|^2] \\
 &\quad - \left\{ \delta^2 |Dy|^2 + \frac{1}{\delta^2} |(2x_N k_{0N} + 2c_N)\dot{y} + (-2x_N c_N + 2k_{1N})D\dot{y}|^2 \right\} \\
 &\quad - \frac{x_N}{E_N I_N} (c_N \dot{y} + k_{1N} D\dot{y})^2 - m_N x_N |\dot{y}|^2 \\
 &\geq 2(1-\varepsilon)t[k_{0N}|\dot{y}|^2 + k_{1N}|D\dot{y}|^2] \\
 &\quad - \frac{2}{\delta^2} \{ |2x_N k_{0N} + 2c_N|^2 |\dot{y}|^2 + |-2x_N c_N + 2k_{1N}|^2 |D\dot{y}|^2 \} \\
 &\quad - \frac{2x_N}{E_N I_N} \{ c_N^2 |\dot{y}|^2 + k_{1N}^2 |D\dot{y}|^2 \} - m_N x_N |\dot{y}|^2 - \delta^2 |Dy|^2.
 \end{aligned}$$

The last term can be absorbed into $\int_{J_N} R_N dx$ for small δ . If either

$$(3.27a) \quad k_{0N} > 0, \quad c_N = 0, \quad k_{1N} \geq 0, \quad \text{or}$$

$$(3.27b) \quad k_{0N} > 0, \quad c_N \neq 0, \quad k_{1N} > 0$$

then for t large enough, the sum of the remaining terms is nonnegative.

At intermediate nodes, under the assumption that

$$(3.28) \quad k_{0i} \geq 0, \quad c_i = 0, \quad k_{1i} = 0, \quad m_i \leq m_{i+1}, \quad E_i I_i \geq E_{i+1} I_{i+1}, \quad 1 \leq i \leq N-1,$$

we have

$$\begin{aligned}
 S_{i1}(t) &= 2(1-\varepsilon)t\{\dot{y}[E_i I_i D^3 y(x_i^-, t) - E_{i+1} I_{i+1} D^3 y(x_i^+, t)] \\
 &\quad - D\dot{y}[E_i I_i D^2 y(x_i^-, t) - E_{i+1} I_{i+1} D^2 y(x_i^+, t)]\} \\
 &= 2(1-\varepsilon)tk_{0i}|\dot{y}|^2 \geq 0, \\
 S_{i2}(t) &= \frac{1}{2}(-m_i + m_{i+1})x_i|\dot{y}(x_i, t)|^2 \geq 0, \\
 S_{i3}(t) &= 2x_i D\dot{y}(x_i, t)[E_i I_i D^3 y(x_i^-, t) - E_{i+1} I_{i+1} D^3 y(x_i^+, t)] \\
 &= 2x_i D\dot{y}(x_i, t)k_{0i}\dot{y}(x_i, t), \\
 S_{i4}(t) &= 2D\dot{y}(x_i, t)[-E_i I_i D^2 y(x_i^-, t) + E_{i+1} I_{i+1} D^2 y(x_i^+, t)] = 0, \\
 S_{i5}(t) &= x_i[-E_i I_i |D^2 y(x_i^-, t)|^2 + E_{i+1} I_{i+1} |D^2 y(x_i^+, t)|^2] \\
 &= \left(\frac{E_{i+1} I_{i+1}}{E_i I_i} \right) x_i [E_i I_i - E_{i+1} I_{i+1}] |D^2 y(x_i, t)|^2 \geq 0.
 \end{aligned}$$

$S_{i3}(t)$ can be absorbed into $S_{i1}(t)$ and $\int_{J_i} R_i dx$.

So we have shown that (3.19) is satisfied. This concludes the proof of the theorem. \square

An immediate consequence of the uniform exponential stabilizability result above is the exact controllability result (see [6], [9]).

Remark 3.2. Condition (3.2) has an interesting physical interpretation: as

$$\frac{E_{i+1}I_{i+1}}{m_{i+1}} \leq \frac{E_i I_i}{m_i}, \quad i = 1, 2, \dots, N-1,$$

the beam is “more flexible” toward the right end.

4. Miscellaneous discussions: Other cases and boundary conditions.

4.1. Change of the conservative left end condition. The proof of Theorem 3.1 was carried out with the conservative clamped left end condition (2.16). Other types of conservative left end boundary conditions can be used; they are given below:

- (i) $y(0, t) = 0, \quad \frac{\partial^2 y}{\partial x^2}(0, t) = 0$ (simply supported or pinned left end),
- (ii) $\frac{\partial^2 y}{\partial x^2}(0, t) = 0, \quad \frac{\partial^3 y}{\partial x^3}(0, t) = 0$ (free left end),
- (iii) $\frac{\partial y}{\partial x}(0, t) = 0, \quad \frac{\partial^3 y}{\partial x^3}(0, t) = 0$ (shear hinge left end),
- (iv) $\frac{\partial y}{\partial t}(0, t) = 0, \quad \frac{\partial y}{\partial x}(0, t) = 0,$
- (v) $\frac{\partial y}{\partial t}(0, t) = 0, \quad \frac{\partial^2 y}{\partial x^2}(0, t) = 0,$
- (vi) $\frac{\partial^2 y}{\partial t \partial x}(0, t) = 0, \quad \frac{\partial^3 y}{\partial x^3}(0, t) = 0.$

For any of the above, very minor modifications of the proof will enable Theorem 3.1 to go through, and the exponential decay result holds.

Note that for the above boundary conditions, the elastic energy functional E no longer serves as a norm. A new underlying Hilbert space $V \times H$ must be formulated which usually excludes one or two steady states in $H^2 \times L^2$.

4.2. Stabilization at one end for a beam with varying mass density $m(x)$ and flexural rigidity $I(x)E(x)$. Consider the case of a single beam

$$(4.1) \quad m(x) \frac{\partial^2 y}{\partial t^2}(x, t) + \frac{\partial^2}{\partial x^2} \left[E(x) I(x) \frac{\partial^2 y}{\partial x^2}(x, t) \right] = 0, \quad 0 < x < L, \quad t > 0,$$

which is clamped, simply supported or free at the left end $x = 0$. At the right end $x = L$, assume dissipative boundary conditions

$$(4.2) \quad \begin{aligned} (EI y_{xx})_x &= k_{0N} y_t, & k_{0N} &> 0, \\ -EI y_{xx} &= k_{1N} y_{xt}, & k_{1N} &\geq 0. \end{aligned}$$

Under what conditions on $m(x)$ and $E(x)I(x)$ do we have uniform exponential decay

$$(4.3) \quad \begin{aligned} & \frac{1}{2} \int_0^L \left[m(x) \frac{\partial y(x, t)^2}{\partial t} + E(x) I(x) \left(\frac{\partial^2 y(x, t)}{\partial x^2} \right)^2 \right] dx \\ & \leq \frac{1}{2} \mu e^{-\omega t} \int_0^L \left[m(x) \frac{\partial y^2}{\partial t}(x, 0) + E(x) I(x) \left(\frac{\partial^2 y}{\partial x^2}(x, 0) \right)^2 \right] dx \end{aligned}$$

as in § 3?

We return to (3.2). The hypothesis (3.2) seems to signify that the beam (4.1), (4.2) should be more massive and less rigid toward the right end, i.e.,

$$(4.4) \quad m(x_2) \geq m(x_1), \quad E(x_2)I(x_2) \leq E(x_1)I(x_1) \quad \text{for } x_1, x_2 \in [0, L], \quad x_2 > x_1$$

in order for a proof for (4.3) to go through. A closer look indicates otherwise.

An energy identity for (4.1) can be given as

$$\begin{aligned} & \frac{d}{dt} \int_0^L [(1-\varepsilon)t(my^2 + EI|D^2y|^2) + m\eta y Dy] dx \\ &= \{2(1-\varepsilon)t[-D(EID^2y)\dot{y} + EID^2yD\dot{y}] \\ & \quad + \frac{1}{2}m\eta\dot{y}^2 - \eta D(EID^2y)Dy + EID^2y(D\eta Dy + \eta D^2y) - \frac{1}{2}EI|D^2y|^2\}_0^L \\ & \quad + \int_0^L \left\{ \frac{1}{2}[2(1-\varepsilon)m - (Dm\eta + mD\eta)]\dot{y}^2 \right. \\ & \quad \left. + (EI[(1-\varepsilon) - 2D\eta] + \frac{1}{2}D(EI\eta))|D^2y|^2 - EID^2\eta DyD^2y \right\} dx. \end{aligned}$$

Thus if there exist $\varepsilon > 0$, $\delta > 0$ and a twice differentiable function $\eta(x)$ on $[0, L]$ such that

$$\begin{aligned} & \eta(0) = 0, \\ (4.5) \quad & 2(1-\varepsilon)m - (Dm\eta + mD\eta) < -\delta < 0, \\ & EI[(1-\varepsilon) - 2D\eta] + \frac{1}{2}D(EI\eta) < -\delta < 0, \end{aligned}$$

then (4.3) holds. It is easy to see that when $m(x) = \text{constant}$, $E(x)I(x) = \text{constant}$, we can just choose $\eta(x) = 2x$.

Generally, (4.4) is neither necessary nor sufficient for (4.5) to hold.

4.3. Stabilization at an intermediate node. Consider the following situation: two beams are connected at x_1 by a co-located sensor-stabilizer of the rigid support and angle guide type as in Fig. 2.

$$\begin{aligned} & y(x_1^-, t) = y(x_1^+, t), \\ & \frac{\partial}{\partial x} y(x_1^-, t) = \frac{\partial}{\partial x} y(x_1^+, t), \\ (4.6) \quad & E_1 I_1 \frac{\partial^3 y}{\partial x^3}(x_1^-, t) - E_2 I_2 \frac{\partial^3 y}{\partial x^3}(x_1^+, t) = k_0 \frac{\partial y}{\partial t}(x_1, t), \quad k_0 > 0, \\ & - \left[E_1 I_1 \frac{\partial^2 y}{\partial x^2}(x_1^-, t) - E_2 I_2 \frac{\partial^2 y}{\partial x^2}(x_1^+, t) \right] = k_1 \frac{\partial^2 y}{\partial x \partial t}(x_1, t), \quad k_1 > 0. \end{aligned}$$

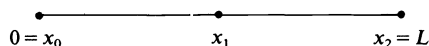


FIG. 2

At the left end x_0 and the right end x_2 , we assume conservative boundary conditions—clamped, hinged, free, etc. Can stabilizing conditions (4.6) lead to a uniform exponential decay result?

This problem seems to have significant application interest as structural engineers often design stabilizers in the middle of the span.

Preliminary results seem to suggest that exponential decay of energy does not hold in general. This problem is still open and requires a careful investigation.

5. Numerical simulation. We consider a single beam with feedback control at the end point.

The chosen beam is characterized by the following figures (MKSA units):

$$(5.1) \quad \begin{aligned} \gamma &= m(\text{mass density}) = 3.58 \text{ kg/m}, \quad L = 44 \text{ m}, \quad EI = 223.20 \text{ kg} \times \text{m}^3/\text{s}^2 \\ &(\text{m} = \text{meter}, \text{s} = \text{sec}, \text{kg} = \text{kilogram}). \end{aligned}$$

It corresponds to beams encountered in the simulation of large flexible space structures. In our model $N = 1$ and the matrix K_1 is chosen of the form

$$(5.2) \quad K_1 = \begin{bmatrix} \alpha & 0 \\ 0 & \beta \end{bmatrix}, \quad \alpha \geq 0, \quad \beta \geq 0.$$

The beam equation was approximated by the finite element method. The spatial domain $[0, L]$ was divided into nine equal intervals. The approximating function was globally C^1 and its restriction to each interval was a cubic polynomial. There are 4 boundary conditions: the two (clamped) conditions at left end are essential conditions and the two (dissipative) conditions at right end are natural conditions. Thus there are a total of $(2 \cdot 9 + 2) - 2 = 18$ elements with $2 \times 18 = 36$ degrees of freedom.

We let the approximate solution be

$$y^h(x, t) = \sum_{i=1}^{18} a_i(t) \psi_i(x),$$

where $\{\psi_i(x) | 1 \leq i \leq 18\}$ is the global finite element basis just mentioned. We use (2.28):

$$\langle m\ddot{y}^h + Ay^h + C^*K_1C\dot{y}^h, \psi_j \rangle_{V' \times V} = 0 \quad \text{for } 1 \leq j \leq 18.$$

This yields a matrix equation

$$(5.3) \quad M_2\ddot{q} + M_1\dot{q} + M_0q = 0,$$

where

$$q = q(t) = \begin{bmatrix} a_1(t) \\ \vdots \\ a_{18}(t) \end{bmatrix},$$

and M_0, M_1, M_2 are 18×18 matrices with entries

$$\begin{aligned} M_0 &= [m_{ij}^0]_{18 \times 18}, & m_{ij}^0 &= \langle A\psi_i, \psi_j \rangle_{V' \times V} = a(\psi_i, \psi_j), \\ M_1 &= [m_{ij}^1]_{18 \times 18}, & m_{ij}^1 &= \langle C^*K_1C\psi_i, \psi_j \rangle_{V' \times V}, \\ M_2 &= [m_{ij}^2]_{18 \times 18}, & m_{ij}^2 &= \gamma \langle \psi_i, \psi_j \rangle_{V' \times V} = \gamma \langle \psi_i, \psi_j \rangle_{L^2}. \end{aligned}$$

All three matrices are symmetric and nonnegative. M_0 and M_2 are strictly positive definite.

To compute eigenfrequencies, let

$$q(t) = e^{\lambda t} q_0, \quad q_0 = \text{a constant 18-vector},$$

and substitute it in (5.3). We get

$$(\lambda^2 M_2 + \lambda M_1 + M_0) q_0 = 0.$$

Thus the following eigenvalue problem is obtained:

$$\left(\begin{bmatrix} 0 & I \\ -M_2^{-1} M_0 & -M_2^{-1} M_1 \end{bmatrix} - \lambda I_{36 \times 36} \right) \begin{bmatrix} q_0 \\ \lambda q_0 \end{bmatrix} = 0.$$

The 36×36 matrix has 36 eigenvalues. We have computed them and provided some graphs here. Due to numerical truncation, the reader should only pay attention to eigenvalues near the real axis. The ones far away from that axis are spurious. In order to make comparisons, the eigenvalues were also computed from the transcendental equation (by separation of variables) using Newton's method (cf. Fig. 7, to be compared with Fig. 6).

Three patterns of spectrum seem to appear, as shown below (Fig. 3):

Case 1. The boundary conditions are

$$\begin{aligned} EI \frac{\partial^3 y}{\partial x^3}(L, t) &= \alpha \frac{\partial y}{\partial t}(L, t), \quad \alpha > 0, \\ -EI \frac{\partial^2 y}{\partial x^2}(L, t) &= 0. \end{aligned}$$

See Fig. 4. As α is increased, the spectrum seems to move uniformly to the left.

Case 2. The boundary conditions are

$$\begin{aligned} EI \frac{\partial^3 y}{\partial x^3}(L, t) &= 0, \\ -EI \frac{\partial^2 y}{\partial x^2}(L, t) &= \beta \frac{\partial^2 y}{\partial x \partial t}(L, t), \quad \beta > 0. \end{aligned}$$

See Fig. 5. As β becomes larger, the spectrum seems to bend toward the negative real axis.

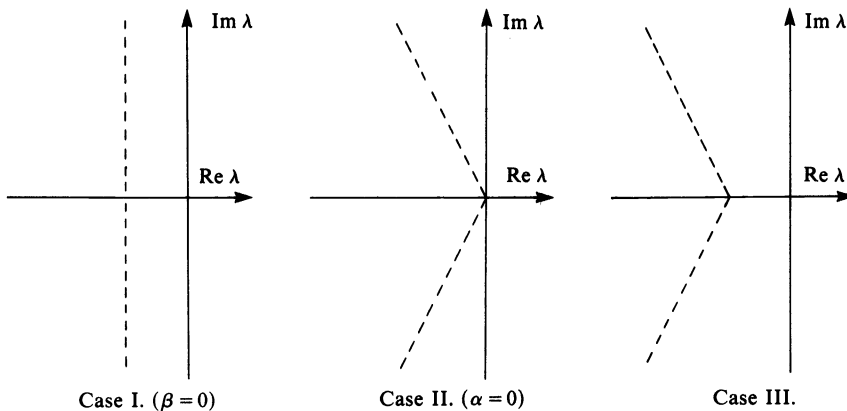
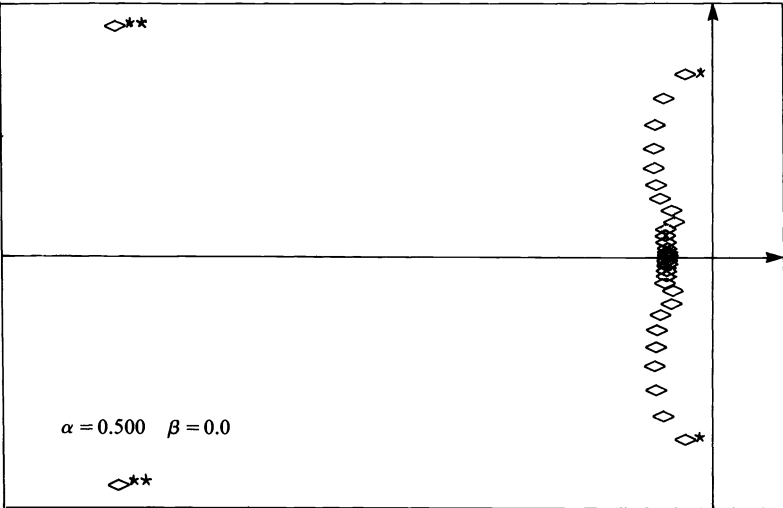


FIG. 3. Locus of the spectrum of the finite element approximation to the spectrum.



Eigenfrequencies					
	Re	Im		Re	Im
1	-.6503D-02	-.1316D-01	19	-.5987D-02	-4.082
2	-.6503D-02	.1316D-01	20	-.5987D-02	4.082
3	-.6295D-02	-.8992D-01	21	-.7481D-02	-5.041
4	-.6295D-02	.8992D-01	22	-.7481D-02	5.041
5	-.6332D-02	-.2540	23	-.7895D-02	-6.235
6	-.6332D-02	.2540	24	-.7895D-02	6.235
7	-.6368D-02	-.4989	25	-.8107D-02	-7.673
8	-.6368D-02	.4989	26	-.8107D-02	7.673
9	-.6427D-02	-.8270	27	-.8178D-02	-9.393
10	-.6427D-02	.8270	28	-.8178D-02	9.393
11	-.6519D-02	-1.241	29	-.8002D-02	-11.42
12	-.6519D-02	1.241	30	-.8002D-02	11.42
13	-.6633D-02	-1.744	31	-.7078D-02	-13.68
14	-.6633D-02	1.744	32	-.7078D-02	13.68
15	-.6668D-02	-2.339	33	-.3944D-02	-15.78
16	-.6668D-02	2.339	34	-.3944D-02	15.78
17	-.5508D-02	-2.994	35	-.8391D-01	-20.00
18	-.5508D-02	2.994	36	-.8391D-01	20.00

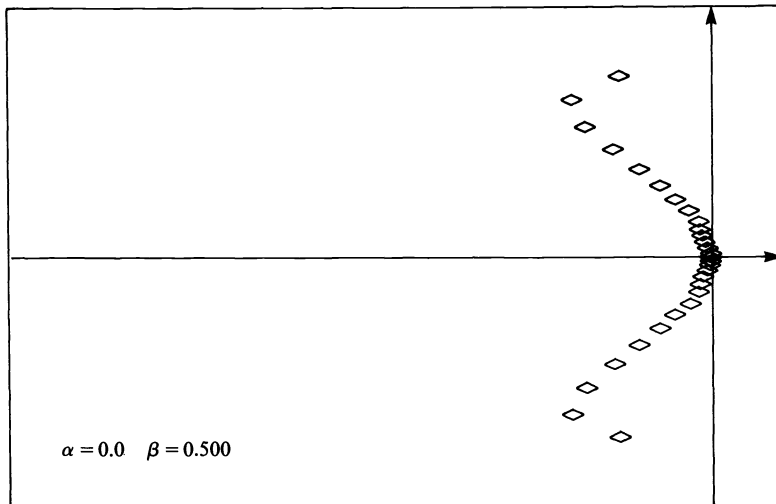
FIG. 4.*: These eigenvalues accompanied with asterisks are spurious in the sense that they distort the real spectral pattern of the damped operator, because by Theorem 3.1, the eigenvalues are uniformly bounded away from the imaginary axis thus they should not bend backward toward the imaginary axis. These are some of the artificial features and difficulties created by finite element discretizations. See also subsequent figures (except Fig. 7). In Figs. 4, 5 and 6, the first $\frac{3}{4}$ of the computed eigenvalues have good accuracy. As a general rule, the accuracy of computed eigenvalues decreases as their moduli increase.

** : These eigenvalues accumulate the largest numerical errors.

Case 3. The boundary conditions are

$$EI \frac{\partial^3 y}{\partial x^3}(L, t) = \alpha \frac{\partial y}{\partial t}(L, t), \quad \alpha > 0,$$
$$-EI \frac{\partial^2 y}{\partial x^2}(L, t) = \beta \frac{\partial^2 y}{\partial x \partial t}(L, t), \quad \beta > 0.$$

See Fig. 6.



Eigenfrequencies

	Re	Im		Re	Im
1	-.6214D-05	-.1450D-01	19	-.3330D-02	-4.083
2	-.6214D-05	.1450D-01	20	-.3330D-02	4.083
3	-.7497D-04	-.9088D-01	21	-.5428D-02	-5.041
4	-.7497D-04	.9088D-01	22	-.5428D-02	5.041
5	-.2023D-03	-.2546	23	-.7626D-02	-6.236
6	-.2023D-03	.2546	24	-.7626D-02	6.236
7	-.3989D-03	-.4993	25	-.1049D-01	-7.673
8	-.3989D-03	.4993	26	-.1049D-01	7.673
9	-.6660D-03	-.8274	27	-.1411D-01	-9.393
10	-.6660D-03	.8274	28	-.1411D-01	9.393
11	-.1013D-02	-1.241	29	-.1808D-01	-11.42
12	-.1013D-02	1.241	30	-.1808D-01	11.42
13	-.1454D-02	-1.744	31	-.2016D-01	-13.68
14	-.1454D-02	1.744	32	-.2016D-01	13.68
15	-.1982D-02	-2.339	33	-.1325D-01	-15.78
16	-.1982D-02	2.339	34	-.1325D-01	15.78
17	-.2138D-02	-2.994	35	-.3576	-19.99
18	-.2138D-02	2.994	36	-.3575	19.99

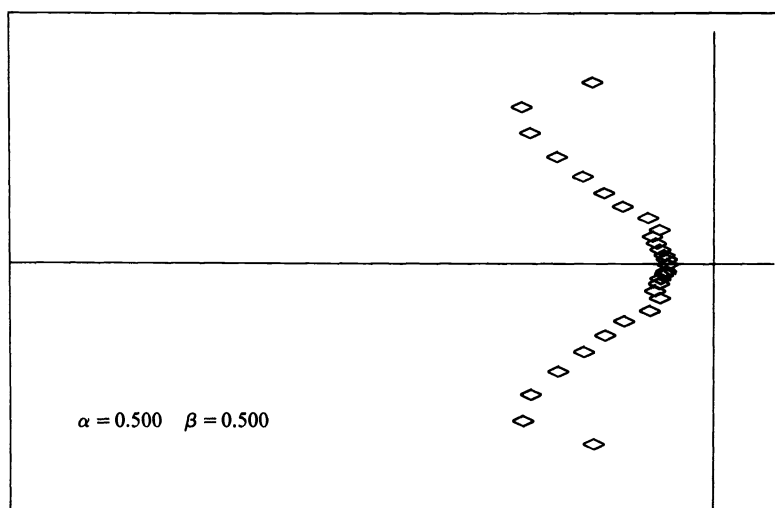
FIG. 5

Note that in Figs. 6, 7 and 8, *structural damping pattern* seems to appear [2]. For computation of spectra of other types of dissipative conditions, see [3].

Note added in proof.

(1) Several researchers have recently studied the asymptotics of eigenfrequencies of a single beam satisfying dissipative boundary conditions as those mentioned in Cases 1-3 in § 5 of this paper:

- [A1] G. CHEN, S. G. KRANTZ, D. W. MA, C. E. WAYNE and H. H. WEST, *The Euler-Bernoulli beam equations with boundary energy dissipation*, in Operator Methods for Optimal Control Problems, S. J. Lee, ed., Lecture Notes in Pure and Applied Mathematics Series, Marcel Dekker, New York, to appear.



Eigenfrequencies

	Re	Im		Re	Im
1	-.6511D-02	-.1316D-01	19	-.9318D-02	-4.083
2	-.6511D-02	.1316D-01	20	-.9318D-02	4.083
3	-.6372D-02	-.8991D-01	21	-.1291D-01	-5.041
4	-.6372D-02	.8991D-01	22	-.1291D-01	5.041
5	-.6536D-02	-.2540	23	-.1553D-01	-6.236
6	-.6536D-02	.2540	24	-.1553D-01	6.236
7	-.6769D-02	-.4989	25	-.1860D-01	-7.673
8	-.6769D-02	.4989	26	-.1860D-01	7.673
9	-.7095D-02	-.8270	27	-.2229D-01	-9.393
10	-.7095D-02	.8270	28	-.2229D-01	9.393
11	-.7535D-02	-1.241	29	-.2609D-01	-11.42
12	-.7535D-02	1.241	30	-.2609D-01	11.42
13	-.8090D-02	-1.744	31	-.2724D-01	-13.68
14	-.8090D-02	1.744	32	-.2724D-01	13.68
15	-.8652D-02	-2.339	33	-.1718D-01	-15.79
16	-.8652D-02	2.339	34	-.1718D-01	15.79
17	-.7648D-02	-2.994	35	-.4413	-19.99
18	-.7648D-02	2.994	36	-.4413	19.99

FIG. 6

[A2] P. RIDEAU, *Contrôle d'un assemblage de poutres flexibles par des capteurs-actionneurs ponctuels: étude du spectre du système*. Thèse, L'Ecole Nationale Supérieure des Mines de Paris, Sophia-Antipolis, France, November, 1985.

[A3] D. L. RUSSELL, *On Mathematical models for the elastic beam with frequency-proportional damping*, to appear.

See these references for details.

(2) The energy multiplier method in this paper does not work for several other types of dissipative boundary conditions. A recent theorem of Professor F. L. Huang of Sichuan University, Chengdu, Sichuan, China enables one to establish uniform exponential decay by estimating the resolvent operator on the imaginary axis. Cf. [A1] as mentioned above.

(3) Mechanical designs and realizations of dissipative boundary conditions studied in this paper have also been given in [A1]. The designs and analysis for several types of dissipative joints will appear in another forthcoming paper.

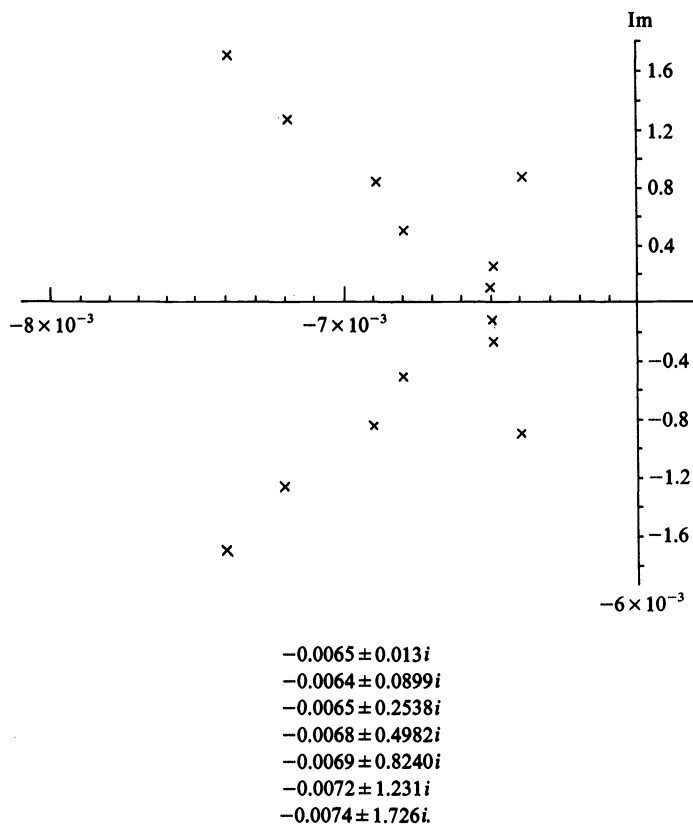


FIG. 7. The above are the first 14 eigenvalues obtained directly from the transcendental equation by Newton's method for the sample example as in Fig. 6, i.e., $\alpha = \beta = 0.5$. The reader can compare them with the first 14 eigenvalues obtained from the finite element method and find that they are very consistent with each other. Note in the graph that except for the six eigenvalues on the right, the spectral pattern appears the same as in Case III, Fig. 3.

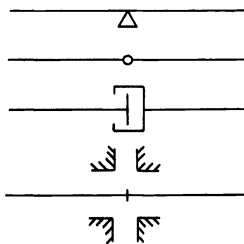


FIG. 8.

REFERENCES

- [1] G. CHEN, *Energy decay estimates and exact boundary value controllability for the wave equation in a bounded domain*, J. Math. Pures Appl., 58 (1979), pp. 249-273.
- [2] G. CHEN AND D. L. RUSSELL, *A mathematical model for linear elastic systems with structural damping*, Quart. Appl. Math., 39 (1981-82), pp. 433-454.
- [3] M. C. DELFOUR, M. P. POLIS AND G. PAYRE, *Sensor and actuator positioning for large flexible space structures*, DOC-CR-SP-83-019, CDT Project P791, Communications Research Centre, Ottawa, Canada, May 1983.

- [4] J. LAGNESE, *Decay of solutions of wave equations in a bounded region with boundary dissipation*, J. Differential Equations, 50 (1983), pp. 163–182.
- [5] J. L. LIONS AND E. MAGENES, *Non-Homogeneous Boundary Value Problems and Applications*, I, Springer-Verlag, New York, 1972.
- [6] J. P. QUINN AND D. L. RUSSELL, *Asymptotic stability and energy decay rates for solutions of hyperbolic equations with boundary damping*, Proc. Roy. Soc. Edinburgh Sect. A, 77 (1977–78), pp. 97–127.
- [7] A. PAZY, *Semigroup of linear operators and application to partial differential equations*, Lecture note #10, Dept. of Math., Univ. of Maryland, College Park, MD, 1974.
- [8] W. D. PILKEY, *Manual for the Response of Structural Members*, Vol. I, IIT Res. Inst. Project J6094, Chicago, IL, 1969.
- [9] D. L. RUSSELL, *Exact boundary value controllability theorems for wave and heat processes in star-complemented regions*, in *Differential Games and Control Theory*, Roxin, Liu, Sternberg, eds., Marcel Dekker, New York, 1974.
- [10] H. H. WEST, *Analysis of Structures*, John Wiley, New York, 1980.

OUTPUT TRACKING FOR NONLINEAR SYSTEMS WITH SINGULAR POINTS*

R. HIRSCHORN† AND J. DAVIS†

Abstract. Some of the well-known linear results on output tracking have been generalized to include nonlinear control systems by removing from the state space a codimension one submanifold of "singular points." These points do not exist in the linear case, but in nonlinear tracking applications the system trajectory can begin at or pass through "singular points." The purpose of this paper is to study nonlinear output tracking with singular points.

Key words. nonlinear systems, output tracking, singularities

AMS(MOS) subject classification. 93B05

1. Introduction. In the output tracking problem one tries to control a system so that its output follows or tracks some desired path. For linear systems this problem was first considered by Brockett and Mesarović [1] in 1965. They used an inverse system to generate the required control. Since then a number of approaches to the linear output tracking problem have been explored (cf. [2]–[4]).

For nonlinear control systems the output tracking problem is more difficult but a number of the well-known linear results have been generalized to the nonlinear case (cf. [5]–[9]). To a large extent this has been possible because the methods used by Silverman [3] in the linear case can be generalized provided one removes from the state space a codimension one submanifold of "singular points" which do not exist in the linear case. When the state of the system approaches a singular point the theory breaks down because the control effort typically becomes unbounded. For rigid manipulators the current theory works well [10], but is inadequate when trying to control flexible manipulators. This is partly due to the existence of "singular points" in the state trajectory for the natural tracking problems. The purpose of this paper is to study output tracking with "singular points." In § 2 singular points in the output tracking problem are introduced. In § 3 the main results, Theorem 3.1 and Theorem 3.2 are proved and the use of these results illustrated.

2. Singular points for output tracking. Consider the single-input affine nonlinear system model

$$(2.1) \quad \begin{aligned} \dot{x}(t) &= f(x(t)) + u(t)g(x(t)), & x(0) &= x_0 \in M, \\ y(t) &= h(x(t)) \end{aligned}$$

where M is a connected C^∞ manifold, f and g are C^∞ (smooth) vector fields on M , h is a C^∞ (smooth) function on M , and the controls $u: [0, \infty) \rightarrow \mathbb{R}$ are continuous functions. For each control u let $x(t, u, x_0)$ denote the corresponding solution to the state differential equation (2.1) and let $y(t, u, x_0)$ denote the corresponding output. The reachable set from x_0 for the system model (2.1) is $\mathcal{R}(x_0) = \{x(t, u, x_0) | t \geq 0, u \text{ admissible}\}$. The system is assumed to have the accessibility property (i.e., $\mathcal{R}(x_0)$ has a nonempty interior in M). If this is not the case, one simply replaces M by the appropriate submanifold (cf. [1], [12]).

The output tracking problem is the identification of functions $y_d(t)$ which can appear as outputs for the system (2.1), i.e., $y_d(\cdot) = y(\cdot, u, x_0)$ for some admissible

* Received by the editors October 7, 1985, and in revised form January 3, 1986.

† Department of Mathematics and Statistics, Queen's University, Kingston, Ontario, Canada K7L 3N6.

control u , and the construction of a control u_d for which $y_d(\cdot) = y(\cdot, u_d, x_0)$. To avoid keeping track of degrees of differentiability, y_d is assumed to be infinity differentiable (i.e., C^∞ or smooth).

The basic idea in both linear and nonlinear tracking is to solve for the control, u_d , as a function of the desired output, y_d , and the state, x , of the system. The following example shows how this approach can break down when "singular points" are present.

Example 2.1. Consider the nonlinear system model with state equation

$$(2.2) \quad \begin{aligned} \dot{x}_1(t) &= u(t), & x_1(0) &= 1, \\ \dot{x}_2(t) &= x_1^2(t), & x_2(0) &= 0, \end{aligned}$$

and with output $y(t) = x_2(t)$. Here $\dot{y}(t) = \dot{x}_2(t) = x_1^2(t)$ and $\ddot{y}(t) = 2x_1(t)u(t)$ so that $u(t)$ can be expressed as a function of y and the states. Also $y(0) = 0$ and $y'(0) = 1$ so that a necessary condition for the output of this system to track a desired smooth path y_d (i.e. $y \equiv y_d$) is that $y_d(0) = 0$, $y'_d(0) = 1$. To see that this is sufficient, suppose that $y_d(0) = 0$, $y'_d(0) = 1$ and set $u_d(t, x) = y''_d(t)/2x_1$, a control using state feedback, and $y''_d(t)$ as a feedforward term. Using this control the output y satisfies $y(0) = 0 = y_d(0)$, $y'(0) = 1 = y'_d(0)$ and $y''(t) = 2x_1(t)u_d(t) = 2x_1(t)y''_d(t)/2x_1(t) = y''_d(t)$ for $t \geq 0$. This implies that $y \equiv y_d$.

In particular for $y_d(t) = t + t^2$ one has $y_d(0) = 0$, $y'_d(0) = 1$ and therefore $u_d(t) = y''_d(t)/2x_1(t) = 1/x_1(t)$ should make $y(t)$ "track" $y_d(t)$. Solving (2.2) with $u \equiv u_d$ yields $x_1(t) = (1 + 2t)^{1/2}$ for all $t \geq 0$ and $y(t) = x_2(t) = t + t^2 = y_d(t)$ for all $t \geq 0$. The resulting trajectory $x(t)$ is shown in Fig. 1.1.

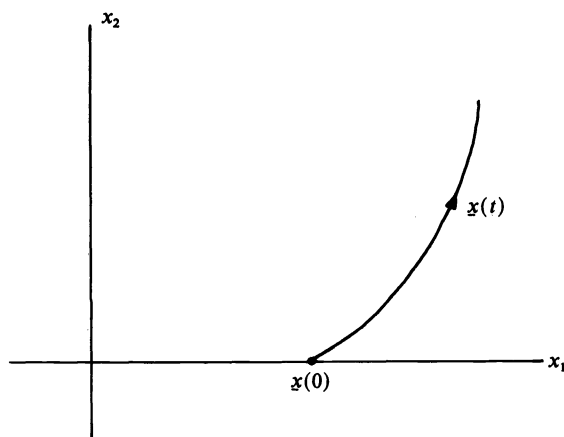


FIG. 1.1

Replacing y_d with $y_d(t) = t - t^2$, one has $y_d(0) = 0$, $y'_d(0) = 1$ as required, and the control $u_d(t) = y''_d(t)/2x_1(t) = -1/x_1(t)$ will cause y to track y_d . In particular, solving (2.1), one sees that $x_1(t) = (1 - 2t)^{1/2}$ for $0 \leq t \leq \frac{1}{2}$ and $y(t) = t - t^2 = y_d(t)$ for $0 \leq t < \frac{1}{2}$. There is a problem, however, when $x_1 = 0$, namely $u_d \rightarrow -\infty$ as $x_1 \rightarrow 0$ (or as $t \rightarrow \frac{1}{2}$). Such points will be called "singular points" for the tracking problem. On first consideration, they seem to create a barrier to the state trajectory (see Fig. 1.2).

On the other hand, if $y_d(t) = t - t^2 + t^3/3$ then $y_d(0) = 0$, $y'_d(0) = 1$, and using $u_d(t) = -1$ for all $t \geq 0$ in (2.1) yields $x_1(t) = 1 - t$ for all $t \geq 0$ and $y = x_2(t) = t - t^2 + t^3/3 = y_d(t)$ for all $t \geq 0$. Here the state trajectory has no problem passing through the set of "singular points" (see Fig. 1.3).

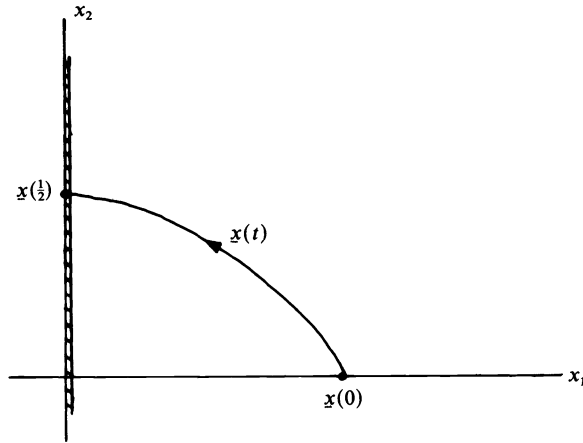


FIG. 1.2

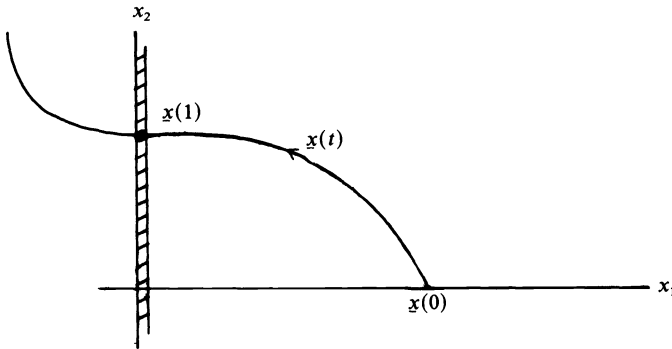


FIG. 1.3

When the state trajectory passes through a singular point, extra restrictions are placed on the class of outputs that the system can track. To gain a better understanding of output tracking in the neighborhood of these "singular points" one should study (2.1), where the initial state is a singular point—for example, $x_0 = (0, 1)$. Before doing this a precise definition for singular points will be given.

DEFINITION 2.1. The *relative order* α of the nonlinear system (2.1) is the least nonnegative integer k such that $gf^{k-1}h \neq 0$ on M or $\alpha = \infty$ if $gf^k h \equiv 0$ for all $k \geq 0$.

If $\alpha = \infty$ the input-output corresponding to system (2.1) is essential trivial; that is, varying the control has little effect on the output. In particular, Theorem 3.1 of [5] asserts that when M, f, g, h are real analytic $\alpha = \infty$ implies that $y(t, u, x_0) = y(t, 0, x_0)$ for all $x_0 \in M$ and for all real analytic controls u .

If $\alpha < \infty$ then $gf^{\alpha-1}h \neq 0$ on M and

$$\begin{aligned}
 y(t) &= h(x(t)) \\
 y^{(1)}(t) &= fh(x(t)) \\
 &\vdots \\
 y^{(\alpha-1)}(t) &= f^{\alpha-1}h(x(t)) \\
 y^{(\alpha)}(t) &= f^\alpha h(x(t)) + u(t)gf^{\alpha-1}h(x(t)).
 \end{aligned}
 \tag{2.3}$$

Thus a necessary condition for the system to track y_d is that $y_d(0) = h(x_0)$, $y_d^{(1)}(0) = fh(x_0)$, \dots , $y_d^{(\alpha-1)}(0) = f^{\alpha-1}h(x_0)$. If $gf^{\alpha-1}h(x_0) \neq 0$ then this is sufficient. One can use

$$u_d(t, x) = \frac{y^{(\alpha)}(t) - f^{\alpha-1}h(x)}{gf^{\alpha-1}h(x)}$$

and show $y^{(\alpha)}(t, u_d, x_0) = y_d^{(\alpha)}(t)$ for t in some *nbhd* of 0, and hence that $y(t, u_d, x_0) = y_d(t)$ in view of the initial condition requirements.

In the event that $gf^{\alpha-1}h(x_0) = 0$, higher derivatives of y_d play a role in determining if y_d can be tracked by the system output.

DEFINITION 2.2. A state x_0 is called a *singular point for output tracking* if $gf^{\alpha-1}h(x_0) = 0$.

Remark 2.1. The input-output behavior of a system with singular initial state x_0 can be quite different from that of a linear system or a nonlinear system with a nonsingular initial state. At a nonsingular initial state the input-output map is injective and so an inverse system exists. In particular, suppose that y is the output to the system (2.1) which results from a control u , and the input to the inverse system is set equal to $y^{(\alpha)}$. Then the output of the inverse system will be u (cf. [5]). Thus if $y_d^{(\alpha)}$, the α th derivative of a desired C^∞ output, is used as an input then the inverse system will generate a unique C^∞ function u_d which can be used for forcing the system (2.1) to track y_d . Thus a desired C^∞ output is always generated by a unique C^∞ input. At a singular initial state x_0 one can have a noninvertible input output map (e.g., Example 2.1 and Remark 3.1) and C^∞ outputs which are generated by inputs which are continuous but not differentiable (e.g., Example 3.5).

To facilitate the computation of higher derivatives of y , set $a(x) = f^\alpha h(x)$ and $b(x) = gf^{\alpha-1}h(x)$, so that $a, b \in C^\infty(M)$, the ring of smooth real valued functions on M . Then

$$y^{(\alpha)}(t) = a(x(t)) + u(t)b(x(t))$$

and

$$y^{(\alpha+1)}(t) = \frac{d}{dt}a(x(t)) + u(t)\frac{d}{dt}b(x(t)) + u^{(1)}(t)b(x(t)).$$

As before,

$$\frac{d}{dt}a(x(t)) = fa(x(t)) + uga(x(t)) = [f + u(t)g]a(x(t))$$

where

$$[f + u(t)g]a(x(t)) = (da)_{x(t)}(f(x(t)) + u(t)g(x(t))),$$

and

$$\begin{aligned} \frac{d^2}{dt^2}a(x(t)) &= [f + u(t)g]^2a(x(t)) + \dot{u}(t)ga(x(t)) \\ &= f^2a(x(t)) + u(t)(gfa(x(t)) + fga(x(t))) + u^2(t)g^2a(x(t)) \\ &\quad + \dot{u}(t)ga(x(t)) \\ &= \left(\frac{\partial}{\partial x} \left(\frac{da}{dt}(x(t)) \right) \right) (f(x(t)) + u(t)g(x(t))) + \left(\frac{\partial}{\partial u} \left(\frac{da}{dt}(x(t)) \right) \right) \dot{u}(t). \end{aligned}$$

This leads to the following definitions: for $r_1, \dots, r_k \in R$ set

$$\begin{aligned}
 a_0(x) &\triangleq a(x) \\
 a_1(x, r_1) &\triangleq \left(\frac{\partial}{\partial x} a_0(x) \right) (f(x) + r_1 g(x)) \\
 (2.4) \quad a_2(x, r_1, r_2) &\triangleq \left(\frac{\partial}{\partial x} a_1(x, r_1) \right) (f(x) + r_1 g(x)) + \left(\frac{\partial}{\partial r_1} a_1(x, r_1) \right) r_2 \\
 &\vdots \\
 a_k(x, r_1, \dots, r_k) &\triangleq \left(\frac{\partial}{\partial x} a_{k-1} \right) (f(x) + r_1 g(x)) + \left(\frac{\partial}{\partial r_1} a_{k-1} \right) r_2 + \dots + \left(\frac{\partial}{\partial r_{k-1}} a_{k-1} \right) r_k.
 \end{aligned}$$

Note that for x fixed $a_k(x, r_1, \dots, r_k)$ is a polynomial in r_1, \dots, r_k of degree $\leq k$. It follows that for $x(t) = x(t, u, x_0)$ one has

$$\begin{aligned}
 (2.5) \quad \frac{d^k}{dt^k} a(x(t)) &= a_k(x(t), u(t), u^{(1)}(t), \dots, u^{(k-1)}(t)), \text{ and} \\
 \frac{d^k}{dt^k} b(x(t)) &= b_k(x(t), u(t), u^{(1)}(t), \dots, u^{(k-1)}(t))
 \end{aligned}$$

where $b_k(x, r_1, \dots, r_k)$ is defined as above with $b_0(x) = b(x)$ (i.e., using the recursion formula (2.4)).

It turns out that the number of further restrictions placed on y_d at a singular point x_0 (i.e. $b(x_0) = 0$) depends on how many derivatives of $b(x(t))$ must be taken before $(d^k/dt^k)b(x(t))|_{t=0} \neq 0$. We formalize this in the following definition.

DEFINITION 2.3. The *degree of singularity of a state* $x_0 \in M$, $\beta(x_0)$, is the least nonnegative integer k such that the polynomial $(r_1, \dots, r_k) \rightarrow b_k(x_0, r_1, \dots, r_k)$ is not the zero polynomial, or $\beta(x_0) = \infty$ if $b_k(x_0, r_1, \dots, r_k) = 0$ for all $r_1, \dots, r_k \in R$ and for all $k \geq 0$.

Remark 2.2. When $k < \beta(x_0)$ it follows that $b_k(x_0, r_1, \dots, r_k) = 0$ for all $r_i \in R$. For $k = \beta(x_0)$ there exists $s_1, \dots, s_{\beta(x_0)}$ such that $b_{\beta(x_0)}(x_0, s_1, \dots, s_{\beta(x_0)}) \neq 0$. This means that any sufficiently differentiable control u such that $u^{(k)}(0) = s_{k+1}$ for $0 \leq k \leq \beta(x_0) - 1$ has the property that $(d^k/dt^k)b(x(t, u, x_0))|_{t=0} = b_k(x_0, s_1, \dots, s_k) = 0$ for $k < \beta(x_0)$ and $(d^{\beta(x_0)}/dt^{\beta(x_0)})b(x(t, u, x_0))|_{t=0} = b_{\beta(x_0)}(x_0, s_1, \dots, s_{\beta(x_0)}) \neq 0$. In particular, if x_0 is a singular point the trajectory $t \rightarrow x(t, u, x_0)$ leaves the set of singular points $\{x \in M | b(x) = 0\}$.

Remark 2.3. If x_0 is *not* a singular point then $b_0(x_0) \neq 0$ and the degree of singularity of x_0 is zero.

Example 2.1 (continued). Consider the system described by (2.2). Here $f(x_1, x_2) = (0, x_1^2)$, $g(x_1, x_2) = (1, 0)$, $h(x_1, x_2) = x_2$, $fh(x) = dh_x f(x) = x_1^2$,

$$gh(x) = dh_x g(x) = \begin{bmatrix} 0 & 1 \end{bmatrix} \begin{bmatrix} 1 \\ 0 \end{bmatrix} \equiv 0,$$

and

$$ghf(x) = (d(fh))_x g(x) = \begin{bmatrix} 2x_1 & 0 \end{bmatrix} \begin{bmatrix} 1 \\ 0 \end{bmatrix} = 2x_1 \neq 0 \quad \text{on } M = \mathbb{R}^2.$$

Thus the relative order of this system is $\alpha = 2$. With the initial condition $x_0 = (1, 0)$ considered previously, $gh(x_0) = 2 \neq 0$ so that the degree of singularity of x_0 is $\beta(x_0) = 0$ and x_0 is not a singular point, and the existing theory on tracking applies. If the system has the initial state $\hat{x}_0 = (0, 1)$ then $gh(\hat{x}_0) = 0$ so that \hat{x}_0 is a singular point. Here

$$\begin{aligned} y^{(\alpha)}(t) &= y^{(2)}(t) = a(x(t)) + u(t)b(x(t)) = f^2h(x) + ugh(x) \\ &= 0 + u(2x_1) \end{aligned}$$

so that $b(x) = 2x_1$. Using (2.3)

$$\begin{aligned} b_0(x) &= b(x) = 2x_1, \\ b_1(x, r_1) &= \left(\frac{\partial}{\partial x} b_0(x) \right) r_1 g(x) = [2 \quad 0] \begin{bmatrix} r_1 \\ 0 \end{bmatrix} = 2r_1, \end{aligned}$$

and at \hat{x}_0 the polynomial $r_1 \rightarrow b_1(\hat{x}_0, r_1) = 2r_1$ is not the zero polynomial while $b_0(\hat{x}_0) = 2(0) = 0$. This means that the degree of singularity of \hat{x}_0 is $\beta(\hat{x}_0) = 1$. Note that at \hat{x}_0 the input output map is not one-to-one as $u(t)$ and $-u(t)$ produce the same output.

For a system with finite relative order (i.e., a nontrivial input-output map) the set of singular points $S = \{x | b(x) = 0\}$ is basically a codimension one submanifold of M . If the system has the accessibility property and if x_0 is a singular point (i.e., $b(x_0) = 0$) then most controls u will steer the system out S . For such a control $b(x(t, u, x_0)) \neq 0$ and one would expect that, for some k , $(d^k/dt^k)b(x(t, u, x_0))|_{t=0} \neq 0$. From Remark 2.2 this implies $\beta(x_0) < \infty$. Also for nonsingular points $\beta(x_0) = 0$. Thus one expects that $\beta(x_0) < \infty$ for all $x \in M$ iff the system is nontrivial (i.e., $\alpha < \infty$). For real analytic systems this is easy to prove.

THEOREM 2.1. *Consider the nonlinear system (2.1) where f , g , h and M are real analytic. Then $\alpha < \infty$ if and only if $\beta(x_0) < \infty$ for all $x_0 \in M$.*

Proof. Suppose that $\beta(x_0) < \infty$ for all $x_0 \in M$. Then $\alpha < \infty$ by definition. Suppose that $\alpha < \infty$ but $\beta(x_0) = \infty$ for some $x_0 \in M$. Then x_0 must be a singular point and thus the real analytic function $x \rightarrow b(x) = gf^{\alpha-1}h(x)$ vanishes at x_0 . Now $b \neq 0$ as $\alpha < \infty$ and by real analyticity the set $\{x \in M | b(x) \neq 0\}$ must be an open dense subset of M . Since (2.1) has the accessibility property the reachable set from x_0 has an open interior in M and thus there exists a real analytic control $u_0(t)$ such that $b(x(t, u_0, x_0)) \neq 0$. If $b^{(k)}(x(t, u_0, x_0))|_{t=0} = 0$ for all $k \geq 0$ (i.e., $\beta(x_0) = \infty$) then the Taylor series for $b(x(t, u_0, x_0))$ is 0 and hence $b(x(t, u_0, x_0)) \equiv 0$, a contradiction. Thus $\beta(x_0) < \infty$ and the proof is complete.

3. Output tracking at a singular point. Suppose that $y(t) = y(x(t, u, x_0))$ is the output of system (2.1) corresponding to the input u . If $\alpha < \infty$ then from (2.3)

$$y^{(k)}(t) = f^k h(x(t)) \quad \text{for } 0 \leq k < \alpha,$$

and a necessary condition for tracking a path $y_d(t)$ is that

$$y_d^{(k)}(0) = f^k h(x_0) \quad \text{for } 0 \leq k < \alpha.$$

It is well known that this is sufficient when x_0 is not a singular point (c.f. [5], [6]) but this is not the case when x_0 is a singular point. For example, the system described by (2.2) with $x_0 = (0, 1)$, a singular point, and $y_d(t) = 1 - t^2$ satisfies the above necessary condition but can never be tracked since $y(t) \geq 0$ for all t and $y^{(2)}(0) = 0$ for all u .

The additional requirements placed on y_d when x_0 is a singular point are easily described. For each $x_0 \in M$ (2.4) defines inductively a sequence of polynomials

$$\{a_k(x_0, r_1, \dots, r_k)\} \quad \text{with } a_0(x_0) = a(x_0) = f^\alpha h(x_0),$$

and

$$\{b_k(x_0, r_1, \dots, r_k)\} \quad \text{with } b_0(x_0) = b(x_0) = gf^{\alpha-1}h(x_0).$$

For $\beta > 0$ consider the mapping

$$F_{x_0}^{\alpha, \beta} : (r_1, \dots, r_\beta) \rightarrow (a_0(x_0), a_1(x_0, r_1), \dots, a_{\beta-1}(x_0, r_1, \dots, r_{\beta-1}), a_\beta(x_0, r_1, \dots, r_\beta) + r_1 b_\beta(x_0, r_1, \dots, r_\beta))$$

of R^β into $R^{\beta+1}$.

THEOREM 3.1. *Consider the nonlinear system (2.1) with $\beta = \beta(x_0) < \infty$. A necessary condition for $y_d \in C^\infty(R)$ to be tracked by the output (using an input which is $(\beta - 1)$ -times differentiable at 0) is that*

$$y_d^{(k)}(0) = f^k h(x_0) \quad \text{for } 0 \leq k < \alpha$$

and $(y_d^{(\alpha)}(0), y_d^{(\alpha+1)}(0), \dots, y_d^{(\alpha+\beta)}(0)) \in \text{Range } F_{x_0}^{\alpha, \beta}$ if $\beta > 0$.

Proof. If $\beta(x_0) = 0$ the proof is immediate. If $\beta(x_0) > 0$ then as in §2

$$y^{(\alpha)}(t) = a(x(t)) + u(t)b(x(t))$$

where

$$a(x) = f^\alpha h(x) \quad \text{and} \quad b(x) = gf^{\alpha-1}h(x),$$

$$\begin{aligned} y^{(\alpha+1)}(t) &= \frac{d}{dt} a(x(t)) + u(t) \frac{d}{dt} b(x(t)) + \left(\frac{d}{dt} u(t) \right) b(x(t)) \\ &= a^{(1)}(x(t)) + u(t)b^{(1)}(x(t)) + u^{(1)}(t)b(x(t)) \\ y^{(\alpha+2)}(t) &= a^{(2)}(x(t)) + u(t)b^{(2)}(x(t)) + 2u^{(1)}(t)b^{(1)}(x(t)) + u^{(2)}(t)b(x(t)) \\ &\vdots \\ y^{(\alpha+k)}(t) &= a^{(k)}(x(t)) + u(t)b^{(k)}(x(t)) + ku^{(1)}(t)b^{(k-1)}(x(t)) + \dots + u^{(k)}(t)b(x(t)) \end{aligned}$$

for $k \geq 0$. Using (2.5) to evaluate $a^{(k)}(x(t))$, $b^{(k)}(x(t))$ and the fact that $b^{(k)}(x(0)) = 0$ for $k < \beta(x_0)$, one sees that

$$\begin{aligned} y^{(\alpha)}(0) &= a_0(x_0) \\ y^{(\alpha+1)}(0) &= a_1(x_0, u(0)) \\ &\vdots \\ y^{(\alpha+\beta(x_0)-1)}(0) &= a_{\beta(x_0)-1}(x_0, u(0), \dots, u^{(\beta(x_0)-2)}(0)) \\ y^{(\alpha+\beta(x_0))}(0) &= a_{\beta(x_0)}(x_0, u(0), \dots, u^{(\beta(x_0)-1)}(0)) \\ &\quad + u(0)b_{\beta(x_0)}(x_0, u(0), \dots, u^{(\beta(x_0)-1)}(0)). \end{aligned}$$

This says that

$$(y_d^{(\alpha)}(0), \dots, y_d^{(\alpha+\beta(x_0))}(0)) \in \text{Range } F_{x_0}^{\alpha, \beta(x_0)}.$$

This completes the proof.

Theorem 3.1 comes close to identifying those functions y_d that can be tracked by the output of a nonlinear system, and by slightly reducing the class of functions defined by Theorem 3.1 a control u_d which generates y_d can be described. In particular, for x_0 fixed, the polynomial

$$(r_1, \dots, r_\beta) \rightarrow b(x_0, r_1, \dots, r_\beta)$$

is nonzero on an open dense subset of R^β , namely $D_{x_0}^\beta = \{(r_1, \dots, r_\beta) \mid b_\beta(x_0, r_1, \dots, r_\beta) \neq 0\}$. Let $F_{x_0}^{\alpha, \beta} \upharpoonright D_{x_0}^\beta$ denote the restriction of $F_{x_0}^{\alpha, \beta}$ to $D_{x_0}^\beta$.

THEOREM 3.2. *Consider the nonlinear system (2.1) with $\beta = \beta(x_0) < \infty$. A sufficient condition for $y_d \in C^\infty(R)$ to be tracked by the output is that*

$$y_d^{(k)}(0) = f^k h(x_0) \quad \text{for } 0 \leq k < \alpha,$$

and

$$(y_d^{(\alpha)}(0), y_d^{(\alpha+1)}(0), \dots, y_d^{(\alpha+\beta)}(0)) \in \text{Range } F_{x_0}^{\alpha, \beta} \upharpoonright_{D_{x_0}^{\beta_0}} \quad \text{for } \beta > 0,$$

where $D_{x_0}^{\beta_0} = \{r \in R^{\beta_0} | b_\beta(x_0, r) \neq 0\} \subseteq R^{\beta_0}$.

Proof. Suppose that $y_d(t)$ is a C^∞ function of t that satisfies the hypothesis of Theorem 3.2. If $\beta(x_0) = 0$ then Theorem 3.2 reduces to Theorem 3.1 of [5]. If $\beta(x_0) > 0$ set $\beta = \beta(x_0)$ and choose $(r_1, \dots, r_\beta) \in D_{x_0}^{\beta_0}$ such that

$$(3.1) \quad (y_d^{(\alpha)}(0), \dots, y_d^{(\alpha+\beta)}(0)) = F_{x_0}^{\alpha, \beta}(r_1, \dots, r_\beta)$$

and $b_\beta(x_0, r_1, \dots, r_\beta) \neq 0$. Let $u_*(t)$ be any smooth control with the property that $u_*^{(k)}(0) = r_{k+1}$ for $0 \leq k \leq \beta - 1$. Set $x_*(t) = x(t, u_*, x_0)$ so that $(d^k/dt^k)b(x_*(t))|_{t=0} = 0$ for $0 \leq k < \beta$ and $(d^\beta/dt^\beta)b(x_*(t))|_{t=0} = b_\beta(x_0, r_1, \dots, r_\beta) \neq 0$. Now choose $\varepsilon > 0$ such that $b(x_*(s)) \neq 0$ for $0 < |s| < \varepsilon$. For each such value of s the system (2.1) with $t_0 = s$, $x_0 = x_*(s)$ and $u(t, x) = (y_d^{(\alpha)}(t) - a(x))/b(x)$ has a smooth solution $x(t) = x(t, u, x_*(s))$. Since $b(x_*(s)) \neq 0$ the initial condition is not a singular point. Thus for $t > s$ the output $y \equiv y_d$, and initially the control is

$$\lim_{t \rightarrow s} u(t, x(t)) = u(s, x(s)) = u(s, x_*(s)) = \frac{y_d^{(\alpha)}(s) - a(x_*(s))}{b(x_*(s))}.$$

The effort here is to show that this method still works when x_0 is singular. By showing that $u(s, x_*(s))$ has a finite limit as s tends to 0 (i.e., as $x_*(s) \rightarrow x_0$), it follows from the continuity of solutions of differential equations with respect to parameters that $u(t, x(t))$ is continuous on $[0, \infty)$ when $\dot{x}(t) = f(x(t)) + u(t, x)g(x(t))$ and $x(0) = x_0$.

To show that $\lim_{s \rightarrow 0^+} u(s, x_*(s))$ has a limit, l'Hôpital's rule is used β times. By definition $\lim_{s \rightarrow 0^+} u(s, x_*(s)) = \lim_{s \rightarrow 0^+} (y_d^{(\alpha)}(s) - a(x_*(s))/b(x_*(s)))$. Here $\lim_{s \rightarrow 0^+} b(x_*(s)) = b(x_0) = b_0(x_0) = 0$ as $\beta > 0$ and from equation (3.1) $y_d^{(\alpha)}(0)$ is the first component of $F_{x_0}^{\alpha, \beta}(r_1, \dots, r_\beta)$. In particular $y_d^{(\alpha)}(0) = a_0(x_0)$ so that $\lim_{s \rightarrow 0^+} y_d^{(\alpha)}(s) - a(x_*(s)) = a_0(x_*(0)) - a(x_*(0)) = a_0(x_0) - a_0(x_0)$. Thus by l'Hôpital's rule

$$\lim_{s \rightarrow 0^+} u(s, x_*(s)) = \lim_{s \rightarrow 0^+} \frac{y_d^{(\alpha+1)}(s) - a^{(1)}(x_*(s))}{b^{(1)}(x_*(s))}$$

(provided the limit on the right exists). The above argument can be repeated. If $\beta > 1$ then $b^{(1)}(x_0) = 0$ and from (3.1) $y_d^{(\alpha+1)}(0) = a_1(x_0, r_1)$ and $a^{(1)}(x_*(0)) = a_1(x_0, u_*(0)) = a_1(x_0, r_1)$ by definition of a_1 . In this way l'Hôpital's rule can be repeated until we attempt to calculate

$$\begin{aligned} \lim_{s \rightarrow 0^+} u(s, x_*(s)) &= \frac{y_d^{(\alpha+\beta)}(0) - a^{(\beta)}(x_*(0))}{b^{(\beta)}(x_*(0))} \\ &= \frac{y_d^{(\alpha+\beta)}(0) - a_\beta(x_0, r_1, \dots, r_\beta)}{b_\beta(x_0, r_1, \dots, r_\beta)}. \end{aligned}$$

From (3.1) $y_d^{(\alpha+\beta)}(0) = a_\beta(x_0, r_1, \dots, r_\beta) + r_1 b_\beta(x_0, r_1, \dots, r_\beta)$ so that $\lim_{s \rightarrow 0^+} u(s, x_*(s)) = r_1 = u_*(0)$. In particular, if one sets

$$u_d(t, x) = \begin{cases} \frac{y_d^{(\alpha)}(t) - a(x)}{b(x)} & \text{for } t > 0, \\ r_1 & \text{for } t = 0, \end{cases}$$

and $x_d(t)$ solves $\dot{x}_d = f(x_d) + u_d g(x_d)$ ($x_d(0) = x_0$), then $u_d(t, x_d(t))$ will be continuous. In fact, the resulting output $y(t)$ has the property that

$$y^{(\alpha)} = a(x_d) + u_d(t, x_d)b(x_d) = a(x_d) + \frac{y_d^{(\alpha)} - a(x_d)}{b(x_d)}b(x_d) = y_d^{(\alpha)},$$

so that $y \equiv y_d$. This completes the proof.

COROLLARY 3.1. *Suppose that $y_d \in C^\infty(R)$ satisfies the hypothesis of Theorem 3.2 so that*

$$(y_d^{(\alpha)}(0), y_d^{(\alpha+1)}(0), \dots, y_d^{(\alpha+\beta)}(0)) = F_{x_0}^{\alpha, \beta}(r_1, r_2, \dots, r_\beta)$$

for (r_1, \dots, r_β) in the open dense subset $D_{x_0}^\beta$ of R^β . Then the output y of system (2.1) can be controlled so that $y \equiv y_d$ by using

$$u_d(t, x) = \begin{cases} r_1 & \text{for } t = 0, \\ \frac{y_d^{(\alpha)}(t) - a(x)}{b(x)} & \text{for } t > 0. \end{cases}$$

Proof. This follows directly from the proof of Theorem 3.2.

COROLLARY 3.2. *Consider the nonlinear system (2.1) with $\beta = \beta(x_0) < \infty$, and suppose that $D_{x_0}^\beta = R^\beta$. Then any $y_d \in C^\infty(R)$ that satisfies the hypothesis of Theorem 3.1 can be tracked.*

Proof. When $D_{x_0}^\beta = R^\beta$ the necessary conditions of Theorem 3.1 and the sufficient conditions of Theorem 3.2 agree.

Remark 3.1. From Theorem 3.2 one can see that the input-output map for the system (2.1) at a singular point will not be 1-1 when $F_{x_0}^{\alpha, \beta} \upharpoonright_{D_{x_0}^\beta}$ fails to be injective.

Example 3.1 (Example 2.1 continued). Consider the system described by (2.2) with $x_0 = (0, 1)$ so that $\alpha = 2$, $\beta = \beta(x_0) = 1$. Here $a(x) \equiv 0$, $b(x) = 2x_1$, $b_0(x) = 2x_1$, $b_1(x, r_1) = 2r_1$ and $a_k \equiv 0$ for all $k \geq 0$. Here

$$F_{x_0}^{\alpha, \beta}(r_1, \dots, r_\beta) = F_{x_0}^{1, 2}(r_1) = (a_0(x_0), a_1(x_0, r_1) + r_1 b_1(x_0, r_1)) = (0, 2r_1^2),$$

$$D_{x_0}^\beta = D_{x_0}^1 = \{r_1 | b_1(x_0, r_1) \neq 0\} = \{r_1 | 2r_1 \neq 0\} = (-\infty, 0) \cup (0, \infty).$$

Theorem 3.1 asserts that a necessary condition to track y_d is that $y_d(0) = 1$, $y_d'(0) = 0$ and $(y_d''(0), y_d'''(0)) \in \text{Range } F_{x_0}^{1, 2} = \{0\} \times R$. Theorem 3.2 asserts that y_d can be tracked if $(y_d''(0), y_d'''(0)) \in \text{Range } F_{x_0}^{1, 2} \upharpoonright_{D_{x_0}^1} = \{0\} \times (R \sim \{0\})$. Thus $y_d(t) = 1 + t^3 z(t)$ can be tracked for all $z \in C^\infty(R)$ with $z(0) \neq 0$, and if y_d can be tracked, then $y_d(t) = 1 + t^3 z(t)$ for $z \in C^\infty(R)$ by Theorem 3.1. Notice that in Example 2.1 in § 2 the trajectory is at a singular point at time $\frac{1}{2}$ when $y_d(t) = t - t^2$ and if $t_0 = \frac{1}{2}$ then $y_d''(t_0) = -2 \notin \text{Range } F_{x_0}^{1, 2}$ (here $x_0 = x(t_0) = x(\frac{1}{2})$). When $y_d(t) = t - t^2 + t^3/3$, then $x(1)$ is singular and for $t_0 = 1$, $y_d''(t_0) = 0$, $y_d'''(t_0) = 2 \neq 0$ as required by Theorem 3.2.

Example 3.2. (This is a nontrivial application of Theorem 3.2). Consider the system (2.1) with $x \in R^3$,

$$f(x) = (x_1 x_2, x_1, x_2 e^{x_1}), \quad g(x) = (x_2^2, 0, 0), \quad h(x) = e^{x_3}.$$

Since the system is to have the accessibility property, let $M = R^3 \sim (\{0, 0\} \times R) = \{(x_1, x_2, x_3) \in R^3 | x_1^2 + x_2^2 \neq 0\}$. By direct computation $gh \equiv 0$, $gfh(x) = x_2^3 e^{x_1 + x_3}$ so that $\alpha = 2$, and $y^{(\alpha)} = y^{(2)} = f^2 h(x) + ugfh(x) = a(x) + ub(x)$ where $a(x) = (x_1 + x_2^2(x_1 + e^{x_1})) e^{x_1 + x_3}$ and $b(x) = x_2^3 e^{x_1 + x_3}$. Thus the set of singular points are those x_0 with $b(x_0) = 0$, i.e., x_2 -coordinate zero.

Let $x_0 = (x_{01}, 0, x_{03})$ be a singular point. To compute $\beta(x_0)$, note that

$$\begin{aligned} b^{(1)}(x(t)) &= \frac{d}{dt}(x_2^3 e^{x_1+x_3}) = 3x_2^2 \dot{x}_2 e^{x_1+x_3} + x_2^3(\dot{x}_1 + \dot{x}_3) e^{x_1+x_3} \\ &= 3x_2^2 x_1 e^{x_1+x_3} + x_2^4(x_1 + x_2 u + e^{x_1}) e^{x_1+x_3} = b_1(x, u) \end{aligned}$$

and thus $b_1(x_0, r_1) = 0$ for all r_1 . Similarly $b_2(x_0, r_1, r_2) = 0$, but $b_3(x_0, r_1, r_2, r_3) = 6x_{01}^3 e^{x_{01}+x_{03}} \neq 0$ since $(0, 0, x_{03}) \notin M$. Thus $\beta(x_0) = 3$ for all singular x_0 . To find the map $F_{x_0}^{\alpha, \beta} = F_{x_0}^{2,3}: R^3 \rightarrow R^4$ one can compute a_0, a_1, a_2 and a_3 at x_0 . Here $a_0(x_0) = x_{01} e^{x_{01}+x_{03}}$, $a_1(x_0, r_1) = 0$, $a_2(x_0, r_1, r_2) = x_{01}^2(1 + 3x_{01} + 3e^{x_{01}}) e^{x_{01}+x_{03}}$, $a_3(x_0, r_1, r_2, r_3) = r_1 b_3(x_0, r_1, r_2, r_3) = 2x_{01}^2 e^{x_{01}+x_{03}}(1 + 4x_{01})r_1$. Thus $F_{x_0}^{2,3}(r_1, r_2, r_3) = (x_{01} e^{x_{01}+x_{03}}, 0, x_{01}^2(1 + 3x_{01} + 3e^{x_{01}}) e^{x_{01}+x_{03}}, 2r_1(1 + 4x_{01})x_{01}^2 e^{x_{01}+x_{03}})$ and since $b_3(x_0, r_1, r_2, r_3) \neq 0$ for all $(r_1, r_2, r_3) \in R^3$, $D_{x_0}^3 = R^3$ and Corollary 3.2 of Theorem 3.2 says that y_d can be tracked by this system if

$$\begin{aligned} y_d(0) &= h(x_0) = e^{x_{03}}, \\ y_d^{(1)}(0) &= fh(x_0) = 0, \end{aligned}$$

and

$$\begin{bmatrix} y_d^{(2)}(0) \\ y_d^{(3)}(0) \\ y_d^{(4)}(0) \\ y_d^{(5)}(0) \end{bmatrix} \in \text{Range } F_{x_0}^{2,3}.$$

If $q_0 = x_{01} e^{x_{01}+x_{03}}$, $q_2 = x_{01}^2(1 + 3x_{01} + 3e^{x_{01}}) e^{x_{01}+x_{03}}$ then this means $y_d^{(2)}(0) = q_0$, $y_d^{(3)}(0) = 0$, $y_d^{(4)}(0) = q_2$ and $y_d^{(5)}(0)$ is free if $1 + 4x_{01} \neq 0$ and is zero if $1 + 4x_{01} = 0$. Thus for $x_0 = (1, 0, -1)$ y_d can be tracked if

$$y_d(0) = e^{-1}, \quad y_d^{(1)}(0) = 0, \quad y_d^{(2)}(0) = 1, \quad y_d^{(3)}(0) = 0, \quad y_d^{(4)}(0) = 4 + 3e.$$

A control function u which works is given by

$$\begin{aligned} u_d(t, x) &= \frac{y_d^{(2)}(t) - a(x)}{b(x)} \\ &= \frac{y_d^{(2)}(t) - (x_1 + x_2^2(x_1 + e^{x_1})) e^{x_1+x_3}}{x_2^3 e^{x_1+x_3}} \quad \text{for } t > 0. \end{aligned}$$

The appropriate limiting initial value is $u_d(0, x_0) = r_1$, where r_1 solves

$$y_d^{(5)}(0) = a_3 + r_1 b_3 = 2r_1(1 + 4x_{01})x_{01}^2 e^{x_{01}+x_{03}} = 10r_1 \quad \text{or} \quad r_1 = \frac{y_d^{(5)}(0)}{10}.$$

Example 3.3. This example illustrates that $y_d \in C^\infty(R)$ does not imply $u_d \in C^\infty(R)$ if x_0 is singular. Consider the system

$$\begin{aligned} \dot{x}_1 &= 1, & a(x) &= x_2, \\ \dot{x}_2 &= -2u, & b(x) &= x_1, \\ \dot{x}_3 &= x_2 + x_1 u, \\ y &= x_3, \end{aligned}$$

where $\alpha = 1$ and $b^{(1)}(x(0)) = \dot{x}_1 = 1$ so that $\beta(x_0) = 1$ for x_0 singular (i.e., $x_0 = (0, x_{02}, x_{03})$). Here $D'_{x_0} = R^2$ so Corollary 3.2 of Theorem 3.2 applies.

If $x_0 = (0, 1, 1)$ then y_d can be tracked if $y_d(0) = 1$, and $(y_d^{(1)}(0), y_d^{(2)}(0)) \in \text{Range } F_{x_0}^{1,1} = \{(1, -r_1) | r_1 \in R\}$ or equivalently, $y_d(0) = 1$ and $y_d^{(1)}(0) = 1$. This means that

$$y_d(t) = 1 + t + \frac{t^3}{3}$$

can be tracked using

$$u_d(t, u) = \begin{cases} r_1 & \text{for } t = t_0 = 0, \\ \frac{y_d'(t) - a(x)}{b(x)} & \text{for } t > 0 \end{cases}$$

where $y_d^{(2)}(0) = 0 = -r_1$. Thus

$$u_d(t, x) = \begin{cases} 0 & \text{for } t = 0, \\ \frac{1 + t^2 - x_2}{x_1} & \text{for } t > 0. \end{cases}$$

In this example the state equations can be solved explicitly to find u as a function of t alone. In fact, solving the state equations with $u = u_d$ yields $x_1(t) = t$, $x_2(t) = -2t^2 \ln t + 1$, $x_3(t) = 1 + t + (t^3/3) = y(t) = y_d(t)$ as required. Here y_d is smooth, and u_d is continuous at $t = 0$ but not differentiable, since

$$u_d(t) = u_d(t, x_d(t)) = \frac{1 + t^2 - x_2(t)}{x_1(t)} = \frac{1 + t^2 + 2t^2 \ln t - 1}{t} = t(1 + 2 \ln t)$$

so

$$\lim_{t \rightarrow 0^+} u_d(t) = 0 = u_d(0, x_0) \quad \text{and} \quad \lim_{t \rightarrow 0^+} u_d^{(1)}(t) = \lim_{t \rightarrow 0^+} (3 + 2 \ln t) = -\infty.$$

REFERENCES

- [1] R. W. BROCKETT AND M. D. MESAROVIC, *The reproducibility of multivariable systems*, J. Math. Anal. Appl., 11 (1965), pp. 548-563.
- [2] M. K. SAIN AND J. L. MASSEY, *Invertibility of linear time-invariant dynamical systems*, IEEE Trans. Automat. Control, AC-14 (1969), pp. 141-149.
- [3] L. M. SILVERMAN, *Inversion of multivariable linear systems*, IEEE Trans. Automat. Control, AC-14 (1969), pp. 270-276.
- [4] A. S. WILLSKY, *On the invertibility of linear systems*, IEEE Trans. Automat. Control, AC-19 (1974), pp. 272-274.
- [5] R. M. HIRSCHORN, *Invertibility of nonlinear control systems*, this Journal, 17 (1979), pp. 289-297.
- [6] ———, *Output tracking in multivariable nonlinear systems*, IEEE Trans. Automat. Control, AC-26 (1981), pp. 593-595.
- [7] S. N. SINGH, *Reproducibility in nonlinear systems using dynamic compensation and output feedback*, IEEE Trans. Automat. Control, AC-27 (1982), pp. 955-958.
- [8] ———, *Generalized functional reproducibility condition for nonlinear systems*, IEEE Trans. Automat. Control, AC-27 (1982), pp. 958-960.
- [9] H. NIJMEIJER, *Invertibility of affine nonlinear control systems: a geometric approach*, Systems Control Lett., 2 (1982), pp. 163-168.
- [10] S. N. SINGH AND A. A. SCKY, *Invertibility and robust nonlinear control of robotic systems*, Proc. of 23rd CDC, Las Vegas, 1984.
- [11] H. SUSSMANN AND V. JURDJEVIC, *Controllability of nonlinear systems*, J. Differential Equations, 12 (1972), pp. 95-116.
- [12] H. SUSSMANN, *Existence and uniqueness of minimal realizations of nonlinear systems*, Math. Systems Theory, 1979.

ASYMPTOTICALLY OPTIMAL ADAPTIVE CONTROL WITH CONSISTENT PARAMETER ESTIMATES*

H. F. CHEN† AND L. GUO†

Abstract. For the discrete-time linear stochastic systems with unknown coefficients we give an adaptive control by which both strong consistency of the parameter estimates and asymptotic optimality for the tracking system are achieved simultaneously. This is done by disturbing the signal that is to be tracked, and the disturbance consists of a sequence of random vectors with covariance matrices tending to zero. The main result is essentially based on some criteria for consistency of parameter estimate for the system without monitoring, which are also demonstrated in the paper. The existence of adaptive control is also discussed.

Key words. stochastic system, parameter estimate, strong consistency, adaptive tracking, asymptotic optimality

AMS(MOS) subject classification. 93C40

1. Introduction. Since Åström and Wittenmark [1] introduced the self-tuning regulator, much work has been devoted in recent years to the parameter-adaptive control and to the related parameter estimation problem. For the adaptive tracking problem, Goodwin, Ramadge and Caines [15] and Sin and Goodwin [21] have established the global convergence of the system and the asymptotic optimality of the tracking error by use of the stochastic gradient and the modified least squares algorithms, respectively. On the other hand, for linear stochastic systems without monitoring there are different conditions guaranteeing the strong consistency of estimates for the unknown system coefficients by invoking various approaches such as the probabilistic method (Ljung [18], Moore [20], Solo [22]), the ordinary differential equation method (Ljung [19], Kushner and Clark [17]) and the combined treatment (Chen [5], [6], [8]). But the crucial point in these different conditions is almost the same fact—the persistent excitation condition, which means that for the matrix $\sum_{i=1}^n \varphi_i \varphi_i^T$ consisting of the stochastic regressors φ_i the ratio of its maximum to minimum eigenvalues is bounded. Unfortunately, it does not always take place for the system with asymptotically optimal adaptive control given in Goodwin, Ramadge and Caines [15] and Sin and Goodwin [21], as shown in Becker, Kumar and Wei [2].

In order to get the consistent estimate for unknown parameters the adaptive control law is disturbed by a random noise introduced artificially (see Caines and Lafortune [3], Chen [7], Chen and Caines [9]). With such a treatment it turns out that the estimate is strongly consistent but the tracking error differs from its minimal value by an additional term caused by the random noise added to the adaptive control law.

However, all these facts do not mean that there is no adaptive control law forcing the long run average of the tracking errors to be minimal and, at the same time, making the parameter estimate strongly consistent, since the asymptotically optimal adaptive control law is not unique.

In this paper, we first give an adaptive control by which both strong consistency of the estimates and optimality for the tracking system are achieved simultaneously. The main idea is that the asymptotically optimal adaptive control is disturbed by a random vector sequence with vanishing covariance matrices, in contrast to the work of Caines and Lafortune [3], Chen and Caines [9] and Chen [7], where the disturbance

* Received by the editors March 27, 1985; accepted for publication (in revised form) March 5, 1986.

† Institute of Systems Science, Academia Sinica, Beijing, People's Republic of China.

is of constant covariance matrix. As a result the matrix $\sum_{i=1}^n \varphi_i \varphi_i^T$ mentioned above is ill-conditioned; hence no persistent excitation-like condition can be applied to guarantee consistency for estimates. However, recently the authors have obtained some new results (Chen and Guo [10], [11], [12]), establishing the strong consistency of parameter estimates for systems with $\sum_{i=1}^n \varphi_i \varphi_i^T$ ill-conditioned, and it appears that they are suitable to the analysis of the case of adaptive control with vanishing disturbances and make the system asymptotically optimal and the parameter estimates strongly consistent.

2. Statement of the problem. Let (Ω, \mathcal{F}, P) be a probability space with a family $\{\mathcal{F}_n\}$ of nondecreasing sub- σ -algebras. Consider the following stochastic control system:

$$(2.1) \quad y_n + A_1 y_{n-1} + \cdots + A_p y_{n-p} = B_1 u_{n-1} + \cdots + B_q u_{n-q} + w_n + C_1 w_{n-1} + \cdots + C_r w_{n-r}$$

where y_n , u_n and w_n are the m -, l - and m -dimensional output, input and driven noise, respectively, and $p \geq 1$, $q \geq 1$, $y_n = w_n = 0$, $u_n = 0$ for $n < 0$. A_i , B_j , C_k ($i = 1 \dots p$, $j = 1 \dots q$, $k = 1 \dots r$) are the unknown matrices.

Assume that u_n and w_n are \mathcal{F}_n -measurable and

$$(2.2) \quad E(w_n | \mathcal{F}_{n-1}) = 0, \quad E(\|w_n\|^2 | \mathcal{F}_{n-1}) \leq c_0 r_{n-1}^\varepsilon$$

with constants $c_0 > 0$, $\varepsilon \in [0, 1)$ and r_{n-1} defined later on by (2.9).

Let z be the shift-back operator and set

$$(2.3) \quad A(z) = I + A_1 z + \cdots + A_p z^p,$$

$$(2.4) \quad B(z) = B_1 + B_2 z + \cdots + B_q z^{q-1},$$

$$(2.5) \quad C(z) = I + C_1 z + \cdots + C_r z^r,$$

$$(2.6) \quad \theta^T = [-A_1 \cdots -A_p B_1 \cdots B_q C_1 \cdots C_r].$$

Denote by θ_n the n th estimate for θ , and let θ_n be given by

$$(2.7) \quad \theta_{n+1} = \theta_n + \frac{\varphi_n}{r_n} (y_{n+1}^T - \varphi_n^T \theta_n)$$

with

$$(2.8) \quad \varphi_n^T = [y_n^T, y_{n-1}^T, \dots, y_{n-p+1}^T, u_n^T \cdots u_{n-q+1}^T, y_n^T - \varphi_{n-1}^T \theta_{n-1}, \dots, y_{n-r+1}^T - \varphi_{n-r}^T \theta_{n-r}],$$

$$(2.9) \quad r_n = 1 + \sum_{i=1}^n \|\varphi_i\|^2, \quad r_0 = 1.$$

The initial values θ_0 and φ_0 are arbitrarily chosen.

Under reasonable conditions Goodwin, Ramadge and Caines [15] proved the global convergence and asymptotical optimality of the tracking system with u_n defined from

$$(2.10) \quad \theta_n^T \varphi_n = y_{n+1}^*,$$

i.e.,

$$(2.11) \quad \overline{\lim}_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n \|u_i\|^2 < \infty, \quad \overline{\lim}_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n \|y_i\|^2 < \infty \quad \text{a.s.}$$

$$(2.12) \quad \lim_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n (y_i - y_i^*)(y_i - y_i^*)^T = R \quad \text{a.s.}$$

However, in this case the estimate θ_n may be inconsistent (Becker, Kumar and Wei [2]). It can be easily explained by the following example. Let $y_n^* \equiv 0$ and $\theta_0^T \theta_0 > \theta^T \theta$. Then we have

$$\theta_n^T \varphi_n \equiv 0, \quad \theta_n^T (\theta_{n+1} - \theta_n) = \theta_n^T \frac{\varphi_n}{r_n} y_{n+1}^* \equiv 0;$$

hence

$$\begin{aligned} \theta_n^T \theta_n &= \theta_{n-1}^T \theta_{n-1} + (\theta_n - \theta_{n-1})^T (\theta_n - \theta_{n-1}) \\ &= \theta_0^T \theta_0 + \sum_{i=1}^n (\theta_i - \theta_{i-1})^T (\theta_i - \theta_{i-1}) \geq \theta_0^T \theta_0 > \theta^T \theta. \end{aligned}$$

In order to achieve strongly consistent parameter estimates, Caines and Lafortune [3], Chen [7] and Chen and Caines [9] added a disturbance with covariance matrix $R_1 > 0$ to the reference sequence $\{y_n^*\}$. In this case θ_n tends to θ but the long run average of the tracking errors differs from its minimum value R by an additional term R_1 :

$$\lim_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n (y_i - y_i^*)(y_i - y_i^*)^T = R + R_1 \quad \text{a.s.}$$

It is natural to ask: Is it possible to achieve simultaneously both asymptotic optimality of the adaptive tracking system and the strong consistency of parameter estimates? To answer this question is the topic of the paper.

3. Main result. We first define adaptive control for the tracking system. Let $\{\varepsilon_i\}$ be an m -dimensional i.i.d. sequence which is independent of $\{w_n\}$ with properties $E\varepsilon_i \varepsilon_i^T = I$, $E\|\varepsilon_i\|^5 < \infty$.

Without loss of generality we assume $\mathcal{F}_n = \sigma\{w_i, i \leq n; \varepsilon_j, j \leq n\}$.

Unlike (2.10) we define adaptive control from the equation

$$(3.1) \quad \theta_n^T \varphi_n = y_{n+1}^* + v_n$$

where $\{y_n^*\}$ is a bounded deterministic reference sequence and

$$(3.2) \quad v_1 = 0, \quad v_n = \frac{\varepsilon_n}{\log^{1/8} n} \quad \forall n \geq 2.$$

(The existence of u_n satisfying (3.1) or (2.10) is discussed in Appendix 1.)

The disturbance v_n in (3.1) is designed to have a vanishing covariance matrix in order to make tracking error asymptotically minimal, but for this the system loses the persistent excitation property which is of crucial importance in the analysis of Caines and Lafortune [3], Chen [7] and Chen and Caines [9]. To overcome this difficulty is the main task of the present paper.

We need the following conditions:

(A₁) $C(z) - \frac{1}{2}I$ is strictly positive real;

(A₂) B_1 if of full rank and zeros of $\det B_1^+ B(z)$ lie outside the closed unit disk;

(A₃) $B_1^+ A(z)$ and $B_1^+ B(z)$ are left-coprime and $B_1^+ B_q$ is of full rank;

(A₄) $\{w_i\}$ is a mutually independent sequence with $Ew_i = 0$; $\sup_i E\|w_i\|^{4+\delta} < \infty$ for some $\delta > 0$ and

$$(3.3) \quad \lim_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n w_i w_i^T = R > 0 \quad \text{a.s.}$$

THEOREM 1. For system (2.1)–(2.2) let the parameter estimate be given by (2.7)–(2.9) and let the control be defined by (3.1) and (3.2). If conditions A_1 – A_4 are fulfilled, then the tracking system is asymptotically optimal and the estimate is strongly consistent, i.e., (2.11) and (2.12) take place and $\theta_n \xrightarrow{n \rightarrow \infty} \theta$ a.s.

The proof of this theorem is given in § 5. For this we need some criteria for strong consistency of parameter estimate for systems without monitoring.

4. Parameter estimation for systems without monitoring. In contrast to φ_n we define an estimate-free vector φ_n^0 :

$$(4.1) \quad \varphi_n^0 = [y_n^\tau \cdots y_{n-p+1}^\tau, u_n^\tau \cdots u_{n-q+1}^\tau, w_n^\tau \cdots w_{n-r+1}^\tau]^\tau$$

and set

$$(4.2) \quad \xi_n = y_n - w_n - \theta_{n-1}^\tau \varphi_{n-1},$$

$$(4.3) \quad \varphi_n^\xi = [0 \cdots 0, 0 \cdots 0, \xi_n^\tau \cdots \xi_{n-r+1}^\tau]^\tau.$$

Then we have

$$(4.4) \quad \varphi_n = \varphi_n^0 + \varphi_n^\xi,$$

$$(4.5) \quad y_{n+1} = \theta^\tau \varphi_n^0 + w_{n+1},$$

and

$$\begin{aligned} \theta_{n+1} &= \theta_n + \frac{\varphi_n}{r_n} (\varphi_n^{0\tau} \theta + w_{n+1}^\tau - \varphi_n^\tau \theta_n) \\ &= \theta_n + \frac{\varphi_n}{r_n} (\varphi_n^\tau \theta - \varphi_n^{\xi\tau} \theta + w_{n+1}^\tau - \varphi_n^\tau \theta_n); \end{aligned}$$

hence

$$(4.6) \quad \tilde{\theta}_{n+1} = \left(I - \frac{\varphi_n \varphi_n^\tau}{r_n} \right) \tilde{\theta}_n + \frac{\varphi_n \varphi_n^{\xi\tau}}{r_n} \theta - \frac{\varphi_n}{r_n} w_{n+1}^\tau$$

with

$$(4.7) \quad \tilde{\theta}_n = \theta - \theta_n.$$

Let the matrix $\Phi(n, i)$ be recursively defined by

$$(4.8) \quad \Phi(n+1, i) = \left(I - \frac{\varphi_i \varphi_i^\tau}{r_i} \right) \Phi(n, i), \quad \Phi(i, i) = I.$$

Then from (4.6) it follows that

$$(4.9) \quad \tilde{\theta}_{n+1} = \Phi(n+1, 0) \tilde{\theta}_0 + \sum_{j=0}^n \Phi(n+1, j+1) \frac{\varphi_j \varphi_j^{\xi\tau}}{r_j} \theta - \sum_{j=0}^n \Phi(n+1, j+1) \frac{\varphi_j}{r_j} w_{j+1}^\tau,$$

from which we see that the behavior of $\Phi(n, 0)$ is of great importance for consistency of parameter estimates.

LEMMA 1. For the system and algorithm defined by (2.1)–(2.2) and (2.7)–(2.9) if condition A_1 holds, then

$$(4.10) \quad \sum_{n=0}^{\infty} \frac{\|\xi_{n+1}\|^2}{r_n} < \infty \quad \text{a.s.};$$

moreover, if conditions A_2 and A_4 hold and (2.10) or (3.1) is satisfied, then $r_n \rightarrow \infty$, and

$$(4.11) \quad \frac{1}{n} \sum_{i=0}^n \|\xi_{i+1}\|^2 \xrightarrow{n \rightarrow \infty} 0.$$

Proof. In Chen and Caines [9] and Chen [7], (4.10) and (4.11) are proved for v_n with the constant covariance matrix, but they can be verified by the same argument used there. \square

LEMMA 2. For the system and algorithm defined by (2.1)–(2.2) and (2.7)–(2.9) if condition A_1 holds then $\Phi(n, 0) \xrightarrow{n \rightarrow \infty} 0$ implies $\theta_n \xrightarrow{n \rightarrow \infty} \theta$ a.s. for any initial value θ_0 . For the special case of $r = 0$, the converse assertion is also true, i.e., if $\theta_n \xrightarrow{n \rightarrow \infty} \theta$ a.s. for any θ_0 then $\Phi(n, 0) \xrightarrow{n \rightarrow \infty} 0$ a.s.

Proof. The first step is to show

$$(4.12) \quad \sum_{i=0}^{n-1} \frac{\|\Phi(n, i+1)\varphi_i\|^2}{r_i} \leq d$$

for any vector sequence $\{\varphi_n\}$ with $\Phi(n, i)$ and r_n related by (2.9) and (4.8), where d is the dimension of φ_n .

Then by (2.2) and (4.12) we can prove that the last term of (4.9) goes to zero if $\Phi(n, 0) \xrightarrow{n \rightarrow \infty} 0$. Finally, by (4.10) and (4.12) the second term on the right-hand side of (4.9) also converges to zero if $\Phi(n, 0) \xrightarrow{n \rightarrow \infty} 0$. This is just a sketch proof for the first conclusion. For detailed proof we refer to Chen and Guo [12]. The second conclusion can be easily seen from (4.9). \square

LEMMA 3. If $r_n \xrightarrow{n \rightarrow \infty} \infty$, $\overline{\lim}_{n \rightarrow \infty} r_n / r_{n-1} < \infty$ and there exist quantities N_0 and M possibly depending on ω such that

$$(4.13) \quad \frac{\lambda_{\max}^n}{\lambda_{\min}^n} \leq M(\log r_n)^{1/4} \quad \text{a.s.} \quad \forall n \geq N_0;$$

then $\Phi(n, 0) \xrightarrow{n \rightarrow \infty} 0$, where λ_{\max}^n and λ_{\min}^n denote the maximum and minimum eigenvalue of the matrix $\sum_{i=1}^n \varphi_i \varphi_i^T + (1/d)I$ respectively and d denotes the dimension of φ_n .

Proof. We only give a sketch of the proof and refer readers interested in details to Chen and Guo [10], [11].

The key point is to find a function $m(t)$ such that $m(t) \xrightarrow{t \rightarrow \infty} \infty$ and

$$\|\Phi(m(N + k\alpha), m(N + (k-1)\alpha))\| \leq \sqrt{1 - \frac{\beta^2}{c_1 k}} \quad \forall k \geq 1$$

for some $N, \alpha > 0, \beta > 0$ and $c_1 > 0$. If it has been done, then

$$\begin{aligned} \|\Phi(m(N + k\alpha), 0)\| &\leq \prod_{i=1}^k \|\Phi(m(N + i\alpha), m(N + (i-1)\alpha))\| \cdot \|\Phi(m(N), 0)\| \\ &\leq \left[\prod_{i=1}^k \left(1 - \frac{\beta^2}{c_1 i} \right) \right]^{1/2} \xrightarrow{k \rightarrow \infty} 0; \end{aligned}$$

hence $\Phi(n, 0) \xrightarrow{n \rightarrow \infty} 0$ since $\|\Phi(j+1, j)\| \leq 1$ for all j .

It appears that the following defined function can serve as the desired one:

$$m(t) = \max [n: t_n \leq t],$$

$$t_n \triangleq \sum_{i=2}^{n-1} \frac{\|\varphi_i\|^2}{r_i (\log r_i)^{1/4}}. \quad \square$$

Remark 1. Lemma 3 is a purely algebraic result, namely, it is true for any vector sequence $\{\varphi_n\}$, only if $\Phi(n, 0)$, φ_n and r_n are related by (2.9) and (4.8).

Remark 2. There exists an example (see Chen and Guo [13]) showing that Lemma 3 is no longer true if condition (4.13) is replaced by a more general one:

$$\lambda_{\max}^n / \lambda_{\min}^n \leq M(\log r_n)^{1+a}, \quad a > 0.$$

This means that, in order for the estimate given by (2.7) to be consistent, the condition number of $\sum_{i=1}^n \varphi_i \varphi_i^\tau + (1/d)I$ is allowed to diverge at a rate of $(\log r_n)^{1/4}$, but not faster than $(\log r_n)^{1+a}$.

LEMMA 4. Let $\{\varphi_n^1\}$, $\{\varphi_n^2\}$ and $\{\psi_n\}$ be the vector sequence satisfying conditions $\varphi_n^1 = \varphi_n^2 + \psi_n$ and

$$(4.14) \quad \sum_{n=0}^{\infty} \frac{\|\psi_n\|^2}{r_{1n}} < \infty.$$

Then $\Phi_1(n, 0) \xrightarrow{n \rightarrow \infty} 0$ if and only if $\Phi_2(n, 0) \xrightarrow{n \rightarrow \infty} 0$, where by definition

$$\Phi_i(n+1, 0) = \left(I - \frac{\varphi_n^i \varphi_n^{i\tau}}{r_{in}} \right) \Phi_i(n, 0), \quad \Phi_i(0, 0) = I,$$

$$r_{in} = 1 + \sum_{j=1}^n \|\varphi_j^i\|^2, \quad r_{i0} = 1, \quad i = 1, 2.$$

Proof. Without loss of generality we assume that $\|\varphi_0^1\| \neq 1$.

Suppose $\Phi_1(n, 0) \xrightarrow{n \rightarrow \infty} 0$; then from the following chain of equalities:

$$\begin{aligned} \det \Phi_1(n+1, 0) &= \det \prod_{i=0}^n \Phi_1(i+1, i) = \prod_{i=0}^n \det \left(I - \frac{\varphi_i^1 \varphi_i^{1\tau}}{r_{1i}} \right) \\ &= \prod_{i=1}^n \frac{r_{1i-1}}{r_{1i}} (1 - \|\varphi_0^1\|^2) = \frac{1}{r_{1n}} (1 - \|\varphi_0^1\|^2) \end{aligned}$$

we see that $r_{1n} \xrightarrow{n \rightarrow \infty} \infty$.

By (4.14) and the Kronecker lemma we have

$$(4.15) \quad \frac{r_{2n}}{r_{1n}} = \frac{r_{1n} - 2 \sum_{i=1}^n \varphi_i^{1\tau} \psi_i + \sum_{i=1}^n \|\psi_i\|^2}{r_{1n}} \xrightarrow{n \rightarrow \infty} 1$$

and by (4.14)

$$(4.16) \quad \sum_{n=0}^{\infty} \frac{\|\psi_n\|^2}{r_{2n}} < \infty.$$

We immediately verify that

$$\begin{aligned} \Phi_2(n+1, 0) &= \Phi_1(n+1, 0) + \sum_{j=0}^n \Phi_1(n+1, j+1) \frac{\varphi_j^1 \psi_j^\tau}{r_{1j}} \Phi_2(j, 0) \\ &\quad + \sum_{j=0}^n \Phi_1(n+1, j+1) \frac{\psi_j \varphi_j^{2\tau}}{r_{2j}} \Phi_2(j, 0) \\ &\quad + \sum_{j=0}^n \frac{\Phi_1(n+1, j+1) \varphi_j^1}{r_{1j}^{1/2}} \left(\sqrt{\frac{r_{2j}}{r_{1j}}} - \sqrt{\frac{r_{1j}}{r_{2j}}} \right) \frac{\varphi_j^{2\tau} \Phi_2(j, 0)}{r_{2j}^{1/2}}. \end{aligned}$$

By using (4.12) and (4.14)–(4.16) it is not difficult to conclude that $\Phi_1(n, 0) \xrightarrow{n \rightarrow \infty} 0$ implies $\Phi_2(n, 0) \xrightarrow{n \rightarrow \infty} 0$. The converse implication is proved in a similar way. \square

THEOREM 2. For the system and algorithm defined by (2.1)–(2.2) and (2.7)–(2.9) if condition A_1 holds and if $r_n \xrightarrow{n \rightarrow \infty} \infty$,

$$\overline{\lim}_{n \rightarrow \infty} r_n / r_{n-1} < \infty \text{ and } \lambda_{\max}^n / \lambda_{\min}^n \leq M(\log r_n)^{1/4} \quad \forall n \geq N$$

(or $r_n^0 \xrightarrow{n \rightarrow \infty} \infty$, $\overline{\lim}_{n \rightarrow \infty} r_n^0 / r_{n-1}^0 < \infty$ and $\lambda_{\max}^{0n} / \lambda_{\min}^{0n} \leq M(\log r_n^0)^{1/4}$ for all $n \geq N$) with N and M possibly depending on ω , then

$$\theta_n \xrightarrow{n \rightarrow \infty} \theta \quad \text{a.s.}$$

for any initial value, where λ_{\max}^{0n} , λ_{\min}^{0n} denote the maximum and minimum eigenvalue of $\sum_{i=1}^n \varphi_i^0 \varphi_i^{0\tau} + (1/d)I$, respectively, and $r_n^0 = 1 + \sum_{i=1}^n \|\varphi_i^0\|^2$ with φ_n^0 defined by (4.1).

Proof. Since (4.4), (4.10) and Lemma 4 can be applied with $\varphi_n^1 = \varphi_n$, $\varphi_n^2 = \varphi_n^0$ and $\psi_n = \varphi_n^\varepsilon$, hence $\Phi(n, 0) \xrightarrow{n \rightarrow \infty} 0$ if and only if $\Phi_0(n, 0) \xrightarrow{n \rightarrow \infty} 0$, where $\Phi_0(n, 0)$ is defined by (4.8) with φ_n and r_n replaced by φ_n^0 and r_n^0 , respectively. Then the conclusions of the theorem immediately follow from Lemmas 2 and 3. \square

5. Proof of Theorem 1. To begin with we prove the following lemmas.

LEMMA 5. Let $\{v_n\}$ be defined by (3.2) and let $H_N(z) = \sum_{i=0}^{\infty} H_i(N)z^i$ be the matrix series in shift-back operator z , where the matrix coefficients $H_i(N)$ may depend upon ω , but there are constants (independent of ω) $k_1 > 0$ and $k_2 > 0$ such that

$$\|H_i(N)\| \leq k_1 \exp(-k_2 i) \quad \forall i \quad \forall N \geq 0.$$

Then

$$\begin{aligned} \lim_{N \rightarrow \infty} \frac{\log^{1/4} N}{N} \sum_{n=1}^N (H_N(z)v_n)(H_N(z)v_n)^\tau \\ = \lim_{N \rightarrow \infty} \frac{\log^{1/4} N}{N} \sum_{i=0}^N H_i(N)H_i^\tau(N) \left(\sum_{j=2}^N \frac{1}{\log^{1/4} j} \right). \end{aligned}$$

Proof. See Appendix 2.

LEMMA 6. Let condition A_4 except (3.3) be held and let $H(z) = \sum_{i=0}^{\infty} H_i z^i$ and $G(z) = \sum_{i=0}^{\infty} G_i z^i$ be matrix series in shift-back operator z with $\|H_i\| + \|G_i\| \leq k_1 \exp(-k_2 i)$ for all $i \geq 0$, for some constants $k_1 > 0$, $k_2 > 0$. Then there exists $\gamma \in (0, 1)$ such that for all $l \geq 0$, $m \geq 0$,

$$(5.1) \quad \lim_{N \rightarrow \infty} \frac{1}{N^\gamma} \sum_{n=1}^N (H(z)w_{n+1-l})(G(z)v_{n-m})^\tau = 0 \quad \text{a.s.},$$

$$(5.2) \quad \lim_{N \rightarrow \infty} \frac{1}{N^\gamma} \sum_{n=1}^N (H(z)w_{n+1-l})\eta_n^\tau = 0 \quad \text{a.s.}$$

for any bounded deterministic sequence $\{\eta_n\}$, and

$$(5.3) \quad \overline{\lim}_{N \rightarrow \infty} \frac{1}{N} \sum_{n=1}^N \|H(z)w_{n+1-l}\|^2 < \infty \quad \text{a.s.}$$

Proof. See Appendix 2.

Set

$$(5.4) \quad H_1(z) = [B_1^+ B(z)]^{-1} B_1^+ A(z),$$

$$(5.5) \quad H_2(z) = H_1(z) - [B_1^+ B(z)]^{-1} B_1^+ C(z),$$

$$(5.6) \quad Y_n^* = [y_n^{*\tau} \cdots y_{n-p+1}^{*\tau}, (H_1(z)y_{n+1}^*)^\tau \cdots (H_1(z)y_{n-q+2}^*)^\tau]^\tau$$

and

$$(5.7) \quad Z_n = [v_{n-1}^\tau \cdots v_{n-p}^\tau, (H_1(z)v_n)^\tau \cdots (H_1(z)v_{n-q+1})^\tau]^\tau.$$

In the following, by $\lambda_{\min}(X)$ ($\lambda_{\max}(X)$) we mean the minimum (maximum) eigenvalue of the matrix X ; we have the following.

LEMMA 7. *Under conditions of Theorem 1 if*

$$(5.8) \quad \lim_{N \rightarrow \infty} \lambda_{\min} \left(\frac{\log^{1/4} N}{N} \sum_{n=1}^N (Y_n^* Y_n^{*\tau} + Z_n Z_n^\tau) \right) \neq 0 \quad \text{a.s.}$$

then

$$\theta_n \xrightarrow[n \rightarrow \infty]{} \theta \quad \text{a.s.}$$

Proof. By Theorem 2 we only need to prove $\Phi_0(n, 0) \xrightarrow[n \rightarrow \infty]{} 0$ a.s. From (2.1) we have

$$(5.9) \quad u_n = [B_1^+ B(z)]^{-1} B_1^+ A(z) y_{n+1} - [B_1^+ B(z)]^{-1} B_1^+ C(z) w_{n+1};$$

then by (3.1), (4.2), (5.4), (5.5) and (5.9), φ_n^0 defined by (4.1) can be written as

$$(5.10) \quad \varphi_n^0 = \varphi_n^1 + \psi_n$$

where

$$(5.11) \quad \psi_n = [\xi_n^\tau \cdots \xi_{n-p+1}^\tau, (H_1(z)\xi_{n+1})^\tau \cdots (H_1(z)\xi_{n-q+2})^\tau 0 \cdots 0]^\tau,$$

$$(5.12) \quad \varphi_n^1 = \varphi_n^2 + \varphi_n^3,$$

$$(5.13) \quad \varphi_n^2 = [w_n^\tau \cdots w_{n-p+1}^\tau, (H_2(z)w_{n+1})^\tau \cdots (H_2(z)w_{n-q+2})^\tau, w_n^\tau \cdots w_{n-r+1}^\tau]^\tau,$$

$$(5.14) \quad \varphi_n^3 = [Y_n^{*\tau} + Z_n^\tau, 0 \cdots 0]^\tau.$$

By (4.4), (4.10), similar to (4.15) we have

$$(5.15) \quad \frac{r_n^0}{r_n} = \frac{r_n - 2 \sum_{i=1}^n \varphi_i^\tau \varphi_i^\xi + \sum_{i=1}^n \|\varphi_i^\xi\|^2}{r_n} \xrightarrow[n \rightarrow \infty]{} 1.$$

Then by the Schwarz inequality it follows that

$$\begin{aligned} \sum_{n=0}^{\infty} \frac{\|H_1(z)\xi_{n+1-l}\|^2}{r_n^0} &\leq \sum_{n=0}^{\infty} \frac{1}{r_n^0} \sum_{i=0}^{\infty} \|H_{1i}\| \sum_{i=0}^{\infty} \|H_{1i}\| \cdot \|\xi_{n+1-l-i}\|^2 \\ &\leq k_0 \sum_{i=0}^{\infty} \|H_{1i}\| \sum_{n=0}^{\infty} \frac{\|\xi_{n+1-l-i}\|^2}{r_n^0} < \infty \end{aligned}$$

where the last inequality is obtained because $\xi_i = 0$ for $i < 0$ and the coefficients in $H_1(z) = \sum_{i=0}^{\infty} H_{1i} z^i$ have the estimates $\|H_{1i}\| \leq k_1 \exp(-k_2 i)$, for all $i \geq 0$, ($k_1 > 0$, $k_2 > 0$) by condition A_2 . Thus we have established

$$(5.16) \quad \sum_{n=0}^{\infty} \frac{\|\psi_n\|^2}{r_n^0} < \infty$$

and by Lemma 4 we conclude that $\Phi_0(n, 0) \xrightarrow[n \rightarrow \infty]{} 0$ iff $\Phi_1(n, 0) \xrightarrow[n \rightarrow \infty]{} 0$.

Next, we prove that

$$(5.17) \quad \lim_{N \rightarrow \infty} \lambda_{\min} \left(\frac{\log^{1/4} N}{N} \sum_{n=1}^N \varphi_n^1 \varphi_n^{1\tau} \right) \neq 0 \quad \text{a.s.}$$

If for some $\omega \in \Omega$ (5.17) were not true, then we would find a subsequence of eigenvectors $\begin{bmatrix} \alpha_{N_k} \\ \beta_{N_k} \end{bmatrix}$ for matrix $(\log^{1/4} N_k / N_k) \sum_{n=1}^{N_k} \varphi_n^1 \varphi_n^{1\tau}$ with $N_k \xrightarrow[k \rightarrow \infty]{} \infty$, $\alpha_{N_k} \in R^{mp+lq}$, $\beta_{N_k} \in R^{mr}$ and

$$(5.18) \quad \|\alpha_{N_k}\|^2 + \|\beta_{N_k}\|^2 = 1$$

such that

$$(5.19) \quad (\alpha_{N_k}^\tau, \beta_{N_k}^\tau) \frac{\log^{1/4} N_k}{N_k} \sum_{n=1}^{N_k} \varphi_n^1 \varphi_n^{1\tau} \begin{pmatrix} \alpha_{N_k} \\ \beta_{N_k} \end{pmatrix} \xrightarrow[k \rightarrow \infty]{} 0.$$

Without loss of generality we always assume that this fixed ω does not belong to a possible exceptional set of probability zero. Obviously, α_{N_k} and β_{N_k} would be ω -dependent but not necessarily measurable.

Utilizing Lemma 6 one can easily be convinced of the fact

$$(5.20) \quad \frac{\log^{1/4} N}{N} \sum_{n=1}^N \varphi_n^2 \varphi_n^{3\tau} \xrightarrow[N \rightarrow \infty]{} 0,$$

then (5.19) is reduced to

$$(5.21) \quad (\alpha_{N_k}^\tau, \beta_{N_k}^\tau) \frac{\log^{1/4} N_k}{N_k} \sum_{n=1}^{N_k} \varphi_n^2 \varphi_n^{2\tau} \begin{pmatrix} \alpha_{N_k} \\ \beta_{N_k} \end{pmatrix} \xrightarrow[k \rightarrow \infty]{} 0,$$

$$(5.22) \quad (\alpha_{N_k}^\tau, \beta_{N_k}^\tau) \frac{\log^{1/4} N_k}{N_k} \sum_{n=1}^{N_k} \varphi_n^3 \varphi_n^{3\tau} \begin{pmatrix} \alpha_{N_k} \\ \beta_{N_k} \end{pmatrix} \xrightarrow[k \rightarrow \infty]{} 0.$$

In view of Lemma 6, (5.8) implies

$$(5.23) \quad \lim_{N \rightarrow \infty} \lambda_{\min} \left(\frac{\log^{1/4} N}{N} \sum_{n=1}^N (Y_n^* + Z_n)(Y_n^* + Z_n)^\tau \right) \neq 0.$$

Paying attention to the fact that the last mr elements in φ_n^3 are zeros, by (5.22) and (5.23) we conclude that

$$(5.24) \quad \alpha_{N_k} \xrightarrow[k \rightarrow \infty]{} 0;$$

hence, recalling (5.18) we have

$$(5.25) \quad \|\beta_{N_k}\| \xrightarrow[k \rightarrow \infty]{} 1.$$

Let

$$x_n^1 = [w_n^\tau \cdots w_{n-p+1}^\tau, (H_2(z)w_{n+1})^\tau \cdots (H_2(z)w_{n-q+2})^\tau]^\tau,$$

$$x_n^2 = [w_n^\tau \cdots w_{n-r+1}^\tau]^\tau.$$

Then $\varphi_n^2 = [x_n^1, x_n^2]^\tau$ and (5.21) implies

$$(5.26) \quad \frac{1}{N_k} \sum_{n=1}^{N_k} \|\alpha_{N_k}^\tau x_n^1 + \beta_{N_k}^\tau x_n^2\|^2 \xrightarrow[k \rightarrow \infty]{} 0.$$

Further, we have

$$(5.27) \quad \overline{\lim}_{N \rightarrow \infty} \frac{1}{N} \sum_{n=1}^N \|x_n^1\|^2 < \infty \quad \text{a.s.}$$

by Lemma 6, and

$$(5.28) \quad \lim_{N \rightarrow \infty} \frac{1}{N} \sum_{n=1}^N x_n^2 x_n^{2\tau} = \begin{bmatrix} R & 0 \\ 0 & R \end{bmatrix} > 0$$

by ergodicity.

Thus from (5.24) and (5.26)–(5.28) it follows that

$$\lim_{k \rightarrow \infty} \frac{1}{N_k} \sum_{n=1}^{N_k} \beta_{N_k}^\tau x_n^2 x_n^{2\tau} \beta_{N_k} = 0,$$

which leads to $\beta_{N_k} \xrightarrow{k \rightarrow \infty} 0$ by (5.28). Comparing it with (5.25) we obtain a contradiction, which shows the truth of (5.17). Therefore, there exist $\alpha_0 > 0$, N_0 such that

$$(5.29) \quad \lambda_{\min} \left(\sum_{i=1}^n \varphi_i^1 \varphi_i^{1\tau} \right) \geq \frac{n}{\log^{1/4} n} \alpha_0 \quad \forall n \geq N_0.$$

By (5.12) and Lemma 6 it follows that

$$(5.30) \quad \lim_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n \|\varphi_i^1\|^2 \geq \lim_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n \|w_i\|^2 = \text{tr } R > 0$$

and

$$(5.31) \quad \overline{\lim}_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n \|\varphi_i^1\|^2 < \infty \quad \text{a.s.}$$

From (5.30) and (5.31) it follows that there are positive quantities $\beta \geq \alpha > 0$ such that

$$(5.32) \quad \alpha n \leq r_{1n} \leq \beta n,$$

which, together with (5.29), yields

$$\lambda_{\max}^{1n} / \lambda_{\min}^{1n} \leq M \log^{1/4} r_{1n} \quad \forall n \geq N_0 \quad \text{with some } M > 0$$

where λ_{\max}^{1n} and λ_{\min}^{1n} denote, respectively, the maximum and minimum eigenvalues of $\sum_{i=1}^n \varphi_i^1 \varphi_i^{1\tau} + (1/\alpha)I$. Then we obtain the required assertion $\Phi_1(n, 0) \xrightarrow{n \rightarrow \infty} 0$ by Lemma 3 and Remark 1. \square

Proof of Theorem 1. Since $\sum_{i=2}^n (v_i v_i^\tau - (1/\log^{1/4} i)I)/i$ is a convergent martingale, by the Kronecker lemma it follows that

$$(5.33) \quad \lim_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n v_i v_i^\tau = \lim_{n \rightarrow \infty} \frac{1}{n} \sum_{i=2}^n \frac{1}{\log^{1/4} i} I = 0.$$

Similarly we have

$$(5.34) \quad \lim_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n w_i v_{i-1}^\tau = 0.$$

From (3.1) and (4.2) it follows that

$$(5.35) \quad y_{n+1} = \xi_{n+1} + y_{n+1}^* + w_{n+1} + v_n.$$

Then (2.11) and (2.12) follow immediately from (4.11), (5.9), (5.33)–(5.35) and condition A_2 .

Thus we only need to prove $\theta_n \xrightarrow{n \rightarrow \infty} \theta$ a.s. By Lemma 7 it suffices to verify that

$$(5.36) \quad \lim_{N \rightarrow \infty} \lambda_{\min} \left(\frac{\log^{1/4} N}{N} \sum_{n=1}^N Z_n Z_n^\tau \right) \neq 0 \quad \text{a.s.}$$

If (5.36) were not true, then there would exist a subsequence of eigenvectors $\alpha_{N_k} \in R^{mp+lq}$ for matrix $(\log^{1/4} N_k / N_k) \sum_{n=1}^{N_k} Z_n Z_n^\tau$ with $N_k \xrightarrow[k \rightarrow \infty]{} \infty$ and

$$(5.37) \quad \|\alpha_{N_k}\| = 1 \quad \forall k \geq 1$$

such that

$$(5.38) \quad \alpha_{N_k}^\tau \frac{\log^{1/4} N_k}{N_k} \sum_{n=1}^{N_k} Z_n Z_n^\tau \alpha_{N_k} \xrightarrow[k \rightarrow \infty]{} 0.$$

Without loss of generality we suppose $\alpha_{N_k} \xrightarrow[k \rightarrow \infty]{} \alpha$. Write α_{N_k} and α in the component form

$$\alpha_{N_k} = [\alpha_1^\tau(N_k) \cdots \alpha_{p+q}^\tau(N_k)]^\tau, \quad \alpha = [\alpha_1^\tau \cdots \alpha_{p+q}^\tau]^\tau$$

with $\alpha_i(N_k)$, α_i being m -dimensional and $\alpha_{p+j}(N_k)$, α_{p+j} l -dimensional vectors, $i = 1 \cdots p$, $j = 1 \cdots q$.

Set

$$(5.39) \quad \begin{aligned} H_{N_k}(z) &= \alpha_1^\tau(N_k)z + \cdots + \alpha_p^\tau(N_k)z^p \\ &\quad + \alpha_{p+1}^\tau(N_k)H_1(z) + \cdots + \alpha_{p+q}^\tau(N_k)H_1(z)z^{q-1} \\ &\triangleq \sum_{i=0}^{\infty} h_i^\tau(N_k)z^i, \end{aligned}$$

$$(5.40) \quad \begin{aligned} H(z) &= \alpha_1^\tau z + \cdots + \alpha_p^\tau z^p + \alpha_{p+1}^\tau H_1(z) + \cdots + \alpha_{p+q}^\tau H_1(z)z^{q-1} \\ &\triangleq \sum_{i=0}^{\infty} h_i^\tau z^i. \end{aligned}$$

We note that α_{N_k} and hence $H_{N_k}(z)$ may depend on ω , but by condition A₂ and (5.37) it is clear that there are constants $c_1 > 0$, $c_2 > 0$ such that $\|h_i(N_k)\| \leq c_1 \exp(-c_2 i)$ for all $i \geq 0$, for all $k \geq 0$. Then Lemma 5 can be applied, and from (5.38), (5.39) we have

$$\begin{aligned} 0 &= \lim_{k \rightarrow \infty} \frac{\log^{1/4} N_k}{N_k} \sum_{n=1}^{N_k} [\alpha_1^\tau(N_k)v_{n-1} + \cdots + \alpha_p^\tau(N_k)v_{n-p} + \alpha_{p+1}^\tau(N_k)H_1(z)v_n + \cdots \\ &\quad + \alpha_{p+q}^\tau(N_k)H_1(z)v_{n-q+1}]^2 \\ &= \lim_{k \rightarrow \infty} \frac{\log^{1/4} N_k}{N_k} \sum_{n=1}^{N_k} [(\alpha_1^\tau(N_k)z + \cdots + \alpha_p^\tau(N_k)z^p + \alpha_{p+1}^\tau(N_k)H_1(z) + \cdots \\ &\quad + \alpha_{p+q}^\tau(N_k)H_1(z)z^{q-1})v_n]^2 \\ &= \lim_{k \rightarrow \infty} \frac{\log^{1/4} N_k}{N_k} \sum_{n=1}^{N_k} (H_{N_k}(z)v_n)(H_{N_k}(z)v_n)^\tau \\ &= \lim_{k \rightarrow \infty} \frac{\log^{1/4} N_k}{N_k} \sum_{i=0}^{N_k} h_i^\tau(N_k) \sum_{n=2}^{N_k} \frac{1}{\log^{1/4} n} h_i(N_k), \end{aligned}$$

which implies that

$$\lim_{k \rightarrow \infty} \sum_{i=0}^{N_k} \|h_i(N_k)\|^2 = 0$$

and hence $\sum_{i=0}^{\infty} \|h_i\|^2 = 0$ by the dominated convergence theorem; therefore $H(z) = 0$. Setting $z = 0$ and paying attention to the fact that $H_1(0)(= B_1^+)$ is of full row rank we see $\alpha_{p+1}^\tau B_1^+ = 0$ and so $\alpha_{p+1} = 0$. Then it follows directly from (5.40) that

$$(5.41) \quad (\alpha_1^\tau + \alpha_2^\tau z + \cdots + \alpha_p^\tau z^{p-1}) = -(\alpha_{p+2}^\tau + \cdots + \alpha_{p+q}^\tau z^{q-2})H_1(z).$$

In view of condition A_3 , applying Lemma 6.6-1 of Kailath [16] to (5.41) we know that there exists a polynomial with vector coefficients $f(z)$:

$$f(z) = f_1 + f_2 z + \cdots + f_s z^{s-1}, \quad s \geq 1,$$

such that

$$(5.42) \quad (\alpha_{p+2}^\tau + \cdots + \alpha_{p+q}^\tau z^{q-2}) = f^\tau(z) B_1^+ B(z) \\ = (f_1^\tau + \cdots + f_s^\tau z^{s-1})(B_1^+ B_1 + \cdots + B_1^+ B q z^{q-1}).$$

From here it is easy to conclude that $f_i = 0$ ($1 \leq i \leq s$) since $B_1^+ B_q$ is of full rank by condition A_3 , then $\alpha_{p+j} = 0$ by (5.42), and then $\alpha_j = 0$ by (5.41) ($1 \leq j \leq q$, $1 \leq i \leq p$). Thus $\alpha = 0$, and $\alpha_{N_k} \xrightarrow{k \rightarrow \infty} 0$, thus contradicting (5.37). Hence (5.36) holds. \square

6. Concluding discussion. In order to get optimality in both tracking and estimating we have added to $\{y_n^*\}$ a random disturbance with covariance matrix tending to zero, but, intuitively, the disturbance may harm the tracking if time is bounded. However, all assertions of Theorem 1 can remain valid for u_n defined from (3.1) with v_n deleted (i.e. from (2.10)) if the reference signal y_n^* itself is "complicated" enough in the sense that

$$(6.1) \quad \lim_{N \rightarrow \infty} \lambda_{\min} \left(\frac{\log^{1/4} N}{N} \sum_{n=1}^N Y_n^* Y_n^{*\tau} \right) \neq 0.$$

This remark can easily be seen from Lemma 7.

For the single-input and single-output system it is easy to show that for (6.1) it suffices to require condition A_3 and

$$(6.2) \quad \lim_{N \rightarrow \infty} \lambda_{\min} \left(\frac{\log^{1/4} N}{N} \sum_{n=1}^N [y_n^* \cdots y_{n-p-q+1}^*]^\tau [y_n^* \cdots y_{n-p-q+1}^*] \right) \neq 0.$$

Recently, for multidimensional and random $\{y_n^*\}$ we have obtained conditions similar to (6.2) in order that all conclusions of Theorem 1 hold by applying u_n defined from (2.10). It will be published elsewhere.

Appendix 1. Existence of adaptive control.

LEMMA. (1) Let A and B be two matrices of dimensions $m \times n$ and $n \times m$, respectively. Then the following equality takes place

$$\det(I_m + AB) = \det(I_n + BA)$$

where I_n means the $n \times n$ identity matrix.

(2) Provided x_1 and x_2 are independent random variables, then

$$\sup_{a \in R^1} P(x_1 + x_2 = a) \leq \min \left\{ \sup_{a \in R^1} P(x_1 = a), \sup_{a \in R^1} P(x_2 = a) \right\}.$$

Proof. (1) By taking determinants for both sides of the following matrix identity:

$$\begin{bmatrix} I_m & -A \\ 0 & BA + I_n \end{bmatrix} = \begin{bmatrix} I_m & 0 \\ -B & I_n \end{bmatrix} \cdot \begin{bmatrix} I_m + AB & -A \\ 0 & I_n \end{bmatrix} \cdot \begin{bmatrix} I_m & 0 \\ B & I_n \end{bmatrix},$$

the desired equality is immediately verified.

(2) Denote by $F_1(x)$, $F_2(x)$, $F_{12}(x)$ the distributions of x_1 , x_2 , $x_1 + x_2$, respectively. Clearly we have

$$F_{12}(x) = \int_{-\infty}^{\infty} F_1(x-y) dF_2(y)$$

and

$$F_{12}(x+) = \int_{-\infty}^{\infty} F_1((x-y)+) dF_2(y)$$

by the dominated convergence theorem.

Then for any $a \in R^1$

$$\begin{aligned} P(x_1 + x_2 = a) &= F_{12}(a+) - F_{12}(a) = \int_{-\infty}^{\infty} [F_1((a-y)+) - F_1(a-y)] dF_2(y) \\ &= \int_{-\infty}^{\infty} P(x_1 = a-y) dF_2(y) \leq \sup_{a \in R^1} P(x_1 = a) \int_{-\infty}^{\infty} dF_2(y) \\ &= \sup_{a \in R^1} P(x_1 = a). \end{aligned}$$

Similarly, we have

$$P(x_1 + x_2 = a) \leq \sup_{a \in R^1} P(x_2 = a)$$

and thus the desired result follows. \square

THEOREM. Assume $m \leq l$ and $\{w_n\}$ and $\{v_n\}$ are two sequences of mutually independent random vectors and the components of w_n are independent and with continuous distribution functions. Then for any $n \geq 1$ there exists u_n satisfying (3.1) if the initial values are appropriately chosen. Further, this u_n is unique if and only if $m = l$.

Proof. Let $A_{in}, B_{jn}, C_{kn}, i = 1 \cdots p, j = 1 \cdots q, k = 1 \cdots r$ be the matrix components of θ_n , i.e.,

$$\theta_n^\tau = [-A_{1n} \cdots -A_{pn} B_{1n} \cdots B_{qn} C_{1n} \cdots C_{rn}].$$

Set

$$\bar{\theta}_n^\tau = [-A_{1n} \cdots -A_{pn} 0 B_{2n} \cdots B_{qn} C_{1n} \cdots C_{rn}],$$

and

$$\bar{\varphi}_n = [y_n^\tau \cdots y_{n-p+1}^\tau, 0, u_{n-1}^\tau \cdots u_{n-q+1}^\tau, y_n^\tau - \varphi_{n-1}^\tau \theta_{n-1}, \cdots, y_{n-r+1}^\tau - \varphi_{n-r}^\tau \theta_{n-r}]^\tau.$$

Equation (3.1) is equivalent to

$$(A.1) \quad B_{1n} u_n = y_{n+1}^* + v_n - \bar{\theta}_n^\tau \bar{\varphi}_n.$$

First let $m = l$. For this case we only need to prove that B_{1n} is invertible a.s. In fact, if this is true, then from (A.1) u_n is uniquely defined by $u_n = B_{1n}^{-1}(y_{n+1}^* + v_n - \bar{\theta}_n^\tau \bar{\varphi}_n)$, which obviously is \mathcal{F}_n -measurable. (In adaptive tracking cases we take $\mathcal{F}_n \triangleq \sigma\{w_i, v_i, i \leq n\}$.)

From (2.7) and (4.2) we obtain

$$(A.2) \quad B_{1n+1} = B_{1n} + \frac{1}{r_n}(\xi_{n+1} + w_{n+1})u_n^\tau.$$

It is easy to take initial values φ_0, θ_0 such that B_{11} is invertible; for example, take $u_0 = 0$ and B_{10} invertible.

We now inductively prove that B_{1n} is nondegenerate for any $n \geq 0$. Assuming B_{1n} is nonsingular a.s., we show that B_{1n+1} is also. In other words, we need to prove that $P(N) = 0$ implies $P(DN^c) = 0$, where

$$N \triangleq \{\omega \mid \det B_{1n} = 0\}, \quad D \triangleq \{\omega \mid \det B_{1n+1} = 0\}.$$

Suppose that the opposite were true, i.e., $P(N) = 0$, but $P(DN^c) > 0$.

From (A.2) we have

$$\det \left(B_{1n} + \frac{1}{r_n} (\xi_{n+1} + w_{n+1}) u_n^T \right) = 0 \quad \forall \omega \in DN^c$$

but $\det B_{1n} \neq 0$ for $\omega \in DN^c$; hence

$$\det \left(I + \frac{1}{r_n} B_{1n}^{-1} (\xi_{n+1} + w_{n+1}) u_n^T \right) = 0 \quad \forall \omega \in DN^c$$

or

$$\det \left(1 + \frac{1}{r_n} u_n^T B_{1n}^{-1} (\xi_{n+1} + w_{n+1}) \right) = 0 \quad \forall \omega \in DN^c$$

by part (1) of the lemma.

Then we have

$$(A.3) \quad u_n^T B_{1n}^{-1} (\xi_{n+1} + w_{n+1}) = -r_n \quad \forall \omega \in DN^c$$

and consequently,

$$(A.4) \quad u_n^T B_{1n} \neq 0 \quad \forall \omega \in DN^c$$

since $r_n \geq 1$.

We denote by $\alpha_i(\omega)$ and $w_{n+1,i}$ the components of $u_n^T B_{1n}^{-1}$ and w_{n+1} respectively, i.e.,

$$(A.5) \quad u_n^T B_{1n}^{-1} = [\alpha_1(\omega), \dots, \alpha_m(\omega)],$$

$$(A.6) \quad w_{n+1} = [w_{n+1,1}, \dots, w_{n+1,m}]^T.$$

Then from (A.3), (A.5) and (A.6) we have

$$(A.7) \quad \sum_{i=1}^m \alpha_i(\omega) w_{n+1,i} + r_n + u_n^T B_{1n}^{-1} \xi_{n+1} = 0 \quad \forall \omega \in DN^c.$$

From (A.4) and the assumption $P(DN^c) > 0$ we would have some $\alpha_i(\omega)$ and a subset $D_1 \subset DN^c$ such that

$$(A.8) \quad \alpha_i(\omega) \neq 0 \quad \forall \omega \in D_1, \quad P(D_1) > 0.$$

Without loss of generality, we assume $i = 1$, and define the random variable $z(\omega)$:

$$z(\omega) = \begin{cases} \frac{1}{\alpha_1(\omega)} \left[\sum_{i=2}^m \alpha_i(\omega) w_{n+1,i} + r_n + u_n^T B_{1n}^{-1} \xi_{n+1} \right], & \omega \in D_1, \\ 0, & \omega \in D_1^c, \end{cases}$$

which is clearly independent of $w_{n+1,1}$. By part (2) of the lemma, it follows that

$$(A.9) \quad P(w_{n+1,1} + z(\omega) = 0) = 0.$$

However, (A.7) and (A.8) would yield

$$(A.10) \quad P(w_{n+1,1} + z(\omega) = 0) \geq P(D_1) > 0.$$

The contradiction obtained proves $P(DN^c) = 0$, and hence the nonsingularity of $B_{1,n+1}$ a.s.

Now assume $m < l$.

Let

$$B_{1n} \triangleq [\overbrace{B_{1n}^1}^m, \overbrace{B_{1n}^2}^{l-m}]m, \quad u_n^\tau \triangleq [\overbrace{u_n^{1\tau}}^m, \overbrace{u_n^{2\tau}}^{l-m}].$$

From (A.2) we see

$$B_{1n+1}^1 = B_{1n}^1 + \frac{1}{r_n}(\xi_{n+1} + w_{n+1})u_n^{1\tau}.$$

By an argument similar to that given for the $m = l$ case we can prove that B_{1n}^1 is invertible a.s. for any $n \geq 1$ if φ_0, θ_0 are adequately chosen. Then (A.1) is equivalent to

$$(A.11) \quad [I, (B_{1n}^1)^{-1}B_{1n}^2]u_n = (B_{1n}^1)^{-1}(y_{n+1}^* + v_n - \bar{\theta}_n^\tau \bar{\varphi}_n)$$

or

$$u_n^1 + (B_{1n}^1)^{-1}B_{1n}^2u_n^2 = (B_{1n}^1)^{-1}(y_{n+1}^* + v_n - \bar{\theta}_n^\tau \bar{\varphi}_n).$$

Obviously, the solution of (A.11) can be expressed by

$$u_n = \begin{bmatrix} (B_{1n}^1)^{-1}(y_{n+1}^* + v_n - \bar{\theta}_n^\tau \bar{\varphi}_n - B_{1n}^2u_n^2) \\ u_n^2 \end{bmatrix} \quad \text{a.s.}$$

with any $(l-m)$ -dimensional and \mathcal{F}_n -measurable u_n^2 . This means that for the case $m < l$ the control u_n satisfying (3.1) exists but it is not unique. \square

Remark. Recently Caines and Meyn [4] also have shown the existence of u_n satisfying (2.10) for a one-dimensional case but under conditions different from those imposed here.

Appendix 2. Proof of lemmas.

Proof of Lemma 5. Due to the assumption $v_n = 0$ for $n < 0$, we have

$$\begin{aligned} \sum_{h=1}^N (H_N(z)v_n)(H_N(z)v_n)^\tau &= \sum_{i,j=0}^{\infty} H_i(N) \left(\sum_{n=1}^N v_{n-i}v_{n-j}^\tau \right) H_j^\tau(N) \\ &= \sum_{i,j=0}^{\infty} H_i(N) \left(\sum_{n=\max(i,j,1)}^N v_{n-i}v_{n-j}^\tau \right) H_j^\tau(N). \end{aligned}$$

Set

$$S_N(i, j) = \sum_{n=\max(i,j,1)}^N [v_{n-i}v_{n-j}^\tau - \delta_{ij}R_{n-i}],$$

$$R_n = Ev_nv_n^\tau = \begin{cases} \frac{1}{\log^{1/4} n} I, & n > 1, \\ 0, & n \leq 1. \end{cases}$$

Clearly, $S_N(i, j)$ is a martingale and by Burkholder inequality (Chow and Teicher [14]), C_r -inequality and Schwarz inequality we have

$$\begin{aligned} E\|S_N(i, j)\|^{2+\delta/2} &\leq c_1 E \left(\sum_{n=\max(i,j,1)}^N \|v_{n-i}v_{n-j}^\tau - \delta_{ij}R_{n-i}\|^2 \right)^{1+(\delta/4)} \\ &\leq c_1 N^{\delta/4} E \sum_{n=\max(i,j,1)}^N \|v_{n-i}v_{n-j}^\tau - \delta_{ij}R_{n-i}\|^{2+(\delta/2)} \\ &\leq c_2 N^{1+\delta/4} \quad \text{for any } i \geq 0, j \geq 0 \text{ and some } c_1 > 0, c_2 > 0. \end{aligned}$$

From here and the Hölder inequality it follows that for any $\varepsilon > 0$ and

$$\gamma \in \left(\frac{2 + (\delta/4)}{2 + (\delta/2)}, 1 \right),$$

$$\begin{aligned} & P \left\{ \sum_{i,j=0}^{\infty} e^{-k_2(i+j)} \|S_N(i,j)\| > N^{\gamma} \cdot \varepsilon \right\} \\ & \leq \frac{1}{\varepsilon^{2+(\delta/2)} N^{\gamma(2+(\delta/2))}} E \left(\sum_{i,j=0}^{\infty} e^{-k_2(i+j)} \|S_N(i,j)\| \right)^{2+(\delta/2)} \\ & \leq c_3 \frac{1}{N^{\gamma(2+(\delta/2))}} E \sum_{i,j=0}^{\infty} (e^{-k_2(i+j)})^{1+(\delta/4)} \|S_N(i,j)\|^{2+(\delta/2)} \\ & \leq c_4 \cdot \frac{1}{N^{\gamma(2+(\delta/2)) - (1+(\delta/4))}} \quad \text{for any } N \geq 1 \text{ and some constants } c_3 > 0, c_4 > 0. \end{aligned}$$

Then by the Borel–Cantelli lemma we see

$$(A.12) \quad \overline{\lim}_{N \rightarrow \infty} \frac{1}{N^{\gamma}} \left\| \sum_{i,j=0}^{\infty} H_i(N) S_N(i,j) H_j^{\tau}(N) \right\| \leq \lim_{N \rightarrow \infty} \frac{k_1^2}{N^{\gamma}} \sum_{i,j=0}^{\infty} e^{-k_2(i+j)} \|S_N(i,j)\| \xrightarrow{N \rightarrow \infty} 0.$$

Finally, we obtain the desired result

$$\begin{aligned} & \lim_{N \rightarrow \infty} \frac{\log^{1/4} N}{N} \sum_{n=1}^N (H_N(z) v_n) (H_N(z) v_n)^{\tau} \\ & = \lim_{N \rightarrow \infty} \frac{\log^{1/4} N}{N} \sum_{i,j=0}^{\infty} (H_i(N) S_N(i,j) H_j^{\tau}(N)) \\ & \quad + \lim_{N \rightarrow \infty} \frac{\log^{1/4} N}{N} \sum_{i,j=0}^{\infty} H_i(N) \sum_{n=\max(i,j,1)}^N \delta_{ij} R_{n-i} H_j^{\tau}(N) \\ & = \lim_{N \rightarrow \infty} \frac{\log^{1/4} N}{N} \sum_{i=0}^N H_i(N) \sum_{n=\max(i,1)}^N R_{n-i} H_i^{\tau}(N) \\ & = \lim_{N \rightarrow \infty} \frac{\log^{1/4} N}{N} \left[H_0(N) \sum_{n=1}^N R_n H_0^{\tau}(N) + \sum_{i=1}^N H_i(N) \sum_{n=0}^{N-i} R_n H_i^{\tau}(N) \right] \\ & = \lim_{N \rightarrow \infty} \frac{\log^{1/4} N}{N} \left[H_0(N) \sum_{n=1}^N R_n H_0^{\tau}(N) \right. \\ & \quad \left. + \sum_{i=1}^N H_i(N) \left(\sum_{n=1}^N R_n + R_0 - \sum_{n=N-i+1}^N R_n \right) H_i^{\tau}(N) \right] \\ & = \lim_{N \rightarrow \infty} \frac{\log^{1/4} N}{N} \sum_{i=0}^N H_i(N) \sum_{n=1}^N R_n H_i^{\tau}(N). \quad \square \end{aligned}$$

Proof of Lemma 6. Set

$$S_N(i,j) \triangleq \sum_{n=1}^N w_{n+1-l-i} v_{n-m-j}^{\tau}.$$

Similar to the proof of (A.12), one can easily be convinced that

$$\lim_{N \rightarrow \infty} \frac{1}{N^{\gamma}} \sum_{i,j=0}^{\infty} H_i S_N(i,j) G_j^{\tau} = 0,$$

which is tantamount to (5.1).

Clearly, (5.2) can be verified in similar fashion.

By setting $H_N(z) \equiv H(z)$ and $v_n \equiv w_{n+1-l}$ in (A.12) we have

$$\lim_{N \rightarrow \infty} \frac{1}{N^\gamma} \sum_{i,j=0}^{\infty} H_i S_N(i,j) H_j^\tau = 0$$

where

$$S_N(i,j) = \sum_{n=\max(i,j,1)}^N [w_{n-l+1-i} w_{n-l+1-j}^\tau - \delta_{ij} R_{n-l+1-i}]$$

and

$$R_{n-l+1-i} \triangleq E w_{n-l+1-i} w_{n-l+1-i}^\tau.$$

Hence by the uniform boundedness of R_n we have

$$\begin{aligned} & \overline{\lim}_{N \rightarrow \infty} \frac{1}{N} \sum_{n=1}^N \|H(z) w_{n+1-l}\|^2 \\ &= \overline{\lim}_{N \rightarrow \infty} \operatorname{tr} \left[\frac{1}{N} \sum_{n=1}^N (H(z) w_{n+1-l})(H(z) w_{n+1-l})^\tau \right] \\ &= \overline{\lim}_{N \rightarrow \infty} \operatorname{tr} \left[\frac{1}{N} \sum_{i,j=0}^{\infty} H_i S_N(i,j) H_j^\tau + \sum_{i,j=0}^{\infty} H_i \frac{1}{N} \sum_{n=\max(i,j,1)}^N \delta_{ij} R_{n-l+1-i} H_j^\tau \right] \\ &= \overline{\lim}_{N \rightarrow \infty} \operatorname{tr} \sum_{i=0}^{\infty} H_i \frac{1}{N} \sum_{n=\max(i,1)}^N R_{n-l+1-i} H_i^\tau < \infty. \end{aligned}$$

This completes the proof of the lemma. \square

REFERENCES

- [1] K. J. ÅSTRÖM AND B. WITTENMARK, *On self-tuning regulators*, Automatica—J. IFAC, 9 (1973), pp. 185–195.
- [2] A. BECKER, P. R. KUMAR AND C. Z. WEI, *Adaptive control with the stochastic approximation algorithm—Geometry and convergence*, IEEE Trans. Automat. Control, AC-30 (1985), pp. 330–338.
- [3] P. E. CAINES AND S. LAFORTUNE, *Adaptive control with recursive identification for stochastic linear systems*, IEEE Trans. Automat. Control, AC-29 (1984), pp. 312–321.
- [4] P. E. CAINES AND S. MEYN, *On the zero divisor problem and singularities occurring in the recursive schemes of stochastic adaptive control*, Systems Control Lett., 6 (1985), pp. 235–238.
- [5] H. F. CHEN, *Quasi-least-squares identification and its strong consistency*, Internat. J. Control, 34 (1981), pp. 921–936.
- [6] ———, *Strong consistency of recursive identification under correlated noise*, J. Systems Sci. Math. Sci., 1 (1981), pp. 34–52.
- [7] ———, *Recursive system identification and adaptive control by use of the modified least squares algorithm*, this Journal, 22 (1984), pp. 758–776.
- [8] ———, *Recursive Estimation and Control for Stochastic Systems*, John Wiley, New York, 1985.
- [9] H. F. CHEN AND P. E. CAINES, *Strong consistency of the stochastic gradient algorithm of adaptive control*, IEEE Trans. Automat. Control, AC-30 (1985), pp. 189–192.
- [10] H. F. CHEN AND L. GUO (1984), *Adaptive control with recursive identification for stochastic linear systems*, in Advances in Control and Dynamic Systems, C. T. Leondes, ed., Vol. 24, Academic Press, New York, to appear.
- [11] ———, *Strong consistency of recursive identification by no use of persistent excitation condition*, Acta Math. Appl. Sinica, to appear.
- [12] ———, *Strong consistency of parameter estimates for discrete-time stochastic systems*, J. Systems Sci. Math. Sci., 5 (1985), pp. 81–93.

- [13] H. F. CHEN AND L. GUO, *The limit of stochastic gradient algorithm for identifying systems excited not persistently*, Kexue Tongbao (Science Bulletin), to appear.
- [14] Y. S. CHOW AND H. TEICHER (1978), *Probability Theory*, Springer, New York.
- [15] G. C. GOODWIN, P. T. RAMADGE AND P. E. CAINES, *Discrete-time stochastic adaptive control*, this Journal, 19 (1981), pp. 829-853.
- [16] T. KAILATH, *Linear Systems*, Prentice-Hall, Englewood Cliffs, NJ, 1980.
- [17] H. J. KUSHNER AND D. S. CLARK (1978), *Stochastic Approximation Methods for Constrained and Unconstrained Systems*, Springer, New York.
- [18] L. LJUNG, *Consistency of the least squares identification method*, IEEE Trans. Automat. Control, AC-22 (1976), pp. 551-575.
- [19] ———, *Analysis of recursive stochastic algorithms*, IEEE Trans. Automat. Control, AC-22 (1977), pp. 551-575.
- [20] J. B. MOORE, *On strong consistency of least squares identification algorithm*, Automatica—J. IFAC, 14 (1978), pp. 505-509.
- [21] K. S. SIN AND G. C. GOODWIN, *Stochastic adaptive control using a modified least squares algorithm*, Automatica—J. IFAC, 18 (1982), pp. 315-321.
- [22] V. SOLO, *The convergence of AML*, IEEE Trans. Automat. Control, AC-24 (1979), pp. 958-962.

OPTIMAL CONTROL OF A SIGNORINI PROBLEM*

A. BERMUDEZ† AND C. SAGUEZ‡

Abstract. In this paper we present a new method to obtain necessary conditions for optimal control problems of variational inequalities. It uses only techniques of classical convex analysis and is based on a transformation of the original problem into another one involving a linear state equation and nonconvex constraints on the state.

Key words. optimal control, variational inequalities, optimality conditions

AMS(MOS) subject classifications. Primary 49B22, 49A29; secondary 35J85, 35R35

1. Introduction. Optimal control problems for variational inequalities have been considered by many authors (J. L. Lions [5], J. P. Yvon [13], C. Saguez [12], F. Mignot [7], V. Barbu [1], [2]). The main difficulties appearing in such problems are to obtain optimality conditions and efficient numerical algorithms. They are due to the non-differentiability of the mapping giving the state from the control.

For the obstacle problem some results have been given by F. Mignot and J. P. Puel [8] using the conical derivative and by V. Barbu [1] using a penalty method. This last method has also been used in C. Saguez [11] and V. Barbu [2] to deal with parabolic problems.

More recently the authors have introduced a new method (A. Bermudez and C. Saguez [4]) to study optimal control problems of variational inequalities. The main idea is to transform the optimal control problem into another one involving a linear state equation and nonconvex constraints on the state. Optimality conditions are then obtained by analyzing the ones corresponding to this new optimal control problem.

The same transformation has been used by F. Mignot and J. P. Puel [9] to study an optimal control problem of a parabolic variational inequality of the obstacle type. For obtaining optimality conditions they use a regularization technique which needs some difficult a priori estimates.

In this paper we introduce another method which consists of considering two directions along which this optimal control problem is convex and then writing the corresponding optimality conditions.

This method which is quite general is applied to solve a boundary control problem for a variational inequality of Signorini type.

2. The optimal control problem. Let Ω be an open bounded subset of \mathbb{R}^N with smooth boundary Γ . Let f be given in $L^2(\Omega)$.

For $v \in L^2(\Gamma)$ we define $y(v)$ the state of the system as the solution of the variational inequality:

$$(2.1) \quad a(y, Z - y) \cong \int_{\Omega} f(Z - y) \, dx + \int_{\Gamma} v(Z - y) \, d\Gamma \quad \forall Z \in K$$

where a is the bilinear form:

$$(2.2) \quad a(y, Z) = \sum_{i=1}^N \int_{\Omega} \frac{\partial y}{\partial X_i} \frac{\partial Z}{\partial X_i} \, dx + \int_{\Omega} yZ \, dx$$

* Received by the editors December 16, 1985; accepted for publication (in revised form) March 11, 1986.

† Dep. Ecuaciones Funcionales, University of Santiago, Spain.

‡ SIMULOG S.A. Av. du Centre-78182 St. Quentin en Yvelines, France.

and K is given by:

$$(2.3) \quad K = \{Z \in V: Z \geq 0 \text{ a.e. on } \Gamma\}$$

with $V = H^1(\Omega)$.

We denote by A the bounded linear operator from V into V' defined by:

$$(2.4) \quad \langle Ay, Z \rangle = a(y, Z) \quad \forall y, Z \in V$$

where \langle, \rangle represents the duality map between V' and V .

Consider the following cost function:

$$(2.5) \quad J(v) = \int_{\Omega} (y(v) - Z_d)^2 dx + \frac{\nu}{2} \int_{\Gamma} v^2 d\Gamma$$

where Z_d is a given function in $L^2(\Omega)$ and ν a strictly positive real number.

The optimal problem is to find $u \in U_{ad}$ such that

$$(2.6) \quad J(u) \leq J(v) \quad \forall v \in L^2(\Gamma).$$

Existence of a solution of this problem is shown in J. L. Lions [5] by using a compacity method.

3. A regularity result. In this paragraph we show some regularity properties which will be used to obtain optimality conditions.

PROPOSITION 3.1. *If y is the solution of (2.1) there exists η in $L^2(\Gamma)$ such that:*

$$(3.1) \quad \frac{\partial y}{\partial n} + \eta = v \quad \text{on } \Gamma,$$

$$(3.2) \quad \eta \leq 0 \quad \text{on } \Gamma,$$

$$(3.3) \quad \int_{\Gamma} y \eta d\Gamma = 0.$$

Proof. Let y_{ε} be the solution of the penalized problem:

$$(3.4) \quad -\Delta y_{\varepsilon} + y_{\varepsilon} = f \quad \text{in } \Omega, \quad \frac{\partial y_{\varepsilon}}{\partial n} - \frac{1}{\varepsilon} y_{\varepsilon}^{-} = v \quad \text{on } \Gamma.$$

Multiplying the first equation by $-(1/\varepsilon)y_{\varepsilon}^{-}$ and using a Green's formula we obtain:

$$(3.5) \quad \frac{1}{\varepsilon} a(y_{\varepsilon}^{-}, y_{\varepsilon}^{-}) + \frac{1}{\varepsilon^2} \int_{\Gamma} |y_{\varepsilon}^{-}|^2 d\Gamma = -\frac{1}{\varepsilon} \int_{\Omega} f y_{\varepsilon}^{-} dx - \frac{1}{\varepsilon} \int_{\Gamma} v y_{\varepsilon}^{-} d\Gamma$$

from which it follows that:

$$(3.6) \quad \frac{1}{\varepsilon} y_{\varepsilon}^{-} \text{ is bounded in } L^2(\Gamma)$$

and then

$$(3.7) \quad y_{\varepsilon} \text{ is bounded in } H^{3/2}(\Omega) \quad (\text{see J. L. Lions and E. Magenes [6]}).$$

By using the compacity of the inclusion $H^{3/2}(\Omega) \subset H^1(\Omega)$, we deduce from (3.7) the existence of $y \in H^{3/2}(\Omega)$, $\eta \in L^2(\Gamma)$ and a sequence $\{\varepsilon_n\} \rightarrow 0$ such that:

$$(3.8) \quad \{y_{\varepsilon_n}\} \rightarrow y \quad \text{in } H^1(\Omega) \text{ strongly,}$$

$$(3.9) \quad \left\{ -\frac{1}{\varepsilon} y_{\varepsilon_n}^{-} \right\} \rightarrow \eta \quad \text{in } L^2(\Gamma) \text{ weakly,}$$

which implies that:

$$(3.10) \quad \begin{aligned} -\Delta y + y &= f, \\ \frac{\partial y}{\partial n} + \eta &= u, \\ y|_{\Gamma} &\geq 0 \quad \eta \leq 0; \end{aligned}$$

on the other hand we have

$$(3.11) \quad 0 \leq \lim_{n \rightarrow \infty} \left\{ -\frac{1}{\varepsilon} \int_{\Gamma} y_{\varepsilon_n}^- y_{\varepsilon_n} d\Gamma \right\} = \int_{\Gamma} \eta y \leq 0,$$

which finishes the proof. \square

PROPOSITION 3.2. *If u is an optimal control of the problem (2.6) then $u \in H^{1/2}(\Gamma)$.*

Proof. We use a regularization method as in C. Saguez [12], F. Mignot and J. P. Puel [9]. Consider the optimal control problem obtained by replacing the state variational inequality (2.1) by:

$$(3.12) \quad -\Delta y_{\varepsilon}^{\delta} + y_{\varepsilon}^{\delta} = f, \quad \frac{\partial y_{\varepsilon}^{\delta}}{\partial n} + \frac{1}{\varepsilon} \beta^{\delta}(y_{\varepsilon}^{\delta}) = v$$

where

$$(3.13) \quad \beta^{\delta}(r) = \begin{cases} r + \frac{\delta}{2} & \text{if } r \leq -\delta, \\ -\frac{1}{2\delta} r^2 & \text{if } -\delta \leq r \leq 0, \\ 0 & \text{if } r \geq 0, \end{cases}$$

and the cost function by the adapted cost function:

$$(3.14) \quad J_{\varepsilon}^{\delta}(v) = \frac{1}{2} \int_{\Omega} (y_{\varepsilon}^{\delta}(v) - Z_d)^2 dx + \frac{\nu}{2} \int_{\Gamma} v^2 d\Gamma + \frac{1}{2} \int_{\Gamma} (v - u)^2 d\Gamma.$$

It is not difficult to prove that for each $\varepsilon > 0$ there exists an optimal control u_{ε}^{δ} satisfying the following optimality conditions:

$$(3.15) \quad -\Delta y_{\varepsilon}^{\delta} + y_{\varepsilon}^{\delta} = f, \quad \frac{\partial y_{\varepsilon}^{\delta}}{\partial n} + \frac{1}{\varepsilon} \beta^{\delta}(y_{\varepsilon}^{\delta}) = u_{\varepsilon}^{\delta},$$

$$(3.16) \quad a(p_{\varepsilon}^{\delta}, Z) + \frac{1}{\varepsilon} \int_{\Gamma} \beta^{\delta'}(y_{\varepsilon}^{\delta}) p_{\varepsilon}^{\delta} Z d\Gamma = \int_{\Omega} (y_{\varepsilon}^{\delta} - Z_d) Z dx, \quad \forall Z \in V,$$

$$(3.17) \quad p_{\varepsilon}^{\delta} + \nu u_{\varepsilon}^{\delta} + u_{\varepsilon}^{\delta} - u = 0.$$

As in the proof of Proposition 3.1 we can show that:

$$(3.18) \quad y_{\varepsilon}^{\delta}(v) \rightarrow y(v) \quad \text{strongly in } V$$

and then

$$(3.19) \quad \limsup_{\varepsilon \rightarrow 0} J_{\varepsilon}^{\delta}(u_{\varepsilon}^{\delta}) \leq J(u)$$

because

$$(3.20) \quad J_{\varepsilon}^{\delta}(u_{\varepsilon}^{\delta}) \leq J_{\varepsilon}^{\delta}(u).$$

Moreover $\{u_\varepsilon^\delta\}$ is bounded in $L^2(\Gamma)$ then there exists a sequence $\{\varepsilon_n\} \rightarrow 0$ such that:

$$(3.21) \quad u_{\varepsilon_n}^\delta \rightarrow \bar{u} \text{ in } H^{-1/2}(\Gamma) \text{ strongly,}$$

from which it follows that:

$$(3.22) \quad y_{\varepsilon_n}^\delta(u_{\varepsilon_n}^\delta) \rightarrow y(\bar{u}) \text{ strongly in } V.$$

Therefore

$$(3.23) \quad \begin{aligned} \liminf_{n \rightarrow +\infty} J_{\varepsilon_n}^\delta(u_{\varepsilon_n}^\delta) &\geq J(\bar{u}) \frac{1}{2} \int_{\Omega} (\bar{u} - u)^2 \\ &\geq J(u) + \frac{1}{2} \int_{\Omega} (\bar{u} - u)^2 d\Gamma \end{aligned}$$

and then $u = \bar{u}$ by using (3.19).

On the other hand, by taking $Z = p_\varepsilon^\delta$ in (3.16) we can show that $\{p_\varepsilon^\delta\}$ is bounded in V .

If $p \in V$ is a weak limit point, then we deduce from (3.17)

$$p + \nu u = 0$$

which completes the proof. \square

4. Optimality conditions. The main result is the following:

THEOREM 4.1. *If u is an optimal control of the problem (2.6), there exists (y, ξ, q, θ) such that $y \in H^{3/2}(\Omega)$, $\xi \in L^2(\Gamma)$, $q \in H^1(\Omega)$, $\theta \in H^{-1/2}(\Gamma)$:*

$$(4.1) \quad \begin{aligned} -\Delta y + y &= f \quad \text{in } \Omega, \\ \frac{\partial y}{\partial n} + \xi &= u \quad \text{on } \Gamma, \\ y &\geq 0, \quad \xi \leq 0, \quad \int_{\Gamma} y \xi d\Gamma = 0. \end{aligned}$$

$$(4.2) \quad a(q, Z) = \int_{\Omega} (y - Z_d) Z dx + \langle \theta, Z \rangle_{H^{-1/2}(\Gamma) - H^{1/2}(\Gamma)} \quad \forall Z \in H^1(\Omega),$$

$$(4.3) \quad \int_{\Gamma} q \eta d\Gamma = 0 \quad \forall \eta \in \tilde{\mathcal{C}}_\xi \quad \text{with} \quad \int_{\Gamma} \eta y(u, \xi) = 0,$$

$$(4.4) \quad \langle \theta, Z \rangle_{H^{-1/2}(\Gamma) - H^{1/2}(\Gamma)} = 0 \quad \forall Z \in \mathcal{C}_y \quad \text{with} \quad \int_{\Gamma} Z \xi = 0$$

where

$$\tilde{\mathcal{C}}_\xi = \{\eta \in L^2(\Gamma) : \exists t > 0, \xi + t\eta \leq 0\},$$

and

$$(4.5) \quad \begin{aligned} \mathcal{C}_y &= \{Z \in H^{1/2}(\Gamma) : \exists t > 0, y + tZ \geq 0\}, \\ q + \nu u &= 0 \quad \text{on } \Gamma. \end{aligned}$$

Proof. (i) From Proposition 3.1 we deduce that if u is an optimal control then there exists $\xi \in L^2(\Gamma)$ such that (u, ξ) is an optimal control of the following constrained optimal control problem:

$$(4.6) \quad \text{minimizing } G(v, \eta) = \frac{1}{2} \int_{\Omega} (y(v, \eta) - Z_d)^2 dx + \frac{\nu}{2} \int_{\Gamma} v^2 d\Gamma \quad (v, \eta) \in (L^2(\Gamma))^2,$$

$$(4.7) \quad \eta \leq 0, y(v, \eta) \geq 0 \quad \text{on } \Gamma, \quad \int_{\Gamma} y(v, \eta) \eta d\Gamma = 0$$

where $y(v, \eta)$ denotes the solution of the (linear) state equation:

$$(4.8) \quad -\Delta y + y = f, \quad \frac{\partial y}{\partial n} + \eta = v.$$

(ii) Denote by χ_E the indicator function of a set E . Let ψ be the function from $H^1(\Gamma) \times L^2(\Gamma)$ into $(-\infty, \infty]$ given by:

$$(4.9) \quad \psi(Z, \eta) = \chi_C(Z) + \chi_{C^0}(\eta) + \chi_{\{0\}}\left(\int_{\Gamma} Z \eta d\Gamma\right)$$

where $C = \{\phi \in H^1(\Gamma), \phi \geq 0 \text{ a.e. on } \Gamma\}$

C^0 denotes the dual cone of C .

Then ψ is convex with respect to each variable separately, and moreover the problem (4.6)–(4.7) is equivalent to minimizing in the space $(L^2(\Gamma))^2$ the function F given by:

$$(4.10) \quad F(v, \eta) = G(v, \eta) + \psi(y(v, \eta)|_{\Gamma}, \eta).$$

(iii) If (u, ξ) is a minimum for F , then u is a minimum for the following convex function:

$$(4.11) \quad v \in L^2(\Gamma) \rightarrow F(v, \xi) \in (-\infty, \infty],$$

which implies:

$$(4.12) \quad \int_{\Omega} (y(u, \xi) - Z_d) \frac{\partial y}{\partial v}(v, \xi) dx + \nu \int_{\Gamma} uv d\Gamma + \left\langle \theta, \frac{\partial y}{\partial v}(v, \xi) \right\rangle_{H^{-1}(\Gamma)H^1(\Gamma)} = 0 \quad \forall v \in L^2(\Gamma),$$

with

$$(4.13) \quad \theta \in \partial_1 \psi(y(u, \xi)|_{\Gamma}, \xi)$$

because the mapping

$$(4.14) \quad v \in L^2(\Gamma) \rightarrow \frac{\partial y}{\partial v}(v, \xi)|_{\Gamma} \in H^1(\Gamma)$$

is surjective (see A. Bermudez [3]).

Similarly, if (u, ξ) is a minimum for F , then $\eta = 0$ is a minimum for the convex function

$$(4.15) \quad \eta \in L^2(\Gamma) \rightarrow F(u + \eta, \xi + \eta) \in (-\infty, \infty],$$

from which we can deduce the existence of p :

$$(4.16) \quad p \in \partial_2 \psi(y(u, \xi)|_{\Gamma}, \xi)$$

such that

$$(4.17) \quad \nu u + p = 0.$$

By using (4.17) in (4.12) we get

$$(4.18) \quad \int_{\Gamma} p \frac{\partial Z}{\partial n} d\Gamma = \int_{\Omega} (y(u, \xi) - Z_d) Z + \langle \theta, Z|_{\Gamma} \rangle_{H^{-1}(\Gamma)H^1(\Gamma)} \quad \forall Z \in H^{3/2}(\Omega).$$

By Proposition 3.2, we have $u \in H^{1/2}(\Gamma)$. Then (4.17) implies the existence of $q \in H^1(\Omega)$ such that $q|_{\Gamma} = p$ and from (4.18) we deduce $\theta \in H^{-1/2}(\Gamma)$ and

$$(4.19) \quad a(q, Z) = \int_{\Omega} (y(u, \xi) - Z_d) Z + \langle \theta, Z \rangle_{H^{-1/2}(\Gamma)H^{1/2}(\Gamma)} \quad \forall Z \in H^1(\Omega),$$

from which the uniqueness of q follows.

(iv) To prove (4.3)–(4.4) notice first that (4.13) and (4.16) are respectively equivalent to:

$$(4.20) \quad \theta \in \partial \chi_{C \cap [\mathbb{R}\xi]^0}(y(u, \xi))$$

and

$$(4.21) \quad p \in \partial \chi_{C^0 \cap [\mathbb{R}y(u, \xi)]^0}(\xi),$$

which can also be written as follows:

$$(4.22) \quad \begin{aligned} \langle \theta, y(u, \xi) \rangle &= 0, \\ \langle \theta, Z \rangle &\leq 0 \quad \forall Z \in C, \quad \int_{\Gamma} Z \xi d\Gamma = 0, \end{aligned}$$

$$(4.23) \quad \begin{aligned} \int_{\Gamma} p \xi d\Gamma &= 0, \\ \int_{\Gamma} p \eta d\Gamma &\leq 0 \quad \forall \eta \in C^0, \quad \int_{\Gamma} \eta y(u, \xi) d\Gamma = 0. \end{aligned}$$

Finally it is easy to see that (4.22) and (4.23) are equivalent to (4.4) and (4.3), respectively. \square

Remark 4.1. The variational equality (4.12) can be interpreted as follows:

$$(4.24) \quad -\Delta q + q = y - Z_d \quad \text{in } \Omega, \quad \frac{\partial q}{\partial n} = \theta \quad \text{on } \Gamma.$$

Remark 4.2. From the Proposition 2.2 in F. Mignot [7] we have

$$(4.25) \quad \bar{\mathcal{C}}_y = (C^0 \cap [\mathbb{R}y]^0)^0,$$

$$(4.26) \quad \bar{\mathcal{C}}_{\xi} = (C \cap [\mathbb{R}\xi]^0)^0.$$

Therefore (4.20) and (4.21) imply the existence of sequences $\{\theta_n\}$, $\{p_n\}$, $\{\tau_n\}$ and $\{t_n\}$ such that

$$(4.27) \quad \begin{aligned} \{\theta_n\} &\rightarrow \theta \quad \text{in } H^{-1/2}(\Gamma), \quad \{p_n\} \rightarrow p \quad \text{in } H^{1/2}(\Gamma), \\ \tau_n &> 0, \quad t_n > 0, \\ \xi + \tau_n \theta_n &\leq 0, \quad y + t_n p_n \geq 0. \end{aligned}$$

Assume that the sequences $\{\tau_n\}$ and $\{t_n\}$ have strictly positive limit points τ and t , respectively. Then, by taking $\lambda = \max \{1/\tau, 1/t\}$ we deduce

$$(4.28) \quad \alpha = \lambda \xi + \theta \leq 0,$$

$$(4.29) \quad \beta = \lambda y + p \geq 0.$$

Now it is easy to see that the equations (4.1), (4.2), (4.5), (4.28) and (4.29) are nothing other than the first order necessary conditions for $((u, \xi), (\alpha, \beta, \lambda))$ to be a stationary point of the following Lagrangian function associated with the optimization problem (4.6), (4.7):

$$(4.30) \quad \begin{aligned} L((v, \eta), (\gamma, \delta, \mu)) = & \frac{1}{2} \int_{\Omega} (y(v, \eta) - Z_d)^2 dx + \nu \int_{\Gamma} v^2 d\Gamma + \langle y(v, \eta), \gamma \rangle_{H^{1/2}(\Gamma), H^{-1/2}(\Gamma)} \\ & + \int_{\Gamma} \eta \delta d\Gamma - \mu \int_{\Gamma} y(v, \eta) \eta d\Gamma, \end{aligned}$$

under the constraints

$$\gamma \leq 0, \quad \delta \geq 0, \quad \mu \geq 0.$$

Remark 4.3. A direct proof of (4.1), (4.2), (4.5), (4.28) and (4.29), by using some well-known results on existence of Lagrange multipliers, is considered in A. Bermudez and C. Saguez [4]. It requires one to replace the functional space V by another one in which the set K has a nonvoid interior.

Moreover, to get nontrivial optimality conditions a “qualification property” has to be proved. This point, or alternatively, the assumption on the sequences $\{\tau_n\}$ and $\{t_n\}$ we have made in the Remark 4.2, are open problems.

REFERENCES

- [1] V. BARBU, *Necessary conditions for distributed control problems governed by parabolic variational inequalities*, this Journal, 19 (1981), pp. 64–86.
- [2] ———, *Optimal Control of Variational Inequalities*, Pitman, London, 1984.
- [3] A. BERMUDEZ, *Elementos finitos mixtos para problemas no lineales*, Proceedings of the 4th CEDYA, University of Sevilla, Spain, 1981, pp. 429–464.
- [4] A. BERMUDEZ AND C. SAGUEZ, *Optimal control of variational inequalities. Optimality conditions and numerical methods*, in *Free boundary: theory and applications*, A. Bossavit, A. Damlamian, M. Frémond, eds., Pitman, London, 1985, pp. 478–487.
- [5] J. L. LIONS, *Contrôle optimal de systèmes gouvernés par des équations aux dérivées partielles*, Dunod, Paris, 1968.
- [6] J. L. LIONS AND E. MAGENES, *Problèmes aux limites non homogènes*, Dunod, Paris, 1969.
- [7] F. MIGNOT, *Contrôle dans les inéquations variationnelles elliptiques*, J. Funct. Anal., 22 (1976), pp. 130–185.
- [8] F. MIGNOT AND J. P. PUEL, *Optimal control in some variational inequalities*, this Journal, 22 (1984), pp. 466–476.
- [9] ———, *Contrôle optimal d'un système gouverné par une inéquation variationnelle parabolique*, C.R. Acad. Sc. Paris, t.298, Série I, 12 (1984), pp. 277–280.
- [10] ———, *Contrôle optimal de systèmes gouvernés par des inéquations variationnelles*, to appear.
- [11] C. SAGUEZ, *Conditions nécessaires d'optimalité pour des problèmes de contrôle optimal associés à des inéquations variationnelles*, LABORIA Report n. 345, INRIA, Rocquencourt, 1979.
- [12] ———, *Contrôle optimal de problèmes à frontière libre*, Thèse d'Etat, University of Technology of Compiègne, 1981.
- [13] J. P. YVON, *Contrôle optimal de systèmes gouvernés par des inéquations variationnelles*, LABORIA Report n. 53, INRIA, Rocquencourt, 1974.

LIPSCHITZ CONTINUITY OF SOLUTIONS OF LINEAR INEQUALITIES, PROGRAMS AND COMPLEMENTARITY PROBLEMS*

O. L. MANGASARIAN† AND T.-H. SHIAU‡

Abstract. It is shown that solutions of linear inequalities, linear programs and certain linear complementarity problems (e.g. those with P -matrices or Z -matrices but *not* semidefinite matrices) are Lipschitz continuous with respect to changes in the right-hand side data of the problem. Solutions of linear programs are *not* Lipschitz continuous with respect to the coefficients of the objective function. The Lipschitz constant given here is a generalization of the role played by the norm of the inverse of a nonsingular matrix in bounding the perturbation of the solution of a system of equations in terms of a right-hand side perturbation.

Key words. linear inequalities, linear programming, linear complementarity problems, Lipschitz continuity, perturbation analysis

AMS (MOS) subject classifications. 15A39, 90C05, 65F35

1. Introduction. The purpose of this work is to show that solutions of linear inequalities, linear programs and certain linear complementarity problems are Lipschitz continuous with respect to changes in the right-hand side of the problem. Speaking in general and in somewhat loose terms, if we denote, by r^1 and r^2 , two distinct right-hand sides, then there exist corresponding solutions x^1 and x^2 such that

$$(1.1) \quad \|x^1 - x^2\| \leq K \|r^1 - r^2\|$$

where the Lipschitz constant K depends only on the matrix defining the problem, but not on the right-hand sides nor the objective function if there is one. A key role in determining the Lipschitz constant K is played by the condition number for linear inequalities, introduced in [11], which is a generalization of the very useful concept of a condition number for a nonsingular square matrix [3]. In [19] Robinson obtained local Lipschitz continuity results for generalized equations which include linear programs, convex quadratic programs and monotone linear complementarity problems. Robinson's Lipschitz constant [19, Thm. 2] involves a bound on the solution set which is assumed to be bounded. By contrast our Lipschitz constants are global, and our solution sets need not be bounded. In [18] Robinson obtained a Lipschitz constant for the perturbation of linear inequalities which is different from our constant (2.5).

We give now a summary of our principal results. Theorem 2.2 deals with a system of linear inequalities and equalities (2.1) and shows that if the system is solvable for right-hand sides r^1 and r^2 , then for each solution x^1 for right-hand side r^1 there exists a solution x^2 for right-hand side r^2 such that (1.1) holds. The Lipschitz constant here plays the same role as the norm of the inverse of a nonsingular matrix does for a system of linear equations. Our Lipschitz constant for the system (2.1) defined by (2.5), is a minor variation of the constant (6) of [11] but is different from Robinson's [18]. Furthermore, the Lipschitz continuity Theorem 2.2 leads in a very elementary way to Theorem 2.2' which is essentially equivalent to Theorem 1 of [11] and to Hoffman's theorem [8], [18] and which gives an estimate of the error in an approximate solution to the systems of linear inequalities and equalities (2.1) in terms of the residual of the

* Received by the editors July 8, 1985; accepted for publication (in revised form) March 11, 1986. This work was sponsored by the United States Army under contract DAAG29-80-C0041. The material is based upon work sponsored by the National Science Foundation under grants MCS-8200632 and MCS-8420963, and by the Air Force Office for Scientific Research under grants AFOSR-ISSA-85-00080 and AFOSR-86-0124.

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

‡ Department of Computer Science, University of Missouri, Columbia, Missouri 65211.

approximate solution and the Lipschitz constant. Again the role played in Theorem 2.2' by the Lipschitz constant is an extension of the same role played by the norm of the inverse of a matrix for a system of linear equations. Computation of the Lipschitz constant (2.5) for the system of linear inequalities and equalities (2.1) is quite difficult, but an important fact is that such a constant exists and is finite. For some special cases such as when we have strongly stable linear inequalities only (that is linear inequalities solvable for all right-hand sides) the Lipschitz constant can be computed by a single linear program as in (2.17) below. By using the Lipschitz constant for linear inequalities and equalities we show in Theorem 2.4 that solutions of linear programs are also Lipschitz continuous with respect to right-hand side perturbations only. Proposition 2.6 shows that our Lipschitz constant (2.20) for the linear program (2.18) is sharper than that of Cook et al. [4, Thm. 5]. By means of a simple example (2.26), we show that solutions of linear programs are *not* Lipschitz continuous with respect to perturbations in the objective function coefficients. Finally in § 3 by using the Lipschitz constant for linear inequalities and equalities we establish in Theorem 3.2 Lipschitz continuity of solutions of linear complementarity problems with respect to right-hand side perturbations that generate unique solutions along the line segment joining perturbed and unperturbed right-hand sides. A simple consequence of this result is Theorem 3.3 which shows that the solution of a linear complementarity problem with a P -matrix (that is a matrix with positive principal minors) is Lipschitz continuous with respect to right-hand side perturbations. Example 3.4 shows that solutions of positive semidefinite linear complementarity problems are not Lipschitz continuous with respect to their right-hand sides. Finally by exploiting the fact that for certain classes of matrices such as Z -matrices (real matrices with nonpositive off-diagonal elements) the linear complementarity problem can be solved as a linear program [10], Lipschitz continuity of solutions of such linear complementarity problems are obtained in Theorem 3.5.

A brief word about notation and some basic concepts employed. For a vector x in the n -dimensional real space R^n , $|x|$ and x_+ will denote the vectors in R^n with components $|x|_i := |x_i|$ and $(x_+)_i := \max\{x_i, 0\}$, $i = 1, \dots, n$, respectively. For a norm $\|x\|_\beta$ on R^n , $\|x\|_{\beta^*}$ will denote the dual norm [9], [16] on R^n , that is $\|x\|_{\beta^*} := \max_{\|y\|_\beta = 1} xy$, where xy denotes the scalar product $\sum_{i=1}^n x_i y_i$. The generalized Cauchy-Schwarz inequality $|xy| \leq \|x\|_\beta \cdot \|y\|_{\beta^*}$, for x and y in R^n , follows immediately from this definition of the dual norm. For $1 \leq p, q \leq \infty$, and $(1/p) + (1/q) = 1$, the p -norm $(\sum_{i=1}^n |x_i|^p)^{1/p}$ and the q -norm are dual norms on R^n [16]. If $\|\cdot\|_\beta$ is a norm on R^n , we shall, with a slight abuse of notation, let $\|\cdot\|_\beta$ also denote the corresponding norm on R^m for $m \neq n$. For an $m \times n$ real matrix A , A_i denotes the i th row, $A_{\cdot j}$ denotes the j th column, $A_I := A_{i \in I}$, and $A_{\cdot J} := A_{\cdot j \in J}$, where $I \subset \{1, \dots, m\}$ and $J \subset \{1, \dots, n\}$. $\|A\|_\beta$ denotes the matrix norm [16], [20] subordinate to the vector norm $\|\cdot\|_\beta$, that is $\|A\|_\beta = \max_{\|x\|_\beta = 1} \|Ax\|_\beta$. The consistency condition $\|Ax\|_\beta \leq \|A\|_\beta \|x\|_\beta$ follows immediately from this definition of a matrix norm. A monotonic norm on R^n is any norm $\|\cdot\|$ on R^n such that for a, b in R^n , $\|a\| \leq \|b\|$ whenever $|a| \leq |b|$ or equivalently if $\|a\| = \||a|\|$ [9, p. 47]. The p -norm for $p \geq 1$ is monotonic [16]. A vector of ones in any real space will be denoted by e . The identity matrix of any order will be denoted by I . The nonnegative orthant in R^n will be denoted by R_+^n . The abbreviation rhs will denote "right-hand side."

2. Linear inequalities and programs. We shall first be concerned with Lipschitz continuity of solutions of the following set of linear inequalities with respect to changes in the right-hand side

$$(2.1) \quad Ax \leq b, \quad Cx = d$$

where b and d are given points in R^m and R^k , respectively, $A \in R^{m \times n}$, that is an $m \times n$ real matrix and $C \in R^{k \times n}$. We shall employ a slight variation of the condition constant introduced in [11, Eq. (6)] for linear inequalities and programs as our Lipschitz constant for the linear inequalities (2.1) and subsequently for the linear program (2.18) and the linear complementarity problem (3.1).

We begin with a simple extension of the fundamental theorem on basic solutions [6, Thm. 2.11] to unrestricted as well as nonnegative variables.

LEMMA 2.1 (Basic solutions). *Let $A \in R^{m \times n}$, $C \in R^{k \times n}$ and $p \in R^n$. The system*

$$(2.2) \quad A^T u + C^T v = p, \quad u \geq 0$$

has a solution $(u, v) \in R^{m+k}$ if and only if it has a basic solution, that is a solution (u, v) such that the rows of $\begin{pmatrix} A \\ C \end{pmatrix}$ corresponding to nonzero components of (u, v) are linearly independent.

Proof. The system (2.2) having a solution (u, v) implies that

$$(2.3) \quad A^T u + \tilde{C}^T v = p, \quad (u, v) \geq 0$$

has a solution where \tilde{C} is obtained from C by multiplying by -1 those rows of C corresponding to negative components of v . It follows from the fundamental theorem on basic solutions [6, Thm. 2.11] that (2.3) has a basic solution and consequently so does (2.2). \square

We proceed now to establish Lipschitz continuity of solutions of (2.1) with respect to right-hand side perturbations. Robinson [18, Cor. 2.2] gives this result with a different Lipschitz constant.

THEOREM 2.2 (Lipschitz continuity of feasible points of linear inequalities and equalities). *Let the linear inequalities and equalities (2.1) have nonempty feasible sets S^1 and S^2 for the right-hand sides (b^1, d^1) and (b^2, d^2) , respectively. For each $x^1 \in S^1$ there exists an $x^2 \in S^2$ closest to x^1 in the ∞ -norm such that*

$$(2.4) \quad \|x^1 - x^2\|_\infty \leq \mu_\beta(A; C) \left\| \begin{pmatrix} b^1 - b^2 \\ d^1 - d^2 \end{pmatrix} \right\|_\beta$$

where $\|\cdot\|_\beta$ is some norm on R^{m+k} and

$$(2.5) \quad \mu_\beta(A; C) := \sup_{u, v} \left\{ \left\| \begin{pmatrix} u \\ v \end{pmatrix} \right\|_{\beta^*} \left| \begin{array}{l} \|uA + vC\|_1 = 1, \quad u \geq 0. \\ \text{Rows of } \begin{pmatrix} A \\ C \end{pmatrix} \text{ corresponding to nonzero} \\ \text{elements of } \begin{pmatrix} u \\ v \end{pmatrix} \text{ are linear independent} \end{array} \right. \right\}.$$

Proof. We note that $\mu_\beta(A; C)$ is finite. For if not, there would exist fixed subsets I and J of $\{1, \dots, m\}$ and $\{1, \dots, k\}$, respectively, and a sequence $\{u_i^I, v_j^J\}$ such that $\{\|u_i^I, v_j^J\|\} \rightarrow \infty$ and the rows of $\begin{pmatrix} A \\ C \end{pmatrix}$ are linearly independent. Hence a subsequence

$$\left\{ \frac{(u_i^I, v_j^J)}{\|u_i^I, v_j^J\|} \right\}$$

converges to (\bar{u}_I, \bar{v}_J) satisfying $\bar{u}_I A_I + \bar{v}_J C_J = 0$, $\|\bar{u}_I, \bar{v}_J\| = 1$, which contradicts the linear independence of the rows of $\begin{pmatrix} A \\ C \end{pmatrix}$.

Now let $x^1 \in S^1$. Choose $x^2 \in S^2$ which is closest to x^1 in the ∞ -norm. Thus x^2 must solve

$$(2.6) \quad \min_x \|x - x^1\|_\infty \quad \text{s.t.} \quad Ax \leq b^2, \quad Cx = d^2$$

which is equivalent to the linear program

$$(2.7) \quad \min_{x, \delta} \delta \quad \text{s.t.} \quad Ax \leq b^2, \quad Cx = d^2, \quad x + e\delta \geq x^1, \quad -x + e\delta \geq -x^1.$$

Hence (x^2, δ^2) and some $(u^2, v^2, r^2, s^2) \in R^{\dot{m}+k+2n}$ satisfy the following Karush-Kuhn-Tucker conditions for (2.7)

$$(2.8) \quad \begin{aligned} Ax^2 &\leq b^2, \quad Cx^2 = d^2, \quad \|x^1 - x^2\|_\infty = \delta^2, \\ u^2(-Ax^2 + b^2) &= 0, \quad r^2(x^2 + e\delta^2 - x^1) = 0, \quad s^2(-x^2 + e\delta^2 + x^1) = 0, \\ -u^2A + v^2C + r^2 - s^2 &= 0, \quad e(r^2 + s^2) = 1, \quad (u^2, r^2, s^2) \geq 0. \end{aligned}$$

Note that if $0 = \delta^2 = \|x^1 - x^2\|_\infty$, then (2.4) is trivially true. So assume that $\delta^2 > 0$. It follows from $\delta^2 > 0$ and $r_j^2(x^2 + e\delta^2 - x^1)_j = 0$ and $s_j^2(-x^2 + e\delta^2 + x^1)_j = 0$ that $r_j^2 s_j^2 = 0$, for $j = 1, \dots, n$. Hence

$$(2.9) \quad -u^2A + v^2C + r^2 - s^2 = 0, \quad e(r^2 + s^2) = 1, \quad r^2 s^2 = 0, \quad (u^2, r^2, s^2) \geq 0.$$

By Lemma 2.1 and $u^2(-Ax^2 + b^2) = 0$ it follows that we may take

$$u^2 = \begin{pmatrix} u_I^2 \\ 0 \end{pmatrix} \geq 0 \quad \text{and} \quad v^2 = \begin{pmatrix} v_J^2 \\ 0 \end{pmatrix}$$

such that the rows of (A_I^J) are linearly independent and $u_I^2(-A_I x^2 + b_I^2) = 0$. Hence (2.9) becomes

$$\|-u_I^2 A_I + v_J^2 C_J\|_1 = \|r^2 - s^2\|_1 = e(r^2 + s^2) = 1, \quad u_I^2 \geq 0,$$

Rows of (A_I^J) linear independent.

Hence by (2.5) we have

$$(2.10) \quad \left\| \frac{u^2}{v^2} \right\|_{\beta^*} \leq \mu_\beta(A; C).$$

We now have

$$(2.11) \quad \begin{aligned} \|x^1 - x^2\|_\infty &= \delta^2 = -b^2 u^2 + d^2 v^2 + x^1(r^2 - s^2) \\ &= -b^2 u^2 + d^2 v^2 + x^1(A^T u^2 - C^T v^2) \\ &= u^2(Ax^1 - b^2 + b^1 - b^1) + v^2(-Cx^1 + d^2 + d^1 - d^1) \\ &\leq u^2(b^1 - b^2) + v^2(d^2 - d^1) \\ &\leq \left\| \frac{u^2}{v^2} \right\|_{\beta^*} \|b^1 - b^2\|_\beta \\ &\leq \mu_\beta(A; C) \left\| \frac{b^1 - b^2}{d^1 - d^2} \right\|_\beta \quad (\text{by (2.10)}). \quad \square \end{aligned}$$

Note that the Lipschitz constant $\mu_\beta(A; C)$ of (2.4) plays the same role as that of the norm of the inverse of a nonsingular matrix of a system of linear equations. This fact can be seen more clearly from the following corollary to Theorem 2.2 applied to systems solvable for all right-hand sides (i.e. strongly stable) systems. Note also that we can get a sharper result by replacing $(b^1 - b^2)$, in (2.4) and (2.11) onward, by $(b^1 - b^2)_+$.

COROLLARY 2.3 (Lipschitz continuity of feasible points of strongly stable linear inequalities). *Let $A \in R^{m \times n}$ and $C \in R^{k \times n}$ be such that*

$$(2.12) \quad \begin{aligned} &\text{Rows of } C \text{ are linearly independent and} \\ &Ax < 0, \quad Cx = 0 \quad \text{has a solution } x. \end{aligned}$$

Then the linear inequalities (2.1) are solvable for all right-hand sides $(b, d) \in R^{m+k}$. For each x^1 in the solution set of (2.1) with right-hand sides (b^1, d^1) , there exists an x^2 in the solution set of (2.1) with right-hand sides (b^2, d^2) such that

$$(2.13) \quad \|x^1 - x^2\|_\infty \leq \bar{\mu}_\beta(A; C) \left\| \begin{matrix} b^1 - b^2 \\ d^1 - d^2 \end{matrix} \right\|_\beta$$

where $\|\cdot\|_\beta$ is some norm on R^{m+k} and

$$(2.14) \quad \bar{\mu}_\beta(A; C) := \max_{(u,v) \in R^{m+k}} \left\{ \left\| \begin{matrix} u \\ v \end{matrix} \right\|_{\beta^*} \mid \begin{matrix} \|uA + vC\|_1 = 1 \\ u \geq 0 \end{matrix} \right\}.$$

Proof. That (2.1) is solvable for any right-hand side (b, d) follows from solving $Cx = d$ for x^d for any given d and then taking as the desired solution $x^d + \lambda \bar{x}$ for sufficiently large positive λ , where \bar{x} solves $Ax < 0$, $Cx = 0$. The rest of the proof of the corollary is similar to the proof of Theorem 2.2, except that u^2 and v^2 are not decomposed into

$$\begin{pmatrix} u_i^2 \\ 0 \end{pmatrix} \quad \text{and} \quad \begin{pmatrix} v_j^2 \\ 0 \end{pmatrix}.$$

The finiteness of $\bar{\mu}_\beta(A; C)$ of (2.14) follows from the boundedness of the feasible region of (2.14). For if it were unbounded, there would exist $\{u^i, v^i\}$ such that $\{\|u^i, v^i\|\} \rightarrow \infty$, and consequently an accumulation point (\bar{u}, \bar{v}) would exist such that

$$(2.15) \quad \bar{u}A + \bar{v}C = 0, \quad \bar{u} \geq 0, \quad (\bar{u}, \bar{v}) \neq 0.$$

This, however, would contradict the linear independence of the rows of C if $\bar{u} = 0$, and if $\bar{u} \neq 0$ would contradict the solvability of $Ax < 0$, $Cx = 0$, because then $0 = \bar{u}Ax + \bar{v}Cx = \bar{u}Ax < 0$. \square

Note that if A is vacuous and C is a nonsingular square matrix, then

$$(2.16) \quad \bar{\mu}_\infty(\phi; C) = \max_{v \in R^k} \{ \|v\|_1 \mid \|vC\|_1 = 1 \} = \|(C^T)^{-1}\|_1 = \|C^{-1}\|_\infty.$$

This was already pointed out in [11, Remark 2]. Note also that (2.14) can be written in the equivalent form

$$(2.14') \quad \bar{\mu}_\beta(A; C) = \max_{(u,v,z) \in R^{m+k+n}} \left\{ \left\| \begin{matrix} u \\ v \end{matrix} \right\|_{\beta^*} \mid \begin{matrix} -z \leq uA + vC \leq z \\ u \geq 0, \quad ez = 1 \end{matrix} \right\}.$$

This is a difficult convex-function maximization problem on a polyhedral set which is closely related to the NP-complete problem of a norm-maximization problem on a polyhedral set for positive integer β^* [12]. However for $\beta^* = \infty$, that is $\beta = 1$, it can be shown, as in [12], that (2.14') is in P. In addition a good bound for $\bar{\mu}_\beta(A; C)$ for any β can be obtained by solving a single linear program [12]. When C is empty and $\beta = \infty$, (2.14') degenerates to the following linear program:

$$(2.17) \quad \bar{\mu}_\infty(A; \phi) = \max_{(u,z) \in R^{m+n}} \{ eu \mid -z \leq uA \leq z, u \geq 0, ez = 1 \}.$$

We note that the Lipschitz constants $\mu_\beta(A; C)$ and $\bar{\mu}_\beta(A; C)$, which play the role of the norm of the inverse of a nonsingular matrix of a system of linear equations, can also be used, just as the norm of the inverse can, to obtain a bound on the error in an approximate solution in terms of the residual. Thus, if we assume for the moment that A is vacuous and that C is $n \times n$ and nonsingular, then $\bar{\mu}_\beta(\phi; C) = \|C^{-1}\|_\infty$ by

(2.16). Thus (2.5) and (2.13) are the extensions to a system of linear inequalities and equalities of the following simple Lipschitz continuity property of $Cx = d$

$$\|x^1 - x^2\|_\infty \leq \|C^{-1}\|_\infty \|d^1 - d^2\|_\infty$$

where $x^1 = C^{-1}d^1$ and $x^2 = C^{-1}d^2$. Since $\|C^{-1}\|_\infty$ can also be used to estimate the error in an approximate solution x to $Cx^1 = d^1$ in terms of its residual $\|Cx - d^1\|_\infty$ as follows:

$$\|x - x^1\|_\infty = \|C^{-1}(Cx - d^1)\|_\infty \leq \|C^{-1}\|_\infty \|Cx - d^1\|_\infty,$$

it follows that the Lipschitz constants $\mu_\beta(A; C)$ and $\bar{\mu}_\beta(A; C)$ can be similarly used to give an estimate on the error in an approximate solution to (2.1) in terms of its residual. In fact this estimate has been given in [11, Thm. 1] and by Hoffman [8], [18] with a different constant. It also follows very easily from Theorem 2.2 above as follows.

THEOREM 2.2' (Error bound for approximate solution of linear inequalities and equalities). *Let the linear inequalities and equalities (2.1) have a nonempty feasible set S^1 for the right-hand side (b^1, d^1) . For each x in R^n there exists an $x^1 \in S^1$ such that*

$$\|x - x^1\|_\infty \leq \mu_\beta(A; C) \left\| \begin{pmatrix} Ax - b^1 \\ Cx - d^1 \end{pmatrix} \right\|_\beta$$

where $\mu_\beta(A; C)$ is defined by (2.5).

Proof. Since for each $x \in R^n$

$$Ax \leq b^1 + (Ax - b^1)_+, \quad Cx = d^1 + (Cx - d^1)$$

it follows by Theorem 2.2 that there exists an $x^1 \in S^1$ such that the conclusion of the theorem holds. \square

A similar error bound holds for strongly stable linear inequalities which is based on (2.13).

It is interesting to note that Theorem 2.2 is stronger than Theorem 2.2' in the sense that the latter follows directly from the former as was demonstrated above, whereas the converse holds with the additional assumption that the norm $\|\cdot\|_\beta$ is a monotonic norm [9], [16]. Thus to obtain Theorem 2.2 from Theorem 2.2', we have from Theorem 2.2' that for each $x^1 \in S^1$ there exists an $x^2 \in S^2$ such that

$$\|x^2 - x^1\|_\infty \leq \mu_\beta(A; C) \left\| \begin{pmatrix} Ax^1 - b^2 \\ Cx^1 - d^2 \end{pmatrix} \right\|_\beta \leq \mu_\beta(A; C) \left\| \begin{pmatrix} b^1 - b^2 \\ d^1 - d^2 \end{pmatrix} \right\|_\beta$$

where the last inequality follows from

$$\begin{aligned} (Ax^1 - b^2)_+ &= (Ax^1 - b^2 + b^1 - b^1)_+ \leq (b^1 - b^2)_+ \leq |b^1 - b^2|, \\ |Cx^1 - d^2| &= |Cx^1 - d^2 + d^1 - d^1| = |d^1 - d^2| \end{aligned}$$

and the monotonicity of the norm $\|\cdot\|_\beta$.

Next we establish the Lipschitz continuity with respect to right-hand side perturbation of solutions of the linear program

$$(2.18) \quad \max_x px \quad \text{s.t.} \quad Ax \leq b, \quad Cx = d$$

where $p \in R^n$ and A, b, C, d are as in (2.1). For the Lipschitz continuity results for linear programs we have to restrict the norms employed to monotonic norms [9], [16] and have to drop $u \geq 0$ from (2.5). Lipschitz continuity results for more general optimization problems are given in [1], [7].

THEOREM 2.4 (Lipschitz continuity of solutions of linear programs with respect to right-hand side perturbation). *Let the linear program (2.18) have nonempty solution sets S^1 and S^2 for right-hand sides (b^1, d^1) and (b^2, d^2) , respectively. For each $x^1 \in S^1$ there exists an $x^2 \in S^2$ such that*

$$(2.19) \quad \|x^1 - x^2\|_\infty \leq \nu_\beta(A; C) \left\| \begin{matrix} b^1 - b^2 \\ d^1 - d^2 \end{matrix} \right\|_\beta$$

where $\|\cdot\|_\beta$ is some monotonic norm on R^{m+k} and

$$(2.20) \quad \nu_\beta(A; C) := \sup_{u, v} \left\{ \left\| \begin{matrix} u \\ v \end{matrix} \right\|_{\beta^*} \left| \begin{matrix} \|uA + vC\|_1 = 1 \\ \text{Rows of } \begin{pmatrix} A \\ C \end{pmatrix} \text{ corresponding to nonzero} \\ \text{elements of } \begin{pmatrix} u \\ v \end{pmatrix} \text{ are linear independent} \end{matrix} \right. \right\}.$$

Proof. Given $x^1 \in S^1$, let

$$A_I x^1 = b_I^1, \quad A_J x^1 < b_J^1$$

where $I \cup J = \{1, 2, \dots, m\}$. Fix any $\bar{x}^2 \in S^2$ and let $I = I_1 \cup I_2$ where

$$I_1 := \{i \in I \mid A_i \bar{x}^2 = b_i^2\}, \quad I_2 := \{i \in I \mid A_i \bar{x}^2 < b_i^2\}.$$

Since $x = \bar{x}^2$ satisfies the system of constraints

$$(2.21) \quad \begin{aligned} & \text{(i)} \quad A_{I_1} x = b_{I_1}^2, \\ & \text{(ii)} \quad A_{I_2} \bar{x}^2 \leq A_{I_2} x, \quad A_{I_2} x \leq b_{I_2}^2, \\ & \text{(iii)} \quad A_J x \leq b_J^2, \quad Cx = d^2, \end{aligned}$$

it follows that (2.21) is nonvacuous. As in Theorem 2.2, let x^2 be a solution of

$$(2.22) \quad \min \|x - x^1\|_\infty \quad \text{s.t. (2.21)}.$$

Since (2.22) is a convex program, x^2 remains optimal after we remove any number of inactive constraints. For each $i \in I_2$, at least one of the two constraints of (2.21)(ii) is inactive because $A_i \bar{x}^2 < b_i^2$. So we can remove one inactive constraint for each $i \in I_2$ thus obtaining

$$(2.23) \quad \|x^2 - x^1\|_\infty = \min \|x - x^1\|_\infty \quad \text{s.t. (2.24)} = \min \|x - x^1\|_\infty \quad \text{s.t. (2.21)}$$

where

$$(2.24) \quad \begin{aligned} & \text{(i)} \quad A_{I_1} x = b_{I_1}^2, \\ & \text{(iia)} \quad A_K \bar{x}^2 \leq A_K x, \\ & \text{(iib)} \quad A_L x \leq b_L^2, \\ & \text{(iic)} \quad A_J x \leq b_J^2, \quad Cx = d^2 \end{aligned}$$

where $K \cup L = I_2$, $K \cap L = \emptyset$. So $I_1 \cup K \cup L \cup J = \{1, 2, \dots, m\}$ and I_1 , K , L and J are all disjoint. On the other hand, since

$$A_K x^1 = b_K^1 - b_K^2 + b_K^2 \geq b_K^1 - b_K^2 + A_K \bar{x}^2,$$

it follows that $x = x^1$ satisfies the following system:

$$(2.24') \quad \begin{aligned} & \text{(i)} \quad A_{I_1} x = b_{I_1}^1, \\ & \text{(iia)} \quad b_K^1 - b_K^2 + A_K \bar{x}^2 \leq A_K x, \\ & \text{(iib)} \quad A_L x \leq b_L^1, \\ & \text{(iic)} \quad A_J x \leq b_J^1, \quad Cx = d^1. \end{aligned}$$

It follows by (2.23), (2.24'), Theorem 2.2 and the norm monotonicity that

$$\begin{aligned} \|x^1 - x^2\|_\infty &\leq \mu_\beta \left\| \begin{pmatrix} -A_K \\ A_L \\ A_J \end{pmatrix}; \begin{pmatrix} A_{I_1} \\ C \end{pmatrix} \right\| \left\| \begin{matrix} -b_K^1 + b_K^2 \\ b_H^1 - b_H^2 \\ d^1 - d^2 \end{matrix} \right\|_\beta \\ &\leq \nu_\beta(A; C) \left\| \begin{matrix} b^1 - b^2 \\ d^1 - d^2 \end{matrix} \right\|_\beta \end{aligned}$$

where $H = I_1 \cup L \cup J$ is the complement of K .

It remains to show that $x^2 \in S^1$. Since $x^1 \in S^1$, we have by the Karush-Kuhn-Tucker optimality conditions that

$$(2.25) \quad A_I^T u_I^1 + C^T v^1 = p \quad \text{for some } u_I^1 \geq 0 \quad \text{and some } v^1.$$

Since both \bar{x}^2 and x^2 satisfy (2.21) it follows that

$$px^2 = u_I^1 A_I x^2 + v^1 C x^2 \geq u_I^1 A_I \bar{x}^2 + v^1 C \bar{x}^2 = p \bar{x}^2$$

and the proof is complete. \square

Remark 2.5. We note that Cook, Gerards, Schrijver and Tardos [4, Thm. 5] have a similar result to Theorem 2.4 for *integer* entries for A but without the equality constraints $Cx = d$. However their Lipschitz constant is bigger than or equal to our Lipschitz constant. In fact their Lipschitz constant $n\Delta(A)$ is only for $\beta = \infty$, where ΔA is the maximum of the absolute values of the determinants of the square submatrices of A . We formalize the relation between the two Lipschitz constants as follows.

PROPOSITION 2.6. *For integer A , $\nu_\infty(A; \phi) \leq n\Delta(A)$.*

Proof. For any u_I for which $\|u_I A_I\|_1 = 1$ and the rows of A_I are linearly independent, we can assume that

$$A_I = [B \quad N]$$

where B is a nonsingular square submatrix.

Let $q := u_I B$, then $\|q\|_1 \leq \|u_I A_I\|_1 = 1$ since $u_I B$ is a subvector of $u_I A_I$. It follows that

$$\|u_I\|_1 = \|(B^T)^{-1} q\|_1 \leq \|(B^T)^{-1}\|_1 \|q\|_1 \leq \|(B^T)^{-1}\|_1 = \max_i \sum_j |h_{ij}|$$

where h_{ij} is the (i, j) entry of B^{-1} [16, p. 22]. Hence

$$h_{ij} = \frac{1}{\det B} (-1)^{i+j} B_{ji}$$

where B_{ij} is the (i, j) cofactor of B which is the determinant of a square submatrix of A . Hence

$$|B_{ji}| \leq \Delta(A).$$

If A is integral $|\det B| \geq 1$ is an integer; hence

$$|h_{ij}| \leq \frac{1}{|\det B|} |B_{ji}| \leq \Delta(A).$$

Consequently,

$$\|u_I\|_1 \leq \|(B^T)^{-1}\|_1 = \max_i \sum_j |h_{ij}| \leq n\Delta(A).$$

Since u_I is arbitrary, we have

$$\nu_\infty(A; \phi) = \sup \{ \|u_I\|_1 \mid \|u_I A_I\|_1 = 1, \text{ rows of } A_I \text{ linear independent} \} \leq n\Delta(A). \quad \square$$

Remark 2.7. Note that it is not true that solutions of linear programs are Lipschitzian with respect to perturbations in the objective function coefficients as evidenced by the following simple example:

$$(2.26) \quad \max (1 + \delta)x_1 + x_2 \quad \text{s.t. } x_1 + x_2 \leq 1, \quad (x_1, x_2) \geq 0.$$

The solution to this problem is:

$$x(\delta) = \begin{cases} (1, 0) & \text{for } \delta > 0, \\ (0, 1) & \text{for } \delta < 0. \end{cases}$$

Hence

$$\lim_{\delta \rightarrow 0+} \frac{\|x(\delta) - x(-\delta)\|}{2\delta} = \infty$$

and hence $x(\delta)$ is not Lipschitzian with respect to δ .

3. Linear complementarity problems. In this section we shall employ the Lipschitz constant $\mu_\beta(A; C)$ developed in Theorem 2.2 for linear inequalities and equalities to obtain a Lipschitz constant for linear complementarity problems with matrices that have positive principal minors [5] or which are hidden Z -matrices [17]. We will show by means of Example 3.4 that solutions of linear complementarity problems with a positive semidefinite matrix are not Lipschitz continuous with respect to right-hand side perturbations.

We consider the linear complementarity problem (M, q) of finding an x in R^n such that

$$(3.1) \quad Mx + q \geq 0, \quad x \geq 0, \quad x(Mx + q) = 0$$

where $M \in R^{n \times n}$ and $q \in R^n$. Note that given $J \subset \{1, \dots, n\}$, any solution of the following system of $2n$ linear inequalities and equalities

$$(3.2) \quad \begin{aligned} M_J x + q_j &\geq 0, & x_j &= 0, & j &\in J, \\ M_J x + q_j &= 0, & x_j &\geq 0, & j &\notin J, \end{aligned}$$

is a solution of (M, q) . For $J \subset \{1, \dots, n\}$ let $Q(J)$ denote the set of all q vectors for which (3.2) has a solution. It is easy to verify that $Q(J)$ is a closed convex cone. In fact it is called a *complementary cone* of (M, q) [14, p. 482]. It is also obvious that $\bigcup Q(J)_{J \subset \{1, \dots, n\}}$ is the set of all q for which (M, q) is solvable. Define

$$(3.3) \quad \sigma_\beta(M) := \max_{J \subset \{1, \dots, n\}} \mu_\beta \left(\begin{pmatrix} -M_J \\ -I_{\bar{J}} \end{pmatrix}; \begin{pmatrix} I_J \\ M_{\bar{J}} \end{pmatrix} \right)$$

where μ_β is defined by (2.5) and \bar{J} is the complement of J in $\{1, \dots, n\}$. We shall prove (Theorem 3.3) that $\sigma_\beta(M)$ will serve as a Lipschitz constant for solutions of (M, q) when M is a P -matrix, that is a matrix with positive principal minors [5], [2], or more generally (Theorem 3.2) for perturbations of q such that the linear complementarity problem is uniquely solvable along the line joining the original q and the perturbed q . We will also establish Lipschitz continuity for solutions of (M, q) when M is a hidden Z -matrix (Theorem 3.5). We begin with a lemma. A related result to this lemma appears in [15].

LEMMA 3.1. *Let q^1 and q^2 be fixed distinct vectors in R^n and let $q(t) := (1-t)q^1 + tq^2$ for $t \in [0, 1]$. Assume that $(M, q(t))$ is solvable for $t \in [0, 1]$. Then there exists a partition $0 = t_0 < t_1 < \dots < t_N = 1$ such that for $1 \leq i \leq N$*

$$(3.4) \quad q(t_{i-1}) \in Q(J_i), \quad q(t_i) \in Q(J_i) \quad \text{for some } J_i \subset \{1, \dots, n\}.$$

Proof. Let

$$T(J) := \{t \mid t \in [0, 1], q(t) \in Q(J)\}$$

for $J \subset \{1, \dots, n\}$. It is easy to see that $T(J)$ is closed and convex and hence it is a closed interval which may degenerate to a single point or to the empty set. Since $(M, q(t))$ is solvable for $t \in [0, 1]$ it follows that

$$[0, 1] \subset \bigcup_{J \subset \{1, \dots, n\}} T(J).$$

Let

$$L := \{[l_1, u_1], \dots, [l_K, u_K]\}$$

be the set of *maximal* intervals in $\{T(J) \mid J \subset \{1, \dots, n\}\}$, that is there is no other interval $T(J)$, $J \subset \{1, \dots, n\}$ that properly contains $[l_i, u_i]$. By removing duplicates from L if needed, we can assume that $[l_i, u_i], \dots, [l_K, u_K]$ are distinct and that $l_i < l_2 < \dots < l_K$. Since each $t \in [0, 1]$ belongs to $T(J)$ (for some $J \subset \{1, \dots, n\}$) which is either in L or contained in some interval of L , we have that

$$[0, 1] \subset \bigcup_{i=1}^K [l_i, u_i].$$

Thus $l_i \leq u_{i-1}$, otherwise (u_{i-1}, l_i) would be an uncovered gap of $[0, 1]$. Also $u_{i-1} < u_i$, otherwise $[l_i, u_i]$ would not be maximal because it would be contained in $[l_{i-1}, u_{i-1}]$.

Hence $l_1 = 0$, $l_{i-1} < l_i \leq u_{i-1} < u_i$ and $u_K = 1$. Let $0 = t_0 < t_1 < \dots < t_N = 1$ be the sorted numbers of $\{l_1, u_1, l_2, u_2, \dots, l_K, u_K\}$ with duplicates removed. Then each interval $[t_{i-1}, t_i]$ is contained in some interval $T(J_i)$ in L and so

$$q(t_{i-1}) \in Q(J_i) \quad \text{and} \quad q(t_i) \in Q(J_i). \quad \square$$

We establish now the Lipschitz continuity of linear complementarity problems with unique solutions along the line segment $q(t) := (1-t)q^1 + tq^2$, $t \in [0, 1]$.

THEOREM 3.2 (Lipschitz continuity of uniquely solvable linear complementarity problems). *Let q^1 and q^2 be points in R^n such that the linear complementarity problem $(M, q(t))$ with $q(t) := (1-t)q^1 + tq^2$ has a unique solution for each $t \in [0, 1]$. Then the unique solutions x^1 of (M, q^1) and x^2 of (M, q^2) satisfy*

$$\|x^1 - x^2\|_\infty \leq \sigma_\beta(M) \|q^1 - q^2\|_\beta$$

where $\sigma_\beta(M)$ is defined by (3.3).

Proof. There exist $0 = t_0 < t_1 < \dots < t_N = 1$ with properties stated in Lemma 3.1. Let $x(t_i)$ be the unique solution of $(M, q(t_i))$. Since for $1 \leq i \leq N$, $q(t_{i-1})$ and $q(t_i)$ belong to $Q(J_i)$ for some $J_i \subset \{1, \dots, n\}$, there exists a solution $y(t_{i-1})$ of $(M, q(t_{i-1}))$ such that by (2.4) and (3.3) it follows that

$$\begin{aligned} \|x(t_i) - y(t_{i-1})\|_\infty &\leq \mu_\beta \begin{pmatrix} -M_{J_i} & I_{J_i} \\ -I_{\bar{J}_i} & M_{\bar{J}_i} \end{pmatrix} \|q(t_i) - q(t_{i-1})\|_\beta \\ (3.5) \qquad &\leq \sigma_\beta(M) (t_i - t_{i-1}) \|q^1 - q^2\|_\beta \end{aligned}$$

where \bar{J}_i is the complement of J_i in $\{1, \dots, n\}$. Summing up for $i = 1, \dots, N$ gives

$$\sum_{i=1}^N \|x(t_i) - y(t_{i-1})\|_\infty \leq \sigma_\beta(M) \|q^1 - q^2\|_\beta.$$

Since $(M, q(t_{i-1}))$ has a unique solution, $y(t_{i-1}) = x(t_{i-1})$. Hence

$$\|x^1 - x^2\|_\infty \leq \sum_{i=1}^N \|x(t_i) - x(t_{i-1})\|_\infty \leq \sigma_\beta(M) \|q^1 - q^2\|_\beta. \quad \square$$

Since for P -matrix M , the linear complementarity problem (M, q) has a unique solution for each $q \in R^n$ [13], the following theorem is an immediate corollary to Theorem 3.2.

THEOREM 3.3 (Lipschitz continuity of solutions of linear complementarity problems with P -matrices). *Let M be a P -matrix. For each q^1 and q^2 in R^n the corresponding unique solutions x^1 and x^2 of (M, q^1) and (M, q^2) , respectively, satisfy*

$$\|x^1 - x^2\|_\infty \leq \sigma_\beta(M) \|q^1 - q^2\|_\beta$$

where $\sigma_\beta(M)$ is defined by (3.3).

The following example shows that solutions of positive semidefinite linear complementarity problems may not be Lipschitzian.

Example 3.4.

$$M = \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix}, \quad q^1 = \begin{pmatrix} -\varepsilon \\ 1 \end{pmatrix}, \quad q^2 = \begin{pmatrix} \varepsilon \\ 1 \end{pmatrix}, \quad \varepsilon > 0,$$

$$q(t) = \begin{pmatrix} -\varepsilon + 2\varepsilon t \\ 1 \end{pmatrix}, \quad t_0 = 0, \quad t_1 = \frac{1}{2}, \quad t_2 = 1,$$

$$J_1 = \phi, \quad J_2 = \{1, 2\},$$

$$q(t_0) \text{ and } q(t_1) \text{ are in } Q(J_1) = \{q \in R^2 \mid q_1 \leq 0, q_2 \geq 0\},$$

$$q(t_1) \text{ and } q(t_2) \text{ are in } Q(J_2) = R_+^2,$$

$$y(t_0) = x(t_0) = \begin{pmatrix} 1 \\ \varepsilon - 2\varepsilon t_0 \end{pmatrix} = \begin{pmatrix} 1 \\ \varepsilon \end{pmatrix},$$

$$x(t_2) = \begin{pmatrix} 0 \\ 0 \end{pmatrix}.$$

In order to satisfy (3.5), $y(t_1)$ must be $\begin{pmatrix} 0 \\ 0 \end{pmatrix}$. However (3.5) also requires that

$$x(t_1) = \begin{pmatrix} 1 \\ 0 \end{pmatrix}.$$

Hence $x(t_1) \neq y(t_1)$ and the proof of Theorem 3.2 fails. In fact, since

$$\lim_{\varepsilon \rightarrow 0} \frac{\|x(t_2) - x(t_0)\|_\infty}{\|q^2 - q^1\|_\infty} = \lim_{\varepsilon \rightarrow 0} \frac{1}{2\varepsilon} = \infty,$$

the solutions of the problem cannot be Lipschitzian.

We conclude by showing that other linear complementarity problems that can be formulated as linear programs [10] have solutions which are Lipschitzian with respect to their right-hand sides as a consequence of Theorem 2.4. In particular if M satisfies the condition of Theorem 2 of [10] with $c = 0$, that is

$$(3.6) \quad MZ_1 = Z_2, \quad rZ_1 + sZ_2 > 0, \quad (r, s) \geq 0$$

for some $n \times n$ Z -matrices Z_1 and Z_2 , and some n -vectors r and s , then a solution to such a linear complementarity problem is obtained by solving the single linear program

$$\min px \quad \text{s.t.} \quad Mx + q \geq 0, \quad x \geq 0$$

where $p = r + M^T s$, and hence p is independent of q . In the terminology of [17], such a matrix M is called a hidden Z -matrix and is a generalization of Z -matrix which

includes such matrices as those with a strictly dominant diagonal, and all matrices of Table 1 in [10] except cases 12 to 14.

THEOREM 3.5 (Lipschitz continuity of solutions of linear complementarity problems with hidden Z -matrices). *Let M be a hidden Z -matrix, that is M satisfies (3.6). For each q^1 and q^2 in R^n for which (M, q^1) and (M, q^2) are solvable, there exist solutions x^1 of (M, q^1) and x^2 of (M, q^2) such as*

$$\|x^1 - x^2\|_\infty \leq \nu_\beta \left(\begin{matrix} M \\ I \end{matrix}; \phi \right) \|q^1 - q^2\|_\beta$$

where $\|\cdot\|_\beta$ is some norm on R^n and ν_β is defined by (2.20).

Proof. By [10], there exist solutions of (M, q^1) and (M, q^2) which are obtained by solving the linear programs

$$\min \{px \mid Mx + q^1 \geq 0, x \geq 0\},$$

$$\min \{px \mid Mx + q^2 \geq 0, x \geq 0\}$$

where p is a fixed vector independent of q^1 and q^2 . The conclusion of the theorem follows immediately from Theorem 2.4. \square

We note that for the case of a strictly diagonally dominant positive definite matrix M , (M, q) is uniquely solvable for each q in R^n , and the Lipschitz continuity of the solution follows also from either Theorem 3.5 or Theorem 3.3.

REFERENCES

- [1] J.-P. AUBIN, *Lipschitz behavior of solutions to convex optimization problems*, Math. Oper. Res., 9 (1984), pp. 87–111.
- [2] A. BERMAN AND R. J. PLEMMONS, *Nonnegative Matrices in the Mathematical Sciences*, Academic Press, New York, 1979.
- [3] S. D. CONTE AND C. DE BOOR, *Elementary Numerical Analysis*, third edition, McGraw-Hill, New York, 1980.
- [4] W. COOK, A. M. H. GERARDS, A. SCHRIJVER AND É. TARDOS, *Sensitivity results in integer linear programming*, Math. Programming, 34 (1986), pp. 251–264.
- [5] R. W. COTTLE AND G. B. DANTZIG, *Complementary pivot theory in mathematical programming*, Linear Algebra Appl., 1 (1968), pp. 103–125.
- [6] D. GALE, *The Theory of Linear Economic Models*, McGraw-Hill, New York, 1960.
- [7] W. W. HAGER, *Lipschitz continuity for constrained processes*, this Journal, 17 (1979), pp. 321–338.
- [8] A. J. HOFFMAN, *On approximate solutions of systems of linear inequalities*, J. Res. Nat. Bur. Standards, 49 (1952), pp. 263–265.
- [9] A. S. HOUSEHOLDER, *The Theory of Matrices in Numerical Analysis*, Blaisdell Publishing, New York, 1964.
- [10] O. L. MANGASARIAN, *Characterization of linear complementarity problems as linear programs*, Math. Programming Stud., 7 (1978), pp. 74–87.
- [11] ———, *A condition number for linear inequalities and linear programs*, in Methods of Operations Research 43, Proceedings of 6. Symposium über Operations Research, Universität Augsburg, September 7–9, 1981, G. Bamberg and O. Opitz, eds., Verlagsgruppe Athenäum/Hain/Scrip-tor/Hanstein, Königstein 1981, pp. 3–15.
- [12] O. L. MANGASARIAN AND T.-H. SHIAU, *A variable-complexity norm maximization problem*, SIAM J. Algebraic Discrete Methods, 7 (1986), pp. 455–461.
- [13] K. G. MURTY, *On the number of solutions of the complementarity problem and spanning properties of complementarity cones*, Linear Algebra Appl., 5 (1972), pp. 65–108.
- [14] ———, *Linear and Combinatorial Programming*, John Wiley, New York, 1976.
- [15] ———, *Linear Complementarity, Linear and Nonlinear Programming*, Heldermann Verlag, West Berlin, 1985.
- [16] J. M. ORTEGA, *Numerical Analysis: A Second Course*, Academic Press, New York, 1972.

- [17] J.-S. PANG, *Hidden Z -matrices with positive principal minors*, Linear Algebra Appl., 23 (1979), pp. 201–215.
- [18] S. M. ROBINSON, *Bounds for error in the solution set of a perturbed linear program*, Linear Algebra Appl., 6 (1973), pp. 69–81.
- [19] ———, *Generalized equations and their solutions, Part I: Basic theory*, Math. Programming Stud., 10 (1979), pp. 128–141.
- [20] G. W. STEWART, *Introduction to Matrix Computations*, Academic Press, New York, 1973.

CHANDRASEKHAR EQUATIONS FOR INFINITE DIMENSIONAL SYSTEMS*

KAZUFUMI ITO[†] AND ROBERT K. POWERS[‡]

Abstract. In this paper we derive the Chandrasekhar equations for linear time invariant systems defined on Hilbert spaces using a functional analytic technique. An important consequence of this is that the solution to the evolutionary Riccati equation is strongly differentiable in time and one can define a “strong” solution of the Riccati differential equation. A detailed discussion on the linear quadratic optimal problem for hereditary differential systems is also included.

Key words. Chandrasekhar equations, Riccati operator, regularity results, infinite dimensional systems

AMS(MOS) subject classification. 49

1. Introduction. The Chandrasekhar equations [14] are an alternative form to the Riccati equations from which the optimal feedback gain operator may be calculated directly. If the system has a small number of inputs and outputs, the Chandrasekhar algorithm offers significant reduction in the computational complexity for determining the optimal feedback gain. As observed in [20], this is much more evident in the infinite dimensional case if the optimal feedback gain operator is calculated numerically using some approximation method. In this case, the number of states grows linearly to the order of approximation.

The purpose of this paper is to derive Chandrasekhar equations for systems defined by evolution equations on Hilbert spaces in which the input and output operators are assumed to be bounded. The form of the Chandrasekhar equations derived immediately implies that the solution of the associated Riccati equation is strongly differentiable in time, and it allows us to define a “strong” solution of the Riccati equation. Another important consequence of this is that the optimal control for the linear quadratic regulator (LQR) problem is continuously differentiable if the initial datum is sufficiently smooth.

The Chandrasekhar equations for infinite dimensional systems have been discussed in [4] and [7] using a Lions-type framework [17]. However, the equations derived in [4] and [7] are satisfied in the distributional sense. In [22], Sorine derived a set of Chandrasekhar equations satisfied in a strong sense for parabolic systems. Sorine’s derivation relied on the analyticity of the semigroup and thus does not apply to general systems. Our approach differs from those above in that it uses an approximation technique. A sequence of approximating optimal control problems is chosen for which the Chandrasekhar equations may be derived as in the finite dimensional case (see [6], [14], and [16]). Convergence is then established and the appropriate equations are shown to be satisfied. In this paper, our considerations are restricted to the LQR problem, but the results are also applicable to the Kalman filtering problem [8].

* Received by the editors January 22, 1985; accepted for publication (in revised form) March 11, 1986. This research was supported by the National Aeronautics and Space Administration under NASA contract NAS1-17070 while the authors were in residence at the Institute for Computer Application in Science and Engineering, NASA Langley Research Center, Hampton, Virginia 23665.

[†] Division of Applied Mathematics, Brown University, Providence, Rhode Island 02912.

[‡] Institute for Computer Applications in Science and Engineering, NASA Langley Research Center, Hampton, Virginia 23665.

A number of test computations have been successfully carried out using hereditary differential systems [20] and viscous damped cantilevered Euler-Bernoulli beam equations with tip mass as our model examples. Our expectations on the computational reduction using the Chandrasekhar algorithm were fully realized. Also, see [5] for the numerical approximating scheme for the solution of Chandrasekhar equations in infinite dimensional spaces. We are currently studying the use of the Chandrasekhar algorithm for computing the optimal steady state feedback gain operator combining it with the Kleinman-Newton algorithm. Details of this and other numerical studies will be reported elsewhere.

The contents of the paper are as follows. Section 2 briefly recalls the linear quadratic problem and characterizes the optimal control (see [2], [9], and [17] for a survey of the literature). In § 3 a characterization of the Riccati operator is derived and used to obtain the Chandrasekhar equations. Regularity results for the Riccati operator and optimal control are discussed in § 4. As a specific example we discuss in § 5 the linear quadratic optimal control problem for hereditary differential systems in which the input and output spaces are finite dimensional. Because of the smoothing property of the solution semigroup, results stronger than those of the general problem are obtained.

The notation used in this paper is standard. The symbol $\langle \cdot, \cdot \rangle$ stands for the inner product in a Hilbert space where the underlying space will be understood from the context. Also, $\| \cdot \|$ denotes the norm for elements of a Banach space and for operators between Banach spaces, while $|\cdot|$ denotes the Euclidean norm. The adjoint of a densely defined operator \mathcal{A} from one Hilbert space to another is denoted by \mathcal{A}^* .

2. Riccati equations. Let Z , U , and Y be Hilbert spaces. We consider the evolution equation on Z

$$(2.1) \quad \begin{aligned} \frac{d}{dt}z(t) &= \mathcal{A}z(t) + \mathcal{B}u(t), & t \geq 0, \\ z(0) &= z \in Z \end{aligned}$$

where $u(\cdot)$ is a U -valued, square integrable (control) function and \mathcal{A} is the infinitesimal generator of a strongly continuous semigroup $S(t)$ on Z . The Y -valued (observation) function y is given by

$$(2.2) \quad y(t) = \mathcal{C}z(t), \quad t \geq 0.$$

We assume that $\mathcal{B} \in \mathcal{L}(U, Z)$ and $\mathcal{C} \in \mathcal{L}(Z, Y)$.

For any $T \geq 0$, if u is differentiable almost everywhere on $[0, T]$, $\dot{u} \in L_1(0, T; U)$ and $z \in \mathcal{D}(\mathcal{A})$, then the initial value problem (2.1) has a unique “strong” solution [19, Cor. 2.10] in the sense that z is differentiable almost everywhere (a.e.) on $[0, T]$ with $\dot{z} \in L_1(0, T; Z)$ and (2.1) holds a.e. on $[0, T]$. It follows from Corollary 2.2 in [19] that (2.1) has at most one solution and if it has a solution, this solution is given by

$$(2.3) \quad z(t) = S(t)z + \int_0^t S(t-s)\mathcal{B}u(s) ds,$$

which we shall call the mild solution of (2.1). Moreover, the mild solution satisfies the “weak” differential equation

$$\frac{d}{dt}\langle z(t), x \rangle = \langle z(t), \mathcal{A}^*x \rangle + \langle \mathcal{B}u(t), x \rangle \quad \text{for all } x \in \mathcal{D}(\mathcal{A}^*).$$

Consider the linear quadratic optimal control problem on a finite time interval: for given initial data $z \in Z$, choose the control $u \in L_2(0, T; \mathbb{R}^m)$ that minimizes the cost functional

$$(2.4) \quad J(u, [0, T]) = \int_0^T (\|y(t)\|^2 + \|u(t)\|^2) dt + \langle Gz(T), z(T) \rangle_Z$$

where G is a nonnegative (definite), self-adjoint operator on Z and z is the mild solution to (2.1). The next theorem, which characterizes the optimal control, follows from [2], [9] and [23].

THEOREM 2.1. *The optimal control u^0 of (2.4) is given by*

$$(2.5) \quad u^0(t) = -\mathcal{B}^* \Pi(t) z^0(t), \quad t \geq 0$$

where $\Pi(t)$, $t \leq T$, is strongly continuous on Z . Moreover, $\Pi(t)$ is the unique solution within the class of nonnegative self-adjoint operators for which $\langle \Pi(t)z, z \rangle$ is absolutely continuous for $z \in \mathcal{D}(\mathcal{A})$, and satisfies the “weak differential” Riccati equation

$$(2.6) \quad \frac{d}{dt} \langle \Pi(t)z, z \rangle + 2\langle \mathcal{A}z, \Pi(t)z \rangle - \langle \mathcal{B}^* \Pi(t)z, \mathcal{B}^* \Pi(t)z \rangle + \langle \mathcal{C}z, \mathcal{C}z \rangle = 0, \quad \text{for all } z \in \mathcal{D}(\mathcal{A}).$$

$$\Pi(T) = G$$

If $\mathcal{U}(\cdot, \cdot)$ denotes the perturbed evolution operator of the semigroup $S(t)$ by $-\mathcal{B}\mathcal{B}^*\Pi$, then for $z \in Z$

$$(2.7) \quad \mathcal{U}(s, t)z = S(s-t)z - \int_t^s S(s-\sigma)\mathcal{B}\mathcal{B}^*\Pi(\sigma)\mathcal{U}(\sigma, t)z d\sigma,$$

$\Pi(t)$ satisfies

$$(2.8) \quad \Pi(t)z = S^*(T-t)G\mathcal{U}(T, t)z + \int_t^T S^*(\sigma-t)\mathcal{C}^*\mathcal{C}\mathcal{U}(\sigma, t)z d\sigma$$

and

$$z^0(t) = \mathcal{U}(t, 0)z.$$

3. Chandrasekhar equations. From here on, we assume that $Gz \in \mathcal{D}(\mathcal{A}^*)$ for all $z \in Z$. By the closed graph theorem \mathcal{A}^*G is then a bounded operator on Z . Let us define a bounded self-adjoint operator Q on Z by

$$(3.1) \quad \langle Qx, y \rangle = \langle \mathcal{A}^*Gx, y \rangle + \langle x, \mathcal{A}^*Gy \rangle - \langle \mathcal{B}^*Gx, \mathcal{B}^*Gy \rangle + \langle \mathcal{C}x, \mathcal{C}y \rangle \quad \text{for all } x, y \in Z.$$

The main result of this paper is given in the following theorem.

THEOREM 3.1. *If $\Pi(t)$, $t \leq T$ is the solution to the Riccati equation (2.6), then for $z \in Z$*

$$(3.2) \quad \Pi(t)z = Gz + \int_t^T \mathcal{U}^*(T, s)Q\mathcal{U}(T, s)z ds.$$

Conversely, if $\Pi(t)$ satisfies (3.2) with (2.7), then it satisfies the Riccati equation (2.6).

Proof. If \mathcal{A} is a bounded linear operator on Z , then

$$S(t) = e^{\mathcal{A}t} = \sum_{n=0}^{\infty} \frac{(\mathcal{A}t)^n}{n!}$$

and $t \rightarrow S(t)$ is differentiable in norm. Hence the same arguments as given in [14] for the finite dimensional system allow us to show that the theorem holds for such a case. Consider the Yosida approximation of \mathcal{A} given by

$$\mathcal{A}_\lambda = \lambda \mathcal{A} (\lambda I - \mathcal{A})^{-1} \quad \text{for } \lambda \in \mathbb{R}^+ \cap \rho(\mathcal{A}).$$

Then \mathcal{A}_λ is a bounded linear operator on Z and from Theorem 5.5 in [19]

$$e^{\mathcal{A}_\lambda t} z \rightarrow S(t)z \quad \text{as } \lambda \rightarrow \infty (\text{strongly}), \quad z \in Z$$

uniformly on bounded t -intervals. Note that

$$\mathcal{A}_\lambda^* = \lambda \mathcal{A}^* (\lambda I - \mathcal{A}^*)^{-1}.$$

Indeed, for $x \in \mathcal{D}(\mathcal{A})$ and $y \in Z$

$$\langle \mathcal{A}_\lambda x, y \rangle = \langle \lambda (\lambda I - \mathcal{A})^{-1} \mathcal{A} x, y \rangle = \langle x, \lambda \mathcal{A}^* (\lambda I - \mathcal{A}^*)^{-1} y \rangle.$$

But since $\mathcal{D}(\mathcal{A})$ is dense in Z , this shows that $\mathcal{A}_\lambda^* = \lambda \mathcal{A}^* (\lambda I - \mathcal{A}^*)^{-1}$. Thus Theorem 5.5 in [19] again implies that

$$e^{\mathcal{A}_\lambda^* t} z \rightarrow S^*(t)z, \quad z \in Z$$

uniformly on bounded t -intervals.

Consider the approximate problem $(\mathcal{A}_\lambda, \mathcal{B}, \mathcal{C})$ for which the theorem holds. If $\Pi_\lambda(t)$ and $\mathcal{U}_\lambda(\cdot, \cdot)$ denote the solution of the Riccati equation and the perturbed evolution operator corresponding to the perturbation of $e^{\mathcal{A}_\lambda t}$ by $-\mathcal{B}\mathcal{B}^*\Pi_\lambda(t)$, respectively, then

$$\Pi_\lambda(t)z = Gz + \int_t^T \mathcal{U}_\lambda^*(T, s) Q_\lambda \mathcal{U}_\lambda(T, s)z \, ds \quad \text{for } z \in Z$$

where

$$Q_\lambda = \mathcal{A}_\lambda^* G + G \mathcal{A}_\lambda - G \mathcal{B} \mathcal{B}^* G + \mathcal{C}^* \mathcal{C}.$$

It follows from Theorem 6.1 in Gibson [13] that $\Pi_\lambda(t)$ converges strongly to $\Pi(t)$ for $t \leq T$, and the convergence is uniform on bounded t -intervals. Moreover, statement (6.14) in [13] implies that

$$\mathcal{U}_\lambda(t, s)z \rightarrow \mathcal{U}(t, s)z, \quad z \in Z, \quad 0 \leq s \leq t \leq T$$

where the convergence is uniform in t and s . Hence, for all $x \in Z$

$$\begin{aligned} \langle \Pi(t)x, x \rangle &= \lim_{\lambda \uparrow \infty} \langle \Pi_\lambda(t)x, x \rangle = \langle Gx, x \rangle + \lim_{\lambda \uparrow \infty} \int_t^T \langle Q_\lambda \mathcal{U}_\lambda(T, s)x, \mathcal{U}_\lambda(T, s)x \rangle \, ds \\ (3.3) \quad &= \langle Gx, x \rangle + \lim_{\lambda \uparrow \infty} \int_t^T \{ 2 \langle \mathcal{A}^* G \mathcal{U}_\lambda(T, s)x, J_\lambda \mathcal{U}_\lambda(T, s)x \rangle \\ &\quad + \langle (\mathcal{C}^* \mathcal{C} - G \mathcal{B} \mathcal{B}^* G) \mathcal{U}_\lambda(T, s)x, \mathcal{U}_\lambda(T, s)x \rangle \} \, ds \end{aligned}$$

where $J_\lambda = \lambda (\lambda I - \mathcal{A})^{-1}$, $\lambda \in \rho(\mathcal{A})$. Note that

$$J_\lambda \mathcal{U}_\lambda(T, s)x = (J_\lambda - I) \mathcal{U}(T, s)x + J_\lambda (\mathcal{U}_\lambda(T, s) - \mathcal{U}(T, s))x + \mathcal{U}(T, s)x$$

converges strongly to $\mathcal{U}(T, s)x$ for $s \leq T$ since J_λ converges strongly to the identity operator I on Z (see [19]). Since the integrand appearing in (3.3) is uniformly bound in λ and s , the dominated convergence theorem allows us to obtain that for $x \in Z$

$$\begin{aligned} \langle \Pi(t)x, x \rangle &= \langle Gx, x \rangle + \int_t^T \{2\langle \mathcal{A}^*G\mathcal{U}(T, s)x, \mathcal{U}(T, s)x \rangle \\ &\quad + \langle (\mathcal{C}^*\mathcal{C} - G\mathcal{B}\mathcal{B}^*G)\mathcal{U}(T, s)x, \mathcal{U}(T, s)x \rangle\} ds \\ &= \left\langle \left(G + \int_t^T \mathcal{U}^*(T, s)Q\mathcal{U}(T, s) ds \right) x, x \right\rangle, \end{aligned}$$

which completes the proof of the first statement since the operators appearing in both sides of this equation are self-adjoint. Conversely, suppose that $\Pi(t)$, $t \leq T$ satisfies (3.2). Then $t \rightarrow \Pi(t)x$ is continuously differentiable for $x \in Z$ and

$$\begin{aligned} \frac{d}{dt} \langle \Pi(t)x, y \rangle &= \left\langle \frac{d}{dt} \Pi(t)x, y \right\rangle \\ &= \langle \mathcal{U}(T, t)x, Q\mathcal{U}(T, t)y \rangle \end{aligned}$$

for all $x, y \in Z$. Thus, for $x \in \mathcal{D}(\mathcal{A})$

$$\begin{aligned} \langle Qx, x \rangle - \left\langle \frac{d}{dt} \Pi(t)x, x \right\rangle &= \int_t^T \frac{d}{ds} \langle \mathcal{U}(T, s)x, Q\mathcal{U}(T, s)x \rangle ds \\ &= -2 \int_t^T \langle \mathcal{U}(T, s)(\mathcal{A} - \mathcal{B}\mathcal{B}^*\Pi(s))x, Q\mathcal{U}(T, s)x \rangle ds \\ &= -2 \left[\int_t^T \langle \mathcal{U}(T, s)\mathcal{A}x, Q\mathcal{U}(T, s)x \rangle ds \right. \\ &\quad \left. - \int_t^T \langle \mathcal{U}(T, s)\mathcal{B}\mathcal{B}^*\Pi(s)x, Q\mathcal{U}(T, s)x \rangle ds \right] \\ &= -2 \int_t^T \frac{d}{ds} \langle \Pi(s)\mathcal{A}x, x \rangle ds + \int_t^T \frac{d}{ds} \|\mathcal{B}^*\Pi(s)x\|^2 ds \\ &= -2\langle G, \mathcal{A}x \rangle + \langle \mathcal{B}^*Gx, \mathcal{B}^*Gx \rangle \\ &\quad + 2\langle \Pi(t), \mathcal{A}x \rangle + \langle \mathcal{B}^*\Pi(t)x, \mathcal{B}^*\Pi(t)x \rangle. \end{aligned}$$

By the definition of Q , this implies that $\Pi(t)$ satisfies (2.6). Q.E.D.

Remark 3.2. Important in applications is the case $G \equiv 0$. If this occurs, then $Q = \mathcal{C}^*\mathcal{C}$ and

$$\Pi(t)z = \int_t^T L^*(s)L(s)z ds, \quad z \in Z$$

where $L(s) \equiv \mathcal{C}\mathcal{U}(T, s)$. Define the gain operator by $K(t) = \mathcal{B}^*\Pi(t)$. Then (see Gibson [12])

$$(3.4) \quad \mathcal{U}(T, t)z = S(T-t)z - \int_t^T \mathcal{U}(T, s)\mathcal{B}K(s)S(s-t)z ds, \quad z \in Z$$

and the operators $K(t)$ and $L(t)$ jointly satisfy

$$\begin{aligned} K(t)z &= \int_t^T \mathcal{B}L^*(s)L(s)z ds, \\ L(t)z &= \mathcal{C}S(T-t)z - \int_t^T L(s)\mathcal{B}K(s)S(s-t)z ds \end{aligned}$$

for all $z \in Z$, which are the infinite dimensional Chandrasekhar equations in integral form. Since $K(t)z$ and $L(t)x$ are continuously differentiable for $z \in Z$ and $x \in \mathcal{D}(\mathcal{A})$, $K(t)$ and $L(t)$ also satisfy

$$\frac{d}{dt}K(t)z = -\mathcal{B}^*L^*(t)L(t)z, \quad z \in Z,$$

$$K(T) = 0,$$

$$\frac{d}{dt}L(t)x = -L(t)[\mathcal{A} - \mathcal{B}K(t)]x, \quad x \in \mathcal{D}(\mathcal{A}),$$

$$L(t) = \mathcal{C}.$$

Note that these Chandrasekhar differential equations correspond to those derived for finite dimensional systems [14].

4. Strong differential Riccati equation. An important consequence of (3.2) is the following theorem.

THEOREM 4.1. *If $GZ \subset \mathcal{D}(\mathcal{A}^*)$ and $\Pi(t)$, $t \leq T$, is the solution to the Riccati equation (2.6), then for $z \in \mathcal{D}(\mathcal{A})$, $\Pi(t)z$ is the unique strong solution to the Riccati equation (2.6) in the sense that $\Pi(t)z$ is continuously differentiable on $[0, T]$, $\Pi(t)z \in \mathcal{D}(\mathcal{A}^*)$ for $0 \leq t \leq T$ and the strong differentiable Riccati equation*

$$\left(\frac{d}{dt}\Pi(t) + \mathcal{A}^*\Pi(t) + \Pi(t)\mathcal{A} - \Pi(t)\mathcal{B}\mathcal{B}^*\Pi(t) + \mathcal{C}^*\mathcal{C} \right) z = 0 \quad \text{for all } z \in \mathcal{D}(\mathcal{A}),$$

$$\Pi(T) = G$$

is satisfied on $[0, T]$.

Proof. From (2.6), we have for all $x, y \in \mathcal{D}(\mathcal{A})$

$$\langle \Pi(t)x, \mathcal{A}y \rangle = - \left(\frac{d}{dt} \langle \Pi(t)x, y \rangle + \langle \Pi(t)\mathcal{A}x, y \rangle - \langle \mathcal{B}\Pi(t)x, \mathcal{B}\Pi(t)y \rangle + \langle \mathcal{C}x, \mathcal{C}y \rangle \right).$$

From Theorem 3.1, $\Pi(t)x$ is continuously differentiable for $x \in Z$,

$$\langle \Pi(t)x, \mathcal{A}y \rangle = - \left\langle \left(\frac{d}{dt}\Pi(t) + \Pi(t)\mathcal{A} - \Pi(t)\mathcal{B}\mathcal{B}^*\Pi(t) + \mathcal{C}^*\mathcal{C} \right) x, y \right\rangle$$

for all $x, y \in \mathcal{D}(\mathcal{A})$ and the right-hand side of this expression is defined for all $y \in Z$. Thus, since \mathcal{A} is a densely defined closed operator on Z , this implies that for each $t \leq T$, $\Pi(t)x \in \mathcal{D}(\mathcal{A}^*)$ for $x \in \mathcal{D}(\mathcal{A})$ and

$$\left\langle \left(\frac{d}{dt}\Pi(t) + \mathcal{A}^*\Pi(t) + \Pi(t)\mathcal{A} - \Pi(t)\mathcal{B}\mathcal{B}^*\Pi(t) + \mathcal{C}^*\mathcal{C} \right) x, y \right\rangle = 0$$

$$\text{for all } x \in \mathcal{D}(\mathcal{A}) \quad \text{and} \quad y \in Z,$$

which completes the proof. Q.E.D.

The following lemma is concerned with the strong differentiability of $t \rightarrow \mathcal{U}(t, s)x$ on Z . Here we shall give a proof using the standard constructive argument although it can be considerably simplified using the contraction mapping theorem.

LEMMA 4.2. Suppose that $\mathcal{B}(t)$ is an operator on Z such that for $z \in Z$, $\mathcal{B}(t)z$ is continuously differentiable on $[0, T]$. Then $\mathcal{A} + \mathcal{B}(t)$ generates a perturbed evolution operator $V(t, s)$, of the semigroup $S(t)$ on Z and for $z \in \mathcal{D}(\mathcal{A})$ $V(t, s)z \in \mathcal{D}(\mathcal{A})$, $0 \leq s \leq t \leq T$, $V(t, s)z$ is strongly differentiable in t , and

$$(4.1) \quad \frac{\partial}{\partial t} V(t, s)z = (\mathcal{A} + \mathcal{B}(t))V(t, s)z$$

is satisfied for $0 \leq s \leq t \leq T$. Moreover, the derivative $(\partial/\partial t)V(t, s)z$ for $z \in \mathcal{D}(\mathcal{A})$ is jointly continuous in t and s .

Proof. Consider a class Ω of evolution operators on Z as follows: Ω consists of bounded linear operators $V(t, s)$, $0 \leq s \leq t \leq T$ on Z such that

- (i) $V(s, s) = I$, $V(t, r)V(r, s) = V(t, s)$ for $0 \leq s \leq r \leq t \leq T$;
- (ii) $(t, s) \rightarrow V(t, s)$ is strongly continuous for $0 \leq s \leq t \leq T$;
- (iii) for $z \in \mathcal{D}(\mathcal{A})$, $V(t, s)z \in \mathcal{D}(\mathcal{A})$ is strongly differentiable in t and the derivative $(\partial/\partial t)V(t, s)z$ is strongly continuous in t and s for $0 \leq s \leq t \leq T$.

Note that $V^{(0)}(t, s)z = S(t-s)z$, $z \in Z$ belongs to Ω . Define a sequence of evolution operators $V^{(k)}(t, s)$ by

$$(4.2) \quad V^{(k+1)}(t, s)z = S(t-s)z + \int_s^t S(t-\sigma)\mathcal{B}(\sigma)V^{(k)}(\sigma, s)z d\sigma$$

for $z \in Z$ and $0 \leq s \leq t \leq T$.

It then follows from [9], [19] that

$$V^{(k)}(t, s)z \rightarrow V(t, s)z \quad \text{for } z \in Z \text{ and } 0 \leq s \leq t \leq T$$

where the convergence is uniform in t and s . If $V^{(k)}(t, s)$ belongs to the class Ω , then for $z \in \mathcal{D}(\mathcal{A})$, $\mathcal{B}(t)V^{(k)}(t, s)z$ is continuously differentiable in t and

$$\frac{\partial}{\partial t}(\mathcal{B}(t)V^{(k)}(t, s)z) = \dot{\mathcal{B}}(t)V^{(k)}(t, s)z + \mathcal{B}(t)\frac{\partial}{\partial t}V^{(k)}(t, s)z.$$

It now follows from (4.2) and [15, p. 487] that for $z \in \mathcal{D}(\mathcal{A})$, $V^{(k+1)}(t, s)z$ is continuously differentiable in t and satisfies

$$\frac{\partial}{\partial t}V^{(k+1)}(t, s)z = \mathcal{A}V^{(k+1)}(t, s)z + \mathcal{B}(t)V^{(k)}(t, s)z$$

or

$$(4.3) \quad \begin{aligned} \frac{\partial}{\partial t}V^{(k+1)}(t, s)z &= S(t-s)(\mathcal{A} + \mathcal{B}(t))z + \int_s^t S(t-\sigma)\dot{\mathcal{B}}(\sigma)V^{(k)}(\sigma, s)z d\sigma \\ &\quad + \int_s^t S(t-\sigma)\mathcal{B}(\sigma)\frac{\partial}{\partial \sigma}V^{(k)}(\sigma, s)z d\sigma \quad \text{for } 0 \leq s \leq t \leq T. \end{aligned}$$

Hence, by induction, $V^{(k)}(t, x)z$ belongs to Ω for $k \geq 0$. From (4.3)

$$(4.4) \quad \begin{aligned} &\frac{\partial}{\partial t}V^{(k+1)}(t, s)z - \frac{\partial}{\partial t}V^{(k)}(t, s)z \\ &= \int_s^t S(t-\sigma)\dot{\mathcal{B}}(\sigma)(V^{(k)}(\sigma, s) - V^{(k-1)}(\sigma, s))z d\sigma \\ &\quad + \int_s^t S(t-\sigma)\mathcal{B}(\sigma)\left(\frac{\partial}{\partial \sigma}V^{(k)}(\sigma, s)z - \frac{\partial}{\partial \sigma}V^{(k-1)}(\sigma, s)z\right) d\sigma. \end{aligned}$$

By induction on k one easily verifies the estimate

$$\|V^{(k)}(t, s) - V^{(k-1)}(t, s)\| \leq C_1 M_1^k \frac{(t-s)^k}{k!}$$

where

$$C_1 = \max_{0 \leq s \leq T} \|S(s)\| \quad \text{and} \quad M_1 = \max_{0 \leq s \leq t \leq T} \|S(t-s)\mathcal{B}(s)\|.$$

Since $\mathcal{B}(t)z$ is continuous for each $z \in Z$, $\|\mathcal{B}(t)\|$ is uniformly bounded on $[0, T]$. Thus, from (4.4)

$$\begin{aligned} & \left\| \frac{\partial}{\partial t} V^{(k+1)}(t, s)z - \frac{\partial}{\partial t} V^{(k)}(t, s)z \right\| \\ & \leq C_1 M_2 M_1^k \frac{(t-s)^{k+1}}{(k+1)!} \|z\| + M_1 \int_s^t \left\| \frac{\partial}{\partial \sigma} V^{(k)}(\sigma, s)z - \frac{\partial}{\partial \sigma} V^{(k-1)}(\sigma, s)z \right\| d\sigma \end{aligned}$$

where

$$M_2 = \max_{0 \leq s \leq t \leq T} \|S(t-s)\mathcal{B}(s)\|.$$

By induction on k one obtains

$$\begin{aligned} \left\| \frac{\partial}{\partial t} V^{(k)}(t, s)z - \frac{\partial}{\partial t} V^{(k-1)}(t, s)z \right\| & \leq \frac{(t-s)^k}{(k-1)!} M_2 M_1^{k-1} C_1 \|z\| \\ & \quad + \frac{(t-s)^k}{k!} M_1^k C_1 \|\mathcal{A}z\|. \end{aligned}$$

Hence, $(\partial/\partial t)V^{(k)}(t, s)z$ converges to a function of $C(s, T; Z)$ for $0 \leq s \leq t \leq T$ and $z \in \mathcal{D}(\mathcal{A})$ where the convergence is uniform in t and s . Note that the differential operator $(\partial/\partial t)$ on $C(s, T; Z)$ is closed. These facts, when combined with the convergence of $V^{(k)}(t, s)z$ to $V(t, s)z$ in $C(s, T; Z)$, show that for $z \in \mathcal{D}(\mathcal{A})$ $V(t, s)z$ is continuously differentiable in t , and the derivative is jointly continuous in t and s . Since

$$V(t, s)z = S(t-s)z + \int_s^t S(t-\sigma)\mathcal{B}(\sigma)V(\sigma, s)z d\sigma$$

$$\text{for } z \in Z \quad \text{and} \quad 0 \leq s \leq t \leq T,$$

it now follows from [15, p. 487] that for $z \in \mathcal{D}(\mathcal{A})$, $V(t, s)z \in \mathcal{D}(\mathcal{A})$ and

$$\frac{\partial}{\partial t} V(t, s)z = (\mathcal{A} + \mathcal{B}(t))V(t, s)z$$

for $0 \leq s \leq t \leq T$. Q.E.D.

LEMMA 4.3. *If $GZ \subset \mathcal{D}(\mathcal{A}^*)$ and the initial data $z \in \mathcal{D}(\mathcal{A})$, then the optimal control u^0 to (2.4) is continuously differentiable on $[0, T]$ and*

$$(4.5) \quad \frac{d}{dt} u^0(t) = \mathcal{B}^*(\mathcal{A}^* \Pi(t) + \mathcal{C}^* \mathcal{C})z^0(t).$$

Proof. Since $\mathcal{B}^* \Pi(t)x$ is continuously differentiable for $x \in Z$ on $[0, T]$, it follows from Lemma 4.2 that $z^0(t) = \mathcal{U}(t, 0)z$ is continuously differentiable for $z \in \mathcal{D}(\mathcal{A})$. Thus, from (2.5) u^0 is continuously differentiable on $[0, T]$ and

$$\frac{d}{dt} u^0(t) = -\mathcal{B}^* \left(\frac{d}{dt} \Pi(t)z^0(t) + \Pi(t)\dot{z}^0(t) \right).$$

Equation (4.5) now follows from Theorem 4.1. Q.E.D.

5. Hereditary differential system. In this section we discuss the hereditary differential system

$$(5.1) \quad \begin{aligned} \frac{d}{dt}x(t) &= \int_{-r}^0 d\mu(\theta)x(t+\theta) + Bu(t), \quad t \geq t_0, \\ x(t_0) &= \eta \quad \text{and} \quad x(t_0 + \theta) = \phi(\theta), \quad -r \leq \theta < 0 \end{aligned}$$

where $\mu(\cdot)$ is an $n \times n$ matrix valued function of bounded variation which vanishes at $\theta = 0$ and is left continuous on $(-r, 0)$. Without loss of generality we can assume that for $\theta \geq 0$, $\mu(\theta) = 0$ and for $\theta \leq -r$, $\mu(\theta) = \mu(-r)$. B is an $n \times m$ matrix. The observation y is given by

$$(5.2) \quad y(t) = Cx(t)$$

where C is a $p \times n$ matrix. We will denote by Z the product space $\mathbb{R}^n \times L_2(-r, 0; \mathbb{R}^n)$ in this section. Given an element $z \in Z$, $\eta \in \mathbb{R}^n$ and $\phi \in L_2$ denote the two coordinates of z : $z = (\eta, \phi)$. It is well known [3], [11] that for $(\eta, \phi) \in Z$ and u locally square integrable, (5.1) admits a unique solution $x \in L_2(t_0 - r, T; \mathbb{R}^n) \cap H^1(t_0, T; \mathbb{R}^n)$ for any $T \geq t_0$. If $t_0 = 0$, then (5.1) can be formulated as an evolution equation on Z ,

$$(5.3) \quad \frac{d}{dt}z(t) = \mathcal{A}z(t) + \mathcal{B}u(t), \quad t \geq 0$$

where $z(t) = (x(t), x(t + \cdot)) \in Z$, $t \geq 0$ and $\mathcal{B}u = (Bu, 0) \in Z$ for $u \in \mathbb{R}^m$. The infinitesimal generator \mathcal{A} is then defined by

$$\mathcal{D}(\mathcal{A}) = \{(\eta, \phi) \in Z \mid \eta = \phi(0) \text{ and } \dot{\phi} \in L_2\}$$

and for $(\phi(0), \phi) \in \mathcal{D}(\mathcal{A})$

$$\mathcal{A}(\phi(0), \phi) = \left(\int_{-r}^0 d\mu(\theta)\phi(\theta), \dot{\phi} \right),$$

and generates the strong continuous semigroup $S(t)$: $S(t)(\eta, \phi) = (x(t), x(t + \cdot))$, $t \geq 0$, where x is the solution of (5.1) with $t_0 = 0$ and $u \equiv 0$. Within this framework, the observation equation (5.2) is written as

$$(5.4) \quad y(t) = \mathcal{C}z(t), \quad t \geq 0$$

where $\mathcal{C}(\eta, \phi) = C\eta \in \mathbb{R}^p$ for $(\eta, \phi) \in Z$. Thus the system (5.1)–(5.2) is formulated as the model system (2.1)–(2.2) in which $Z = \mathbb{R}^n \times L_2$, $U = \mathbb{R}^m$ and $Y = \mathbb{R}^p$.

The following lemma (see [10], [21]) gives two important properties of the hereditary differential system which shall be used extensively in the subsequent development.

LEMMA 5.1. (i) If X denotes the Hilbert space $\mathcal{D}(\mathcal{A})$ equipped with the graph norm, then $\int_0^t S(t-s)\mathcal{B}u(s) ds$ is an X -valued function continuous in t for each $u \in L_2(0; T; \mathbb{R}^m)$ and continuous in u for each $t \in [0, T]$.

(ii) If γ is a $p \times n$ matrix-valued function of bounded variation on $[-r, 0]$ and \mathcal{K} denotes an operator defined by $\mathcal{K}(\eta, \phi) = \int_{-r}^0 d\gamma(\theta)\phi(\theta)$ for $(\eta, \phi) \in Z$, then there exists a nondecreasing function $M(\cdot): [0, \infty] \rightarrow \mathbb{R}^+$ such that for $z \in Z$

$$(5.5) \quad \int_0^T |\mathcal{K}S(t)z|^2 dt \leq M(T)\|z\|^2.$$

Remark. In (5.5) the expression $\mathcal{H}S(\dot{t})z$ only makes sense when $z \in \mathcal{D}(\mathcal{A})$. However, because of (ii), we will use the expression $\mathcal{H}S(t)z$, $0 \leq t \leq T$ to denote the function in $L_2(0, T; \mathbb{R}^p)$ which is obtained by a continuous extension of the operator:

$$z \in \mathcal{D}(\mathcal{A}) \rightarrow \mathcal{H}S(t)z \in L_2(0, T; \mathbb{R}^p).$$

Let us consider the linear quadratic optimal control problem: for given $(\eta, \phi) \in Z$ choose the control $u \in L_2(t_0, T; \mathbb{R}^m)$ that minimizes the cost functional

$$(5.6) \quad J(u, [t_0, T]) = \int_{t_0}^T (|Cx(t)|^2 + |u(t)|^2) dt + \langle G_0 x(T), x(T) \rangle_{\mathbb{R}^n}$$

where G_0 is a nonnegative, symmetric matrix on \mathbb{R}^n and $x(\cdot)$ is the solution to (5.1). Note that (5.6) can be equivalently written as

$$J(u, [t_0, T]) = \int_{t_0}^T (|\mathcal{G}z(t)|^2 + |u(t)|^2) dt + \langle Gz(T), z(T) \rangle_Z$$

where G is a nonnegative, self-adjoint operator on Z defined by $G(\eta, \phi) = (G_0\eta, 0) \in Z$ for $(\eta, \phi) \in Z$ and $z(\cdot)$ is given by

$$z(t) = S(t - t_0)(\eta, \phi) + \int_{t_0}^t S(t - s)\mathcal{B}u(s) ds, \quad t \geq t_0.$$

Hence Theorem 2.1 applies to the minimization problem (5.6).

It follows from [13], [24] that if $(y, \psi) \in \mathcal{D}(\mathcal{A}^*)$ then

$$\psi(\theta) - (\mu(\theta) - \mu(-r))^T y \in H^1(-r, 0)$$

and

$$\psi(-r) = (\mu((-r)^+) - \mu(-r))^T y.$$

Obviously $Gz \notin \mathcal{D}(\mathcal{A}^*)$ in general. So, Theorem 3.1 does not apply for (5.6) unless $G_0 = 0$. However, as a result of Lemma 5.1 one can extend the results in §§ 3 and 4 to this case. We will discuss such an extension later and for the present consider the case $G_0 = 0$.

If $G_0 = 0$, then the solution $\Pi(t)$ to the Riccati equation (2.6) is given by

$$(5.7) \quad \Pi(t)z = \int_t^T S^*(\sigma - t)\mathcal{C}^*\mathcal{C}\mathcal{U}(\sigma, t)z d\sigma.$$

Let \mathcal{A}_T be the infinitesimal generator on Z defined by $\mathcal{D}(\mathcal{A}_T) = \mathcal{D}(\mathcal{A})$ and for $\phi \in H^1$

$$\mathcal{A}_T(\phi(0), \phi) = \left(\int_{-r}^0 d\mu^T(\theta)\phi(\theta), \dot{\phi} \right)$$

and let $S_T(t)$ denote the C_0 -semigroup generated by \mathcal{A}_T . Define the structural operator \mathcal{F} on Z by

$$\mathcal{F}(\eta, \phi) = \left(\eta, \int_{-r}^{\theta-} d\mu(\xi)\phi(\xi - \theta) \right) \quad \text{for } (\eta, \phi) \in Z.$$

Then, the following result has been proven by Manitius [18].

THEOREM 5.2.

- (i) $\mathcal{F}S(t) = S_T^*(t)\mathcal{F}$, $\mathcal{F}^*S_T(t) = S^*(t)\mathcal{F}^*$, $t \geq 0$.
- (ii) If $z \in \mathcal{D}(\mathcal{A})$, then $\mathcal{F}z \in \mathcal{D}(\mathcal{A}_T^*)$ and $\mathcal{A}_T^*\mathcal{F}z = \mathcal{F}\mathcal{A}z$.
- (iii) If $z \in \mathcal{D}(\mathcal{A}_T)$, then $\mathcal{F}^*z \in \mathcal{D}(\mathcal{A}^*)$ and $\mathcal{A}^*\mathcal{F}^*z = \mathcal{F}^*\mathcal{A}_Tz$.

Since $\mathcal{C}^* = \mathcal{F}^*\mathcal{C}^*$, it follows from (5.7) and Theorem 5.2 that

$$\Pi(t)z = \mathcal{F}^* \int_t^T S_T(\sigma - t)\mathcal{C}^*\mathcal{C}\mathcal{U}(\sigma, t)z d\sigma.$$

Note that $\mathcal{C}^*y = (C^Ty, 0) \in Z$ for $y \in \mathbb{R}^p$. Thus from (i) of Lemma 5.1 and (iii) of Theorem 5.2, $\Pi(t)z \in \mathcal{D}(\mathcal{A}^*)$ for $z \in Z$. Moreover, since the evolution operator $\mathcal{U}(\sigma, t)$ is jointly continuous for $0 \leq t \leq \sigma \leq T$, $\mathcal{A}^*\Pi(t)z$ is strongly continuous in Z for $z \in Z$, and hence $\Pi(t)\mathcal{A}$ has a bounded extension to all of Z . The next result now follows from Theorem 4.1 and Lemma 4.3 (see [10], [21] for a different derivation of this result).

THEOREM 5.3. *If $G_0 = 0$, then for $z \in Z$, $\Pi(t)z$ is a unique strong solution to the Riccati equation in the sense that $\Pi(t)z$ is continuously differentiable on $[0, T)$, $\Pi(t)z \in \mathcal{D}(\mathcal{A}^*)$ for $0 \leq t \leq T$, and the strong differential Riccati equation*

$$\left(\frac{d}{dt} \Pi(t) + \mathcal{A}^* \Pi(t) + \Pi(t) \mathcal{A} - \Pi(t) \mathcal{B} \mathcal{B}^* \Pi(t) + \mathcal{C}^* \mathcal{C} \right) z = 0 \quad \text{for all } z \in Z$$

is satisfied on $[0, T)$. Moreover, the optimal control $u^0(\cdot)$ to (5.6) is continuously differentiable on $(0, T]$ for $(\eta, \phi) \in Z$.

Proof. From (4.8), if $z = (\eta, \phi) \in \mathcal{D}(\mathcal{A})$, then u^0 is continuously differentiable and

$$\dot{u}^0(t) = \mathcal{B}^*(\mathcal{A}^* \Pi(t) + \mathcal{C}^* \mathcal{C}) z^0(t),$$

$$z^0(t) = \mathcal{U}(t, 0)(\eta, \phi).$$

It has been proven that $\mathcal{A}^*\Pi(t)z$ is strongly continuous in Z for $z \in Z$. So, the theorem follows since $\mathcal{D}(\mathcal{A})$ is dense in Z and $\mathcal{U}(t, 0)$ is continuous on Z for $t \geq 0$. Q.E.D.

Let us turn to the case $G_0 \neq 0$. Consider the λ th approximate problem to (5.6) in which the cost functional is given by

$$(5.8) \quad J^\lambda(u, [t_0, T]) = \int_{t_0}^T (|\mathcal{C}z(t)|^2 + |u(t)|^2) dt + \langle G_\lambda z(T), z(T) \rangle_Z$$

where $G_\lambda = J_\lambda^* G J_\lambda$ and $J_\lambda = \lambda(\lambda I - \mathcal{A})^{-1}$ for $\lambda \in \rho(\mathcal{A})$. Note that $G_\lambda Z \subset \mathcal{D}(\mathcal{A}^*)$ and $G_\lambda \rightarrow G$ in trace norm since G has a finite rank. If $\Pi_\lambda(t)$, $t \leq T$ denotes the solution of the Riccati equation associated with the problem (5.8), then it follows from Theorem 3.1 that

$$\Pi_\lambda(t)z = G_\lambda z + \int_t^T \mathcal{U}_\lambda^*(T, s) Q_\lambda \mathcal{U}_\lambda(T, s) z ds, \quad z \in Z$$

where Q_λ is a self-adjoint operator on Z defined by

$$Q_\lambda = \mathcal{A}^* G_\lambda + G_\lambda \mathcal{A} - G_\lambda \mathcal{B} \mathcal{B}^* G_\lambda + \mathcal{C}^* \mathcal{C}.$$

Such a representation for Q_λ exists since $G_\lambda \mathcal{A}$ can be extended to all Z via (3.1). If we denote the optimal control for the original problem (5.6) by u^0 and the optimal control for the λ th approximate problem (5.8) by u_λ , it follows from [13, pp. 114–115] that u_λ converges strongly to u^0 in $L_2(t_0, T; \mathbb{R}^m)$, and the convergence is uniform in t_0 for $0 \leq t_0 \leq T$.

The following three results are essential to discuss the extension of Theorem 3.1 and Lemma 4.3 to the case when $G_0 \neq 0$.

LEMMA 5.4. *$G\mathcal{A}$ has a bounded extension to all elements $z \in Z$ of the form $z = (\phi(0), \phi)$ with $\phi \in C(-r, 0; \mathbb{R}^n)$ and there exists a nondecreasing function $M(\cdot): [0, \infty) \rightarrow \mathbb{R}^+$ such that for $z \in Z$*

$$(5.9) \quad \int_0^T |G\mathcal{A}S(t)z|^2 dt \leq M(T) \|z\|^2.$$

Proof. For $z = (\phi(0), \phi) \in \mathcal{D}(\mathcal{A})$

$$\begin{aligned} |G\mathcal{A}z| &= \left| G \left(\int_{-r}^0 d\mu(\theta) \phi(\theta), \dot{\phi} \right) \right| = \left| G_0 \int_{-r}^0 d\mu(\theta) \phi(\theta) \right| \\ &\leq |G_0| \int_{-r}^0 |d\mu| \|\phi\|_{C(-r, 0; \mathbb{R}^n)}. \end{aligned}$$

Since $H^1(-r, 0; \mathbb{R}^n)$ is dense in $C(-r, 0; \mathbb{R}^n)$, $G\mathcal{A}$ has the prescribed extension. Upon identifying $G\mathcal{A}$ with \mathcal{H} of Lemma 5.1, (5.9) follows. Q.E.D.

LEMMA 5.5. For $x, y \in Z$

$$\langle G_\lambda \mathcal{A} \mathcal{U}_\lambda(T, t)x, \mathcal{U}_\lambda(T, t)y \rangle \rightarrow \langle G\mathcal{A} \mathcal{U}(T, t)x, \mathcal{U}(T, t)y \rangle \quad \text{in } L_2(t_0, T).$$

Remark. To be precise, Lemma 5.4 only extends $G_\lambda \mathcal{A}$ and $G\mathcal{A}$ to $x = (\phi(0), \phi)$ such that $\phi \in C(-r, 0; \mathbb{R}^n)$. However, as functions in $L_2(t_0, T)$, the inner products may be extended to all Z .

Proof. First note that $\mathcal{U}_\lambda(T, t)z$ converges strongly to $\mathcal{U}(T, t)z$ for $z \in Z$ and the convergence is uniform in t , and that

$$\mathcal{U}_\lambda(T, t)z = S(T-t)z + \int_t^T S(T-s)\mathcal{B}u_\lambda(s) ds$$

where $u_\lambda(\cdot)$ is the optimal control for the λ th approximate (5.8) on the time interval $[t, T]$ with given initial condition $z \in Z$. Since J_λ converges strongly to I as $\lambda \rightarrow \infty$ on $X = \mathcal{D}(\mathcal{A})$, it follows from (i) of Lemma 5.1 and the fact that $u_\lambda \rightarrow u^0$ in $L_2(t, T; \mathbb{R}^m)$ that for $t \leq T$

$$f_\lambda(t) = J_\lambda \int_t^T S(T-s)\mathcal{B}u_\lambda(s) ds$$

converges strongly to

$$f^0(t) = \int_t^T S(T-s)\mathcal{B}u_0 \rightarrow u^0(s) ds$$

in X . Since $\|f_\lambda(t)\|_X$ is uniformly bounded in λ and $t \in [t_0, T]$, by the dominated convergence theorem, $f_\lambda(t)$ converges strongly to $f^0(t)$ in $L_2(t_0, T; X)$. Hence $\langle G\mathcal{A}f_\lambda(t), J_\lambda \mathcal{U}_\lambda(T, t)y \rangle$ converges strongly to $\langle G\mathcal{A}f^0(t), \mathcal{U}(T, t)y \rangle$ in $L_2(t_0, T)$ for $y \in Z$. The remainder of the proof is to show that for $z, y \in Z$

$$(5.10) \quad \langle G\mathcal{A}J_\lambda S(T-t)z, J_\lambda \mathcal{U}_\lambda(T, t)y \rangle \rightarrow \langle G\mathcal{A}S(T-t)z, \mathcal{U}(T, t)y \rangle \quad \text{in } L_2(t_0, T).$$

As in Lemma 5.4, it can be shown that

$$\int_{t_0}^T |G\mathcal{A}J_\lambda S(T-t)z|^2 dt \leq M \|z\|^2, \quad z \in Z,$$

since $\|J_\lambda\|$ is bounded uniformly in λ . The desired result follows from direct applications of the triangle inequality and the dominated convergence theorem.

LEMMA 5.6. There exists a finite rank (\tilde{p}) operator \mathcal{H} on Z and a nonsingular diagonal matrix Λ on $\mathbb{R}^{\tilde{p}}$ such that

$$2\langle G\mathcal{A}z, z \rangle + \langle (\mathcal{C}^* \mathcal{C} - G\mathcal{B}\mathcal{B}^* G)z, z \rangle = \langle \Lambda \mathcal{H}z, \mathcal{H}z \rangle_{\mathbb{R}^{\tilde{p}}} \quad \text{for } z \in \mathcal{D}(\mathcal{A})$$

and \mathcal{H} can be continuously extended to all elements $z \in Z$ of the form $z = (\phi(0), \phi)$ with $\phi \in C(-r, 0; \mathbb{R}^n)$.

Proof. Let X' denote the strong dual space of X . We identify Z with its dual, so that $X \subset Z \subset X'$. If j is the canonical injection from X into Z : $j\phi = (\phi(0), \phi) \in Z$, $\phi \in X$, then j is an embedding from X into Z ; i.e., j is injective and $j(X)$ is dense in Z ; thus it follows from Prop. 4 in [1, p. 65] that j' from Z to X' and $j'j$ from X to X' are embeddings:

$$X \xrightarrow{j} Z \xrightarrow{j'} X'$$

and the bilinear form $(x, y)_{X', X}$ on $X' \times X$ is the unique extension by continuity of the scalar product (x, y) of Z restricted to $Z \times X$. Here (\cdot) stands for dual operators. Let us define an operator $Q \in \mathcal{L}(X, X')$ by

$$Q = \mathcal{A}'G + j'G\mathcal{A} - j'G\mathcal{B}\mathcal{B}^*Gj + j'\mathcal{C}^*\mathcal{C}j.$$

If i is the norm-preserving canonical map from X' into X , then iQ is a self-adjoint operator on X . Indeed,

$$\langle iQx, y \rangle_X = \langle Qx, y \rangle_{X', X} = \langle x, Qy \rangle_{X, X'} = \langle x, iQy \rangle_X.$$

Since G and \mathcal{C} have finite rank, Q has a finite rank, and so iQ does also. Suppose $\text{rank}(iQ) = \tilde{p}$. Then there exists an operator \mathcal{H} on X and a nonsingular diagonal matrix Λ on $\mathbb{R}^{\tilde{p}}$ such that

$$iQz = \mathcal{H}^* \Lambda \mathcal{H}z \quad \text{for all } z \in X = \mathcal{D}(\mathcal{A}).$$

It now follows that for $z \in X$

$$\langle Qz, z \rangle_{X', X} = \langle iQz, z \rangle_X = \langle \mathcal{H}^* \Lambda \mathcal{H}z, z \rangle_X = \langle \Lambda \mathcal{H}z, \mathcal{H}z \rangle_{\mathbb{R}^{\tilde{p}}}.$$

The proof is completed if we note that

$$\langle Qz, z \rangle_{X', X} = 2\langle G\mathcal{A}z, z \rangle_Z + \langle (\mathcal{C}^*\mathcal{C} - G\mathcal{B}\mathcal{B}^*G)z, z \rangle_Z$$

and that from Lemma 5.4 the right-hand side of this equality is continuous on $\phi \in C(-r, 0; \mathbb{R}^n)$. Q.E.D.

The next theorem gives the extension of Remark 3.2 and Theorem 3.1 to the case $G_0 \neq 0$.

THEOREM 5.7. *If $\Pi(t)$, $t \leq T$, is the solution of the Riccati equation (2.6) with $G(\eta, \phi) = (G_0\eta, 0)$ for $(\eta, \phi) \in Z$, then for $z \in Z$*

$$\Pi(t)z = Gz + \int_t^T (\mathcal{H}\mathcal{U}(T, s))^* \Lambda \mathcal{H}\mathcal{U}(T, s)z \, ds$$

where \mathcal{H} and Λ are defined in Lemma 5.6.

Proof. Recall that for $t \leq T$ and $z \in Z$

$$\Pi_\lambda(t)z = G_\lambda z + \int_t^T \mathcal{U}_\lambda^*(T, s)Q_\lambda \mathcal{U}_\lambda(T, s)z \, ds.$$

Since $\Pi_\lambda(t)$, $t \leq T$ converges strongly to $\Pi(t)$, uniformly on bounded t -intervals, for $t \leq T$ and $x, y \in Z$

$$\begin{aligned} \langle \Pi(t)x, y \rangle &= \lim_{\lambda \uparrow \infty} \langle \Pi_\lambda(t)x, y \rangle \\ &= \lim_{\lambda \uparrow \infty} \left(\langle G_\lambda x, y \rangle + \int_t^T \{ \langle \mathcal{A}^* G_\lambda \mathcal{U}_\lambda(T, s)x, \mathcal{U}_\lambda(T, s)y \rangle \right. \\ &\quad \left. + \langle \mathcal{U}_\lambda(T, s)x, \mathcal{A}^* G_\lambda \mathcal{U}_\lambda(T, s)y \rangle \right. \\ &\quad \left. + \langle (\mathcal{C}^*\mathcal{C} - G_\lambda \mathcal{B}\mathcal{B}^*G_\lambda) \mathcal{U}_\lambda(T, s)x, \mathcal{U}_\lambda(T, s)y \rangle \} \, ds \right). \end{aligned}$$

Hence, from Lemma 5.5 and the fact that $\mathcal{U}_\lambda(T, s)z$ converges strongly to $\mathcal{U}(T, s)z$ for $z \in Z$ and the convergence is uniform in s , the dominated convergence theorem allows us to show that for $t \leq T$ and $x, y \in Z$

$$\begin{aligned} \langle \Pi(t)x, y \rangle = & \langle Gx, y \rangle + \int_t^T \{ \langle G\mathcal{A}\mathcal{U}(T, s)x, \mathcal{U}(T, s)y \rangle + \langle \mathcal{U}(T, s)x, G\mathcal{A}\mathcal{U}(T, s)y \rangle \\ & + \langle (\mathcal{C}^*\mathcal{C} - G\mathcal{B}\mathcal{B}^*G)\mathcal{U}(T, s)x, \mathcal{U}(T, s)y \rangle \} ds. \end{aligned}$$

Since $\mathcal{U}(T, t)z \in \mathcal{D}(\mathcal{A})$ for $z \in \mathcal{D}(\mathcal{A})$, it follows from Lemma 5.6 that for $x, y \in \mathcal{D}(\mathcal{A})$

$$(5.11) \quad \langle \Pi(t)x, y \rangle = \langle Gx, y \rangle + \int_t^T \langle \Lambda \mathcal{H}\mathcal{U}(T, s)x, \mathcal{H}\mathcal{U}(T, s)y \rangle ds.$$

But since \mathcal{H} can be continuously extended to all elements z of the form $z = (\phi(0), \phi)$ with $\phi \in C(-r, 0; \mathbb{R}^n)$, it follows from (ii) of Lemma 5.1 and the arguments in the proof of Lemma 5.5 that (5.11) holds for all $x, y \in Z$. Q.E.D.

The following results are concerned with the differentiability of the optimal control $u^0(\cdot)$ of (5.8).

THEOREM 5.8. *For $z \in Z$ the optimal control $u^0(\cdot)$ of (5.8) is absolutely continuous on $[t_0, T]$ with $\dot{u}^0 \in L_2(t_0, T; \mathbb{R}^m)$.*

Proof. It follows from Theorem 4.3 that if $z = (\eta, \phi) \in \mathcal{D}(\mathcal{A})$, then $u_\lambda(\cdot)$ is continuously differentiable on $[t_0, T]$ and is given by

$$\frac{d}{dt} u_\lambda(t) = \mathcal{B}^*(\mathcal{A}^*\Pi_\lambda(t) + \mathcal{C}^*\mathcal{C})\mathcal{U}_\lambda(t, t_0)z.$$

From (2.8)

$$\Pi_\lambda(t)z = S^*(T-t)G_\lambda\mathcal{U}_\lambda(T, t)z + \int_t^T S^*(\sigma-t)\mathcal{C}^*\mathcal{C}\mathcal{U}_\lambda(\sigma, t)z d\sigma \quad \text{for } z \in Z.$$

Note that $G_\lambda Z \subset \mathcal{D}(\mathcal{A}^*)$. Hence, using the same arguments as those in the proof of Lemma 4.3, one can show that $\Pi_\lambda(t)z \in \mathcal{D}(\mathcal{A}^*)$ for $z \in Z$ and $t \leq T$, and moreover, that $\mathcal{A}^*\Pi_\lambda(t)$ is strongly continuous on $[t_0, T]$. Since $\mathcal{U}_\lambda(t, \cdot)$ is strongly continuous on Z , this fact along with the closedness of the differential operator (d/dt) on $C(t_0, T; \mathbb{R}^m)$, shows that for $z \in Z$ $u_\lambda(t)$ is continuously differentiable on $[t_0, T]$. We now note that for $t \leq T$

$$\begin{aligned} \mathcal{B}^*\mathcal{A}^*S^*(T-t)G_\lambda &= \mathcal{B}^*\mathcal{A}^*S^*(T-t)J_\lambda^*GJ_\lambda \\ &= \mathcal{B}^*\mathcal{A}^*S^*(T-t)J_\lambda^*\mathcal{F}^*GJ_\lambda \quad (\text{using } \mathcal{F}^*G = G) \\ &= \mathcal{B}^*\mathcal{A}^*S^*(T-t)\mathcal{F}^*I_\lambda GJ_\lambda \\ &= \mathcal{B}^*\mathcal{F}^*\mathcal{A}_T S_T(T-t)I_\lambda GJ_\lambda \\ &= \mathcal{B}^*\mathcal{A}_T S_T(T-t)I_\lambda GJ_\lambda \quad (\text{using } \mathcal{B}^*\mathcal{F}^* = \mathcal{B}^*) \end{aligned}$$

where $I_\lambda = \lambda(\lambda I - \mathcal{A}_T)^{-1}$, $\lambda \in \rho(\mathcal{A})$ and we have used Theorem 5.2 successively. Since $\mathcal{B}^*(\eta, \phi) = B^T\eta$, the arguments, as in the proof of Lemma 5.4, yield that $\mathcal{B}^*\mathcal{A}_T$ has a bounded extension to all elements $z \in Z$ of the form $z = (\phi(0), \phi)$ with $\phi \in C(-r, 0; \mathbb{R}^n)$ and

$$(5.12) \quad \int_0^T |\mathcal{B}^*\mathcal{A}_T S_T(T-t)z|^2 dt \leq M(T)\|z\|^2$$

for $M(\cdot):[0, \infty) \rightarrow \mathbb{R}^+$ nondecreasing. Hence one obtains

$$\begin{aligned} \frac{d}{dt} u_\lambda(t) &= \mathcal{B}^* \mathcal{A}^* \{S^*(T-t) G_\lambda u_\lambda(T, t) \\ &\quad + \int_t^T S^*(\sigma-t) \mathcal{C}^* \mathcal{C} u_\lambda(\sigma, t) d\sigma\} u_\lambda(t, t_0) z + \mathcal{B}^* \mathcal{C}^* \mathcal{C} u_\lambda(t, t_0) z \\ &= \mathcal{B}^* \mathcal{A}_T S_T(T-t) I_\lambda G J_\lambda u_\lambda(T, t_0) z \\ &\quad + \mathcal{B}^* \mathcal{A}_T \int_t^T S_T(\sigma-t) \mathcal{C}^* \mathcal{C} u_\lambda(\sigma, t_0) z d\sigma + \mathcal{B}^* \mathcal{C}^* \mathcal{C} u_\lambda(t, t_0) z. \end{aligned}$$

Note that $u_\lambda(t, t_0)z$ converges strongly to $u(t, t_0)z$ for $z \in Z$, and the convergence is uniform on $[t_0, T]$. Since I_λ and J_λ converge strongly to the identity operator on Z , $I_\lambda G J_\lambda u_\lambda(T, t_0)z$ converges strongly to $G u(T, t_0)z$ for $z \in Z$. It now follows from (i) of Lemma 5.1 and (5.12) that $\{(d/dt)u_\lambda(\cdot)\}$ is a convergent sequence in $L_2(t_0, T; \mathbb{R}^m)$, which completes the proof when combined with the closedness of the differential operator (d/dt) on $L_2(t_0, T; \mathbb{R}^m)$. Q.E.D.

This last corollary establishes the Chandrasekhar equations for hereditary differential systems.

COROLLARY 5.9. *Define the operator $L(\cdot)$ on Z by $L(t)z = \mathcal{H}u(T, t)z$, $0 \leq t \leq T$ for all $z \in Z$, and the gain operator $K(t) = \mathcal{B}^* \Pi(t)$, $t \leq T$. Then $K(t)z$ and $L(t)x$ are absolutely continuous on $[0, T]$ for $z \in Z$ and $x \in \mathcal{D}(\mathcal{A})$, and they also satisfy*

$$\frac{d}{dt} K(t)z = -\mathcal{B}^* L^*(t) \Lambda L(t)z, \quad z \in Z,$$

$$K(T) = \mathcal{B}^* G$$

and

$$\frac{d}{dt} L(t)x = -L(t)(\mathcal{A} - \mathcal{B}K(t))x, \quad x \in \mathcal{D}(\mathcal{A}),$$

$$L(T)x = \mathcal{H}x.$$

Proof. From (2.7),

$$L(t)\mathcal{B}v = \mathcal{H}S(T-t)\mathcal{B}v - \mathcal{H} \int_t^T S(T-s)\mathcal{B}K(s)u(s, t)\mathcal{B}v ds \quad \text{for } v \in \mathbb{R}^m.$$

Here note that $\mathcal{H}S(\tau)\mathcal{B}v = \mathcal{H}(x(\tau), x(\tau + \cdot))$, $\tau \geq 0$ where x is the homogeneous solution of (5.1) with initial condition $\mathcal{B}v$ and $x \in BV(-r, T; \mathbb{R}^n)$ for any $T \geq 0$. This means that $L(t)\mathcal{B} \in \mathbb{R}^{\tilde{p} \times m}$ exists for each t , and it is not difficult to show that $L(t)\mathcal{B}$ is of bounded variation on $[0, T]$. So $\mathcal{B}^* L^*(\cdot) = (L(\cdot)\mathcal{B})^* \in \mathbb{R}^{m \times \tilde{p}}$. It then follows from Theorem 5.7 that for $z \in Z$

$$K(t)z = \mathcal{B}^* Gz + \int_t^T \mathcal{B}^* L^*(s) \Lambda L(s)z ds,$$

and hence $K(t)z$ is absolutely continuous on $[0, T]$.

Since $\mathcal{H}u(T, t)\mathcal{B} = L(t)\mathcal{B}$ is integrable, it follows from Lemma 5.1 and (3.4) that for $z \in Z$

$$L(t)z = \mathcal{H}S(t-t)z - \int_t^T L(s)\mathcal{B}K(s)S(s-t)z ds,$$

and thus for $z \in \mathcal{D}(\mathcal{A})$, $L(t)z$ is absolutely continuous on $[0, T]$ with square integrable derivative. Q.E.D.

Acknowledgments. The authors would like to thank Professor John A. Burns for motivating this work, and the reviewers, whose comments were most helpful.

REFERENCES

- [1] J. P. AUBIN, *Applied Functional Analysis*, John Wiley, New York, 1979.
- [2] A. V. BALAKRISHNAN, *Applied Functional Analysis, Second Edition*, Springer-Verlag, Berlin, 1981.
- [3] H. T. BANKS AND J. A. BURNS, *Hereditary control problems: numerical methods based on averaging approximations*, this Journal, 16 (1978), pp. 169–208.
- [4] J. S. BARAS AND D. G. LAINIOTIS, *Chandrasekhar algorithms for linear time varying distributed systems*, Inform. Sci., 17 (1979), pp. 153–167.
- [5] J. A. BURNS, K. ITO AND R. K. POWERS, *Chandrasekhar equations and computational algorithms for distributed parameter systems*, Proc. 23rd IEEE Decision and Control Conf., December 1984, Las Vegas, NE.
- [6] J. CASTI, *Dynamical Systems and Their Applications: Linear Theory*, Academic Press, New York, 1977.
- [7] J. CASTI AND L. LJUNG, *Some new analytic and computational results for operator Riccati equations*, this Journal, 13 (1975), pp. 817–826.
- [8] R. F. CURTAIN, *A survey of infinite-dimensional filtering*, SIAM Rev., 17 (1975), pp. 395–411.
- [9] R. F. CURTAIN AND A. J. PRITCHARD, *Infinite dimensional linear system theory*, in Lecture Notes in Control and Information Sciences 8, Springer-Verlag, Berlin, 1978.
- [10] M. C. DELFOUR, *The linear quadratic optimal control problem with delays in the state and control variables: a state space approach*, Université de Montreal, CRMA-1012, 1981.
- [11] M. C. DELFOUR AND S. K. MITTER, *Hereditary differential systems with constant delays: I. General case*, J. Differential Equations, 12 (1972), pp. 213–235.
- [12] J. S. GIBSON, *The Riccati integral equations for optimal control problems on Hilbert spaces*, this Journal, 17 (1979), pp. 537–565.
- [13] ———, *Linear quadratic optimal control of hereditary differential systems: infinite dimensional Riccati equations and numerical approximations*, this Journal, 21 (1983), pp. 95–139.
- [14] T. KAILATH, *Some Chandrasekhar-type algorithms for quadratic regulators*, Proc. IEEE Decision and Control Conf., New Orleans, 1972, pp. 219–223.
- [15] T. KATO, *Perturbation Theory for Linear Operators*, Springer-Verlag, Berlin, 1966.
- [16] A. LINDQUIST, *Optimal filtering of continuous-time stationary processes by means of the backward innovation process*, this Journal, 12 (1974), pp. 747–754.
- [17] J. L. LIONS, *Optimal Control of Systems Governed by Partial Differential Equations*, Springer-Verlag, New York, 1971.
- [18] A. MANITIUS, *Completeness and F-completeness of eigenfunctions associated with retarded functional differential equations*, J. Differential Equations, 35 (1980), pp. 1–29.
- [19] A. PAZY, *Semigroups of Linear Operators and Applications to Partial Differential Equations*, Springer-Verlag, New York, 1983.
- [20] R. POWERS, *Chandrasekhar equations for distributed parameter systems*, Ph.D. thesis, Virginia Polytechnic Institute and State University, Blacksburg, VA, 1984.
- [21] A. J. PRITCHARD AND D. SALAMON, *The linear quadratic control problem for infinite dimensional systems with unbounded input and output operators*, Mathematics Research Center, University of Wisconsin, TSR #2624, 1984.
- [22] M. SORINE, *Sur le semi-groupe nonlinéaire associé à l'équation de Riccati*, INRIA Report No. 167, October 1982.
- [23] R. B. VINTER, *Filter stability of stochastic evolution equations*, this Journal, 15 (1977), pp. 465–485.
- [24] ———, *On the evolution of the state of linear differential delay equations in M^2 : properties of the generator*, J. Inst. Math. Appl., 21 (1978), pp. 13–23.

ALGEBRAIC RICCATI EQUATION ARISING IN BOUNDARY CONTROL PROBLEMS*

FRANCO FLANDOLI†

Abstract. An algebraic Riccati equation is studied, with application to the optimal control of deterministic and stochastic parabolic systems with boundary control. The control function can act on the boundary through Dirichlet or Neumann conditions.

Key words. algebraic Riccati equation, Dirichlet boundary control, deterministic and stochastic dynamics

AMS(MOS) subject classification. 49A22

1. Introduction. In this paper we are concerned with the optimal control over infinite time horizon of two classes of boundary control systems: the first one is a class of deterministic parabolic systems with boundary control, described by an abstract semigroup model (similar to those considered in [B2] or [L1]) which covers both cases of Dirichlet boundary control and Neumann boundary control; the second one is a class of stochastic parabolic systems with state and control dependent noises, which in a sense generalizes the previous one, along the lines of stochastic systems considered in [I1] (but [I1] is concerned only with the case of distributed controls); also the stochastic model covers both cases of Dirichlet and Neumann boundary control. The work is based on a direct study of a generalized version of the algebraic Riccati equation, connected with the stochastic problem and the deterministic one. The properties obtained include existence, uniqueness, asymptotic behavior for the Riccati equation, synthesis of optimal control, and stability of optimal trajectories. Examples of applications are given in §§ 4.2 and 5.3.

Boundary control problems, above all in the deterministic case, have been extensively studied, in recent years, using analytic semigroups. In particular, the optimal control over finite time horizon of deterministic systems has been studied in [B2], [L2] and [F2]; for similar stochastic problems see [F1], [F3]. A different approach, based on techniques along the lines of [L5], is presented in [S1] and [S2], where problems over infinite time horizon and related algebraic Riccati equations are also considered; in particular, a problem with Neumann boundary control and Dirichlet boundary observation is studied under very general assumptions in [S1], while the case of Dirichlet boundary control and distributed observation (one of the problems studied here) is considered in [S2] assuming stability of the uncontrolled system. We remark that in these works a variational technique is employed.

We present here a different approach from that of [S1] and [S2]: first we use analytic semigroups along the lines of [B2], [L2] and [F2]; second we employ a dynamic programming technique. One of the advantages of this approach is that it allows one to study in an almost unified way the deterministic problem and the stochastic one; in fact the main part of the work, concerning a direct solution to the algebraic Riccati equation, is developed by studying an appropriate Riccati equation which covers the stochastic case as well as the deterministic one. The results on the stochastic problem are new; in the case of the Dirichlet boundary control of determinis-

* Received by the editors November 3, 1983; accepted for publication (in revised form) March 11, 1986.

† Dipartimento di Matematica, Università di Torino, 10123 Torino, Italy.

tic systems, we prove some results on stability of optimal trajectories, and asymptotic behavior of the solution $P(t)$ of a differential Riccati equation (see Theorem 2), which are not included in [S1], [S2], and we prove the existence of a solution P_∞ of the algebraic Riccati equation under the natural assumption of existence of admissible controls.

Following a classical idea of distributed control theory ([L7], [D4]), we derive P_∞ as a limit, as $t \rightarrow \infty$, of $P(t)$; however the usual monotonicity and boundedness results for $P(t)$, which can be proved in a standard way, are not sufficient in the present case. A much stronger convergence of $P(t)$ is needed (see (3.44)) because of the presence of an unbounded operator in the nonlinear term of the Riccati equations (see (1.14) and (3.5)), and in the feedback gain (see (4.1) and (5.1)). To this purpose, the major technical issue to settle is the bound (3.21), which is proved with a technique along the lines of [F2].

The direct solution of the algebraic Riccati equation and the asymptotic behavior of $P(t)$ are the objects of § 3. Synthesis of optimal control, stability of optimal trajectories, and uniqueness of the solution to the algebraic Riccati equation are studied separately in the deterministic case, § 4, and in the stochastic case, § 5 (although with similar methods based on results of § 3). Some examples of applications to concrete boundary control problems are given in § 4.2 and § 5.3; we remark that also the pointwise control of parabolic systems in dimension $n \leq 3$ can be described using the abstract semigroup models (1.1) and (1.8) (see [D2] and [I2]), so that the results of the present paper apply to this problem.

We remark that the assumption of existence of admissible controls (which yields existence of P_∞ and synthesis) is verified for deterministic parabolic systems, with a second order elliptic operator and Dirichlet boundary control, also in the case of unstable free systems, in virtue of recent stabilizability results of [T2] and [L3] (for Neumann boundary control problems see [S1]). As to relations between the present work and stabilizability of boundary control systems, we note also that (assuming detectability) the feedback control (4.1), with the feedback gain involving P_∞ , gives another example of stabilizing boundary control. In the stochastic case stability problems for boundary control systems are new in the literature; in § 5 we extend to this case some typical results of distributed control theory ([I1], [H2]).

We conclude by listing some notation. If X is a Banach (resp. Hilbert) space, then we shall denote its norm by $|\cdot|_X$ (resp. its inner product by $\langle \cdot, \cdot \rangle_X$); if X and Y are Banach spaces, then we shall denote by $L(X, Y)$ the Banach space of all bounded linear operators from X to Y , and by $L(X)$ the space $L(X, X)$.

If X is a Hilbert space and T is a linear operator in X , then we shall denote its domain by $D(T)$, and its adjoint operator (if $D(T)$ is dense in X) by T^* (similar notation if $T \in L(X, Y)$, where X and Y are Hilbert spaces); further, we shall set $\Sigma^+(X) = \{T \in L(X) | T = T^*, \langle Tx, x \rangle_X \geq 0 \text{ for any } x \in X\}$. The domains D_A^γ and $D_{A^*}^\gamma$ will be defined in § 1.1.

Let X be a Banach space, and let $a < b$ be two real numbers; if $p \in [1, \infty[$ then we shall denote by $L^p(a, b; X)$ the Banach space of all functions $f: [a, b] \rightarrow X$ such that $\int_a^b |f(t)|_X^p dt < \infty$ (similarly, with obvious modifications, if $p = +\infty$, and if $b = +\infty$). Further, we shall denote by $C(a, b; X)$ the Banach space of all continuous functions $f: [a, b] \rightarrow X$ (similarly if $b = +\infty$), and we shall set $C^1(a, b; X) = \{f \in C(a, b; X) | (df/dt) \in C(a, b; X)\}$.

Finally, if X and Y are Banach spaces then we shall set $C_s(a, b; L(X, Y)) = \{T(\cdot): [a, b] \rightarrow L(X, Y) | T(\cdot)x \in C(a, b; Y) \text{ for any } x \in X\}$. Similar definitions if $b = +\infty$, and for $C_s(a, b; \Sigma^+(X))$ where X is a Hilbert space.

1.1. Problem formulation and hypotheses. We introduce here the infinite time horizon problems and the related algebraic Riccati equations which are the objects of this paper. We formulate first the deterministic problem, and then the stochastic one.

Let H , U , V be Hilbert spaces; H will be the state space, U the control space, and V the observation space. In order to unify the study of deterministic and stochastic problems, and to use the theory of Hilbert space valued Brownian motions and stochastic integration ([C2], [C3]), it is preferable to assume separability of these Hilbert spaces; moreover this is not restrictive for applications. However, as far as we are concerned with the deterministic case, all results hold in the general case.

Throughout the paper we shall make the following assumption:

- (H1) A is the infinitesimal generator of an analytic semigroup e^{tA} , $t \geq 0$, in H ; $\lambda \geq 0$ is a real number such that $A - \lambda$ is stable, $B \in L(U, D_A^\alpha)$ for some $\alpha \in]0, 1]$, where D_A^α denotes the domain of the fractional power $(\lambda - A)^\alpha$; $C \in L(H, V)$, $N \in \Sigma^+(U)$ and $\langle Nu, u \rangle_U \geq \nu |u|_U^2$ for any $u \in U$ and for some fixed $\nu > 0$.

We remark that the fractional powers $(\lambda - A)^\gamma$ and $(\lambda - A^*)^\gamma$, $\gamma \in \mathbb{R}$, are well defined because $A - \lambda$ is stable ([T1]); we shall denote their domains by D_A^γ and $D_{A^*}^\gamma$, respectively.

In applications to boundary control problems, A is defined by an elliptic operator in a bounded domain of \mathbb{R}^n with homogeneous boundary conditions, and B is the Green mapping of a related elliptic boundary value problem; see Example 1, § 4.2, for more precise definitions. We remark that $\alpha = \frac{1}{4} - \varepsilon$, $\varepsilon > 0$, in the case of second order elliptic operators with Dirichlet boundary conditions, while $\alpha = \frac{3}{4} - \varepsilon$ in the Neumann case ([L1]).

It is well known ([B2], [L1]) that solutions to parabolic boundary value problems can be represented by an abstract semigroup formula involving operators A and B which verify (H1). In the case of problems over infinite time horizon it turns out that the system can be appropriately described by the input-output formula (see Example 1, § 4.2)

$$(1.1) \quad y(t) = e^{tA} y_0 + \int_0^t (A - \lambda) e^{(t-s)A} B u(s) ds$$

where $u(\cdot)$ is the control function and $y(\cdot)$ is the state function. We remark that, given $T > 0$, (1.1) defines a function $y \in L^2(0, T; H)$ for any $u \in L^2(0, T; U)$. To see this, we recall first that analyticity and stability of $e^{t(A-\lambda)}$ imply that, for any $\gamma > 0$, there exists a constant $c(\gamma) > 0$ such that

$$(1.2) \quad |(\lambda - A)^\gamma e^{t(A-\lambda)}|_{L(H)} \leq c(\gamma) e^{-\omega t} / t^\gamma, \quad t > 0,$$

for a suitable constant $\omega > 0$ ([T1, Thm. 3.3.3]); then from (H1) it follows that there exists a constant $c > 0$ such that

$$(1.3) \quad \begin{aligned} |(A - \lambda) e^{tA} B|_{L(U, H)} &= |(\lambda - A)^{1-\alpha} e^{t(A-\lambda)} e^{t\lambda} (\lambda - A)^\alpha B|_{L(U, H)} \\ &\leq c e^{t(\lambda-\omega)} / t^{1-\alpha}, \quad t > 0. \end{aligned}$$

If $u \in L^2(0, T; U)$ then (1.3) and a Young inequality yield $y \in L^2(0, T; H)$ in (1.1).

Given $y_0 \in H$, we will be concerned in § 4 with the following deterministic optimal control problem:

minimize

$$(1.4) \quad J_\infty(u) = \int_0^\infty \{|Cy(t)|_V^2 + \langle Nu(t), u(t) \rangle_U\} dt$$

over all $u \in L^2(0, \infty; U)$, subject to (1.1).

The algebraic Riccati equation arising in this problem is formally (a correct meaning is given by (1.7) below):

$$(1.5) \quad A^*P_\infty + P_\infty A + C^*C - P_\infty(A - \lambda)BN^{-1}B^*(A^* - \lambda)P_\infty = 0;$$

this equation will be studied in § 3, as a particular case of (1.10). In order to give a precise meaning to the nonlinear term of (1.5) we introduce a useful notation. Note that the composition $(A - \lambda)B$ is not well defined in general, while the operator $B^*(A^* - \lambda)$ is well defined on $D(A^*)$, and it can be extended, in a unique way, to a bounded operator from $D_{A^*}^{1-\alpha}$ to U . More precisely,

$$(1.6) \quad \begin{aligned} &\text{there exists a unique } K \in L(D_{A^*}^{1-\alpha}, U) \text{ such that} \\ &Kx = B^*(A^* - \lambda)x \text{ for any } x \in D(A^*). \end{aligned}$$

To see this, it is enough to define $K = -[(\lambda - A)^\alpha B]^*(\lambda - A^*)^{1-\alpha}$, and to note that K is the unique extension because $D(A^*)$ is dense in $D_{A^*}^{1-\alpha}$ ([T1]). Using this notation we can say, for instance, that $P_\infty \in \Sigma^+(H) \cap L(H, D_{A^*}^{1-\alpha})$ is a solution of (1.5) if

$$(1.7) \quad \langle P_\infty x, Ay \rangle_H + \langle Ax, P_\infty y \rangle_H + \langle Cx, Cy \rangle_V - \langle N^{-1}KP_\infty x, KP_\infty y \rangle_U = 0$$

for any $x, y \in D(A)$.

Let us define now the stochastic problem. Assume that (H1) holds, and let X_1 and X_2 be two separable Hilbert spaces. Let $w_1(t)$ and $w_2(t)$, $t \geq 0$, be two independent Brownian motions, defined in a probability space $(\Omega, \mathcal{F}, \mathcal{P})$, with values in X_1 and X_2 respectively; for the definition of Hilbert space valued Brownian motions see [C2]. Let \mathcal{F}_t be the σ -algebra generated by $\{w_1(s), w_2(s); 0 \leq s \leq t\}$; if $0 \leq T < \infty$ (similarly if $T = +\infty$), then we shall denote by $M_w^2(0, T; H)$ the Hilbert space of all strongly measurable stochastic processes $f: [0, T] \times \Omega \rightarrow H$ such that $f(t)$ is \mathcal{F}_t -measurable for any $t \in [0, T]$, and

$$E \int_0^T |f(t)|_H^2 dt < \infty$$

(similarly we define $M_w^2(0, T; U)$ and $M_w^2(0, T; V)$). Further, we shall denote by $C_w(0, T; L^2(\Omega; H))$ the Banach space of all $f \in M_w^2(0, T; H)$ such that $E|f(\cdot)|_H^2$ is a continuous function on $[0, T]$.

Assume that

$$(H2) \quad F_1 \in L(H, L(X_1, H)), \quad F_2 \in L(U, L(X_2, H));$$

then a general class of stochastic parabolic systems with boundary control, and state and control dependent noises, can be described by the following integral equation:

$$(1.8) \quad \begin{aligned} y(t) = & e^{tA}y_0 + \int_0^t (A - \lambda) e^{(t-s)A} Bu(s) ds \\ & + \int_0^t e^{(t-s)A} F_1(y(s)) dw_1(s) + \int_0^t e^{(t-s)A} F_2(u(s)) dw_2(s). \end{aligned}$$

Examples are given in § 5.3 and in [F1]. Similar systems, but with distributed control (i.e. without the term $(A - \lambda)$ in (1.8)), are considered, for instance in [I1]. We remark that, given $T > 0$, if $u \in M_w^2(0, T; U)$ then there exists a unique solution $y \in M_w^2(0, T; H)$ of (1.8). This follows from a standard application ([C2]) of the contraction principle; to this end, the only novelty is due to the first integral term in (1.8), and it is sufficient to note that it defines a stochastic process in $M_w^2(0, T; H)$ for any $u \in M_w^2(0, T; U)$, by virtue of (1.3) and the Young inequality.

In § 5 we shall consider the following optimal control problem:

minimize

$$(1.9) \quad J_{\infty}(u) = E \int_0^{\infty} \{ |Cy(t)|_V^2 + \langle Nu(t), u(t) \rangle_U \} dt$$

over all $u \in M_w^2(0, \infty; U)$, subject to (1.8).

The algebraic Riccati equation arising in this problem is formally (a correct meaning is given by (1.14) below):

$$(1.10) \quad A^*P_{\infty} + P_{\infty}A + C^*C + \phi_1(P_{\infty}) - P_{\infty}(A - \lambda)B\phi_2(P_{\infty})B^*(A^* - \lambda)P_{\infty} = 0$$

where $\phi_1(\cdot)$ and $\phi_2(\cdot)$ are the applications from $\Sigma^+(H)$ to $\Sigma^+(H)$ and $\Sigma^+(U)$, respectively, defined as follows. Let

$$(1.11) \quad W_1 \in \Sigma^+(X_1), \quad W_2 \in \Sigma^+(X_2), \quad W_1 \text{ and } W_2 \text{ trace class operators}$$

be the covariance operators of the Brownian motions $w_1(t)$ and $w_2(t)$, respectively [C2]; then from the properties of trace class operators [K1], and (H2), it follows easily that the applications $\Delta_1(\cdot)$ and $\Delta_2(\cdot)$ defined by

$$(1.12) \quad \begin{aligned} \langle \Delta_1(P)x, y \rangle_H &= \text{trace } F_1(x)^* P F_1(y) W_1, & x, y \in H, \quad P \in \Sigma^+(H), \\ \langle \Delta_2(P)u, v \rangle_U &= \text{trace } F_2(u)^* P F_2(v) W_2, & u, v \in U, \quad P \in \Sigma^+(H), \end{aligned}$$

are well defined and continuous from $\Sigma^+(H)$ to $\Sigma^+(H)$ and $\Sigma^+(U)$, respectively; further, if $P_n x \rightarrow Px$ for any $x \in H$ then $\Delta_i(P_n)x \rightarrow \Delta_i(P)x$ for any $x \in H$ and $i = 1, 2$. Finally, $\phi_1(\cdot)$ and $\phi_2(\cdot)$ are defined by

$$(1.13) \quad \phi_1(P) = \Delta_1(P), \quad \phi_2(P) = (N + \Delta_2(P))^{-1} \quad \text{for any } P \in \Sigma^+(H).$$

A similar problem with distributed control is studied in [11].

Using the operator K , given by (1.6), we can say that $P_{\infty} \in \Sigma^+(H) \cap L(H, D_A^{1-\alpha})$ is a solution to (1.10) if

$$(1.14) \quad \begin{aligned} &\langle P_{\infty}x, Ay \rangle_H + \langle Ax, P_{\infty}y \rangle_H + \langle Cx, Cy \rangle_V + \langle \phi_1(P_{\infty})x, y \rangle_H \\ &\quad - \langle \phi_2(P_{\infty})KP_{\infty}x, KP_{\infty}y \rangle_U = 0 \end{aligned}$$

for any $x, y \in D(A)$.

We remark that (1.5) is a particular case of (1.10), with

$$(1.15) \quad \phi_1(P) = 0, \quad \phi_2(P) = N^{-1} \quad \text{for any } P \in \Sigma^+(H).$$

For this reason we study in § 3 the algebraic Riccati equation (1.10), which covers both the deterministic and the stochastic case. It is to be noted that the direct study of (1.10) does not introduce essential complications with respect to a direct study of (1.5).

2. Preliminaries. We summarize results for (1.1) and (1.8) which are useful in subsequent sections. Hypotheses and notation of § 1.1 are assumed throughout this section.

Let $T > 0$ be given; let us consider the linear mapping

$$(2.1) \quad \mathcal{L} \in L(L^2(0, T; U), L^2(0, T; H))$$

defined by

$$(2.2) \quad \mathcal{L}(u)(t) = \int_0^t (A - \lambda) e^{(t-s)A} Bu(s) ds \quad \text{for a.e. } t \in [0, T],$$

for any $u \in L^2(0, T; U)$. Note that (2.1) follows from (1.3) and a Young inequality. Further (1.3) yields easily

$$(2.3) \quad \mathcal{L} \in L(C(0, T; U), C(0, T; H)).$$

It is useful to introduce an approximation of (2.2); following [Y1], we define the operators

$$(2.4) \quad I_n = n(n - A)^{-1} \quad \text{for any } n \geq \lambda.$$

The following properties hold ([Y1]):

$$(2.5) \quad \begin{aligned} &I_n \in L(H, D(A)), I_n x \rightarrow x \text{ for any } x \in H, \text{ there exists a constant } c > 0 \\ &\text{such that } |I_n|_{L(H)} \leq c, I_n \text{ commutes with } e^{tA}, A, (\lambda - A)^\gamma. \end{aligned}$$

Using the operators I_n we define the approximations $\mathcal{L}_n \in L(C(0, T; U), C(0, T; H))$ of (2.2) by

$$(2.6) \quad \mathcal{L}_n(u)(t) = I_n \mathcal{L}(u)(t) = \int_0^t e^{(t-s)A} (A - \lambda) I_n B u(s) ds,$$

$t \in [0, T]$, for any $u \in C(0, T; U)$ (or also $u \in L^2(0, T; U)$).

LEMMA 1. If $u_n \rightarrow u$ in $C(0, T; U)$ then $\mathcal{L}_n(u_n) \rightarrow \mathcal{L}(u)$ in $C(0, T; U)$; moreover, for any $u \in C(0, T; U)$, we have:

$$(2.7) \quad \mathcal{L}_n(u) \in C(0, T; D(A)) \cap C^1(0, T; H),$$

$$(2.8) \quad \frac{d}{dt} \mathcal{L}_n(u)(t) = A \mathcal{L}_n(u)(t) + (A - \lambda) I_n B u(t), \quad t \in [0, T].$$

Proof. $\mathcal{L}_n(u_n) = I_n \mathcal{L}(u_n)$ by definition; then from (2.3) and (2.5) it follows $\mathcal{L}_n(u_n) \rightarrow \mathcal{L}(u)$ in $C(0, T; H)$. Let $u \in C(0, T; U)$; from the equality

$$\begin{aligned} A \mathcal{L}_n(u)(t) &= (A - \lambda) \mathcal{L}_n(u)(t) + \lambda \mathcal{L}_n(u)(t) \\ &= - \int_0^t (\lambda - A)^{1-\alpha} e^{(t-s)A} (A - \lambda) I_n (\lambda - A)^\alpha B u(s) ds + \lambda \mathcal{L}_n(u)(t), \end{aligned}$$

it follows readily that $\mathcal{L}_n(u) \in C(0, T; D(A))$; similarly one can prove that $\mathcal{L}_n(u) \in C^1(0, T; H)$ and that (2.8) holds. \square

2.1. Deterministic closed loop systems. Let $G(\cdot) \in C_s(0, T; L(H, U))$ be given; we consider the closed loop system (which corresponds to (1.1) with the feedback control $u(t) = -G(t)y(t)$)

$$(2.9) \quad y(t) = e^{tA} y_0 - \int_0^t (A - \lambda) e^{(t-s)A} B G(s) y(s) ds, \quad t \in [0, T],$$

and the approximating system

$$(2.10) \quad y_n(t) = e^{tA} I_n y_0 - \int_0^t e^{(t-s)A} (A - \lambda) I_n B G(s) y_n(s) ds,$$

$t \in [0, T]$ where $y_0 \in H$. From Lemma 1, (1.3) and the contraction principle, Corollary 1 follows immediately.

COROLLARY 1. *There exist unique solutions y and y_n of (2.9) and (2.10), respectively, in $C(0, T; H)$, and $y_n \rightarrow y$ in $C(0, T; H)$. Moreover $y_n \in C(0, T; D(A)) \cap C^1(0, T; H)$ and*

$$(2.11) \quad \begin{aligned} \frac{d}{dt} y_n(t) &= A y_n(t) - (A - \lambda) I_n B G(t) y_n(t), \quad t \in [0, T], \\ y_n(0) &= I_n y_0. \end{aligned}$$

Consider now the case of a constant operator $G \in L(H, U)$ in (2.9) and (2.10). In virtue of Corollary 1, we can define a strongly continuous semigroup $T(t)$ by the integral equation

$$(2.12) \quad T(t) = e^{tA} - \int_0^t (A - \lambda) e^{(t-s)A} B G T(s) ds, \quad t \geq 0$$

(the semigroup property follows by standard arguments, as in [K1]). We define also the approximations $T_n(\cdot) \in C_s(0, \infty; L(H))$ by the equations

$$(2.13) \quad T_n(t) = e^{tA} I_n - \int_0^t e^{(t-s)A} (A - \lambda) I_n B G T_n(s) ds, \quad t \geq 0;$$

note that $T(\cdot)I_n^{-1}$ are analytic semigroups, because they are defined by a bounded perturbation of the analytic semigroup e^{tA} .

PROPOSITION 1. *For any $x \in H$ and $T \geq 0$ we have $T_n(\cdot)x \rightarrow T(\cdot)x$ in $C(0, T; H)$, $T_n(\cdot)x \in C(0, T; D(A)) \cap C^1(0, T; H)$ and*

$$(2.14) \quad \frac{d}{dt} T_n(t)x = (A - (A - \lambda) I_n B G) T_n(t)x, \quad T_n(0)x = I_n x.$$

Moreover, $T(t)$ is an analytic semigroup with generator

$$A_G = (A - \lambda)(1 - BG) - \lambda, \quad D(A_G) = \{x \in H \mid (1 - BG)x \in D(A)\}.$$

Proof. The properties of $T_n(\cdot)$ follow readily from Corollary 1. For the proof of the analyticity of $T(t)$ and of the last part of the proposition, we refer, for instance, to [D2], [L4], or [Z2]. \square

2.2. Stochastic closed loop systems. Let $G \in L(H, U)$ be given. We consider here the stochastic equation

$$(2.15) \quad \begin{aligned} y(t) &= e^{tA} y_0 - \int_0^t (A - \lambda) e^{(t-s)A} B G y(s) ds \\ &\quad + \int_0^t e^{(t-s)A} F_1(y(s)) dw_1(s) - \int_0^t e^{(t-s)A} F_2(G y(s)) dw_2(s), \quad t \geq 0, \end{aligned}$$

which corresponds to (1.8) when the feedback control $u(t) = -G y(t)$ is used. We consider also the approximating equation

$$(2.16) \quad \begin{aligned} y_n(t) &= e^{tA} I_n y_0 - \int_0^t e^{(t-s)A} (A - \lambda) I_n B G y_n(s) ds \\ &\quad + \int_0^t e^{(t-s)A} I_n F_1(y_n(s)) dw_1(s) - \int_0^t e^{(t-s)A} I_n F_2(G y_n(s)) dw_2(s), \quad t \geq 0. \end{aligned}$$

LEMMA 2. Let $T \geq 0$ and $y_0 \in H$ be given. Then there exist unique solutions y and y_n of (2.15) and (2.16) respectively in $C_w(0, T; L^2(\Omega, H))$, $y_n \rightarrow y$ in $C_w(0, T; L^2(\Omega, H))$, and y_n verifies the stochastic differential equation

$$(2.17) \quad \begin{aligned} dy_n(t) &= (A - (A - \lambda)I_n BG)y_n(t) dt + I_n F_1(y_n(t)) dw_1(t) - I_n F_2(Gy_n(t)) dw_2(t), \\ y_n(0) &= I_n y_0. \end{aligned}$$

Moreover y verifies the integral equation

$$(2.18) \quad y(t) = T(t)y_0 + \int_0^t T(t-s)F_1(y(s)) dw_1(s) - \int_0^t T(t-s)F_2(Gy(s)) dw_2(s),$$

where $T(t)$ is the semigroup defined by (2.12).

Proof. It is proved in [C2] that the last two integral terms of (2.15) define continuous linear mappings from $C_w(0, T; L^2(\Omega, H))$ into itself; (2.3) implies that also the first integral term of (2.15) defines a continuous linear mapping in $C_w(0, T; L^2(\Omega, H))$; moreover the same results hold for (2.16). Then a standard application of the contraction principle yields existence and uniqueness of solutions to (2.15) and (2.16) in $C_w(0, T; L^2(\Omega, H))$, and convergence of y_n to y in $C_w(0, T; L^2(\Omega, H))$ (using (2.5)). The regularity of coefficients in (2.16) implies that y_n has a stochastic differential and (2.17) holds, in virtue of [I1, § 2.1]. Finally from (2.17) we have (by using a stochastic Fubini theorem of [C1])

$$(2.19) \quad \begin{aligned} y_n(t) &= T_n(t)I_n y_0 + \int_0^t T_n(t-s)I_n F_1(y_n(s)) dw_1(s) \\ &\quad - \int_0^t T_n(t-s)I_n F_2(Gy_n(s)) dw_2(s), \end{aligned}$$

where $T_n(t)$ is defined by (2.13); (2.19) implies (2.18) as $n \rightarrow \infty$, in virtue of the convergence results of Proposition 1, of the present lemma, and in virtue of (2.5). \square

It is useful to know that the solution to the stochastic “boundary feedback” system (2.15) is also the solution to (2.18), because a large variety of results on systems like (2.18) are already available (see for instance [I1]); in § 5.2 we shall study in this way stability and stabilizability of (2.15) and (1.8), respectively.

3. Direct solution to the Riccati equation. In this section we discuss some properties of the algebraic Riccati equation (1.10), and of the corresponding differential Riccati equation arising in finite time horizon problems. Using the already known results on the differential Riccati equation (§ 3.1) we derive in a standard way a “candidate” solution P_∞ to the algebraic Riccati equation (Proposition 4, § 3.2); however, as we remarked in the introduction, the strong convergence (3.13) is not sufficient in the present situation; then we prove the uniform bounds of § 3.2, which yield the existence and the asymptotic behavior results of § 3.3. The results of this section will be applied in §§ 4 and 5 to the deterministic problem (1.4) and to the stochastic problem (1.9), respectively.

3.1. Review on the differential Riccati equation and the problem over finite time horizon. For proofs and further details concerning the arguments of this subsection we refer to our previous works [F1] and [F2]; similar (and further) results for the deterministic case are to be found in [B2], [L2], [S2]. We restrict the discussion to the stochastic problem; all results, with obvious modifications, hold also in the deterministic case.

Given $T > 0$, $y_0 \in H$, and

$$(3.1) \quad P_0 \in \Sigma^+(H) \cap L(H, D_A^{1-\alpha}),$$

let us consider the problem:

minimize

$$(3.2) \quad J_T(u) = E \int_0^T \{ |Cy(t)|_V^2 + \langle Nu(t), u(t) \rangle_U \} dt + E \langle P_0 y(T), y(T) \rangle_H$$

over all $u \in M_w^2(0, T; U)$, subject to (1.8).

The differential Riccati equation arising in this problem has the form

$$(3.3) \quad \begin{aligned} \frac{dP(t)}{dt} &= A^*P(t) + P(t)A + C^*C + \phi_1(P(t)) \\ &\quad - P(t)(A - \lambda)B\phi_2(P(t))B^*(A^* - \lambda)P(t), \quad t \in [0, T], \\ P(0) &= P_0. \end{aligned}$$

Remark 1. Since y defined by (1.8) is not continuous in mean square, for a general control $u \in M_w^2(0, T; U)$, we have to give a meaning to the term $E \langle P_0 y(T), y(T) \rangle_H$. We recall ([F2, Lemma 3.3]) that (3.1) implies that the operator $(\lambda - A^*)^{((1-\alpha)/2)-\varepsilon} P_0 (\lambda - A)^{((1-\alpha)/2)-\varepsilon}$ has a unique extension $L_\varepsilon \in L(H)$ for any $\varepsilon > 0$. Moreover, for $\varepsilon > 0$ sufficiently small, the operator valued function $(\lambda - A)^{\varepsilon - ((1-\alpha)/2)} (\lambda - A) e^{tA} B = (\lambda - A)^{((1-\alpha)/2)+\varepsilon} e^{tA} (\lambda - A)^\alpha B$ belongs to $L^2(0, T; L(U, H))$; therefore, if y is defined by (1.8), then $(\lambda - A)^{\varepsilon - ((1-\alpha)/2)} y \in C_w(0, T; L^2(\Omega, H))$ for any $u \in M_w^2(0, T; U)$, and for $\varepsilon > 0$ sufficiently small (see also [F1, § 5]). Then, for a fixed $\varepsilon > 0$ sufficiently small, a precise meaning to the term $E \langle P_0 y(T), y(T) \rangle_H$ is given by the expression $E \langle L_\varepsilon - (\lambda - A)^{\varepsilon - ((1-\alpha)/2)} y(T), (\lambda - A)^{\varepsilon - ((1-\alpha)/2)} y(T) \rangle_H$, which is well defined. When the control u is more regular (for instance if $E \int_0^T |u(t)|_U^{(1/\alpha)+\varepsilon} dt < \infty$, $\varepsilon > 0$), so that $y \in C_w(0, T; L^2(\Omega, H))$, this expression reduces to $E \langle P_0 y(T), y(T) \rangle_H$.

PROPOSITION 2. Assume that hypotheses (H1), (H2) and notation of § 1.1 hold, and that (3.1) holds. Then there exists a unique mild solution $P(\cdot) \in C_s(0, T; L(H, D_A^{1-\alpha})) \cap C_s(0, T; \Sigma^+(H))$ of (3.3), in the sense that $P(\cdot)$ verifies the integral equation

$$(3.4) \quad \begin{aligned} P(t) &= e^{tA^*} P_0 e^{tA} + \int_0^t e^{(t-s)A^*} (C^*C + \phi_1(P(s)) \\ &\quad - (KP(s))^* \phi_2(P(s)) KP(s)) e^{(t-s)A} ds, \\ &\quad t \in [0, T], \end{aligned}$$

where K is given by (1.6), or equivalently, the equation

$$(3.5) \quad \begin{aligned} \frac{d}{dt} \langle P(t)x, y \rangle_H &= \langle P(t)x, Ay \rangle_H + \langle Ax, P(t)y \rangle_H + \langle Cx, Cy \rangle_V \\ &\quad + \langle \phi_1(P(t))x, y \rangle_H - \langle \phi_2(P(t))KP(t)x, KP(t)y \rangle_U, \quad t \in [0, T], \\ P(0) &= P_0, \end{aligned}$$

for any $x, y \in D(A)$. Further, there exists a unique optimal control $\bar{u} \in M_w^2(0, T; U)$ for problem (3.2); if (\bar{u}, \bar{y}) is the optimal pair, then

$$(3.6) \quad \bar{u}(t) = -\phi_2(P(T-t))KP(T-t)\bar{y}(t), \quad t \in [0, T].$$

Finally, the following relations hold:

$$(3.7) \quad \langle P(T)y_0, y_0 \rangle_H = J_T(\bar{u}),$$

$$(3.8) \quad \begin{aligned} \langle P(T)y_0, y_0 \rangle_H = J_T(u) - E \int_0^T & |\phi_2^{-1/2}(P(T-t))u(t) \\ & + \phi_2^{1/2}(P(T-t))KP(T-t)y(t)|_U^2 dt \end{aligned}$$

for any $u \in M_w^2(0, T; U)$, y given by (1.8).

This proposition is proved in [F1]; however it could be proved easily using the machinery of the present paper: local and maximal existence and uniqueness for (3.3) follows from a standard application of the contraction principle; the main point of the global existence follows from an a priori estimation for $|(\lambda - A^*)^{1-\alpha}P(t)|_{L(H)}$ which is essentially included in Theorem 1 below, if we replace " $t \geq 0$ " with " $t \in J$ " in every part of Theorem 1, where J is the interval of maximal existence of $P(t)$; the rest of Proposition 2 follows readily from (3.8), which can be proved by evaluating the stochastic differential $d\langle P(T-t)y_n(t), y_n(t) \rangle_H$, where y_n are regular approximations of y along the lines of § 2, and by integrating over $[0, T]$ and taking the limit as $n \rightarrow \infty$.

Remark 2. It is useful for the sequel to recall explicitly the deterministic version of (3.8):

for any $y_0 \in H$, $T > 0$, $P(\cdot)$ mild solution of (3.3) with ϕ_1 and ϕ_2 given by (1.15), and $u \in L^2(0, T; U)$, the relation

$$(3.9) \quad \begin{aligned} \langle P(T)y_0, y_0 \rangle_H = \int_0^T & \{ |Cy(t)|_V^2 + \langle Nu(t), u(t) \rangle_U \} dt + \langle P_0 y(T), y(T) \rangle_H \\ & - \int_0^T |N^{1/2}u(t) + N^{-1/2}KP(T-t)y(t)|_U^2 dt \end{aligned}$$

holds, with y given by (1.1).

We conclude this section with a simple result on monotonicity of $P(t)$. From the arbitrariness of $T > 0$ in Proposition 2, and the uniqueness of solution to (3.3), we have that, under the hypotheses of Proposition 2, there exists a unique mild solution $P(\cdot)$ of (3.3) over $[0, \infty[$. When $P_0 = 0$ the following result holds.

PROPOSITION 3. Assume that the hypotheses of Proposition 2 hold, and let $P(\cdot)$ be the mild solution of (3.3) with $P_0 = 0$. Then $\langle P(T_1)y_0, y_0 \rangle_H \leq \langle P(T_2)y_0, y_0 \rangle_H$ for any $0 \leq T_1 \leq T_2$ and $y_0 \in H$.

Proof. Let $y_0 \in H$ and $T_2 \geq T_1 \geq 0$ be given; let (\bar{u}_2, \bar{y}_2) be the optimal pair of problem (3.2) with $T = T_2$; from (3.7) we have

$$\langle P(T_2)y_0, y_0 \rangle_H = J_{T_2}(\bar{u}_2) = E \int_0^{T_2} \{ |C\bar{y}_2(t)|_V^2 + \langle N\bar{u}_2(t), \bar{u}_2(t) \rangle_U \} dt;$$

then

$$\langle P(T_2)y_0, y_0 \rangle_H \geq E \int_0^{T_1} \{ |C\bar{y}_2(t)|_V^2 + \langle N\bar{u}_2(t), \bar{u}_2(t) \rangle_U \} dt = J_{T_1}(\bar{u}_2);$$

but Proposition 2 (and in particular (3.8)) with $T = T_1$ implies $J_{T_1}(\bar{u}_2) \geq \langle P(T_1)y_0, y_0 \rangle_H$. We conclude $\langle P(T_2)y_0, y_0 \rangle_H \geq \langle P(T_1)y_0, y_0 \rangle_H$. \square

3.2. Uniform bounds. Following a well-known idea of the distributed control theory, we derive a candidate P_∞ for the solution to (1.10) as a limit, as $t \rightarrow +\infty$, of the solution $P(t)$ of (3.3) with $P_0 = 0$. It is useful to introduce the following definition.

DEFINITION 1. Given $y_0 \in H$, we say that $u \in M_w^2(0, \infty; U)$ (resp. $u \in L^2(0, \infty; U)$) is an admissible control for problem (1.9) (resp. (1.4)) if $Cy \in M_w^2(0, \infty; V)$ (resp. $Cy \in L^2(0, \infty; V)$), where y is the corresponding solution of (1.8) (resp. (1.1)).

In the sequel, we shall often assume one of the following hypotheses:

$$(3.10) \quad \begin{array}{l} \text{for any } y_0 \in H \text{ there exists an admissible control} \\ u \in M_w^2(0, \infty; U) \text{ for problem (1.9)} \end{array}$$

(stochastic case);

$$(3.11) \quad \begin{array}{l} \text{for any } y_0 \in H \text{ there exists an admissible control} \\ u \in L^2(0, \infty; U) \text{ for problem (1.4)} \end{array}$$

(deterministic case).

PROPOSITION 4. Assume that (H1), (H2) and (3.10) hold (stochastic case), or that (H1), (1.15) and (3.11) hold (deterministic case); let $P(\cdot)$ be the mild solution of (3.3) over $[0, \infty[$, with $P_0 = 0$. Then

$$(3.12) \quad \sup_{t \geq 0} |P(t)|_{L(H)} < \infty,$$

and consequently there exists an operator $P_\infty \in \Sigma^+(H)$ such that

$$(3.13) \quad P(t)x \rightarrow P_\infty x \quad \text{as } t \rightarrow +\infty \quad \text{for any } x \in H.$$

Proof. We give the proof only in the stochastic case, because the proof in the deterministic case is equal, with obvious formal modifications. Let $y_0 \in H$ and $T > 0$ be given, and let $u \in M_w^2(0, \infty; U)$ be an admissible control; then from Proposition 2 we have

$$\langle P(T)y_0, y_0 \rangle_H \leq J_T(u) = E \int_0^T \{ |Cy(t)|_V^2 + \langle Nu(t), u(t) \rangle_U \} dt \leq J_\infty(u)$$

where $J_\infty(\cdot)$ is defined by (1.9); then $\sup_{T \geq 0} \langle P(T)y_0, y_0 \rangle_H < \infty$, and this implies (3.12) in virtue of the Banach-Steinhaus theorem. (3.13) follows from (3.12), along with Proposition 3 and a standard result on nondecreasing bounded families of selfadjoint operators ([D4]). \square

In the case of distributed control problems this result is sufficient to prove that P_∞ is a solution to the algebraic Riccati equation, and to solve the synthesis. In the present case however we see from (1.14), (3.5), (4.1) and (5.1) that we need some properties like $P_\infty \in L(H, D_{A^*}^{1-\alpha})$ and $KP(t)x \rightarrow KP_\infty x$ for any $x \in H$. To this end the major result is the bound (3.21) below.

Let us introduce the following integral equation for the operator valued function $U(t, s)$, $t \geq s \geq 0$:

$$(3.14) \quad U(t, s) = e^{(t-s)(A-\lambda)} - \int_s^t (A-\lambda) e^{(r-s)(A-\lambda)} B\phi_2(P(r))KP(r)U(t, r) dr$$

formally $U(t, s)$ verifies the differential equation

$$\frac{\partial}{\partial s} U(t, s) = -[A - \lambda - (A - \lambda)B\phi_2(P(s))B^*(A^* - \lambda)P(s)]U(t, s),$$

$$0 \leq s \leq t.$$

$$U(t, t) = 1,$$

We remark that $U(t, s)$ can be defined by other equivalent integral equations, which are more common in some part of the literature; however we choose (3.14) because it can be used more directly than others to derive (3.15) and (3.16) below.

LEMMA 3. Let $P(t)$ be the mild solution of (3.3) over $[0, \infty[$, with $P_0 = 0$. Then for any $t \geq 0$ there exists a unique solution $U(t, \cdot) \in C_s(0, t; L(H))$ to (3.14), and the following integral versions of (3.5) hold:

$$(3.15) \quad P(t) = \int_0^t U^*(t, s) [C^*C + 2\lambda P(s) + (KP(s))^* \phi_2(P(s)) KP(s)] U(t, s) ds,$$

$$(3.16) \quad P(t) = \int_0^t e^{(t-s)(A^*-\lambda)} [C^*C + 2\lambda P(s) + \phi_1(P(s))] U(t, s) ds, \quad t \geq 0.$$

Proof. Let $I_n = n(n - A)^{-1}$ as in § 2; let us consider the approximating equation of (3.14)

$$U_n(t, s) = e^{(t-s)(A-\lambda)} I_n - \int_s^t e^{(t-s)(A-\lambda)} (A-\lambda) \cdot I_n B \phi_2(P(r)) KP(r) U_n(t, r) dr, \quad t \geq s \geq 0.$$

With the notation of Corollary 1, given $t \geq 0$ and $y_0 \in H$, let us set $G(s) = \phi_2(P(t-s)) KP(t-s)$, $y(s) = e^{sA} U(t, t-s)y_0$, $y_n(s) = e^{sA} U_n(t, t-s)$; then y and y_n verify (2.9) and (2.10), respectively, and Corollary 1 applies, yielding the existence of unique solutions $U(t, \cdot)$ and $U_n(t, \cdot)$ in $C_s(0, t; L(H))$, for any $t \geq 0$; further for any $h \in H$ and $t \geq 0$,

$$(3.17) \quad \begin{aligned} U_n(t, \cdot)h &\rightarrow U(t, \cdot)h \text{ in } C(0, t; H), \\ U_n(t, \cdot)h &\in C(0, t; D(A)) \cap C^1(0, t; H), \text{ and} \end{aligned}$$

$$(3.18) \quad \begin{aligned} \frac{\partial}{\partial s} U_n(t, s)h &= -[A - \lambda - (A - \lambda) I_n B \phi_2(P(s)) KP(s)] U_n(t, s)h, \quad 0 \leq s \leq t, \\ U_n(t, t) &= I_n. \end{aligned}$$

Let $h \in H$; since $U_n(t, s)h \in D(A)$, we can use (3.5) with $x = y = U_n(t, s)h$; from (3.5) and (3.18) we have

$$\begin{aligned} &\left\langle \frac{\partial}{\partial s} P(s) U_n(t, s)h, U_n(t, s)h \right\rangle_H \\ &= \frac{\partial}{\partial \sigma} \langle P(\sigma) U_n(t, s)h, U_n(t, s)h \rangle_H \Big|_{\sigma=s} + 2 \operatorname{Re} \left\langle P(s) U_n(t, s)h, \frac{\partial}{\partial s} U_n(t, s)h \right\rangle \\ (3.19) \quad &= 2 \operatorname{Re} \langle P(s) U_n(t, s)h, A U_n(t, s)h \rangle \\ &\quad + \langle [C^*C + \phi_1(P(s)) - (KP(s))^* \phi_2(P(s)) KP(s)] U_n(t, s)h, U_n(t, s)h \rangle \\ &\quad - 2 \operatorname{Re} \langle P(s) U_n(t, s)h, (A - \lambda - (A - \lambda) I_n B \phi_2(P(s)) KP(s)) U_n(t, s)h \rangle \\ &= \langle [C^*C + 2\lambda P(s) + \phi_1(P(s)) + (KP(s))^* \phi_2(P(s)) KP(s)] U_n(t, s)h, U_n(t, s)h \rangle \\ &\quad - 2 \langle K(1 + I_n) P(s) U_n(t, s)h, \phi_2(P(s)) KP(s) U_n(t, s)h \rangle; \end{aligned}$$

integrating on $[0, t]$ with respect to s , and using the convergence (3.17) and (2.5), from (3.19) we have:

$$\begin{aligned} \langle P(t)h, h \rangle &= \int_0^t \langle [C^*C + 2\lambda P(s) + \phi_1(P(s)) \\ &\quad + (KP(s))^* \phi_2(P(s)) KP(s)] U(t, s)h, U(t, s)h \rangle ds. \end{aligned}$$

This implies (3.15) because $P(t)$ is selfadjoint. The proof of (3.16) is similar. \square

THEOREM 1. Assume that (H1), (H2) and (3.12) hold; let $P(\cdot)$ be the mild solution of (3.3) with $P_0 = 0$, and $U(t, s)$, $t \geq s \geq 0$ be the solution to (3.14). Then there exists a constant $c > 0$ such that

$$(3.20) \quad |U(t, s)|_{L(H)} \leq c, \quad t \geq s \geq 0;$$

moreover, for any $\gamma < 1$ there exists a constant $c(\gamma) > 0$ such that

$$(3.21) \quad |(\lambda - A^*)^\gamma P(t)|_{L(H)} \leq c(\gamma), \quad t \geq 0.$$

Proof. Step 1. Let us set $R(t, s) = KP(s)U(t, s)$, for $t \geq s \geq 0$; from (3.14), (3.16) and the evolution property $U(s, r)U(t, s) = U(t, r)$, we have

$$(3.22) \quad U(t, s) = e^{(t-s)(A-\lambda)} - \int_s^t (A-\lambda) e^{(r-s)(A-\lambda)} B \phi_2(P(r)) R(t, r) dr,$$

$$(3.23) \quad R(t, s) = \int_0^s K e^{(s-r)(A^*-\lambda)} (C^*C + 2\lambda P(r) + \phi_1(P(r))) U(t, r) dr,$$

for $0 \leq s \leq t$. From (H1) and (1.13) it follows immediately that $0 \leq \langle \phi_2(P(t))u, u \rangle_U \leq 1/\nu |u|_U^2$, for any $t \geq 0$ and $u \in U$, so that $|\phi_2(P(t))|_{L(U)} \leq 1/\nu$ for any $t \geq 0$; then, from (1.2), (H1) and (3.22), there exists a constant $c_1 > 0$ such that

$$(3.24) \quad |U(t, s)x|_H \leq c_1 e^{-\omega(t-s)} |x|_H + c_1 \int_s^t \frac{e^{-\omega(r-s)}}{(r-s)^{1-\alpha}} |R(t, r)x|_U dr$$

for any $x \in H$ and $t \geq s \geq 0$. Further, from (1.2), (H1), (3.12) and (3.23), there exists a constant $c_2 > 0$ such that

$$(3.25) \quad |R(t, s)x|_U \leq c_2 \int_0^s \frac{e^{-\omega(s-r)}}{(s-r)^{1-\alpha}} |U(t, r)x|_H dr$$

for any $x \in H$ and $t \geq s \geq 0$.

Step 2. If $1 < p \leq \infty$, $p \neq 1/\alpha$, then there exists a constant $k_1(p) > 0$ such that

$$(3.26) \quad |U(t, \cdot)x|_{L^q(0,t;H)} \leq k_1(p) |x|_H + k_1(p) |R(t, \cdot)x|_{L^p(0,t;U)},$$

$$(3.27) \quad |R(t, \cdot)x|_{L^q(0,t;U)} \leq k_1(p) |U(t, \cdot)x|_{L^p(0,t;H)}$$

for any $x \in H$ and $t \geq 0$, where

$$(3.28) \quad q = \begin{cases} p/(1-\alpha p) & \text{if } 1 < p < 1/\alpha, \\ \infty & \text{if } 1/\alpha < p \leq \infty; \end{cases}$$

if $p = 1/\alpha$, then for any $\varepsilon \in]0, \min(\alpha, \omega)[$ there exists a constant $k_2(\varepsilon) > 0$ such that

$$(3.29) \quad |U(t, \cdot)x|_{L^{1/\varepsilon}(0,t;H)} \leq k_2(\varepsilon) |x|_H + k_2(\varepsilon) |R(t, \cdot)x|_{L^{1/\alpha}(0,t;U)},$$

$$(3.30) \quad |R(t, \cdot)x|_{L^{1/\varepsilon}(0,t;U)} \leq k_2(\varepsilon) |U(t, \cdot)x|_{L^{1/\alpha}(0,t;H)},$$

for any $t \geq 0$ and $x \in H$. (3.26) and (3.27), with $1 < p < 1/\alpha$, follow from (3.24) and (3.25), respectively, using a Young inequality ([H1, p. 290]); the case $1/\alpha < p < \infty$ follows readily from the Hölder inequality. As to (3.29) and (3.30), let us note first that, for any $t > 0$, the function

$$\varepsilon \rightarrow \frac{e^{t(-\omega+\varepsilon)}}{t^{1-\alpha+\varepsilon}}$$

is nondecreasing for $\varepsilon \geq 0$ (because it has nonnegative first derivative with respect to ε); then from (3.24) it follows

$$(3.31) \quad |U(t, s)x|_H \leq c_1 e^{-\omega(t-s)}|x|_H + c_1 \int_s^t \frac{e^{(-\omega+\varepsilon)(r-s)}}{(r-s)^{1-\alpha+\varepsilon}} |R(t, r)x|_U dr$$

for any $\varepsilon > 0$, $t \geq 0$ and $x \in H$. If $p = 1/\alpha$ and $0 < \varepsilon < \min(\alpha, \omega)$, then we can apply to (3.31) the above-mentioned Young inequality, and we have (3.29); similarly we can prove (3.30).

Step 3. Let $c_P > 0$ be a constant such that

$$(3.32) \quad |P(t)|_{L(H)} \leq c_P \quad \text{for any } t \geq 0$$

(see (3.12)). From (3.15) we have

$$(3.33) \quad \begin{aligned} & \int_0^t \langle \phi_2(P(s))KP(s)U(t, s)x, KP(s)U(t, s)x \rangle_U \\ & \leq \int_0^t \langle [C^*C + 2\lambda P(s) + \phi_1(P(s)) \\ & \quad + (KP(s))^*\phi_2(P(s))KP(s)]U(t, s)x, U(t, s)x \rangle_H ds \\ & = \langle P(t)x, x \rangle_H \leq c_P|x|_H^2 \quad \text{for any } t \geq 0 \text{ and } x \in H. \end{aligned}$$

Moreover, from (3.32) and (1.12) we have $|\Delta_2(P(s))|_{L(U)} \leq c_3 c_P$ for any $s \geq 0$, and for a suitable constant $c_3 > 0$; then, from definition (1.13), there exists $c_4 > 0$ such that

$$(3.34) \quad \langle \phi_2(P(s))u, u \rangle_U \geq c_4|u|_U^2 \quad \text{for any } s \geq 0 \text{ and } u \in U.$$

(3.33) and (3.34) yield

$$(3.35) \quad \int_0^t |R(t, s)x|_U^2 ds \leq (1/c_4)c_P|x|_H^2 \quad \text{for any } t \geq 0 \text{ and } x \in H.$$

Step 4. Let us start with the uniform bound (3.35), and let us apply repeatedly (3.26) and (3.27), or (3.29) and (3.30) if some exponent p is equal to $1/\alpha$; in a finite number of steps we have (3.20). For instance, in the case of Dirichlet (resp. Neumann) boundary control and second order elliptic operators, it is sufficient to apply (3.26) and (3.27) two times (resp. one time), because $\alpha = \frac{1}{4} - \varepsilon$ (resp. $\alpha = \frac{3}{4} - \varepsilon$) for any $\varepsilon > 0$.

Finally, applying (1.2), (3.32), (3.20) and the Hölder inequality to (3.16), we find (3.21). \square

3.3. Asymptotic behavior and existence for the algebraic Riccati equation. For the sake of simplicity, we prove the strong convergence result (3.38) below under the additional assumption that

$$(3.36) \quad (\lambda - A)^{-1} \text{ is a compact operator.}$$

(3.37) can be proved with a lengthy argument without assuming (3.36); however, (3.36) is verified in boundary control problems, because the operator A is defined by an elliptic operator in a bounded domain of R^n (see Example 1, § 4.2), and consequently (3.36) is not restrictive for applications.

THEOREM 2. *Under hypotheses and notation of Proposition 2, we have*

$$(3.37) \quad P_\infty \in L(H, D_A^{1-\alpha}) \quad \text{and} \quad \langle (\lambda - A^*)^\gamma P(t)x, y \rangle_H \rightarrow \langle (\lambda - A^*)^\gamma P_\infty x, y \rangle_H$$

for any $\gamma < 1$, $x, y \in H$. Further, if (3.36) holds then

$$(3.38) \quad (\lambda - A^*)^\gamma P(t)x \rightarrow (\lambda - A^*)^\gamma P_\infty x \quad \text{for any } x \in H \text{ and } \gamma < 1,$$

and in particular

$$(3.39) \quad KP(t)x \rightarrow KP_\infty x \quad \text{for any } x \in H,$$

where K is defined by (1.6); moreover P_∞ is a solution to (1.10), in the sense that it verifies (1.14), or equivalently that $P_\infty \in \Sigma^+(H) \cap L(H, D_A^{1-\alpha}) \cap L(D(A), D(A^*))$ and

$$(3.40) \quad A^*P_\infty x + P_\infty Ax + C^*Cx + \phi_1(P_\infty)x - (KP_\infty)^*\phi_2(P_\infty)KP_\infty x = 0$$

holds for any $x \in D(A)$. Finally,

$$(3.41) \quad P_\infty \leq \hat{P}_\infty \quad (\text{i.e. } \langle P_\infty x, x \rangle_H \leq \langle \hat{P}_\infty x, x \rangle_H \text{ for any } x \in H)$$

for any solution \hat{P}_∞ of (1.10).

Proof. Let $x \in H$ and $\gamma < 1$ be given; the set $\{(\lambda - A^*)^\gamma P(t)x\}_{t \geq 0}$ is bounded in H , in virtue of (3.21); then, for any sequence $t_n \rightarrow +\infty$, there exists a subsequence t_{n_k} such that

$$(3.42) \quad \lim_{k \rightarrow \infty} \langle (\lambda - A^*)^\gamma P(t_{n_k})x, y \rangle_H = \langle z, y \rangle_H \quad \text{for any } y \in H$$

and for some $z \in H$. But (3.13) implies $\lim_{k \rightarrow \infty} \langle P(t_{n_k})x, (\lambda - A)^\gamma y \rangle_H = \langle P_\infty x, (\lambda - A)^\gamma y \rangle_H$ for any $y \in D_A^\gamma$; then $\langle z, y \rangle_H = \langle P_\infty x, (\lambda - A)^\gamma y \rangle_H$ for any $y \in D_A^\gamma$, whence $P_\infty x \in D_{A^*}^\gamma$ and

$$(3.43) \quad z = (\lambda - A^*)^\gamma P_\infty x.$$

Since z does not depend on t_n and t_{n_k} (because (3.43) holds), (3.37) follows from (3.42) and (3.43).

Let us recall now that $(\lambda - A^*)^{-\varepsilon}$ is compact for any $\varepsilon > 0$ if (3.36) holds ([B1]). If $\gamma + \varepsilon < 1$, then from (3.37) we have that $(\lambda - A^*)^{\gamma+\varepsilon} P(t)x$ converges weakly to $(\lambda - A^*)^{\gamma+\varepsilon} P_\infty x$ for any $x \in H$; then (3.38) follows from the compactness of $(\lambda - A^*)^{-\varepsilon}$, and (3.39) follows from (3.38) and (1.6).

Let us prove that P_∞ verifies (1.14). Let $x, y \in D(A)$ be given. From (3.39) and (3.13) it follows that the right side of (3.5) converges to $\langle P_\infty x, Ay \rangle_H + \langle Ax, P_\infty y \rangle_H + \langle Cx, Cy \rangle_V + \langle \phi_1(P_\infty)x, y \rangle_H - \langle \phi_2(P_\infty)KP_\infty x, KP_\infty y \rangle_U$, whence

$$(3.44) \quad \text{there exists } a_{x,y} = \lim_{t \rightarrow \infty} \frac{d}{dt} \langle P(t)x, y \rangle_H;$$

moreover, from (3.13) we have that

$$(3.45) \quad \text{there exists } \lim_{t \rightarrow \infty} \langle P(t)x, y \rangle_H = \langle P_\infty x, y \rangle_H.$$

In virtue of a simple argument of real analysis, (3.44) and (3.45) yield $a_{x,y} = 0$. This implies that (1.14) holds. Moreover, from (1.14) we have readily $P_\infty \in L(D(A), D(A^*))$ and (3.40).

Finally, let us prove (3.41). We consider first the stochastic case, i.e. we assume that (H1), (H2) and (3.10) hold. Let \hat{P}_∞ be a solution to (1.10), in the sense that \hat{P}_∞ verifies (1.14); then \hat{P}_∞ verifies also (3.5) with $P_0 = \hat{P}_\infty$, i.e. \hat{P}_∞ is the mild solution of (3.3) with $P_0 = \hat{P}_\infty$. Let $y_0 \in H$ be fixed; we define the following feedback control

$$(3.46) \quad \hat{u}(t) = -\phi_2(\hat{P}_\infty)K\hat{P}_\infty \hat{y}(t), \quad t \geq 0.$$

Since \hat{P}_∞ is the mild solution of (3.3) with $P_0 = \hat{P}_\infty$, given $T > 0$, (3.8) holds with

$P(T) = P_0 = \hat{P}_\infty$, and in particular with $u(t) = \hat{u}(t)$; this yields

$$\langle \hat{P}_\infty y_0, y_0 \rangle_H = E \int_0^T \{ |C\hat{y}(t)|_V^2 + \langle N\hat{u}(t), \hat{u}(t) \rangle_U \} dt + E \langle \hat{P}_\infty \hat{y}(T), \hat{y}(T) \rangle_H,$$

whence

$$(3.47) \quad E \int_0^T \{ |C\hat{y}(t)|_V^2 + \langle N\hat{u}(t), \hat{u}(t) \rangle_U \} dt \leq \langle \hat{P}_\infty y_0, y_0 \rangle_H;$$

as $T \rightarrow \infty$ (\hat{u} and \hat{y} are independent on T), (3.47) implies

$$(3.48) \quad J_\infty(\hat{u}) \leq \langle \hat{P}_\infty y_0, y_0 \rangle_H,$$

where $J_\infty(\cdot)$ is defined by (1.9). Given $T > 0$, relation (3.8), when $P(\cdot)$ is the mild solution of (3.3) with $P_0 = 0$, yields

$$(3.49) \quad \langle P(T)y_0, y_0 \rangle_H \leq E \int_0^T \{ |Cy(t)|_V^2 + \langle Nu(t), u(t) \rangle_U \} dt$$

for any $u \in M_w^2(0, T; U)$; as $T \rightarrow \infty$, (3.54) implies

$$(3.50) \quad \langle P_\infty y_0, y_0 \rangle_H \leq J_\infty(u) \quad \text{for any } u \in M_w^2(0, \infty; U).$$

From (3.48), and (3.50) with $u = \hat{u}$ defined by (3.46), we have $\langle P_\infty y_0, y_0 \rangle_H \leq \langle \hat{P}_\infty y_0, y_0 \rangle_H$; this implies (3.41), and the proof of (3.41) is complete in the stochastic case.

In the deterministic case, i.e. when (H1), (1.15) and (3.11) hold, the proof is equal; we point out only the relations which correspond to (3.46), (3.48) and (3.50), because they are useful in § 4. Let \hat{P}_∞ be a solution of (1.5) (in the sense of the present theorem). Given $y_0 \in H$, let \hat{u} be the feedback control defined by

$$(3.51) \quad \hat{u}(t) = -N^{-1}K\hat{P}_\infty \hat{y}(t), \quad t \geq 0;$$

then

$$(3.52) \quad J_\infty(\hat{u}) \leq \langle \hat{P}_\infty y_0, y_0 \rangle_H,$$

where $J_\infty(\cdot)$ is defined by (1.4); moreover, if P_∞ is the solution to (1.5) defined by (3.13), then

$$(3.53) \quad \langle P_\infty y_0, y_0 \rangle_H \leq J_\infty(u) \quad \text{for any } u \in L^2(0, \infty; U). \quad \square$$

4. The deterministic optimal control problem. In this section we apply the results of § 3 (in particular Theorem 2) to the deterministic optimal control problem (1.4). First we solve the synthesis (§ 4.1), and then we give a uniqueness result for the algebraic Riccati equation, and a stability result for the optimal trajectories (§ 4.2). Finally, an example of deterministic boundary control problem is discussed at the end of § 4.2.

4.1. Synthesis.

THEOREM 3. Assume that (H1) and (3.36) hold. Then there exists a solution to the algebraic Riccati equation (1.5) (in the sense of Theorem 2) if and only if (3.11) holds. In this case the results of Theorem 2 hold, and for any $y_0 \in H$ there exists a unique optimal control $\bar{u} \in L^2(0, \infty; U)$ for problem (1.4); moreover, if (\bar{u}, \bar{y}) is the optimal pair and P_∞ is defined by (3.13), then

$$(4.1) \quad \bar{u}(t) = -N^{-1}K P_\infty \bar{y}(t), \quad t \geq 0$$

(K given by (1.6)) and

$$(4.2) \quad \langle P_\infty y_0, y_0 \rangle_H = J_\infty(\bar{u}).$$

Finally, $\bar{u}(\cdot)$ and $\bar{y}(\cdot)$ are analytic functions for $t > 0$, and

$$(4.3) \quad \bar{y}(t) = T(t)y_0, \quad t \geq 0,$$

where $T(t)$ is the analytic semigroup defined by the integral equation

$$(4.4) \quad T(t) = e^{tA} - \int_0^t (A - \lambda) e^{(t-s)A} B N^{-1} K P_\infty T(s) ds, \quad t \geq 0.$$

Proof. If (3.11) holds then P_∞ defined by (3.13) is a solution to (1.5) in virtue of Theorem 2, and the results of Theorem 2 hold; conversely, if \hat{P}_∞ is a solution to (1.5), and $y_0 \in H$ is given, then (3.52) holds, whence \hat{u} (defined by (3.51)) is an admissible control; this proves (3.11).

Given $y_0 \in H$, we can take $\hat{P}_\infty = P_\infty$ in (3.53); then (3.52) and (3.53) imply that (4.2) holds, with $\bar{u} = \hat{u}$ given by (3.51) or equivalently by (4.1); moreover, from (3.53) and (4.2) we see that \bar{u} is an optimal control. It is also unique because $J_\infty(\cdot)$ is a strictly convex functional, by virtue of hypothesis (H1) on N .

Finally, Proposition 1 implies that (4.4) defines an analytic semigroup; (4.3) follows from (4.4), (1.1) and (4.1), and the analyticity of \bar{u} and \bar{y} follows from the analyticity of $T(t)$ and from (4.1) and (4.3). \square

4.2. Stability of optimal trajectories and uniqueness of solution to (1.5). The results of this section extend to boundary control problems some results of [Z1].

We recall that the pair (C, A) is said to be detectable if there exists $S \in L(V, H)$ such that $A - SC$ is stable. The following lemma will be useful in the proof of Theorem 4.

LEMMA 4. Assume that (H1) holds and that (C, A) is detectable. Let $G \in L(H, U)$ be given, and assume that the feedback control $u(t) = -Gy(t)$ is an admissible control (Definition 1) for any $y_0 \in H$; then the trajectories y are exponentially stable for any $y_0 \in H$, and the semigroup $T(t)$ defined by (2.12) is exponentially stable.

Proof. Let $T(\cdot)$ and $T_n(\cdot)$ be defined by (2.12) and (2.13), respectively; let $S \in L(V, H)$ be such that $A - SC$ is stable; then from (2.14) we have

$$(4.5) \quad \begin{aligned} \frac{d}{dt} T_n(t)x &= (A - SC)T_n(t)x + SCT_n(t)x + (SC - A)I_n BGT_n(t)x \\ &\quad + (\lambda - SC)I_n BGT_n(t)x \quad \text{for any } t \geq 0 \text{ and } x \in H \end{aligned}$$

and $T_n(0) = I_n$; then by the variation of constant device we have

$$(4.6) \quad \begin{aligned} T_n(t)x &= e^{t(A-SC)} I_n x + \int_0^t e^{(t-s)(A-SC)} SCT_n(s)x ds \\ &\quad + \int_0^t (SC - A) e^{(t-s)(A-SC)} I_n BGT_n(s)x ds \\ &\quad + \int_0^t e^{(t-s)(A-SC)} (\lambda - SC) I_n BGT_n(s)x ds, \quad t \geq 0, \quad x \in H. \end{aligned}$$

In order to take the limit as $n \rightarrow \infty$ in (4.6) we have to study the limit behavior of the second integral term in (4.6). For any $0 < T \leq \infty$ consider the mapping \mathcal{L}_T on $L^2(0, T; H)$ defined by $\mathcal{L}_T(z)(t) = \int_0^t (SC - A) e^{(t-s)(A-SC)} z(s) ds$, for a.e. $t \in]0, T[$, and for any $z \in L^2(0, T; H)$; since $e^{t(A-SC)}$ is a stable, analytic semigroup, it is known that \mathcal{L}_T is a continuous linear mapping in $L^2(0, T; H)$, for any $0 < T \leq \infty$ (see for instance [D5], [D1] or [L6]), and there exists a constant $c > 0$ such that

$$(4.7) \quad \|\mathcal{L}_T(z)\|_{L^2(0, T; H)} \leq c \|z\|_{L^2(0, T; H)} \quad \text{for any } z \in L^2(0, T; H)$$

and $T \in]0, \infty]$. Therefore, for any fixed $T > 0$, we can take the limit (in the $L^2(0, T; H)$ -topology) in (4.6) as $n \rightarrow \infty$, using also the convergence result of Proposition 1; this yields

$$(4.8) \quad \begin{aligned} T(t)x &= e^{t(A-SC)}x + \int_0^t e^{(t-s)(A-SC)} SCT(s)x \, ds \\ &+ \int_0^t (SC - A) e^{(t-s)(A-SC)} BGT(s)x \, ds \\ &+ \int_0^t e^{(t-s)(A-SC)} (\lambda - SC) BGT(s)x \, ds \end{aligned}$$

for any $t \geq 0$ and $x \in H$. We apply now a Young inequality to the first and the third integral terms of (4.8), and (4.7), with $T = +\infty$, to the second integral term of (4.8); we have

$$(4.9) \quad \begin{aligned} |T(\cdot)x|_{L^2(0, \infty; H)} &\leq c_2|x|_H + c_1|SCT(\cdot)x|_{L^2(0, \infty; H)} + c|BGT(\cdot)x|_{L^2(0, \infty; H)} \\ &+ c_1|(\lambda - SC)BGT(\cdot)x|_{L^2(0, \infty; H)} \end{aligned}$$

for any $x \in H$, where c_1 and c_2 are the norms of $e^{t(A-SC)}$ in $L^1(0, \infty; H)$ and $L^2(0, \infty; H)$, respectively. If u is the feedback control defined by $u(t) = -Gy(t)$, corresponding to the initial trajectory $y(0) = x$, then $y(t) = T(t)x$ and $u(t) = -GT(t)x$ (see (2.12)); since $u(\cdot)$ is assumed to be admissible, this implies $GT(\cdot)x \in L^2(0, \infty; U)$ and $CT(\cdot)x \in L^2(0, \infty; V)$. Therefore, from (4.9) we have

$$(4.10) \quad \int_0^\infty |T(t)x|_H^2 \, dt < \infty \quad \text{for any } x \in H.$$

This implies, by virtue of a result of [D3], that $T(t)$ is a stable semigroup, and also the trajectories $y(t) = T(t)y_0$ are exponentially stable for any $y_0 \in H$. \square

DEFINITION 2. The system (1.1) is said to be exponentially stabilizable if there exists $G \in L(H, U)$ such that (2.12) defines an exponentially stable semigroup $T(t)$.

THEOREM 4. (i) Assume that (H1) and (3.36) hold, and that (C, A) is detectable; then there exists at most one solution of the algebraic Riccati equation (1.5) (in the sense of Theorem 2), and in this case the unique solution is P_∞ defined by (3.13). (ii) Moreover if P_∞ exists, then for any $y_0 \in H$ the optimal control (4.1) of problem (1.4) is a stabilizing control, and, more precisely, system (1.1) is exponentially stabilizable and the optimal trajectories \bar{y} (Theorem 3) are exponentially stable for any $y_0 \in H$.

Proof. Let \hat{P}_∞ be a solution to (1.5); given $y_0 \in H$, let $\hat{u}(\cdot)$ be defined by (3.51). From (3.52) it follows that $\hat{u}(\cdot)$ is an admissible control for problem (1.4); then (3.11) holds, and Theorem 2 implies that P_∞ exists, and

$$(4.11) \quad P_\infty \leq \hat{P}_\infty.$$

Since the optimal control $\bar{u}(\cdot)$ given by (4.1) is an admissible control (and this holds for any $y_0 \in H$), Lemma 4 implies that the optimal trajectories $\bar{y}(\cdot)$ are exponentially stable. This proves part (ii) of the theorem. Taking $P(t) = P_0 = \hat{P}_\infty$ and $u = \bar{u}$, $y = \bar{y}$ in (3.9) (this is possible because \hat{P}_∞ verifies (1.7), and consequently it verifies (3.5) with ϕ_1 and ϕ_2 given by (1.15)), we have

$$(4.12) \quad \langle \hat{P}_\infty y_0, y_0 \rangle_H \leq \int_0^T \{ |\bar{C}\bar{y}(t)|_V^2 + \langle N\bar{u}(t), \bar{u}(t) \rangle_U \} \, dt + \langle \hat{P}_\infty \bar{y}(T), \bar{y}(T) \rangle_H;$$

as $T \rightarrow \infty$, from (4.12) and the stability of $\bar{y}(\cdot)$ we have

$$(4.13) \quad \langle \hat{P}_\infty y_0, y_0 \rangle_H \leq J_\infty(\bar{u}),$$

where $J_\infty(\cdot)$ is defined by (1.4). From (4.11), (4.13) and (4.2) we have $\hat{P}_\infty = P_\infty$, and also the proof of part (i) of the theorem is complete. \square

Finally, applying Theorem 4 to the case $C = 1$, we find that the following classical result ([P1]) holds also for boundary control problems.

COROLLARY 2. *Assume that (H1) and (3.36) hold. Then system (1.1) is stabilizable if and only if for any $y_0 \in H$ there exists a control $u \in L^2(0, \infty; U)$ such that*

$$\int_0^\infty \{|y(t)|_H^2 + |u(t)|_U^2\} dt < \infty,$$

where y is the corresponding solution of (1.1).

We conclude this section by discussing an example of a Dirichlet boundary control problem over infinite time horizon.

Example 1. Let θ be a bounded open domain of R^n with C^∞ boundary Γ , with θ locally on one side of Γ . Let $A(x, D)$ be the elliptic operator

$$(4.14) \quad A(x, D)f = - \sum_{i,j=1}^n D_j a_{ij}(x) D_i f - cf,$$

where $c \in R$, the real coefficients $a_{ij} \in C^\infty(\bar{\theta})$, and

$$\sum_{i,j=1}^n a_{ij}(x) \xi_i \xi_j \geq \eta |\xi|^2 \quad \text{for any } \xi \in R^n, \quad \text{for some } \eta > 0$$

(more general elliptic operators can be considered, but (4.14) simplifies the discussion of Examples 2 and 3 of § 5.3).

We consider the following optimal control problem governed by a parabolic system with boundary control:

minimize

$$(4.15) \quad J_\infty(u) = \int_0^\infty \left\{ \int_\theta y^2 + \int_\Gamma u^2 \right\} dt$$

over all $u \in L^2(0, \infty; L^2(\Gamma))$, subject to the state equation

$$(4.16) \quad \begin{aligned} \frac{\partial y}{\partial t} &= -A(x, D)y \quad \text{in } [0, \infty[\times \theta, \\ y &= u \quad \text{on } [0, \infty[\times \Gamma, \\ y(0, x) &= y_0(x) \quad \text{in } \theta, \end{aligned}$$

where $y_0 \in L^2(\theta)$. Let $\lambda \geq c$; then the elliptic system

$$(4.17) \quad \begin{aligned} (A(x, D) + \lambda)z &= 0 \quad \text{in } \theta, \\ z &= g \quad \text{on } \Gamma, \end{aligned}$$

has a unique solution $z \in H^{1/2}(\theta)$ for any $g \in L^2(\Gamma)$ ([L6, pp. 187-188]) and we can define the Green mapping

$$(4.18) \quad B \in L(L^2(\Gamma), H^{1/2}(\theta)), \quad Bg = z$$

where z is the solution to (4.17). Setting $y_\lambda(t, x) = e^{-\lambda t} y(t, x)$, from (4.16) we have

$$(4.19) \quad \begin{aligned} \frac{\partial y_\lambda}{\partial t} &= -(A(x, D) + \lambda)y_\lambda \quad \text{in } [0, \infty[\times \theta, \\ y_\lambda &= e^{-\lambda t} u \quad \text{on } [0, \infty[\times \Gamma, \\ y_\lambda(0, x) &= y_0(x) \quad \text{in } \theta. \end{aligned}$$

Let A be the linear operator in $L^2(\theta)$ defined by $(Af)(x) = -A(x, D)f(x)$, for any $f \in D(A) = H^2(\theta) \cap H_0^1(\theta)$. A generates an analytic semigroup e^{tA} , $t \geq 0$, in $L^2(\theta)$ ([T1]). Then the solution to (4.19) is given by ([L1])

$$y_\lambda(t) = e^{t(A-\lambda)}y_0 + \int_0^t (A-\lambda) e^{(t-s)(A-\lambda)} B e^{-\lambda s} u(s) ds,$$

whence the solution of (4.16) is given by (1.1). Since $D_A^\alpha = H^{2\alpha}(\theta)$, for any $0 \leq \alpha < \frac{1}{4}$ ([L1]), assumption (H1) on B is verified. The connection between notation of this example and § 1.1 is given by: $H = L^2(\theta)$, $U = L^2(\Gamma)$, $V = L^2(\theta)$, $C = 1$, $N = 1$.

In [L3] or [T2] it is proved that system (4.16) is stabilizable; moreover (C, A) is detectable because C is an isomorphism. Then Theorems 2, 3 and 4 apply to this problem.

5. The stochastic optimal control problem. The results of this section are the stochastic counterpart of the results of § 4. Similar results in the case of distributed control problems are proved in [I1].

Since stability and stabilizability of stochastic boundary control systems appear new arguments in the literature, we devote to them § 5.2.

5.1. Synthesis.

THEOREM 5. Assume that (H1), (H2) and (3.36) hold. Then there exists a solution to the algebraic Riccati equation (1.10) if and only if (3.10) holds. In this case the results of Theorem 2 hold, and for any $y_0 \in H$ there exists a unique optimal control $\bar{u} \in M_w^2(0, \infty; U)$ for problem (1.9); moreover if (\bar{u}, \bar{y}) is the optimal pair and P_∞ is defined by (3.13), then

$$(5.1) \quad \bar{u}(t) = -\phi_2(P_\infty) K P_\infty \bar{y}(t), \quad t \geq 0,$$

$$(5.2) \quad \langle P_\infty y_0, y_0 \rangle_H = J_\infty(\bar{u}),$$

and \bar{y} verifies equation (2.15) with $G = \phi_2(P_\infty) K P_\infty$.

We omit the proof of this theorem because it is an obvious modification of the proof of Theorem 3.

5.2. Stability and stabilizability of stochastic boundary control systems. The results of this section are simple consequences of known results in distributed control theory ([I1], [H2]), and of (2.18). They are useful to prove Lemma 5 and to study Examples 2 and 3 of § 5.3.

We consider only exponential stability, and stabilizability, in mean square.

DEFINITION 3. (i) We say that system (2.15) is exponentially stable in mean square if there exist two constants $c > 0$ and $k > 0$ such that

$$E|y(t)|_H^2 \leq c e^{-kt} |y_0|_H^2 \quad \text{for any } y_0 \in H.$$

(ii) We say that system (1.8) is exponentially stabilizable in mean square if there exists $G \in L(H, U)$ such that the closed loop system (2.15) is exponentially stable in mean square.

Let $y(\cdot)$ be the solution to (2.15); since $y(\cdot)$ is also solution to (2.18) (in virtue of Lemma 2), we can use already available results on equations of the form (2.18); from [I1, Thm. 3.1] we have the following.

PROPOSITION 5. *The following statements are equivalent:*

- (i) *system (2.15) is exponentially stable in mean square;*
- (ii) *$E \int_0^\infty |y(t)|_H^2 dt < \infty$ for any $y_0 \in H$;*
- (iii) *there exists $P \in \Sigma^+(H)$ such that $\langle Px, A_G y \rangle_H + \langle A_G x, Py \rangle_H + \langle \Delta_1(P)x, y \rangle_H + \langle \Delta_2(P)Gx, Gy \rangle_U = -\langle x, y \rangle_H$ for any $x, y \in D(A_G)$, where A_G is defined in Proposition 1, $\Delta_1(\cdot)$ and $\Delta_2(\cdot)$ in (1.12).*

Similarly, from [H2, Thm. 1], we have the following.

PROPOSITION 6. *Assume that the deterministic system (1.1) is exponentially stabilizable (Definition 2, § 4.2) by means of the feedback control $u(t) = -Gy(t)$, and that*

$$(5.3) \quad \left| \int_0^\infty T^*(t) [\Delta_1(I) + G^* \Delta_2(I) G] T(t) dt \right|_H < 1,$$

where $T(t)$ is defined by (2.12), $\Delta_1(\cdot)$ and $\Delta_2(\cdot)$ by (1.12), and I is the identity in H . Then the stochastic system (1.8) is exponentially stabilizable in mean square, by means of the (stochastic) feedback control $u(t) = -Gy(t)$.

In Examples 2 and 3 of § 5.3 we have to prove the existence of an admissible control for any initial trajectory $y_0 \in H$ (assumption (3.10)); in this direction Proposition 6 will be useful, because if (1.8) is exponentially stabilizable in mean square then (3.10) is verified.

5.3. Stability of optimal trajectories and uniqueness of solution to (1.10).

LEMMA 5. *Assume that (H1) and (H2) hold; assume moreover that (C, A) is detectable, and that there exists $S \in L(V, H)$ such that*

$$(5.4) \quad (\text{trace } W_1) |F_1|_{L(H, L(X_1, H))}^2 \int_0^\infty |e^{t(A-SC)}|_{L(H)}^2 dt < 1,$$

where I is the identity in H . Let $G \in L(H, U)$ be given, and assume that the feedback control $u(t) = -Gy(t)$ (y given by (2.15)) is an admissible control for any $y_0 \in H$; then the trajectories $y(\cdot)$ are exponentially stable in mean square for any $y_0 \in H$.

Proof. Let $y_0 \in H$ be given, and let y and y_n be defined by (2.15) and (2.16), respectively; arguing as in the proof of Lemma 4 we find the equation (which corresponds to (4.8))

$$(5.5) \quad \begin{aligned} y(t) &= e^{t(A-SC)} y_0 + \int_0^t e^{(t-s)(A-SC)} SCy(s) ds \\ &\quad + \int_0^t (SC - A) e^{(t-s)(A-SC)} BGy(s) ds + \int_0^t e^{(t-s)(A-SC)} (\lambda - SC) BGy(s) ds \\ &\quad + \int_0^t e^{(t-s)(A-SC)} F_1(y(s)) dw_1(s) - \int_0^t e^{(t-s)(A-SC)} F_2(Gy(s)) dw_2(s) \\ &= \int_0^t e^{(t-s)(A-SC)} F_1(y(s)) dw_1(s) + h(t), \quad t \geq 0, \end{aligned}$$

where we have introduced for simplicity of notation the stochastic process $h(t)$, defined

by (5.5). Let $\varepsilon > 0$; from (5.5) and the inequality $(a+b)^2 \leq (1+\varepsilon)a^2 + (1+1/\varepsilon)b^2$ we have

$$\begin{aligned}
 E|y(t)|_H^2 &\leq (1+\varepsilon)E\left|\int_0^t e^{(t-s)(A-SC)}F_1(y(s))dw_1(s)\right|_H^2 + \left(1+\frac{1}{\varepsilon}\right)E|h(t)|_H^2 \\
 &= (1+\varepsilon)E\int_0^t \text{trace}(e^{(t-s)(A-SC)}F_1(y(s)))^*W_1(e^{(t-s)(A+SC)}F_1(y(s))) \\
 (5.6) \quad &+ \left(1+\frac{1}{\varepsilon}\right)E|h(t)|_H^2 \\
 &\leq (1+\varepsilon)(\text{trace } W_1)|F_1|_{L(H,L(X_1,H))}^2 \int_0^t |e^{(t-s)(A-SC)}|_{L(H)}^2 E|y(s)|_H^2 ds \\
 &+ \left(1+\frac{1}{\varepsilon}\right)E|h(t)|_H^2,
 \end{aligned}$$

where we have used some well-known properties of the trace and of the stochastic integral ([C2]). Let us take $S \in L(V, H)$ such that (5.4) holds, and let us denote by k the constant on the left side of (5.4); if $\varepsilon > 0$ is such that

$$(5.7) \quad (1+\varepsilon)k < 1;$$

then from (5.6) and a Young inequality we have

$$(5.8) \quad E\int_0^\infty |y(t)|_H^2 dt \leq (1+\varepsilon)kE\int_0^\infty |y(t)|_H^2 dt + (1+1/\varepsilon)E\int_0^\infty |h(t)|_H^2 dt.$$

Arguing as in Lemma 4, we have that $h \in M_w^2(0, \infty; H)$; then from (5.8) and (5.7) we have

$$(5.9) \quad E\int_0^\infty |y(t)|_H^2 dt < \infty.$$

Since (5.9) holds for any initial trajectory $y_0 \in H$, from Proposition 5 we have that $y(\cdot)$ is exponentially stable in mean square for any $y_0 \in H$. \square

THEOREM 6. (i) Assume that (H1), (H2) and (3.36) hold, (C, A) is detectable, and (5.4) holds; then there exists at most one solution of the algebraic Riccati equation (1.10), and in this case the unique solution is equal to P_∞ given by (3.13). (ii) If moreover P_∞ exists, then for any $y_0 \in H$ the optimal trajectory \bar{y} of problem (1.9) is exponentially stable in mean square.

We omit the proof because it is an obvious modification of the proof of Theorem 4. Finally, as in § 4.2, we have the following.

COROLLARY 3. Assume that (H1), (H2) and (3.36) hold. Then system (1.8) is exponentially stabilizable in mean square if and only if for any $y_0 \in H$ there exists $u \in M_w^2(0, \infty; U)$ such that

$$E\int_0^\infty \{|y(t)|_H^2 + |u(t)|_U^2\} dt < \infty.$$

This result follows readily from Theorem 6 with $C = 1$; we remark only that (5.4) is verified if we take $S = \mu$ for a sufficiently large $\mu \geq 0$.

We conclude by discussing two examples of stochastic boundary control problems over infinite time horizon.

Example 2. Let $(\Omega, \mathcal{F}, \mathcal{P})$ be a probability space, and let $w(t)$ be a real Brownian motion; with notation and hypotheses of Example 1, we consider the problem of

$$\begin{aligned}
 (5.10) \quad & \text{minimizing} \\
 & J_\infty(u) = E \int_0^\infty \left\{ \int_\theta y^2 + \int_\Gamma u^2 \right\} dt \\
 & \text{over all } u \in M_w^2(0, \infty; L^2(\Gamma)), \quad \text{subject to} \\
 & dy = -A(x, D)y \, dt + c_1 y \, dw(t) \quad \text{in } [0, \infty[\times \theta, \\
 (5.11) \quad & y = u \quad \text{on } [0, \infty[\times \Gamma, \\
 & y(0, x) = y_0(x) \quad \text{in } \theta,
 \end{aligned}$$

where $c_1 \in R$ is a fixed constant. This problem can be studied in the abstract setting of § 1.1 if we put $X_1 = X_2 = R$, $W_1 = 1$, $W_2 = 0$, $F_1(y) = c_1 y$ for any $y \in L^2(\theta)$, and consequently $\phi_1(P) = c_1^2 P$ and $\phi_2(P) = N^{-1}$ for any $P \in \Sigma^+(H)$. Assume that

$$(5.12) \quad c < 0 \text{ (hence } A \text{ stable)} \quad \text{and} \quad -c_1^2/2c < 1,$$

where c is given in (4.14); then the hypotheses of Proposition 6 are verified with $G = 0$; since (C, A) is detectable, we can apply to this example all results of the paper.

In the following example we discuss a problem in which it is possible to avoid the restrictive assumption (5.12).

Example 3. With notation and hypotheses of Example 2 except from (5.12), let us consider the stochastic system with distributed and boundary controls

$$\begin{aligned}
 (5.13) \quad & dy = -A(x, D)y \, dt + c_1 y \, dw(t) + c_2 u_1 \, dt \quad \text{in } [0, \infty[\times \theta, \\
 & y = u_2 \quad \text{on } [0, \infty[\times \Gamma, \\
 & y(0, x) = y_0(x) \quad \text{in } \theta,
 \end{aligned}$$

where u_1 and u_2 are two different controls, and c_2 is a positive constant. Let us consider the problem of minimizing

$$(5.14) \quad J_\infty(u) = E \int_0^\infty \left\{ \int_\theta (y^2 + n_1 u_1^2) + \int_\Gamma n_2 u_2^2 \right\} dt$$

over all pairs $(u_1, u_2) \in M_w^2(0, \infty; L^2(\theta)) \cap M_w^2(0, \infty; L^2(\Gamma))$, subject to (5.13); here n_1 and n_2 are positive real numbers. For “large” n_1 (resp. “small” c_2) we can see (5.13) as a system in which it is possible to implement a distributed control, but its use is strongly penalized by (5.14) (resp. it has a small effect on the system); in a sense, this is a problem controlled “mainly” on the boundary. With $U = L^2(\theta) \times L^2(\Gamma)$, $N \in \Sigma^+(U)$ defined by $N(u_1, u_2) = (n_1 u_1, n_2 u_2)$ for any $(u_1, u_2) \in U$, $\tilde{B} \in L(U, D_A^\alpha)$ defined by $\tilde{B}(u_1, u_2) = (A - \lambda)^{\alpha-1} u_1 + B u_2$ where B is given by (4.18), we see that this problem fits in with the abstract scheme of § 1.1, with \tilde{B} in place of B . Moreover, (5.13) is stable

and hypotheses of Proposition 6 are verified, by taking the feedback control $(u_1, u_2) = (-\mu y, 0)$, for a sufficiently large $\mu \geq 0$ such that $c_1^2/(c_2\mu - c) < 1$. Consequently, the results of the present paper can be applied.

REFERENCES

- [B1] A. V. BALAKRISHNAN, *Fractional powers of closed operators and the semigroup generated by them*, Pacific J. Math., 10 (1960), pp. 419–437.
- [B2] ———, *Boundary control of parabolic equations: L–Q–R theory*, in Theory of Nonlinear Operators, Proc. Fifth Intern. Summer School, Berlin, 1977.
- [C1] R. F. CURTAIN, *Estimation theory for abstract evolution equations excited by general white noise processes*, this Journal, 14 (1976), pp. 1124–1150.
- [C2] R. F. CURTAIN AND P. L. FALB, *Ito's lemma in infinite dimensions*, J. Math. Anal. Appl., 31 (1970), pp. 434–448.
- [C3] ———, *Stochastic differential equations in Hilbert space*, J. Differential Equations, 10 (1971), pp. 412–430.
- [D1] G. DA PRATO AND P. GRISVARD, *Sommes d'opérateurs linéaires et équations différentielles opérationnelles*, J. Math. Pures Appl., Serie IX, 54 (1975), pp. 305–387.
- [D2] G. DA PRATO AND A. ICHIKAWA, *Riccati equations with unbounded coefficients*, Scuola Normale Superiore, May 1984.
- [D3] R. DATKO, *Extending a theorem of A. M. Liapunov to Hilbert space*, J. Math. Anal. Appl., 32 (1970), pp. 610–616.
- [D4] ———, *A linear control problem in an abstract Hilbert space*, J. Differential Equations, 9 (1971), pp. 346–359.
- [D5] L. DE SIMON, *Un'applicazione della teoria degli integrali singolari allo studio delle equazioni differenziali astratte del primo ordine*, Rend. Sem. Mat. Univ. Padova, 34 (1964), pp. 205–223.
- [F1] F. FLANDOLI, *Riccati equation arising in a stochastic optimal control problem with boundary control*, Boll. Un. Mat. Ital. C(6), I(1982), pp. 377–393.
- [F2] ———, *Riccati equation arising in a boundary control problem with distributed parameters*, this Journal, 22 (1984), pp. 76–86.
- [F3] ———, *Direct solution of a Riccati equation arising in a stochastic control problem with control and observation on the boundary*, Appl. Math. Optim., to appear.
- [H1] G. H. HARDY, J. E. LITTLEWOOD AND G. POLYA, *Inequalities*, Cambridge University Press, Cambridge, 1934.
- [H2] U. G. HAUSSMANN, *Asymptotic stability of the linear Ito's equation in infinite dimensions*, J. Math. Anal. Appl., 65 (1978), pp. 219–235.
- [I1] A. ICHIKAWA, *Dynamic programming approach to stochastic evolution equations*, this Journal, 17 (1979), pp. 152–173.
- [I2] ———, *A semigroup model for parabolic equations with boundary and pointwise noise*, Workshop on Stochastic Space-Time Models and Limit Theorems, University of Bremen, November 1983.
- [K1] T. KATO, *Perturbation Theory of Linear Operators*, Springer-Verlag, New York–Berlin, 1966.
- [L1] I. LASIECKA, *Unified theory for abstract parabolic boundary problem a semigroup approach*, Appl. Math. Optim., 6 (1980), pp. 287–333.
- [L2] I. LASIECKA AND R. TRIGGIANI, *Dirichlet boundary control problem for parabolic equations with quadratic cost: analyticity and Riccati's feedback synthesis*, this Journal, 21 (1983) pp. 41–67.
- [L3] ———, *Stabilization and structural assignment of Dirichlet boundary feedback parabolic equations*, this Journal, 21 (1983), pp. 766–803.
- [L4] ———, *Feedback semigroups and cosine operators for boundary feedback parabolic and hyperbolic equations*, J. Differential Equations, 47 (1983), pp. 246–272.
- [L5] J. L. LIONS, *Optimal Control of Systems Governed by Partial Differential Equations*, Springer-Verlag, New York, 1971.
- [L6] J. L. LIONS AND E. MAGENES, *Nonhomogeneous Boundary Value Problems and Applications*, I, Springer-Verlag, Berlin, 1972.
- [L7] D. L. LUKES AND D. L. RUSSEL, *The quadratic criterion for distributed systems*, this Journal, 7 (1967), pp. 101–121.
- [P1] A. J. PRITCHARD AND J. ZABCZYK, *Stability and stabilizability of infinite-dimensional systems*, Control Theory Center Report 70, University of Warwick, November 1977.
- [S1] M. SORINE, *Une resultat d'existence et unicité pour l'équation de Riccati stationnaire*, Rapport INRIA no. 55, 1981.

- [S2] M. SORINE, *Sur le semigroupe non linéaire associé à l'équation de Riccati*, Rapport du CRMA no. 1055, Université Montreal, 1981.
- [T1] H. TANABE, *Equation of Evolution, Monographs and Studies in Mathematics*, Pitman, London, 1979.
- [T2] R. TRIGGIANI, *Boundary feedback stabilizability of parabolic equations*, Appl. Math. Optim., 6 (1980), pp. 201–220.
- [Y1] K. YOSIDA, *Functional Analysis*, 123, Springer-Verlag, Berlin, 1965.
- [Z1] J. ZABCZYK, *Remarks on the algebraic Riccati equation in Hilbert space*, Appl. Math. Optim., 2 (1976), pp. 251–256.
- [Z2] ———, *On decomposition of generators*, this Journal, 16 (1978), pp. 523–534.

ON THE SUPREMAL CONTROLLABLE SUBLANGUAGE OF A GIVEN LANGUAGE*

W. M. WONHAM† AND P. J. RAMADGE‡

Abstract. The concept of controllable language has been shown to play a basic role in the existence theory of supervisory controls for discrete event processes. In this paper the supremal controllable sublanguage S of a given language L is characterized as the largest fixpoint of a monotone operator Ω . In the case where the languages involved are regular it is shown that the fixpoint S can be computed as the limit of the (finite) sequence $\{K_j\}$ given by $K_{j+1} = \Omega(K_j)$, $K_0 = L$. An effective computational algorithm is developed, and three examples are provided for illustration.

Key words. supervisory control, discrete event process, controllable language

AMS(MOS) subject classification. 93B

1. Introduction. The concept of controllable language was introduced in Ramadge and Wonham [1983], where it was shown to play a basic role in the existence theory of supervisory controls for discrete event processes. In this paper we characterize the supremal controllable sublanguage S of a given language L as the largest fixpoint of a certain operator Ω . In the case where the languages involved are regular we show that the fixpoint S can be computed as the limit of the (finite) sequence of languages K_j given by

$$K_{j+1} = \Omega(K_j), \quad K_0 = L.$$

Three computational examples are provided for illustration.

A summary version of this paper has appeared as Wonham and Ramadge [1984].

We recall here only the facts needed in the development to follow. Let Σ be a (nonempty) finite alphabet of elements σ , and as usual denote by Σ^* the set of all finite strings of elements of Σ , including the empty string 1. A subset $L \subset \Sigma^*$ is a *language* over Σ . A string $s \in L$ is a *word* of L . If $s, s', t \in \Sigma^*$ with $s't = s$, then s' is a *prefix* of s ; thus both 1 and s are prefixes of s . The *closure* of L is the language \bar{L} consisting of all the prefixes of words in L ; thus if $L = \emptyset$ then $\bar{L} = \emptyset$, and if $L \neq \emptyset$ then $1 \in \bar{L}$. Clearly $L \subset \bar{L}$; L is *closed* if $L = \bar{L}$. Let M be a fixed language over Σ and let Σ_0 be a fixed subset of Σ . A language $K \subset \Sigma^*$ is *controllable* (with respect to M and Σ_0) if

$$\bar{K}\Sigma_0 \cap M \subset \bar{K}.$$

In the supervisory control theory of Ramadge and Wonham [1983] this condition has the following interpretation: Σ is an alphabet of “events”; $\Sigma - \Sigma_0$ is the subset of “controlled events” that can be enabled or disabled by a “supervisor”; thus Σ_0 is the subset of “uncontrolled” events that cannot be disabled; in the absence of control, M is the (closed) language generated by an automaton (or generator) \mathcal{G} over the alphabet Σ ; K (in this context) is a specified sublanguage of M ; \mathcal{G} can be controlled by a

* Received by the editors August 17, 1984; accepted for publication (in revised form) March 24, 1986.

† Systems Control Group, Department of Electrical Engineering, University of Toronto, Toronto, Ontario, Canada M5S 1A4. The work of this author was partially supported by the Natural Sciences and Engineering Research Council of Canada grant A-7399.

‡ Department of Electrical Engineering and Computer Science, Princeton University, Princeton, New Jersey 08544.

supervisor to generate exactly $\bar{K} \subset M$ just when K is controllable, namely strings in M of the form $t\sigma_0$, with $t \in \bar{K}$ and $\sigma_0 \in \Sigma_0$, are again in \bar{K} . In the present article we need not assume until § 6 that M is closed or that $K \subset M$.

Let $L \subset \Sigma^*$ be arbitrary. Then the *supremal controllable sublanguage* of L is the language

$$\sup C(L) := \bigcup \{K : K \subset L \text{ and } K \text{ is controllable}\}.$$

It was shown in Ramadge and Wonham [1983] that $\sup C(L)$ is well defined (although it may be empty), is controllable, and contains every controllable sublanguage of L .

For a very simple instance of these ideas the reader may glance ahead to Example 3 (§ 7.1). Here the events (labelled) α_1, α_2 are controlled (i.e., can be disabled, viz. prevented from occurring) whereas events β are uncontrolled (i.e., are permanently enabled), namely $\Sigma_0 = \{\beta\}$. Let \mathcal{A} be an automaton whose behavior (without disablements) is the language

$$M = \overline{(\alpha_1\beta^2 + \alpha_2)\beta^*}.$$

With respect to M and Σ_0 the language $N := \{\alpha_1\beta^2\}$ is uncontrollable, since $(\alpha_1\beta^2)\beta \in M$ but $(\alpha_1\beta^2)\beta \notin N$. It is easily seen (cf. § 7.1) that $\sup C(N) = \emptyset$. Intuitively, no mechanism that is capable of disabling only α_1 and α_2 could control \mathcal{A} to generate exactly N . If α_1, α_2 are both disabled then \mathcal{A} generates $\{1\} = \sup C(\bar{N})$.

2. A fixpoint characterization of $\sup C(L)$. Let $L, M \subset \Sigma^*$ be two fixed languages, and let Σ_0 be a fixed subset of Σ . Let \mathcal{L} be the set of all languages over Σ (i.e. \mathcal{L} = power set of Σ^*), and define the operator

$$\Omega : \mathcal{L} \rightarrow \mathcal{L}$$

according to

$$(2.1) \quad \Omega(K) = L \cap \sup \{T : T \subset \Sigma^*, T = \bar{T} \text{ and } T\Sigma_0 \cap M \subset \bar{K}\}, \quad K \in \mathcal{L}.$$

It is easy to check (cf. Ramadge and Wonham [1983, proof of Prop. 7.1]) that the condition in $\{ \}$ of (2.1) is closed under arbitrary unions, so Ω is well defined. Of course Ω depends on the fixed elements L, M and Σ_0 .

If $t \in \Sigma^*$ write \bar{t} for $\overline{\{t\}}$, the set of prefixes of the string t . The following alternative description of $\Omega(K)$ will be useful.

LEMMA 2.1.

$$\Omega(K) = \{t : t \in L \text{ and } \bar{t}\Sigma_0 \cap M \subset \bar{K}\}.$$

Proof. Immediate from (2.1). \square

Let $S := \sup C(L)$, the supremal controllable sublanguage of L (with respect to M and Σ_0). We now have the following.

PROPOSITION 2.1. $S = \Omega(S)$ and $S \supset K$ for every K such that $K = \Omega(K)$.

A language K such that $K = \Omega(K)$ is a *fixpoint* of Ω . Thus Proposition 2.1 describes $\sup C(L)$ as the largest fixpoint of Ω . Of incidental interest is the following.

COROLLARY. If $S := \sup C(L)$ then S is L -closed, namely

$$S = \bar{S} \cap L.$$

Proof of Corollary. By Proposition 2.1 we can write

$$S = \Omega(S) = L \cap T$$

for some closed T . Therefore $T \supset \bar{S}$,

$$L \cap T = S \subset L \cap \bar{S} \subset L \cap T,$$

and so $S = L \cap \bar{S}$. \square

Proof of Proposition 2.1. We have that $S \subset L$ and S is controllable, i.e.,

$$\bar{S}\Sigma_0 \cap M \subset \bar{S},$$

so

$$\bar{S} \subset \sup \{T: T = \bar{T} \text{ and } T\Sigma_0 \cap M \subset \bar{S}\},$$

i.e.,

$$S \subset L \cap \bar{S} \subset \Omega(S).$$

To prove the reverse inclusion, let $t \in \Omega(S)$. By Lemma 2.1

$$t \in L \quad \text{and} \quad \bar{t}\Sigma_0 \cap M \subset \bar{S}.$$

If $S' = S \cup \{t\}$ then $S' \subset L$, and

$$\begin{aligned} \bar{S}'\Sigma_0 \cap M &= (\bar{S} \cup \bar{t})\Sigma_0 \cap M \\ &= (\bar{S}\Sigma_0 \cap M) \cup (\bar{t}\Sigma_0 \cap M) \\ &\subset \bar{S} \\ &\subset \bar{S}'. \end{aligned}$$

Therefore $S' \in \mathbf{C}(L)$ and $S' \supset S$. Because S is supremal, $S' = S$, i.e., $t \in S$ and so $\Omega(S) \subset S$. This gives $S = \Omega(S)$, namely S is a fixpoint of Ω .

Now let K be any fixpoint of Ω . It will be shown that $K \subset S$. We have $K \subset L$, and

$$\begin{aligned} K &\subset \sup \{T: T = \bar{T} \text{ and } \bar{T}\Sigma_0 \cap M \subset \bar{K}\} \\ &=: T^+, \quad \text{say.} \end{aligned}$$

It is enough to check that K is controllable. Now since T^+ is closed

$$\bar{K} \subset T^+$$

and since the defining condition in $\{ \}$ is closed under arbitrary unions

$$T^+\Sigma_0 \cap M \subset \bar{K}.$$

Thus

$$\bar{K}\Sigma_0 \cap M \subset \bar{K}$$

and the proof is complete. \square

In view of Proposition 2.1 it is natural to attempt to compute $S = \sup \mathbf{C}(L)$ by iteration of Ω . For this, define the sequence

$$(2.2) \quad \begin{aligned} K_0 &= L, \\ K_{j+1} &= \Omega(K_j), \quad j = 0, 1, \dots \end{aligned}$$

PROPOSITION 2.2. *The (set-theoretic) limit*

$$K := \lim K_j \quad (j \rightarrow \infty)$$

exists and $S \subset K$.

Proof. It is clear that Ω is monotone, i.e., if $A \subset B \subset \Sigma^*$ then $\Omega(A) \subset \Omega(B)$. Then

$$K_1 = \Omega(K_0) \subset L = K_0,$$

and if $K_j \subset K_{j-1}$ we have

$$K_{j+1} = \Omega(K_j) \subset \Omega(K_{j-1}) = K_j.$$

Hence

$$K_0 \supset K_1 \supset K_2 \supset \cdots$$

so that

$$K = \lim K_j = \bigcap_{j=0}^{\infty} K_j$$

exists. Now

$$S = \Omega(S) \subset L = K_0,$$

and $S \subset K_j$ implies

$$S = \Omega(S) \subset \Omega(K_j) = K_{j+1},$$

so $S \subset K$. \square

It is not true in general that $K \subset S$ (see § 4); however, in the regular case, which we discuss next, it is true that $S = K$ and also that our iteration scheme is effective.

3. Computation of $\sup C(L)$ in the regular case. To define the regular case we first recall the definition of Nerode equivalence (e.g., Hopcroft and Ullman [1979, p. 65]). Let $L \subset \Sigma^*$. Strings $s, t \in \Sigma^*$ are *equivalent* (mod L), written

$$s \equiv t \pmod{L} \quad \text{or} \quad s \equiv_L t,$$

if

$$\{s': s' \in \Sigma^* \text{ and } ss' \in L\} = \{t': t' \in \Sigma^* \text{ and } tt' \in L\}.$$

In other words, two strings are equivalent (mod L) if they are each “continuable” in the same ways to form completed words of L . We shall write

$$\|L\| := \text{card}(\Sigma^*/\equiv_L),$$

namely $\|L\|$ is the (possibly denumerable) number of equivalence classes of \equiv_L in Σ^* . For example if $L = \emptyset$ or $L = \Sigma^*$ then any pair of strings are equivalent, so $\|L\| = 1$. If $\Sigma = \{\alpha, \beta\}$ and

$$L = \alpha^* = \{1, \alpha, \alpha\alpha, \dots\},$$

then $\|L\| = 2$; whereas if $\Sigma = \{\alpha, \beta\}$ with

$$L = \{\alpha^n \beta^n; n = 0, 1, 2, \dots\},$$

then $\|L\| = \infty$. Note that an inclusion $K \subset L$ does not imply any particular ordering of the “norms” $\|K\|, \|L\|$.

The language L is *regular* if $\|L\| < \infty$. In that case, it is well known that $\|L\|$ is the minimal cardinality of the state set of an automaton that “recognizes” L ; so $\|L\|$ roughly measures the size of “memory” in L .

We now state the main result of this section.

THEOREM 3.1. *In case the languages L and M are regular, the sequence of languages K_j defined by*

$$K_0 = L, \quad K_{j+1} = \Omega(K_j), \quad j \geq 0$$

converges after a finite number of terms to $S(= \sup C(L))$. Furthermore, S is a regular language, with

$$\|S\| \leq \|L\| \|M\| + 1.$$

For the proof of Theorem 3.1 we shall require the following three lemmas.

LEMMA 3.1. Let $A, B, C \subset \Sigma^*$. Suppose that for all $s, t \in \Sigma^*$ the conditions $s, t \in \bar{C}$; $s \equiv_A t$; $s \equiv_B t$ together imply $s \equiv_C t$. Then

$$\|C\| \leq \|A\| \|B\| + 1.$$

Proof. For each equivalence relation π on Σ^* we write

$$\|\pi\| = \text{card}(\Sigma^*/\pi).$$

If ρ is also an equivalence relation on Σ^* , we say that π refines ρ if, for all $s, t \in \Sigma^*$, $s \equiv t \pmod{\pi}$ implies $s \equiv t \pmod{\rho}$. If π refines ρ then obviously $\|\pi\| \geq \|\rho\|$.

Now define the equivalence relation π on Σ^* according to

$$s \equiv t \pmod{\pi} \quad \text{iff} \quad s \equiv_A t \quad \text{and} \quad s \equiv_B t.$$

Let π_C be the restriction of π to $\bar{C} \subset \Sigma^*$: namely if we identify π with $\Pi \subset \Sigma^* \times \Sigma^*$ then π_C is identified with $\Pi_C = \Pi \cap (\bar{C} \times \bar{C})$. Let ρ denote \equiv_C and let ρ_C denote the restriction of ρ to \bar{C} . Then the hypothesis states that π_C refines ρ_C , so we have

$$\|\rho_C\| \leq \|\pi_C\| \leq \|\pi\| \leq \|A\| \|B\| + 1.$$

The subset $\Sigma^* - \bar{C} \subset \Sigma^*$ is either empty or exactly one equivalence class of ρ , and therefore

$$\|\rho\| \leq \|\rho_C\| + 1 \leq \|A\| \|B\| + 1$$

as claimed. \square

LEMMA 3.2. For $m = 1, 2, \dots$, let

$$\mathcal{L}_m = \{L: L \subset \Sigma^* \text{ and } \|L\| \leq m\}.$$

Then $\text{card}(\mathcal{L}_m) < \infty$.

Proof. In light of the fact about recognizers cited earlier, the lemma states that only a finite number of distinct languages can be defined over a fixed finite alphabet by means of a recognizer of a priori fixed, finite state cardinality m . This conclusion is immediate from the standard description of a finite state recognizer (Hopcroft and Ullman [1979]) and elementary combinatorics. \square

LEMMA 3.3. Let K_j ($j = 0, 1, 2, \dots$) be a sequence of regular languages over Σ , such that

- (i) $K_0 \supset K_1 \supset K_2 \supset \dots$, and
- (ii) there is a finite constant m for which

$$\|K_j\| \leq m$$

for all j .

Then there is an index k such that

$$K_j = K_k, \quad j \geq k.$$

The lemma states that a monotone, “norm”-bounded sequence of regular languages is finitely convergent.

Proof. In the notation of Lemma 3.2 we have by (ii) that $K_j \in \mathcal{L}_m$ for all j . By the finiteness of \mathcal{L}_m there is $K \in \mathcal{L}_m$ such that K occurs in the sequence $\{K_j\}$ infinitely often. The result is then immediate by the monotonicity (i). \square

We are now ready to discuss the computation of $\sup C(L)$. Reverting to the notation of § 2, we make the assumption:

(3.1) The languages L, M are regular.

Our key result is the following.

LEMMA 3.4. For the sequence $\{K_j\}$ defined by (2.2),

$$\|K_j\| \leq \|L\| \|M\| + 1, \quad j \geq 0.$$

Proof. By Lemma 2.1 we have, for $j \geq 1$,

$$\begin{aligned} K_j &= \Omega(K_{j-1}) \\ &= \{t: t \in L \text{ and } t\Sigma_0 \cap M \subset \bar{K}_{j-1}\}. \end{aligned}$$

It will be shown that, for all $i \geq 0$, the conditions

$$(3.2)_i \quad s, t \in \bar{K}_i, \quad s \equiv_L t, \quad s \equiv_M t,$$

together imply

$$(3.3)_i \quad s \equiv t \pmod{K_i}.$$

For this we fix $j \geq 1$ and make the inductive assumption that

$$(3.2)_i \text{ implies } (3.3)_i \text{ for } 0 \leq i < j.$$

Since $K_0 = L$, the assertion for $j = 1$ is trivial. Observe next that if $(3.2)_j$ is true for a pair $s, t \in \Sigma^*$ then, because $K_j \subset K_{j-1}$, we have that $(3.2)_{j-1}$ is also true for s, t ; and therefore $(3.3)_{j-1}$ is true.

For the inductive step assume that $(3.2)_j$ is true for a pair $s, t \in \Sigma^*$, and suppose that $sq \in K_j$ for some $q \in \Sigma^*$. Then

$$sq \in L \text{ and } \overline{sq}\Sigma_0 \cap M \subset \bar{K}_{j-1}.$$

Since $s \equiv_L t$ we have $tq \in L$. Next, suppose t' is a prefix of tq . If t' is a prefix of t then

$$t'\Sigma_0 \cap M \subset \bar{K}_{j-1}$$

because $t \in \bar{K}_j$. If $t' = tp$ for some prefix p of q , and if $tp\sigma \in M$ for some $\sigma \in \Sigma_0$, then $s \equiv_M t$ implies that $sp\sigma \in M$; but sp is a prefix of sq , so that $sp\sigma \in \bar{K}_{j-1}$, namely $sp\sigma r \in K_{j-1}$ for some $r \in \Sigma^*$. By the inductive hypothesis, $s \equiv t \pmod{K_{j-1}}$, so that $tp\sigma r \in K_{j-1}$, namely $tp\sigma \in \bar{K}_{j-1}$. We have now shown that

$$\overline{tq}\Sigma_0 \cap M \subset \bar{K}_{j-1},$$

and so $tq \in K_j$. Thus for $q \in \Sigma^*$,

$$sq \in K_j \text{ implies } tq \in K_j.$$

By symmetry the reverse implication is also true, and therefore $s \equiv t \pmod{K_j}$ as claimed.

We have established that $(3.2)_i$ implies $(3.3)_i$ for all i . We now apply Lemma 3.1 with

$$A = L, \quad B = M, \quad C = K_i$$

and the proof is complete. \square

Finally we are ready for the proof of Theorem 3.1.

Proof of Theorem 3.1. By the proof of Proposition 2.2 the sequence of languages $\{K_j\}$ is monotone decreasing, and by Lemma 3.4 the numerical sequence $\|K_j\|$ is

bounded by $\|L\| \|M\| + 1$. Thus by Lemma 3.3, the sequence K_j is finitely convergent to a limit K . But this implies that K is a fixpoint of Ω . Thus $K \subset S$. The result now follows by Proposition 2.2. \square

4. Counterexamples. It was shown in the previous section that if L and M are regular, then $\sup \mathbf{C}(L) = \bigcap_{j=0}^{\infty} K_j$, i.e., that $\bigcap_{j=0}^{\infty} K_j$ is a fixpoint of the operator Ω . At the moment it remains an open question as to whether this result is also true for more general classes of languages. It is certainly not true for arbitrary languages L, M as the following example illustrates.

4.1. Example 1. Let

$$\begin{aligned}\Sigma &= \{\alpha, \beta, \gamma\}, \quad \Sigma_0 = \{\alpha, \gamma\}, \quad \text{and} \\ L &= \alpha\beta + \alpha\{\gamma^{n+1}\beta\gamma^m : 0 \leq m \leq n, n \geq 0\}, \\ M &= \overline{\alpha\beta + \alpha\gamma\gamma^*\beta\gamma^*}.\end{aligned}$$

An infinite tree recognizer for L is shown in Fig. 4.1. Recalling that

$$\Omega(K_j) = \{t : t \in L \text{ and } \bar{t}\Sigma_0 \cap M \subset \bar{K}_j\}$$

and using Fig. 4.1, we have

$$K_j = \alpha\beta + \alpha\gamma^j\{\gamma^{n+1}\beta\gamma^m : 0 \leq m \leq n, n \geq 0\}.$$

Thus

$$\bigcap_{j=0}^{\infty} K_j = \alpha\beta.$$

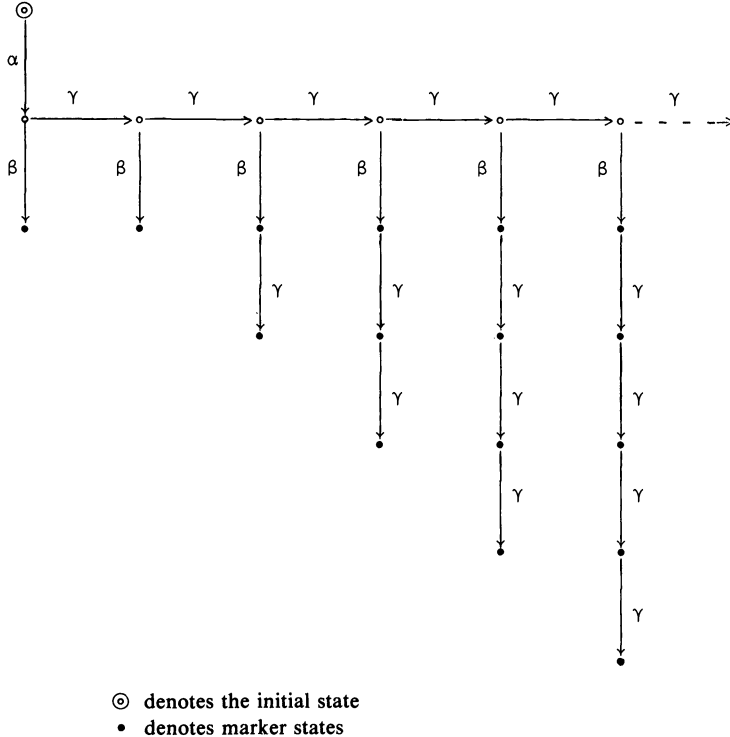


FIG. 4.1. Example 1. Tree recognizer for L .

By inspection of Fig. 4.1 it is evident that $\sup C(L) = \emptyset$. Thus

$$\sup C(L) \neq \bigcap_{j=0}^\infty K_j.$$

It was shown in § 3 that if L and M are regular languages, then $\sup C(L)$ is also a regular language. The conclusion may be false, however, if either L or M fails to be regular. This is illustrated in Example 2.

4.2. Example 2. Let

$$\Sigma = \{\alpha, \beta, \gamma\}, \quad \Sigma_0 = \{\alpha, \gamma\}, \quad \text{and}$$

$$L = \alpha\alpha^*\beta\beta^*,$$

$$M = L + \{\alpha^n\beta^{n+1}\gamma: n \geq 1\}.$$

The infinite tree recognizer for M is shown in Fig. 4.2. Recalling that

$$\Omega(L) = \{t: t \in L \text{ and } \bar{t}\Sigma_0 \cap M \subset \bar{L}\}$$

and using Fig. 4.2 we have

$$\Omega(L) = \{\alpha^n\beta^m: 1 \leq m \leq n, n \geq 1\}$$

and

$$\Omega^2(L) = \Omega(L).$$

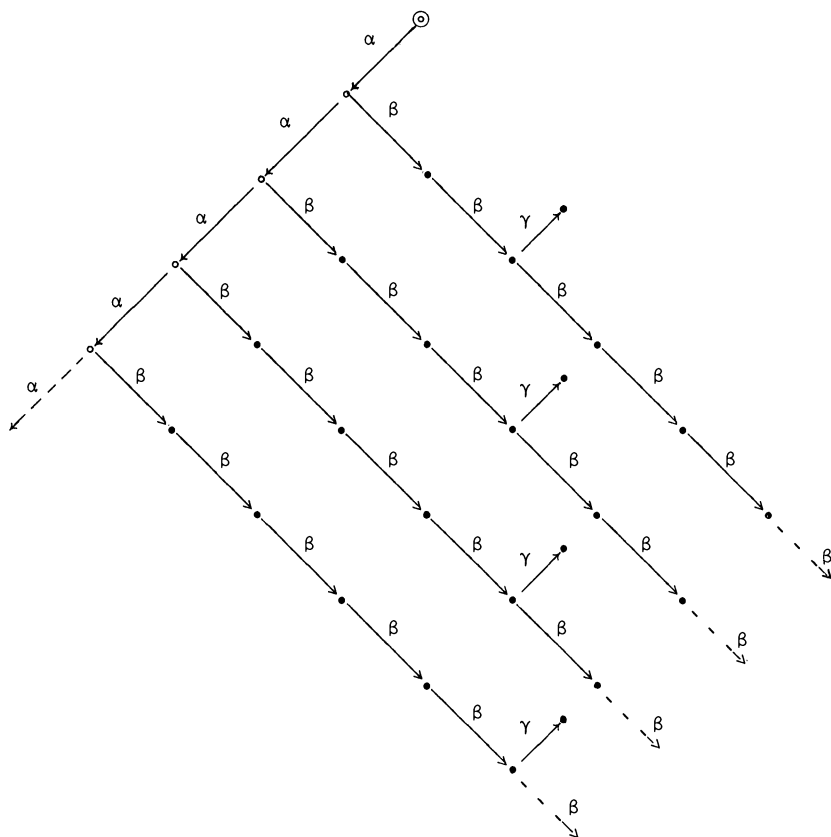


FIG. 4.2. Example 2. Tree recognizer for M .

Thus in this case

$$\sup \mathbf{C}(L) = \bigcap_{j=1}^{\infty} K_j,$$

but $\sup \mathbf{C}(L)$ is not a regular language.

5. Effective computability of the operator Ω . We return to the case when L and M are regular languages and show that our proposed iteration scheme yields an effective procedure for the computation of $\sup \mathbf{C}(L)$. For this it will be sufficient, in view of Theorem 3.1, to show that the operator Ω can be effectively computed.

We begin by reviewing some results from the theory of regular languages. Recall that a language $K \subset \Sigma^*$ is regular iff there exists a finite automaton

$$\mathcal{A} = (Q, \Sigma, \delta, q_0, Q_m)$$

with

$$K = \{s : s \in \Sigma^* \text{ and } \delta(s, q_0) \in Q_m\}.$$

In this case we write $|\mathcal{A}| = K$. The automaton \mathcal{A} provides an effective means of specifying the regular language K . Henceforth when we say that a regular language $K \subset \Sigma^*$ is given, we shall understand that a finite automaton \mathcal{A} is supplied with $|\mathcal{A}| = K$.

Let $\mathcal{O} : \mathcal{L}^n \rightarrow \mathcal{L}$ be an operator which preserves regularity, i.e., K_1, \dots, K_n regular implies $\mathcal{O}(K_1, \dots, K_n)$ regular. We say that \mathcal{O} is *effectively computable* if from each n -tuple (K_1, \dots, K_n) of regular languages we can effectively construct a finite automaton \mathcal{A} with $|\mathcal{A}| = \mathcal{O}(K_1, \dots, K_n)$. For example it is well known (see e.g., Eilenberg [1974]) that each of the four standard operators: closure, complement, union and intersection, preserves regularity and is effectively computable.

Define $H : \mathcal{L} \rightarrow \mathcal{L}$ by

$$H(K) := \sup \{T : T \subset K \text{ and } T = \bar{T}\}$$

and for fixed $\Sigma_0 \subset \Sigma$ define $B_0 : \mathcal{L} \rightarrow \mathcal{L}$ by

$$B_0(K) := \{w : w \in \Sigma^* \text{ and } w\sigma \in K \text{ for all } \sigma \in \Sigma_0\}.$$

LEMMA 5.1. *The operators H and B_0 preserve regularity and are effectively computable.*

Proof. Let $\mathcal{A} = (Q, \Sigma, \delta, q_0, Q_m)$ be a finite automaton with $|\mathcal{A}| = K$.

(1) If $q_0 \notin Q_m$, then clearly $H(K) = \emptyset$. If $q_0 \in Q_m$, then define

$$X = Q_m,$$

$$x_0 = q_0,$$

$$X_m = X,$$

and $\xi : \Sigma \times X \rightarrow X$ by

$$\xi(\sigma, x) = \begin{cases} \delta(\sigma, x) & \text{if } \delta(\sigma, x) \in Q_m, \\ \text{undefined} & \text{otherwise.} \end{cases}$$

Let

$$\mathcal{B} := (X, \Sigma, \xi, x_0, X).$$

Then

$$\begin{aligned} w \in |\mathcal{B}| & \quad \text{iff } \bar{w} \in |\mathcal{A}|, \\ & \quad \text{iff } w \in H(K). \end{aligned}$$

(2) Let

$$X_m = \{q: q \in Q \text{ and } \delta(\sigma, q) \in Q_m \text{ for all } \sigma \in \Sigma_0\},$$

$$\mathcal{B} = (Q, \Sigma, \delta, q_0, X_m).$$

Then

$$w \in |\mathcal{B}| \quad \text{iff } w\sigma \in |\mathcal{A}| \text{ for all } \sigma \in \Sigma_0,$$

$$\text{iff } w \in B_0(K).$$

□

The following lemma is the key to an effective computation of Ω .

LEMMA 5.2. *For each $K \subset \Sigma^*$*

$$\Omega(K) = L \cap H \cdot B_0(\bar{K} \cup M^c).$$

Proof. For each $K \subset \Sigma^*$

$$\Omega(K) = \{w \mid w \in L \text{ and } \bar{w}\Sigma_0 \cap M \subset \bar{K}\}.$$

Hence

$$\begin{aligned} w \in \Omega(K) & \quad \text{iff } w \in L \text{ and } \bar{w}\Sigma_0 \cap M \subset \bar{K}, \\ & \quad \text{iff } w \in L \text{ and } \bar{w}\Sigma_0 \subset \bar{K} \cup M^c, \\ & \quad \text{iff } w \in L \text{ and } \bar{w} \subset B_0(\bar{K} \cup M^c), \\ & \quad \text{iff } w \in L \text{ and } w \in H \cdot B_0(\bar{K} \cup M^c), \\ & \quad \text{iff } w \in L \cap H \cdot B_0(\bar{K} \cup M^c). \end{aligned}$$

□

Finally we have our desired result.

PROPOSITION 5.1. *The operator Ω is effectively computable.*

Proof. The assertion is immediate from Lemmas 5.1 and 5.2 and the effective computability of closure, complement, union and intersection. □

6. Alternative computation of Ω . In the applications so far considered (Ramadge and Wonham [1983]) the language M is closed (i.e., $M = \bar{M}$) and represents the uncontrolled, “physically possible” behavior of a discrete event process; while $L \subset M$ is the “legal” language describing behavior that is both possible and acceptable. For this special case we present a more efficient computation of the operator Ω . This method will be used in the presentation of examples in § 7.

Recall that for the finite automaton

$$\mathcal{A} = (Q, \Sigma, \delta, q_0, Q_m)$$

the transition function $\delta: \Sigma \times Q \rightarrow Q$ is in general only a partial function, meaning that, for each $q \in Q$, $\delta(\sigma, q)$ is only defined for a subset $\Sigma(q) \subset \Sigma$ that depends on q . We call $\Sigma(q)$ the *active set* at the state q . Define the *empty automaton* over Σ , denoted \emptyset_Σ , by

$$\emptyset_\Sigma := (\emptyset, \Sigma, \emptyset, \emptyset, \emptyset).$$

Clearly $|\emptyset_\Sigma| = \emptyset$.

For the automaton \mathcal{A} define

$$Q_{ac} := \{q: q \in Q \text{ and } \delta(w, q_0) = q \text{ for some } w \in \Sigma^*\},$$

$$Q_{co} := \{q: q \in Q \text{ and } \delta(w, q) \in Q_m \text{ for some } w \in \Sigma^*\}.$$

Q_{ac} is called the *accessible set* of \mathcal{A} and Q_{co} the *coaccessible set* of \mathcal{A} . We say that \mathcal{A} is *accessible* if $Q_{ac} = Q$, *coaccessible* if $Q_{co} = Q$ and *trim* if $Q = Q_{ac} = Q_{co}$.

Let

$$Q_{tr} := Q_{ac} \cap Q_{co}, \quad \delta_{tr} := \delta|(\Sigma \times Q_{tr}).$$

Then the *trim component* of \mathcal{A} is defined by

$$\text{Tr}(\mathcal{A}) := \begin{cases} (Q_{tr}, \Sigma, \delta_{tr}, q_0, Q_{tr}) & \text{if } Q_{tr} \neq \emptyset, \\ \emptyset_{\Sigma} & \text{otherwise.} \end{cases}$$

It is clear that $|\text{Tr}(\mathcal{A})| = |\mathcal{A}|$ (Eilenberg [1974, p. 23]) and that $\text{Tr}(\mathcal{A})$ is effectively constructible.

Now let $\mathcal{A} = (Q, \Sigma, \delta, q_0, Q_m)$ and $\mathcal{B} = (X, \Sigma, \xi, X_0, X_m)$ be trim finite automata with $|\mathcal{B}| \subset |\mathcal{A}|$. We say that \mathcal{B} *refines* \mathcal{A} if

$$\text{for all } s, t \in \overline{|\mathcal{B}|}, \quad \xi(s, x_0) = \xi(t, x_0) \text{ implies } \delta(s, q_0) = \delta(t, q_0).$$

If \mathcal{B} refines \mathcal{A} then it is easily shown that there exists a unique function $h: X \rightarrow Q$ satisfying

$$(6.1) \quad h \cdot \xi(s, x_0) = \delta(s, q_0), \quad s \in \overline{|\mathcal{B}|}.$$

We are now ready to examine the operator Ω . For this let $L, M \subset \Sigma^*$ be regular languages with $L \subset M$ and $M = \bar{M}$, and let K_j be the sequence of regular languages defined by

$$K_0 = L, \quad K_{j+1} = \Omega(K_j), \quad j \geq 0.$$

LEMMA 6.1. *Let*

$$\mathcal{A} = (Q, \Sigma, \delta, q_0, Q), \quad \mathcal{C}_j = (X, \Sigma, \xi, x_0, X_m)$$

be trim automata such that

$$|\mathcal{A}| = M, \quad |\mathcal{C}_j| = K_j,$$

and \mathcal{C}_j refines \mathcal{A} . Let $h: X \rightarrow Q$ be the unique map satisfying (6.1). Then $w \in \Omega(K_j)$ iff $w \in K_j$ and

$$\text{for each } u \in \bar{w}, \quad x = \xi(u, x_0) \text{ implies } \Sigma(h(x)) \cap \Sigma_0 \subset \Sigma(x).$$

Proof. Recall that

$$w \in \Omega(K_j)$$

$$\text{iff } w \in K_j \text{ and } \bar{w}\Sigma_0 \cap M \subset \bar{K}_j,$$

$$\text{iff } w \in K_j \text{ and } (\forall u \in \bar{w}) u\Sigma_0 \cap M \subset \bar{K}_j,$$

$$\text{iff } w \in K_j \text{ and } (\forall u \in \bar{w}) (\forall \sigma \in \Sigma_0) (u\sigma \in M \text{ implies } u\sigma \in \bar{K}_j),$$

$$\text{iff } w \in K_j \text{ and } (\forall u \in \bar{w}) (\forall \sigma \in \Sigma(\delta(u, q_0)) \cap \Sigma_0) u\sigma \in \bar{K}_j,$$

$$\text{iff } w \in K_j \text{ and } (\forall u \in \bar{w}) \Sigma(\delta(u, q_0)) \cap \Sigma_0 \subset \Sigma(\xi(u, x_0)),$$

$$\text{iff } w \in K_j \text{ and } (\forall u \in \bar{w}) (x = \xi(u, x_0) \text{ implies } \Sigma(h(x)) \cap \Sigma_0 \subset \Sigma(x)). \quad \square$$

In view of Lemma 6.1 it is natural to attempt to construct an automaton \mathcal{C}_{j+1} with $|\mathcal{C}_{j+1}| = \Omega(K_j)$ simply by removing those states of \mathcal{C}_j that fail to satisfy the *active event constraint*

$$\Sigma(h(x)) \cap \Sigma_0 \subset \Sigma(x).$$

More formally let \mathcal{A} and \mathcal{C}_j be defined as in Lemma 6.1, let

$$\begin{aligned} X' &= \{x: x \in X \text{ and } \Sigma(h(x)) \cap \Sigma_0 \subset \Sigma(x)\}, \\ X'_m &= X_m \cap X', \end{aligned}$$

and let $\xi': \Sigma \times X' \rightarrow X'$ be the function given by

$$\xi'(\sigma, x) = \begin{cases} \xi(\sigma, x) & \text{if } \xi(\sigma, x) \in X', \\ \text{undefined} & \text{otherwise.} \end{cases}$$

Then define

$$(6.2) \quad \mathcal{C}_{j+1} := \begin{cases} \text{Tr}((X', \Sigma, \xi', x_0, X'_m)) & \text{if } x_0 \in X', \\ \emptyset_{\Sigma} & \text{otherwise.} \end{cases}$$

PROPOSITION 6.1. *Let \mathcal{A} and \mathcal{C}_j be as defined in the statement of Lemma 6.1, and let \mathcal{C}_{j+1} be the automaton defined by (6.2). Then*

$$|\mathcal{C}_{j+1}| = \Omega(K_j) = K_{j+1}$$

and \mathcal{C}_{j+1} refines \mathcal{A} .

Proof.

$$w \in |\mathcal{C}_{j+1}|$$

$$\text{iff } w \in |\mathcal{C}_j| \text{ and } (\forall u \in \bar{w}) \xi(u, x_0) \in X',$$

$$\text{iff } w \in K_j \text{ and } (\forall u \in \bar{w}) x = \xi(u, x_0) \text{ implies } \Sigma(h(x)) \cap \Sigma_0 \subset \Sigma(x),$$

$$\text{iff } w \in \Omega(K_j) \text{ (by Lemma 6.1).}$$

That \mathcal{C}_{j+1} refines \mathcal{A} follows immediately from the fact that \mathcal{C}_j refines \mathcal{A} . \square

It only remains to show that, given trim automata $\mathcal{A} = (Q, \Sigma, \delta, q_0, Q_m)$ and $\mathcal{B} = (Z, \Sigma, \alpha, z_0, Z_m)$ with $|\mathcal{A}| = M$ and $|\mathcal{B}| = L$, we can effectively construct an automaton \mathcal{C}_0 such that $|\mathcal{C}_0| = L$, and \mathcal{C}_0 refines \mathcal{A} .

Define the intersection $\mathcal{A} \cap \mathcal{B}$ of \mathcal{A} and \mathcal{B} by

$$\mathcal{A} \cap \mathcal{B} := (Q \times Z, \Sigma, \langle \delta, \alpha \rangle, (q_0, z_0), Q_m \times Z_m)$$

where $\langle \delta, \alpha \rangle(\sigma, q, z)$ is defined iff both $\delta(\sigma, q)$ and $\alpha(\sigma, z)$ are defined and is given by

$$\langle \delta, \alpha \rangle(\sigma, q, z) = (\delta(\sigma, q), \alpha(\sigma, z)).$$

Clearly $|\mathcal{A} \cap \mathcal{B}| = |\mathcal{A}| \cap |\mathcal{B}|$ (Eilenberg [1974, p. 17]).

PROPOSITION 6.2. *Let \mathcal{A} and \mathcal{B} be as defined above and let $\mathcal{C}_0 = \text{Tr}(\mathcal{A} \cap \mathcal{B})$. Then $|\mathcal{C}_0| = L$ and \mathcal{C}_0 refines \mathcal{A} .*

Proof.

$$|\mathcal{C}_0| = |\mathcal{A}| \cap |\mathcal{B}| = M \cap L = L.$$

For $s, t \in \bar{L}$

$$\langle \delta, \alpha \rangle(s, q_0, z_0) = \langle \delta, \alpha \rangle(t, q_0, z_0)$$

implies

$$\delta(s, q_0) = \delta(t, q_0).$$

Hence \mathcal{C}_0 refines \mathcal{A} . \square

The method of this section has been implemented in Pascal by Lin [1984].

7. Illustrations. The computations in this section are intended mainly to illustrate the result of Theorem 3.1. In the examples that follow, $L \subset M$ and $M = \bar{M}$, and so we employ the computation scheme developed in § 6.

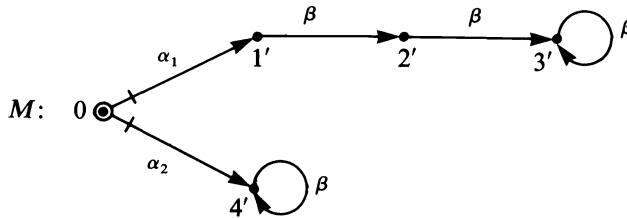
It is often convenient for purposes of computation to represent a finite automaton $\mathcal{A} = (Q, \Sigma, \delta, q_0, Q_m)$ by its transition matrix. This is a finite square matrix Δ whose columns and rows are indexed by the state set of \mathcal{A} and whose entries $\Delta(p, q)$ are the subsets of Σ given by

$$\Delta(p, q) = \{\sigma : \sigma \in \Sigma \text{ and } q = \delta(\sigma, p)\}.$$

7.1. Example 3. Let $\Sigma = \{\alpha_1, \alpha_2, \beta\}$, $\Sigma_0 = \{\beta\}$; and in the notation of regular expressions (Hopcroft and Ullman [1979]) let

$$L = \overline{\alpha_1\beta^2} + \alpha_2\beta^*, \quad M = \overline{(\alpha_1\beta^2 + \alpha_2)}\beta^*.$$

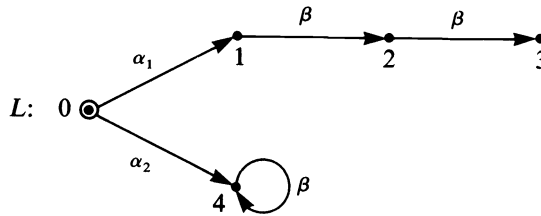
An automaton \mathcal{A} for M (i.e., a trim automaton with $|\mathcal{A}| = M$) is displayed below (\odot denotes the initial state, \cdot denotes marker states).



Note that while the string $\alpha_1\beta^2$ is legal, its successor $\alpha_1\beta^3$ is not; but since $\beta \in \Sigma_0$ this implies $\alpha_1\beta^2 \notin K_1$. Proceeding in this informal manner it is intuitively clear that

$$\sup C(L) = 1 + \alpha_2\beta^*.$$

To verify this by formal computation consider the automaton \mathcal{B} for L shown below.



It is clear that \mathcal{B} refines \mathcal{A} . To the transition matrix of \mathcal{B} we adjoin two columns: one listing $\Sigma(h(x)) \cap \Sigma_0$, the other listing $\Sigma(x)$. This yields the following tableau.¹

	0	1	2	3	4	$\Sigma(h(x)) \cap \Sigma_0$	$\Sigma(x)$
0		α_1			α_2		$\alpha_1\alpha_2$
1			β			β	β
2				β		β	β
3						β	
4					β	β	β

¹ We denote the empty set simply by a blank entry in the tableau.

Since

$$\Sigma(3') \cap \Sigma_0 \not\subset \Sigma(3),$$

we remove state 3 from the tableau and arrange that the resulting automaton is trim. The result (shown below) is an automaton for the language K_1 .

	0	1	2	4	$\Sigma(h(x)) \cap \Sigma_0$	$\Sigma(x)$
$K_1:$	0	α_1		α_2		$\alpha_1 \alpha_2$
	1		β		β	β
	2				β	
	4			β	β	β

$$K_1 = \overline{\alpha_1 \beta} + \alpha_2 \beta^*$$

Iterating this procedure yields the following sequence of tableaux:

	0	1	4	$\Sigma(h(x)) \cap \Sigma_0$	$\Sigma(x)$
$K_2:$	0	α_1	α_2		$\alpha_1 \alpha_2$
	1			β	
	4		β	β	β

$$K_2 = \overline{\alpha_1} + \alpha_2 \beta^*$$

	0	4	$\Sigma(h(x)) \cap \Sigma_0$	$\Sigma(x)$
$K_3:$	0	α_2		$\alpha_1 \alpha_2$
	4	β	β	β

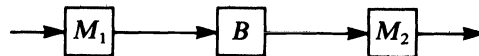
$$K_3 = 1 + \alpha_2 \beta^*$$

Since each state of the last tableau satisfies the active event constraint, our iteration scheme has converged. We conclude that

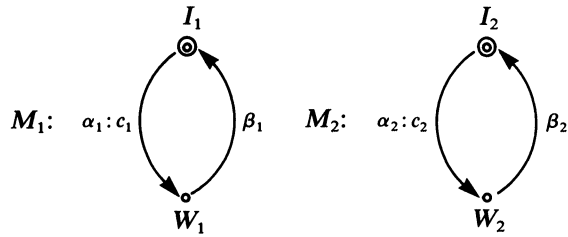
$$\sup C(L) = 1 + \alpha_2 \beta^*$$

as expected. \square

7.2. Example 4. Consider a simplified version of the example of (Ramadge and Wonham [1983, § 12]), to which the reader is referred for any unexplained terminology below. Two machines M_1 and M_2 are connected in tandem, separated by a buffer B .



The M_i are controlled discrete-event processes (CDEPs) with state diagrams as shown.

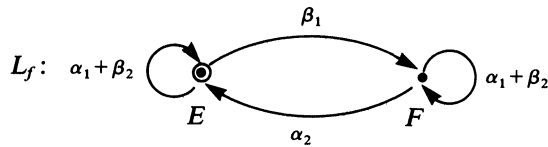


Here the defining condition simply maintains the content of the buffer at zero or one (workpiece) at all times. Finally we define

$$L = L_f \cap M.$$

Evidently L_f , and therefore L , are closed.

To compute $\sup C(L)$ we first obtain an automaton for L . By inspection, L_f consists of those strings of Σ^* in which β_1, α_2 occur in strict alternation. An automaton for L_f is therefore as shown below.



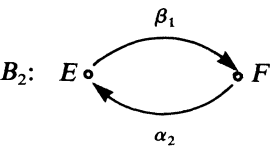
To compute an automaton for L we must form the product of the automata for L_f and M . We omit the simple computation, merely displaying the result.

	0E	0F	1E	1F	2E	2F	3E	3F
0E					α_1			
0F			α_2			α_1		
1E	β_2						α_1	
1F		β_2						α_1
2E		β_1						
2F							α_2	
3E				β_1	β_2			
3F						β_2		

initial state: 0E
marker states: all states

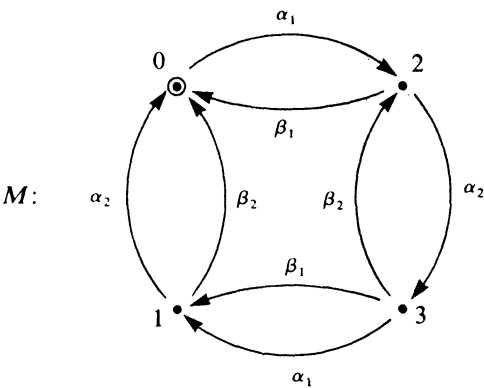
Evidently this automaton refines the automaton for M . Hence we can immediately write down the following tableau.

In state I_i , M_i is “idle”; in state W_i it is “working.” The buffer B has a single slot and is modeled by an automaton with the two states E (empty) and F (full).



Initially the system is in state (I_1, I_2, E) . A work cycle $\alpha_1\beta_1$ of M_1 “deposits a workpiece in B ” driving B to F ; M_2 can then begin its work cycle by “removing the workpiece from B ” thereby resetting B to E . The controlled events are α_1, α_2 ; namely $\Sigma_0 = \{\beta_1, \beta_2\}$.

For M we adopt the closure of the shuffle language generated by M_1 and M_2 . M has the automaton displayed below.



To define the legal language L we impose the “safety” requirement determined by the capacity of the buffer. To formalize this let $\sigma \in \Sigma = \{\alpha_1, \beta_1, \alpha_2, \beta_2\}$ and for a string $s \in \Sigma^*$ write

$$|\sigma|(s) = \text{number of occurrences of } \sigma \text{ in } s.$$

Define the “safety language” $L_f \subset \Sigma^*$ according to

$$L_f = \{s: s \in \Sigma^* \text{ and } |\alpha_2|(t) \leq |\beta_1|(t) \leq |\alpha_2|(t) + 1 \text{ for all prefixes } t \text{ of } s\}.$$

	0E	0F	1E	1F	2E	2F	3E	3F	$\Sigma(h(x) \cap \Sigma_0$	$\Sigma(x)$
K_0 :	0E				α_1					α_1
	0F		α_2			α_1				$\alpha_2\alpha_1$
	1E	β_2					α_1		β_2	$\beta_2\alpha_1$
	1F		β_2					α_1	β_2	$\beta_2\alpha_1$
	2E		β_1						β_1	β_1
	2F						α_2		β_1	α_2
	3E			β_1	β_2				$\beta_1\beta_2$	$\beta_1\beta_2$
	3F					β_2			$\beta_1\beta_2$	β_2

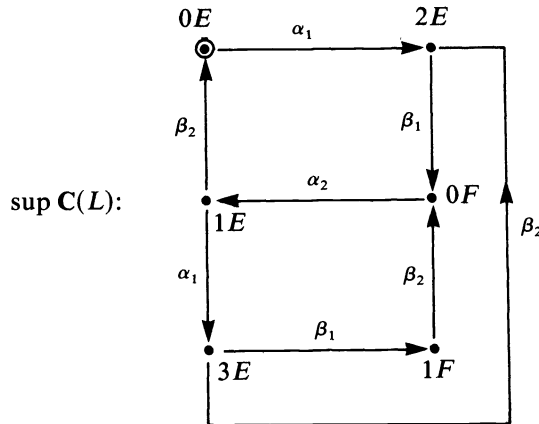
Both the states $2F$ and $3F$ fail to satisfy the active event constraint and hence must be removed from the tableau. This yields

	$0E$	$0F$	$1E$	$1F$	$2E$	$2F$	$\Sigma(h(x)) \cap \Sigma_0$	$\Sigma(x)$
$0E$					α_1			α_1
$0F$			α_2					α_2
$1E$	β_2					α_1	β_2	$\beta_2\alpha_1$
$1F$		β_2					β_2	β_2
$2E$		β_1					β_1	β_1
$3E$				β_1	β_2		$\beta_1\beta_2$	$\beta_1\beta_2$

initial state: $0E$

marker states: all states

Since each state of this trim automaton satisfies the active event constraint it is an automaton for $\text{sup } C(L)$. The state transition graph is shown below.



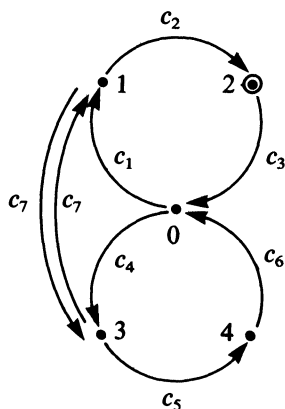
This result is in agreement with Ramadge and Wonham [1983], where our automaton transition graph for $\text{sup } C(L)$ in fact appears as a subgraph of the automaton transition graph of Ramadge and Wonham [1983, Fig. 12.4].

7.3. Example 5. A cat and a mouse are placed in the maze shown in Fig. 7.1, with the cat initially in room 2 and the mouse initially in room 4. Each doorway in the maze must be traversed in the direction indicated and is either for the exclusive use of the cat (displayed as $\neg\uparrow-$) or for the exclusive use of the mouse (displayed as $\neg\downarrow-$). In addition each door, with the exception of c_7 , can be opened or closed as required in order to control the movement of the cat and the mouse.

Our objective is to find the control scheme that permits the cat and the mouse the greatest possible freedom of movement but also guarantees that

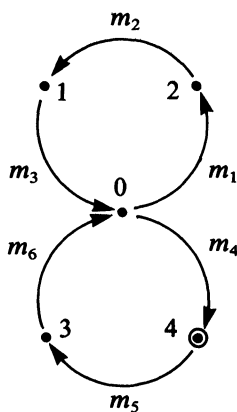
- (a) The cat and the mouse never occupy the same room simultaneously;
- (b) It is always possible for the cat and the mouse to return to the initial state, i.e., the state in which the cat is in room 2 and the mouse is in room 4.

Let $\Sigma = \{c_i, m_j: 1 \leq i \leq 7, 1 \leq j \leq 6\}$. We model the movement of the cat in the maze by the automaton \mathcal{G}_c over Σ shown below.



Here state i corresponds to room i and a transition $i \xrightarrow{c_k} j$ corresponds to traversing the door c_k between rooms i and j .

Similarly we model the movement of the mouse in the maze by the automaton \mathcal{G}_m shown below.



Now for our joint model of the cat and the mouse we adopt the automaton \mathcal{G} constructed by shuffling the automata \mathcal{G}_c and \mathcal{G}_m . The states of \mathcal{G} are ordered pairs (i, j) where i is a state of \mathcal{G}_c and j is a state of \mathcal{G}_m , and the transitions of \mathcal{G}_c are either of the form

$$(i, j) \xrightarrow{c_k} (i', j)$$

where $i \xrightarrow{c_k} i'$ is a transition in \mathcal{G}_c , or of the form

$$(i, j) \xrightarrow{m_k} (i, j')$$

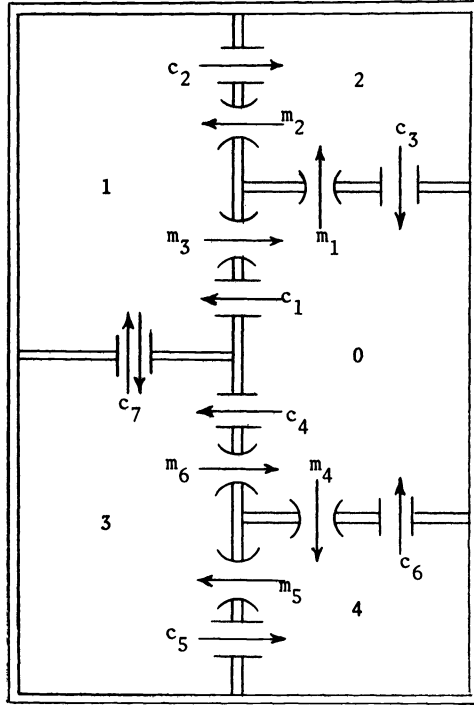


FIG. 7.1. Example 5. Maze for cat and mouse.

where $j \rightarrow^{m_k} j'$ is a transition in \mathcal{G}_m . The automaton \mathcal{G} is displayed in Fig. 7.2.² We let $M = |\mathcal{G}|$ and Σ_0 be the set of labels of uncontrolled doors, i.e., $\Sigma_0 = \{c_7\}$.

To construct an automaton for the legal language L we remove those states (i, j) of \mathcal{G} for which $i = j$, and take the trim component of the resultant automaton. Clearly this automaton satisfies constraint (a). To satisfy constraint (b) we must ensure that every state is coaccessible to the state 24. Hence we designate 24 as the only marker state and again take the trim component of the resultant automaton. These computations yield the automaton \mathcal{H} for L shown on the left-hand side of Fig. 7.3.

Having specified M , L and Σ_0 we now proceed to compute $\sup C(L)$. By the construction of \mathcal{H} it is clear that \mathcal{H} refines \mathcal{G} . Hence we can immediately fill in the right-hand columns of Fig. 7.3 to form our initial tableau. Since the states 13 and 31 fail to satisfy the active event constraint, these are removed and we compute the trim component of the remaining automaton. The resultant automaton for K_1 is shown in Fig. 7.4.

It is clear from Fig. 7.4 that K_1 is controllable. Hence $\sup C(L) = K_1$. The automaton for $\sup C(L)$ is displayed by its state transition graph in Fig. 7.5.

Using the results of Ramadge and Wonham [1983] we can now construct a quotient supervisor which synthesizes the language $\sup C(L)$. This is shown in Fig. 7.6. The "optimal" control strategy which this supervisor implements can be summarized as follows. If the cat and the mouse occupy their respective initial rooms, then both are

² For convenience we often abbreviate (i, j) to simply ij , e.g., $(2, 3)$ becomes 23, $(1, 3)$ becomes 13.

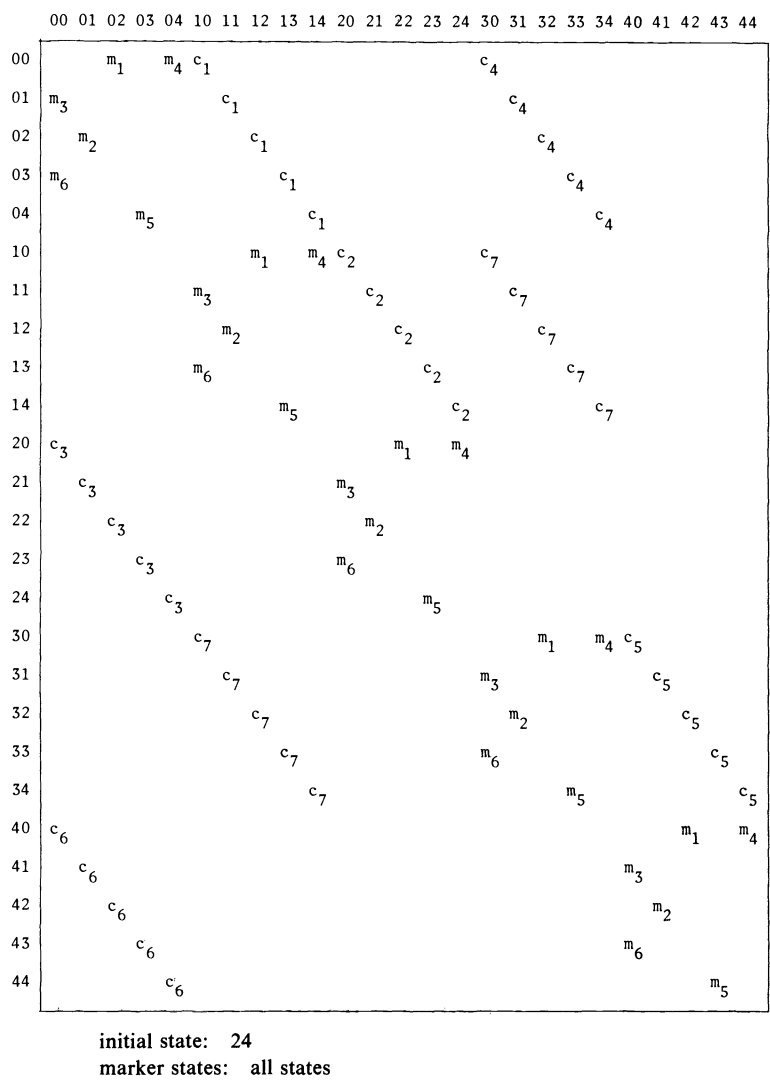


FIG. 7.2. Example 5. Shuffled automaton *G*.

given the opportunity to move to a new room, i.e., c_3 and m_5 are open. If the cat leaves room 2, then the mouse is isolated in room 4, i.e., m_5 and c_5 are closed, and the cat is free to roam the rest of the maze. Similarly, if the mouse leaves room 4, then the cat is isolated in room 2, and the mouse is permitted access to those rooms from which it can return to room 4, i.e., rooms 0, 3, 4.

8. Conclusion. The characterization of supremal controllable language given in this paper is both of theoretical interest and provides a basis for computation in the regular case. It would be of considerable interest to find convergence criteria that could be used to extend Theorem 3.1 to broader classes of languages, as well as to study the issue of computational efficiency.

Acknowledgments. It is a pleasure to acknowledge the helpful contributions of Lin Feng and John Thistle to the development of Theorem 3.1.

01 02 03 04 10 12 13 14 20 23 24 30 31 32 34 40 41 42														$\Sigma(h(x)) \cap \Sigma_0$	$\Sigma(x)$
01															c_4
02	m_2														$m_2 c_1 c_4$
03															c_1
04															$m_5 c_1 c_4$
10		m_5													$m_1 m_4 c_2 c_7$
12															c_7
13															$m_5 c_2$
14															$m_5 c_2 c_7$
20															m_4
23															$c_3 m_6$
24															$c_3 m_5$
30															$c_7 m_1 m_4 c_5$
31															$m_3 c_5$
32															$c_7 m_2 c_5$
34															$c_7 c_5$
40															m_1
41															$c_6 m_3$
42															$c_6 m_2$

initial state: 24
marker state: 24

FIG. 7.3. Example 5. Initial tableau.

04 14 20 23 24 34						$\Sigma(h(x)) \cap \Sigma_0$	$\Sigma(x)$
04							$c_1 c_2$
14						c_7	$c_2 c_7$
20							m_4
23							m_6
24							$c_3 m_5$
34						c_7	c_7

initial state: 24
marker state: 24

FIG. 7.4. Example 5. Tableau for K_1 .

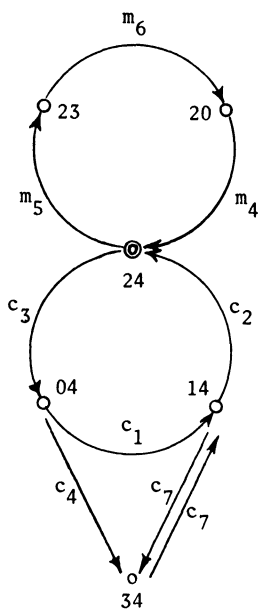
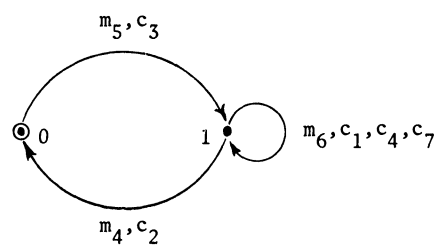


FIG. 7.5. Example 5. Automaton for $\text{sup } C(L)$.



	c_1	c_2	c_3	c_4	c_5	c_6	m_1	m_2	m_3	m_4	m_5	m_6
0	-	-	1	-	-	-	-	-	-	-	1	-
1	1	1	0	1	0	-	0	-	-	1	0	1

(-) may be assigned arbitrarily

FIG. 7.6. Example 5. Quotient supervisor.

REFERENCES

- S. EILENBERG, 1974, *Automata, Languages and Machines Volume A*, Academic Press, New York.
- F. LIN, 1984, *Supervisor synthesis for discrete event processes*, M.A.Sc. thesis, Department of Electrical Engineering, University of Toronto.
- J. E. HOPCROFT AND J. D. ULLMAN, 1979, *Introduction to Automata Theory, Languages, and Computation*, Addison-Wesley, Reading, MA.
- P. J. RAMADGE AND W. M. WONHAM, 1983, *Supervisory control of a class of discrete event processes*, Technical Rept. No. 8311, Systems Control Group, Department of Electrical Engineering, University of Toronto, October; this Journal, 25 (1987). For a summary under the same title see Proc. Sixth International Conference Anal. and Optim. of Systems, Nice, June, 1984; in A. Bensoussan and J. L. Lions, eds., *Analysis and Optimization of Systems*, LNCIS, Vol. 63, Springer-Verlag, Berlin, New York, 1984, Part 2, pp. 477-498.
- W. M. WONHAM AND P. J. RAMADGE, 1984, *On the supremal controllable sublanguage of a given language*, Proc. 23rd IEEE Conference on Decision and Control, IEEE Control Systems Society, New York, pp. 1073-1080.

DUAL TECHNIQUES FOR MINIMAX*

WILLIAM W. HAGER† AND DWAYNE L. PRESLER‡

Abstract. A dual formulation for a convex minimax problem is presented and the notion of reducibility is introduced. For nonconvex problems, Rockafellar's augmented Lagrangian is used to close the duality gap. We show how mathematical programming algorithms can be applied to the minimax problem and we develop a special algorithm for reducible minimax problems.

Key words. minimax, duality, augmented Lagrangians

AMS(MOS) subject classifications. 65K05, 90C30

1. Introduction. A dual minimax problem is formulated and the concept of reducibility is introduced. Since the dual problem may not solve the primal problem when the cost lacks convexity, § 3 develops augmented Lagrangian techniques to bridge the duality gap. In particular, an analogue of Rockafellar's augmented Lagrangian is applied to the minimax problem. Abstractly, the minimax problem is a constrained optimization problem with special structure. In § 4 we show how mathematical programming algorithms such as those in [18] can be embedded into an algorithm to solve the minimax problem. Section 5 presents a special scheme for solving reducible minimax problems. Other algorithms that have been proposed for minimax problems include those in [3], [4], [8], [9], [11]–[14], [20], [21], [24], [26] and [34]. Most of these algorithms are either "primal" in nature or the algorithm addresses problems where the maximization phase of the minimax problem is restricted to a finite set. Our methods, on the other hand, are dual methods derived from an augmented Lagrangian and the maximization can be performed over an infinite set. Another approach to minimax problems utilizes nondifferentiable optimization techniques. This family of methods is described in [23] by Kiwiel and in [32] by Shor. As discussed later, perhaps the algorithm [26] of Murray and Overton is the closest to the approach proposed in § 4. Our algorithm for reducible minimax problems seems to be distinct from other algorithms for the minimax problem.

2. Abstract dual. Given sets X and Y and given a real-valued function f defined on $X \times Y$, we consider the problem

$$(2.1) \quad \underset{x \in X}{\text{minimize}} \quad \underset{y \in Y}{\text{maximum}} \quad f(x, y).$$

In other words, if $\Phi: X \rightarrow \mathbb{R}$ is the real-valued function defined by

$$\Phi(x) = \supremum \{f(x, y): y \in Y\},$$

we are concerned with the problem

$$(2.2) \quad \underset{x \in X}{\text{minimize}} \quad \{\Phi(x): x \in X\}.$$

To derive a dual to (2.1), let us write (2.2) in the form

$$(2.3) \quad \begin{aligned} &\text{minimize } \rho \\ &\text{subject to } \Phi(x) - \rho \leq 0, \quad x \in X, \quad \rho \in \mathbb{R}. \end{aligned}$$

* Received by the editors July 16, 1985; accepted for publication (in revised form) March 27, 1986. This work was supported by National Science Foundation grants MCS-8101892, DMS-8401758 and DMS-8520926 and by Air Force Office of Scientific Research grant AFOSR-ISSA-860091.

† Department of Mathematics, The Pennsylvania State University, University Park, Pennsylvania 16802.

‡ INTER-NATIONAL Research Institute, Newport News, Virginia 23602.

The collection \mathbf{P} of functionals that are nonnegative everywhere on Y is a cone. When g is a functional defined on Y , we write $g \geq 0$ if $g \in \mathbf{P}$. Similarly the relation $g \leq 0$ means that $-g \in \mathbf{P}$. Since $f(x, y) - \rho$ is a real-valued function of y for each fixed $x \in X$, and for each fixed $\rho \in R$, (2.3) can be expressed as follows:

$$(2.4) \quad \text{minimize } \{ \rho: f(x, \cdot) - \rho \mathbf{1}(\cdot) \leq 0, x \in X, \rho \in R \}$$

where $\mathbf{1}: Y \rightarrow R$ is the function defined by $\mathbf{1}(y) = 1$ for every $y \in Y$. That is, $\mathbf{1}$ is identically equal to one on Y .

Let \mathbf{Y} be a normed vector space consisting of functionals defined on Y and suppose that \mathbf{Y} contains $f(x, \cdot) - \rho \mathbf{1}(\cdot)$ for every $x \in X$ and $\rho \in R$. The dual problem associated with (2.4) involves the dual space \mathbf{Y}^* . Given a linear functional $\lambda \in \mathbf{Y}^*$, let $\langle \lambda, p \rangle$ denote the value of λ at p and write $\lambda \geq 0$ if $\langle \lambda, p \rangle \geq 0$ for every $p \in \mathbf{P} \cap \mathbf{Y}$. Then the dual to (2.4) is

$$(2.5) \quad \text{maximize } \{ \mathbf{I}(\lambda): \lambda \in \mathbf{Y}^*, \lambda \geq 0 \}$$

where the dual functional \mathbf{I} is defined by

$$(2.6) \quad \mathbf{I}(\lambda) = \infimum \{ \rho + \langle \lambda, f(x, \cdot) - \rho \mathbf{1}(\cdot) \rangle: x \in X, \rho \in R \}.$$

Since

$$\rho + \langle \lambda, f(x, \cdot) - \rho \mathbf{1}(\cdot) \rangle = \langle \lambda, f(x, \cdot) \rangle + \rho(1 - \langle \lambda, \mathbf{1} \rangle),$$

it follows that $\mathbf{I}(\lambda)$ is $-\infty$ unless $\langle \lambda, \mathbf{1} \rangle = 1$. Moreover, if $\langle \lambda, \mathbf{1} \rangle = 1$, then the dual functional can be expressed

$$\mathbf{I}(\lambda) = \infimum \{ \langle \lambda, f(x, \cdot) \rangle: x \in X \}.$$

Now let us consider the standard question in duality theory: When does there exist a solution to (2.5) and when is the maximum in (2.5) equal to the minimum in (2.2)? Applying [17, Thm. A.1], we have the following.

THEOREM 2.1. *Whenever ρ and x are feasible for (2.3) and λ is feasible for (2.5), we have $\mathbf{I}(\lambda) \leq \rho$. Moreover, if ρ^* and x^* are feasible for (2.3), λ^* is feasible for (2.5), and $\mathbf{I}(\lambda^*) = \rho^*$, then ρ^* and x^* are optimal in (2.3), λ^* is optimal in (2.5), and the complementary slackness condition $\langle \lambda^*, f(x^*, \cdot) \rangle = \rho^*$ holds. On the other hand, suppose that X is a convex subset of a vector space and for each fixed $y \in Y$, $f(x, y)$ is a convex function of $x \in X$. If there exists a ball $\mathbf{B} \subset \mathbf{Y}$ with center at the origin such that*

$$(2.7) \quad \sup_{\phi \in \mathbf{B}} \inf_{x \in X} \sup_{y \in Y} \{ f(x, y) - \phi(y) \} < \infty,$$

then (2.5) has a solution λ^* and

$$\mathbf{I}(\lambda^*) = \inf_{x \in X} \sup_{y \in Y} f(x, y).$$

For illustration, suppose that Y is the finite set $\{1, \dots, m\}$ and let $f_i(x)$ denote the function $f(x, i)$. In this case, both \mathbf{Y} and \mathbf{Y}^* can be identified with R^m , the space of m -tuples of real numbers, and the functional $\langle \cdot, \cdot \rangle$ is the usual dot product in R^m . Hence, the feasibility condition " $\lambda \in \mathbf{Y}^*$, $\lambda \geq 0$, and $\langle \lambda, \mathbf{1} \rangle = 1$ " associated with the dual problem (2.5) is equivalent to saying that $\lambda \in R^m$, $\lambda_i \geq 0$ for $i = 1, \dots, m$, and $\lambda_1 + \dots + \lambda_m = 1$. Observe that in this finite-dimensional framework, assumption (2.7) is satisfied trivially. If both X and the f_i are convex, then Theorem 2.1 tells us that there exists an optimal solution λ^* to (2.5) and

$$\mathbf{I}(\lambda^*) = \inf \left\{ \sum_{i=1}^m \lambda_i^* f_i(x): x \in X \right\} = \inf_{x \in X} \max \{ f_i(x): i = 1, \dots, m \}.$$

For a second illustration, let us consider the case where Y is a compact subset of a normed vector space, \mathbf{Y} is the space of continuous real-valued functions defined on Y , and the norm $\|\cdot\|$ for \mathbf{Y} is defined by

$$\|\phi\| = \text{maximum } \{|\phi(y)|: y \in Y\}.$$

Again (2.7) is satisfied trivially. Furthermore, if Y is an interval $[a, b] \subset \mathbf{R}$, then \mathbf{Y}^* is the space of functions with bounded variation on $[a, b]$ and the complementary slackness condition can be expressed using a Stieltjes integral:

$$(2.8) \quad \int_a^b (f(x^*, y) - \rho^*) d\lambda^*(y) = 0.$$

In the proof of [17, Lemma 5.2], we note that when λ has bounded variation, the inequality $\lambda \geq 0$ is equivalent to saying that $\lambda(y)$ is a nondecreasing function of y . Hence, if ρ^* and x^* are feasible for (2.3) and λ^* is feasible for (2.5), then $f(x^*, y) - \rho^* \leq 0$ for every $y \in Y$ and (2.8) implies that $\lambda^*(\cdot)$ is constant on each subinterval of $[a, b]$ where $f(x^*, \cdot) < \rho^*$.

In many applications, one discovers that a solution pair (ρ^*, x^*) for (2.3) has the property that $f(x^*, y) < \rho^*$ except for y in a finite set $\{y_1, \dots, y_m\} \subset Y$. In this case, the complementary slackness condition (2.8) tells us that $\lambda^*(\cdot)$ is constant on each open interval (y_i, y_{i+1}) and $\lambda^*(y_i^+) \geq \lambda^*(y_i^-)$. If this jump set is known in advance and if we restrict our attention to λ 's which are constant on each interval (y_i, y_{i+1}) , then the dual functional can be expressed

$$I(\lambda) = \inf \left\{ \sum_{i=1}^m \lambda_i f(x, y_i): x \in X \right\}$$

where $\lambda_i = \lambda(y_i^+) - \lambda(y_i^-)$ is the jump at y_i . Generally, the y_i are unknown and the dual functional must be maximized over both the y_i and the λ_i . To summarize, if Y is the interval $[a, b]$, \mathbf{Y} is the space of continuous real-valued functions defined on $[a, b]$, and there exists both a solution ρ^* and x^* to (2.3) and a solution λ^* to (2.5) such that $\rho^* = I(\lambda^*)$ and $f(x^*, y) = \rho^*$ for finitely many y , then the continuous dual problem (2.5) can be replaced by the discrete dual

$$(2.9) \quad \begin{aligned} &\text{maximize } I(\lambda_1, \dots, \lambda_N, y_1, \dots, y_N) \\ &\text{subject to } \lambda_i \geq 0 \quad \text{and} \quad y_i \in Y \quad \text{for } i = 1, \dots, N, \quad \sum_{i=1}^N \lambda_i = 1, \end{aligned}$$

where

$$I(\lambda_1, \dots, \lambda_N, y_1, \dots, y_N) = \inf \left\{ \sum_{i=1}^N \lambda_i f(x, y_i): x \in X \right\}$$

and where N is any integer greater than or equal to the number of y for which $f(x^*, y) = \rho^*$. The optimal λ_i and y_i in (2.9) correspond to the size and the location of the jumps in λ^* . The discrete dual (2.9) will now be studied in a general setting.

The generalized finite sequence space of Charnes, Cooper and Kortanek (see [5]–[7]) is a natural setting for the discrete dual. Let Λ denote a vector space of real-valued functions defined on Y where $\lambda \in \Lambda$ if and only if $\lambda(y) = 0$ for all but a finite number of y in Y . Instead of writing $\lambda(y)$, we write λ_y and we consider λ_y the y th component of λ . Given $\lambda \in \Lambda$, the collection of y in Y for which $\lambda_y \neq 0$ is the

support of λ . In [5] the space Λ is called a *generalized finite sequence space*. Given $\lambda \in \Lambda$, let us define the dual functional

$$I(\lambda) = \inf \left\{ \sum_{y \in Y} \lambda_y f(x, y) : x \in X \right\}.$$

The dual problem is

$$(2.10) \quad \text{maximize } \left\{ I(\lambda) : \lambda \in \Lambda, \lambda \geq 0, \sum_{y \in Y} \lambda_y = 1 \right\}.$$

THEOREM 2.2. *If λ^* is feasible for (2.10), $x^* \in X$, and*

$$(2.11) \quad I(\lambda^*) = \sup \{ f(x^*, y) : y \in Y \},$$

then λ^ is optimal in (2.10), x^* is optimal in (2.2), and $f(x^*, y) = I(\lambda^*)$ for each y in the support of λ^* . Moreover, letting S denote the support of λ^* , we have*

$$(2.12) \quad \min_{x \in X} \max_{y \in S} f(x, y) = \min_{x \in X} \max_{y \in Y} f(x, y).$$

Proof. If the components of λ are nonnegative and sum to 1 and if $\Gamma \subset Y$ is any finite subset which contains the support of λ , then the following relations hold for any $x \in X$:

$$\begin{aligned} I(\lambda) &= \inf \left\{ \sum_{y \in \Gamma} \lambda_y f(z, y) : z \in X \right\} \\ &\leq \sum_{y \in \Gamma} \lambda_y f(x, y) \\ (2.13) \quad &\leq \left(\sum_{y \in \Gamma} \lambda_y \right) \max \{ f(x, y) : y \in \Gamma \} \\ &= \max \{ f(x, y) : y \in \Gamma \} \\ &\leq \sup \{ f(x, y) : y \in Y \} \\ &= \Phi(x). \end{aligned}$$

(The last equality is the definition of $\Phi(x)$.) Hence, $I(\lambda) \leq \Phi(x)$ whenever λ is feasible for (2.10) and $x \in X$. Since $I(\lambda^*) = \Phi(x^*)$ by (2.11), we conclude from (2.13) that λ^* is optimal in (2.10) and x^* is optimal in (2.2). If S is the support of λ^* , then (2.13) also tells us that

$$(2.14) \quad I(\lambda^*) \leq \sum_{y \in S} \lambda_y^* f(x^*, y) \leq \Phi(x^*).$$

But $I(\lambda^*) = \Phi(x^*)$ and the inequalities in (2.14) are equalities. Since $\lambda_y^* > 0$ and $f(x^*, y) \leq \Phi(x^*)$ for every $y \in S$ and since the components of λ^* sum to one, it follows from (2.14) that

$$(2.15) \quad I(\lambda^*) = \Phi(x^*) = f(x^*, y) \quad \text{for every } y \in S.$$

Finally, let us consider (2.12). Relation (2.13) implies that

$$(2.16) \quad I(\lambda^*) \leq \max \{ f(x, y) : y \in S \}$$

for every x in X . Taking the infimum over $x \in X$, (2.16) tells us that

$$(2.17) \quad I(\lambda^*) \leq \inf_{x \in X} \max_{y \in S} f(x, y).$$

Combining (2.15) and (2.17), we have

$$\Phi(x^*) = \max_{y \in S} f(x^*, y) = \min_{x \in X} \max_{y \in S} f(x, y),$$

which completes the proof of (2.12). \square

For any $S \subset Y$, we have the trivial relation

$$\min_{x \in X} \max_{y \in S} f(x, y) \leq \min_{x \in X} \max_{y \in Y} f(x, y).$$

Theorem 2.2 tells us that if the value of the dual problem (2.10) is equal to the value of the primal problem (2.2), then there exists a finite set $S \subset Y$ such that (2.12) holds. Therefore, one strategy for solving the minimax problem (2.1) is to start with a finite set S and adjust it until (2.12) is satisfied. The simplified problem

$$\text{minimize}_{x \in X} \text{maximum}_{y \in S} f(x, y)$$

is often easier to solve than the original problem (2.1). This idea is developed further in § 5. When there exists a finite set S satisfying (2.12), we say that the minimax problem is *reducible*.

How often is a minimax problem reducible? Let us consider the following example:

$$(2.18) \quad \text{minimize}_{x \in R} \text{maximum}_{y \in R} \beta x^2 + 2\alpha xy + y^2.$$

Since the maximum is attained at $y = \alpha x$, the function Φ is given by

$$\Phi(x) = (\alpha^2 + \beta)x^2.$$

When $\alpha^2 + \beta$ is nonnegative, $\Phi(x)$ achieves its minimum at $x = 0$ and when x is 0, the maximizing value of y in (2.18) is $y = 0$. On the other hand, suppose that $S = \{0\}$ and let us consider the restricted minimax problem

$$\text{minimize}_{x \in R} \text{maximum}_{y=0} \beta x^2 + 2\alpha xy - y^2,$$

which is equivalent to

$$\text{minimize}_{x \in R} \beta x^2.$$

For $\beta \geq 0$, the minimum is attained at $x = 0$ while for $\beta < 0$, βx^2 has no minimum. In summary, for $\beta < -\alpha^2$, Φ has no minimum. For $\beta \in [-\alpha^2, 0)$, no set S satisfies (2.12). And for $\beta \geq 0$, the minimax problem is reducible with $S = \{0\}$.

Now suppose that $f: R^2 \rightarrow R$ is an arbitrary twice continuously differentiable function. We assume that there exists a solution x^* to the minimax problem (2.1) and there exists y^* such that

$$f(x^*, y^*) = \text{maximum}_{y \in R} \{f(x^*, y)\}$$

and

$$(2.19) \quad \frac{\partial^2 f}{\partial y^2}(x^*, y^*) < 0.$$

Then for x in a neighborhood of x^* , the implicit function theorem gives us a differentiable function $y(\cdot)$ such that $y(x^*) = y^*$ and

$$(2.20) \quad \frac{\partial f}{\partial y}(x, y(x)) = 0.$$

By (2.19), $y(x)$ is a local maximizer of $f(x, \cdot)$ for x near x^* . Let us assume that $y(x)$ is the global maximizer of $f(x, \cdot)$. Since x^* minimizes $\Phi(x)$, we know that $\Phi''(x^*) \geq 0$. Applying the chain rule to $\Phi(x) = f(x, y(x))$ and utilizing (2.20), it can be shown that

$$(2.21) \quad \frac{\partial^2 \Phi}{\partial x^2}(x^*) = \frac{\partial^2 f}{\partial x^2}(x^*, y^*) - \frac{((\partial^2 f / \partial x \partial y)(x^*, y^*))^2}{(\partial^2 f / \partial y^2)(x^*, y^*)}.$$

If X is restricted to a neighborhood of x^* , then (2.12) holds for $S = \{y^*\}$ provided

$$(2.22) \quad \frac{\partial^2 f}{\partial x^2}(x^*, y^*) > 0.$$

On the other hand, by (2.21), the inequality $\Phi''(x^*) \geq 0$ only guarantees that

$$(2.23) \quad \frac{\partial^2 f}{\partial x^2}(x^*, y^*) \geq \frac{((\partial^2 f / \partial x \partial y)(x^*, y^*))^2}{(\partial^2 f / \partial y^2)(x^*, y^*)}.$$

In other words, relation (2.22) is sufficient for the minimax problem to be reducible when X is a neighborhood of x^* while the fact that x^* minimizes Φ only implies (2.23).

Returning to example (2.18), observe that the range of α and β for which the minimax problem is reducible is larger than the range for which the minimax problem is not reducible. Consequently, if α and β are chosen randomly and if the minimax problem (2.18) has a solution, then the problem is probably reducible. The problem (see [19]) of optimally coating a surface to minimize the maximum reflection associated with incoming waves is an example of a very complicated nonconvex problem that is reducible even though $I(\lambda^*) < \Phi(x^*)$. Based on these observations, we feel that algorithms which search for a set S satisfying (2.12) will apply to a broad class of problems.

Although a general minimax problem is not necessarily reducible, a convex finite-dimensional minimax problem is always reducible. Dem'yanov and Malozemov [13] establish this fact in the following setting: X is a closed convex subset of R^n , Y is a compact subset of R^m , $f(x, y)$ is continuous and continuously differentiable with respect to x on $\tilde{X} \times Y$ where \tilde{X} is an open set containing X , $f(\cdot, y)$ is convex for each $y \in Y$, and there exists a solution to (2.1). One can also establish (2.12) under weaker assumptions using Clarke's result [10, Thm. 2.1] and properties of subgradients found in Rockafellar's book [29]. The analysis of Charnes, Cooper and Kortanek [7] also appears applicable. For completeness, we now derive (2.12) using Theorem 2.1 and results from [29]. First, let us consider the case where Y is a finite set.

LEMMA 2.3. *Suppose that Y is a finite set, X is a nonempty convex subset of R^n , and $f(\cdot, y)$ is convex for each $y \in Y$. If there exists a solution x^* to the primal problem (2.2), then there exists a solution λ^* to the dual problem (2.10) and $I(\lambda^*) = \Phi(x^*)$. Moreover, λ^* can be chosen so that its support has at most $n + 1$ elements.*

(It follows from Theorem 2.2 that under the hypotheses of Lemma 2.3, (2.12) holds for some set S which has at most $n + 1$ elements.)

Proof. By Theorem 2.1 and by the observations that follow the theorem, there exists a solution λ^* to the dual problem (2.10) and $I(\lambda^*) = \Phi(x^*)$. Let m denote the number of elements in Y and assume for convenience that $Y = \{1, \dots, m\}$. The equality $I(\lambda^*) = \Phi(x^*)$ combined with (2.13) tell us that

$$(2.24) \quad I(\lambda^*) = \text{minimum}_{x \in X} \sum_{i=1}^m \lambda_i^* f_i(x) = \sum_{i=1}^m \lambda_i^* f_i(x^*)$$

where $f_i(\cdot)$ denotes $f(\cdot, i)$. Let $\partial f_i(x)$ denote the collection of subgradients of f_i at x .

By [29, Thm. 27.4] and by (2.24), there exists $g_i \in \partial f_i(x^*)$ such that

$$(2.25) \quad \left\langle \sum_{i=1}^m \lambda_i^* g_i, x - x^* \right\rangle \geq 0 \quad \text{for every } x \in X.$$

Define the set $P = \{i \in [1, m]: \lambda_i^* > 0\}$. Since $\lambda^* \geq 0$ and $\lambda_1^* + \cdots + \lambda_m^* = 1$, it follows from [29, Thm. 17.1] that there exists nonnegative scalars μ_i for $i \in P$ such that the support of μ has at most $n+1$ elements,

$$\sum_{i \in P} \mu_i g_i = \sum_{i=1}^m \lambda_i^* g_i \quad \text{and} \quad \sum_{i \in P} \mu_i = 1.$$

Hence, (2.25) yields

$$(2.26) \quad \left\langle \sum_{i \in P} \mu_i g_i, x - x^* \right\rangle \geq 0 \quad \text{for every } x \in X.$$

Again by [29, Thm. 27.4] and by (2.26), we have

$$(2.27) \quad \sum_{i \in P} \mu_i f_i(x^*) = \text{minimum}_{x \in X} \sum_{i \in P} \mu_i f_i(x).$$

The identity $\mathbf{l}(\lambda^*) = \Phi(x^*)$ combined with Theorem 2.2 imply that $f_i(x^*) = \Phi(x^*)$ for every $i \in P$. Since the μ_i sum to one, it follows from (2.27) that

$$\Phi(x^*) = \text{minimum}_{x \in X} \sum_{i \in P} \mu_i f_i(x) = \mathbf{l}(\mu).$$

By Theorem 2.2, μ is a solution to the dual problem. \square

THEOREM 2.4. *Suppose that Y is a nonempty compact subset of a normed space, X is a nonempty compact, convex subset of \mathbb{R}^n , and f is a real-valued function defined on $\overset{\circ}{X} \times Y$ where $\overset{\circ}{X}$ is a relatively open set containing X . If $f(\cdot, y)$ is convex and lower semicontinuous for each y in Y and $f(x, \cdot)$ is continuous for each x in $\overset{\circ}{X}$, then there exists a solution x^* to (2.2), there exists a solution λ^* to (2.10), and $\mathbf{l}(\lambda^*) = \Phi(x^*)$. Moreover, λ^* can be chosen so that its support has at most $n+1$ elements.*

Proof. Let $\{y_1, y_2, \dots\}$ be a dense subset of Y and define $\Phi^N: X \rightarrow \mathbb{R}$ by

$$\Phi^N(x) = \text{maximum} \{f(x, y_i): i = 1, \dots, N\}.$$

Since $f(\cdot, y)$ is lower semicontinuous, Φ^N is lower semicontinuous. Thus the compactness of X guarantees the existence of $x^N \in X$ such that

$$\Phi^N(x^N) = \text{minimum}_{x \in X} \Phi^N(x).$$

In addition, the compactness of X implies that a subsequence of the x^N converges to some $x^* \in X$. By [29, Thm. 10.8] and the assumption that $f(x, \cdot)$ is continuous and $\{y_1, y_2, \dots\}$ is a dense subset of Y , Φ^N converges to Φ uniformly on X . Consequently, we have

$$\Phi(x^*) = \text{minimum} \{\Phi(x): x \in X\}.$$

Referring to Lemma 2.3, there exists a set $Y^N \subset \{y_1, \dots, y_N\}$ where Y^N has at most $n+1$ elements and there exists a corresponding set of nonnegative scalars $\{\lambda_y^N: y \in Y^N\}$ which sum to one and which satisfy the relation

$$(2.28) \quad \text{minimum}_{x \in X} \sum_{y \in Y^N} \lambda_y^N f(x, y) = \Phi^N(x^N).$$

Since the sets Y^N and the scalars λ_y^N lie in compact sets, we can extract convergent subsequences. Assume for convenience that x^N converges to x^* , λ^N converges to λ^* and Y^N converges to Y^* . For any $x \in X$,

$$(2.29) \quad \lim_{N \rightarrow \infty} \sum_{y \in Y^N} \lambda_y^N f(x, y) = \sum_{y \in Y^*} \lambda_y^* f(x, y)$$

since $f(x, \cdot)$ is continuous. Minimizing the left side of (2.29) over $x \in X$ yields

$$\lim_{N \rightarrow \infty} I(\lambda^N) \leq \sum_{y \in Y^*} \lambda_y^* f(x, y)$$

for every $x \in X$. Since $I(\lambda^N)$ is equal to $\Phi(x^N)$ and since $\Phi(x^N)$ approaches $\Phi(x^*)$ as N tends to infinity, we conclude that

$$(2.30) \quad \Phi(x^*) \leq \sum_{y \in Y^*} \lambda_y^* f(x, y)$$

for every $x \in X$. Minimizing the right side of (2.30) over x in X yields the relation $\Phi(x^*) \leq I(\lambda^*)$ and by (2.13), λ^* is a solution to the dual problem (2.10). \square

In Theorem 2.4, we can replace the assumption that X is compact with the assumption that X is closed, however, the existence of x^* is lost.

COROLLARY 2.5. *Suppose that Y is a nonempty compact subset of a normed space, X is a nonempty closed, convex subset of R^n and f is a real-valued function defined on $\hat{X} \times Y$ where \hat{X} is a relatively open set containing X . If $f(\cdot, y)$ is convex and lower semicontinuous for each y in Y and $f(x, \cdot)$ is continuous for each x in \hat{X} , then there exists a solution λ^* to (2.10), and*

$$I(\lambda^*) = \inf_{x \in X} \Phi(x).$$

Moreover, λ^* can be chosen so that its support has at most $n+1$ elements.

Proof. Given an integer N , define the set

$$X^N = \{x \in X: \|x\| \leq N\}$$

where $\|\cdot\|$ is any norm for R^n . We assume that N is large enough that X^N is nonempty. By Theorem 2.4, there exists $\lambda^N \in \Lambda$ such that

$$I^N(\lambda^N) = \text{minimum}_{x \in X^N} \Phi(x)$$

where I^N is defined by

$$I^N(\lambda) = \text{minimum}_{x \in X^N} \sum_{y \in Y} \lambda_y f(x, y)$$

and where the support Y^N of λ^N has at most $n+1$ elements. From (2.10), the components of λ^N are nonnegative and sum to one. Since Y^N and the components of λ^N lie in compact sets, there exists a subsequence of the λ^N which converges to some λ^* . Let Y^* denote the support of λ^* and assume for convenience that the entire sequence $\{\lambda^N\}$ converges to λ^* . For each $x \in X$, the continuity of $f(x, \cdot)$ implies that

$$(2.31) \quad \lim_{N \rightarrow \infty} \sum_{y \in Y^N} \lambda_y^N f(x, y) = \sum_{y \in Y^*} \lambda_y^* f(x, y).$$

We minimize the left side of (2.31) over $x \in X^N$ to obtain

$$(2.32) \quad \lim_{N \rightarrow \infty} I^N(\lambda^N) \leq \sum_{y \in Y^*} \lambda_y^* f(x, y).$$

Minimizing the right side of (2.32) over $x \in X$ yields

$$I(\lambda^*) \geq \lim_{N \rightarrow \infty} I^N(\lambda^N),$$

and since $\Phi(x^N)$ is equal to $I^N(\lambda^N)$, we have

$$I(\lambda^*) \geq \lim_{N \rightarrow \infty} \Phi(x^N) = \inf_{x \in X} \Phi(x).$$

Finally, (2.13) tells us that λ^* is a solution to the dual problem (2.10). \square

3. Augmented Lagrangians. A nonconvex problem often has a duality gap and the value of the dual problem is strictly less than the value of the primal problem. A strategy for bridging this gap emanates from work of Arrow and Solow [1], Hestenes [22], and Powell [28]. The basic idea is to augment the ordinary Lagrangian with a penalty term. To introduce this penalized dual approach, we first consider a finite-dimensional mathematical program with equality constraints:

$$(3.1) \quad \begin{array}{ll} \text{minimize} & f(x) \\ \text{subject to} & h(x) = 0, \quad x \in R^n \end{array}$$

where $f: R^n \rightarrow R$ and $h: R^n \rightarrow R^m$. (Mathematical programs in a Hilbert space setting are studied in [16] and [27].) The ordinary Lagrangian corresponding to (3.1) is $I(\lambda, x) = f(x) + \lambda^T h(x)$. Letting r be a positive scalar and letting $|\cdot|$ denote the Euclidean norm, the augmented Lagrangian corresponding to the penalty term $r|h(x)|^2$ is

$$(3.2) \quad L(\lambda, x) = f(x) + \lambda^T h(x) + r|h(x)|^2.$$

To illustrate the type of results that can be proved about the augmented Lagrangian, we state the following theorem which is extracted from Bertsekas [2]. In stating this theorem, our convention is that the gradient ∇ is a row vector and the gradient ∇h of the vector valued function h is a $m \times n$ matrix with i th row ∇h_i for $i = 1$ to m . Also, we let ∇^2 denote the Hessian matrix of second partial derivatives and the phrase “ x^* is a local minimizer for (3.1)” means that $h(x^*) = 0$ and $f(x^*) \leq f(x)$ whenever $h(x) = 0$ and x is near x^* .

THEOREM 3.1. *Suppose that x^* is a local minimizer for (3.1), both f and h are twice continuously differentiable in a neighborhood of x^* , and the rows of $\nabla h(x^*)$ are linearly independent. If $\lambda = \lambda^*$ is the solution to the equation*

$$(3.3) \quad \nabla f(x^*) + \lambda^{*T} \nabla h(x^*) = 0$$

and $\nabla_x^2 I(\lambda^, x^*)$ is positive definite in the null space of $\nabla h(x^*)$, then there exists a parameter s and a neighborhood N of x^* such that the problem*

$$\text{minimize } \{L(\mu, x): x \in N\}$$

has a unique minimizer $x_{\mu,r}$ whenever $r \geq s$ and $|\lambda^ - \mu| \leq r/s$. Moreover, there exists a constant c , independent of r and μ , such that*

$$(3.4) \quad |x_{\mu,r} - x^*| + |\lambda_{\mu,r} - \lambda^*| \leq c|\mu - \lambda^*|/r$$

where $\lambda_{\mu,r} := \mu + 2rh(x_{\mu,r})$.

Now let us consider the inequality constrained problem

$$(3.5) \quad \begin{array}{ll} \text{minimize} & f(x) \\ \text{subject to} & g(x) \leq 0, \quad x \in R^n \end{array}$$

where $g: R^n \rightarrow R^l$. Rockafellar's augmented Lagrangian (see [30]) is obtained by converting the inequality constraints to equality constraints using slack variables,

forming the augmented Lagrangian corresponding to these equality constraints, and minimizing over the slack variables to obtain

$$(3.6) \quad L(\lambda, x) = f(x) + \sum_{i \in I_+} (\lambda_i g_i(x) + r g_i(x)^2) - \frac{1}{4r} \sum_{i \in I_-} \lambda_i^2$$

where the sets I_+ and I_- are defined by

$$I_+ = \{i \in [1, l]: 2r g_i(x) + \lambda_i \geq 0\} \quad \text{and} \quad I_- = \{i \in [1, l]: 2r g_i(x) + \lambda_i < 0\}.$$

Thus the part of the Lagrangian (3.6) corresponding to indices $i \in I_+$ resembles the equality Lagrangian (3.2) while the part of the Lagrangian corresponding to indices $i \in I_-$ is locally independent of x . Theorem 3.1 also applies to inequality constrained problems since an inequality can be converted to an equality using Valentine's device (see [2] and [33]).

Augmented Lagrangians are now applied to the minimax problem (2.1). Let us consider the case where the set Y connected with the maximization is the integers $\{1, 2, \dots\}$ and at some solution x^* to (2.2), we have

$$f(x^*, i) \geq f(x^*, i+1)$$

for each i . We assume that $f(x^*, i) = \Phi(x^*)$ for $i = 1, \dots, N$ while $f(x^*, i) < \Phi(x^*)$ for $i > N$. Recall from § 2 that the set $S = \{1, \dots, N\}$ is usually the support of a dual multiplier λ^* . If a good estimate for x^* is known, then S is known and, at least locally, x^* is a solution to the equality constrained problem

$$(3.7) \quad \text{minimize } \{\rho: f(x) = \rho \mathbf{1}, x \in X, \rho \in R\}$$

where $\mathbf{1}$ denotes the vector in R^N with every component equal to 1 and $f(x)$ denotes the vector-valued function with i th component $f_i(x)$ equal to $f(x, i)$ for $i = 1, \dots, N$. The corresponding augmented Lagrangian is

$$(3.8) \quad L(\lambda, x, \rho) = \rho + \lambda^T (f(x) - \rho \mathbf{1}) + r |f(x) - \rho \mathbf{1}|^2.$$

In practice, the support set S for the minimax problem is not known, and we must use the inequality constrained formulation

$$\text{minimize } \{\rho: f(x) \leq \rho \mathbf{1}, x \in X, \rho \in R\}.$$

Since this formulation is equivalent to

$$(3.9) \quad \text{minimize } \{\rho: f(x) + z = \rho \mathbf{1}, z \geq 0, x \in X, \rho \in R, z \in R^N\},$$

the analogue of Rockafellar's augmented Lagrangian is

$$(3.10) \quad L(\lambda, x, \rho) = \text{minimum } \{\rho + \lambda^T (f(x) + z - \rho \mathbf{1}) + r |f(x) + z - \rho \mathbf{1}|^2: z \geq 0, z \in R^N\}.$$

Since the extremum in (3.10) is a strictly convex function of z , there exists a unique z which attains the minimum, and by (3.6), the augmented Lagrangian (3.10) can be expressed:

$$L(\lambda, x, \rho) = \rho + \sum_{i \in I_+} \{\lambda_i (f_i(x) - \rho) + r (f_i(x) - \rho)^2\} + \frac{1}{4r} \sum_{i \in I_-} \lambda_i^2$$

where

$$I_+ = \{i \in [1, N]: 2r (f_i(x) - \rho) + \lambda_i \geq 0\} \quad \text{and} \quad I_- = \{i \in [1, N]: 2r (f_i(x) - \rho) + \lambda_i < 0\}.$$

When using an augmented Lagrangian to solve the constrained optimization problem (3.9), we minimize $L(\lambda, x, \rho)$ over x in X and ρ in R to obtain the dual

functional $L(\lambda)$. Then the dual functional is maximized over λ to obtain a solution λ^* to the dual problem

$$\text{maximize } \{L(\lambda): \lambda \in R^N\}.$$

As we will show shortly, the minimum of $L(\lambda, x, \rho)$ over ρ can be computed explicitly. Let $L(\lambda, x)$ denote the partly minimized functional defined by

$$(3.11) \quad L(\lambda, x) = \text{minimum } \{L(\lambda, x, \rho): \rho \in R\}.$$

LEMMA 3.2. *There exists a unique ρ , which attains the minimum in (3.11).*

Proof. Defining the parameter

$$\rho_1 = \text{maximum } \{f_i(x) + \lambda_i/2r: i = 1, \dots, N\},$$

observe that $L(\lambda, x, \rho) = \rho$ plus a constant (independent of ρ) for $\rho \geq \rho_1$. Thus the minimum of $L(\lambda, x, \cdot)$ occurs on the interval $[-\infty, \rho_1]$. By [17, Cor. A6], the derivative of $L(\lambda, x, \rho)$ with respect to ρ is a Lipschitz continuous function of ρ on bounded intervals. Since the second derivative of $L(\lambda, x, \cdot)$ is at least $2r$ on $(-\infty, \rho_1]$, we conclude that $L(\lambda, x, \cdot)$ is strictly convex on $(-\infty, \rho_1]$ and there exists a unique minimum. \square

To compute the minimum for $L(\lambda, x, \cdot)$, we define the parameters

$$\rho_i = f_i(x) + \lambda_i/2r$$

for $i = 1, \dots, N$ and we reindex the components of f and λ so that

$$\rho_1 \geq \rho_2 \geq \dots \geq \rho_N.$$

Since $L(\lambda, x, \cdot)$ is strictly convex on $(-\infty, \rho_1]$, the derivative of $L(\lambda, x, \cdot)$ is monotone increasing (with slope at least $2r$). For ρ between ρ_{j+1} and ρ_j , the derivative of $L(\lambda, x, \cdot)$ is given by

$$(3.12) \quad \frac{d}{d\rho} L(\lambda, x, \rho) = 1 + 2rj\rho - \sum_{i=1}^j (\lambda_i + 2rf_i(x)).$$

With the convention that ρ_{N+1} is $-\infty$, there exists an interval $[\rho_{j+1}, \rho_j]$ where the derivative changes sign. Since $L(\lambda, x, \cdot)$ is a quadratic on this interval, the minimizer ρ^* of the quadratic is easily evaluated:

$$\rho^* = \frac{-1 + \sum_{i=1}^j (\lambda_i + 2rf_i(x))}{2rj}.$$

Since the computer time to sort $f_i(x) + \lambda_i/2r$ into decreasing order is proportional to $N \log_2 N$ (see [25]) while the time to evaluate the derivative (3.12) for $\rho = \rho_1$ through $\rho = \rho_N$ is proportional to N , the computer time required to minimize $L(\lambda, x, \cdot)$ is proportional to $N \log_2 N$.

4. General minimax problems. Now let us return to the minimax problem

$$(4.1) \quad \text{minimize}_{x \in X} \text{maximum}_{y \in Y} f(x, y)$$

where f is a real-valued function defined on $X \times Y$. As demonstrated in the proof of Theorem 2.4, one method to solve the minimax problem is to introduce a set $Y^N \subset Y$ with N elements and to consider the approximation

$$\text{minimize}_{x \in X} \text{maximum}_{y \in Y^N} f(x, y).$$

If $x^N \in X$ has the property that

$$\text{maximum}_{y \in Y^N} f(x^N, y) = \text{minimum}_{x \in X} \text{maximum}_{y \in Y^N} f(x, y)$$

where $Y^1 \subset Y^2 \subset Y^3 \subset \dots$ and the union of Y^N over N is a dense subset of Y , then under the hypotheses of Theorem 2.4, every convergent subsequence of $\{x^N\}$ approaches a solution to (4.1). Moreover, defining $\Phi^N: X \rightarrow R$ by

$$\Phi^N(x) = \text{maximum}_{y \in Y^N} f(x, y),$$

Dem'yanov and Malozemov [13] show that if x^N is an extreme point of Φ^N , then every convergent subsequence of $\{x^N\}$ approaches an extreme point of the function

$$\Phi(x) = \text{maximum}_{y \in Y} f(x, y).$$

Their assumptions are that $X \subset R^n$, $Y \subset R^m$, Y is compact, X is closed and convex, and $f(x, y)$ is continuous and continuously differentiable with respect to x on $\tilde{X} \times Y$ where \tilde{X} is an open set containing X .

The principal difficulty involved with primal algorithms for minimax problems is that the function Φ is almost always nondifferentiable at its minimum. Ways to circumvent this lack of smoothness are developed in the algorithms of Dem'yanov [12] and others. Unlike the primal function, the dual function is usually smooth at the solution to the dual problem. Nonetheless, as we now show, the dual problem can be ill conditioned and algorithms for solving the dual problem must deal with this conditioning. In describing the ill conditioning associated with the dual problem, we assume for simplicity that X is R^n . The dual functional corresponding to Y^N is

$$(4.2) \quad \mathbf{I}(\lambda) = \infimum_{x \in X} \mathbf{I}(\lambda, x)$$

where the Lagrangian $\mathbf{I}: R^N \times R^n \rightarrow R$ is defined by

$$\mathbf{I}(\lambda, x) = \sum_{y \in Y^N} \lambda_y f(x, y).$$

Suppose that $x = x^*$ attains the minimum in (4.2) when $\lambda = \lambda^*$, that $f(x, y)$ is twice continuously differentiable with respect to x for every $y \in Y^N$, and that the Hessian

$$(4.3) \quad \sum_{y \in Y^N} \lambda_y^* \nabla_x^2 f(x^*, y)$$

is positive definite. Since x^* attains the minimum in (4.2) when $\lambda = \lambda^*$, the gradient of the Lagrangian with respect to x is zero at $x = x^*$: $\nabla_x \mathbf{I}(\lambda^*, x^*) = 0$. Since the Hessian (4.3) is nonsingular, the implicit function theorem tells us that for λ near λ^* , there exists an $x(\lambda)$ that satisfies the equation

$$(4.4) \quad \nabla_x \mathbf{I}(\lambda, x(\lambda)) = 0,$$

and by the second order sufficiency condition, $x(\lambda)$ is a local minimizer for $\mathbf{I}(\lambda, \cdot)$. Let us assume that $x(\lambda)$ is also a global minimizer for $\mathbf{I}(\lambda, \cdot)$. By the chain rule and (4.4), we have

$$(4.5) \quad \frac{\partial \mathbf{I}}{\partial \lambda_y}(\lambda) = f(x(\lambda), y) + \nabla_x \mathbf{I}(\lambda, x(\lambda)) \frac{\partial x}{\partial \lambda_y}(x(\lambda), \lambda) = f(x(\lambda), y).$$

Differentiating (4.4) with respect to λ_z yields

$$\frac{\partial x}{\partial \lambda_z}(\lambda) = -\nabla_x^2 \mathbf{I}(\lambda, x(\lambda))^{-1} \nabla_x f(x(\lambda), z)^T,$$

and differentiating (4.5) with respect to λ_z gives us

$$\frac{\partial^2 \mathbf{I}}{\partial \lambda_y \partial \lambda_z}(\lambda) = -\nabla_x f(x(\lambda), y) \nabla_x^2 \mathbf{I}(\lambda, x(\lambda))^{-1} \nabla_x f(x(\lambda), z)^T.$$

Hence, the Hessian of the dual functional has the form

$$\frac{\partial^2 \mathbf{I}}{\partial \lambda^2}(\lambda) = -GF^{-1}G^T$$

where F is the $n \times n$ matrix given by

$$F = \sum_{y \in Y^N} \lambda_y \nabla_x^2 f(x(\lambda), y)$$

and G is the $N \times n$ matrix whose y th row is $\nabla_x f(x(\lambda), y)$ for $y \in Y^N$.

Remember that if λ is feasible in the dual problem, then the components of λ sum to one. If P is a $(N-1) \times N$ matrix whose rows are a basis in R^N for the space orthogonal to the vector with every component equal to one, then the convergence speed of steepest ascent applied to the dual functional is related to the distribution of eigenvalues for the Hessian of \mathbf{I} evaluated in the row space of P . The Hessian of $\mathbf{I}(P^T \mu)$ with respect to μ is given by $-(PG)F^{-1}(PG)^T$. Since PG is $(N-1) \times n$ while the Hessian of \mathbf{I} with respect to μ is $(N-1) \times (N-1)$, we conclude that the Hessian is singular whenever $N-1$ is greater than n , or equivalently, whenever N is greater than $n+1$.

Now consider the strategy of Theorem 2.4 where we introduce a set $Y^N \subset Y$ for which

$$\lim_{N \rightarrow \infty} \inf \{|y - z| : z \in Y^N\} = 0$$

for every $y \in Y$. By its structure, the Hessian of $\mathbf{I}(\lambda)$ is singular whenever the support of λ has more than $n+1$ elements. The convergence speed of numerical schemes (like steepest ascent) for solving the dual problem is governed by the ratio between the absolute largest eigenvalue and the absolute smallest eigenvalue of the Hessian, and as the ratio tends to infinity, the convergence speed approaches zero. If the support of λ has more than $n+1$ elements, then the smallest eigenvalue is zero, the ratio is infinity, and convergence is slow. In other words, asymptotically, it is impractical to maximize the dual functional using say steepest descent (or almost any standard algorithm) when N is large.

The augmented Lagrangian is subject to similar instabilities. For the inequality constrained problem (3.5) and the augmented Lagrangian (3.6), Rockafellar shows in an appropriate setting (see [31]) that

$$(4.6) \quad \frac{\partial^2 \mathbf{L}}{\partial \lambda_+^2}(\lambda^*) = -\nabla g_+(x^*) F_r^{-1} \nabla g_+(x^*)^T, \quad \frac{\partial^2 \mathbf{L}}{\partial \lambda_-^2}(\lambda^*) = -\frac{1}{2r} I, \quad \frac{\partial^2 \mathbf{L}}{\partial \lambda_+ \partial \lambda_-}(\lambda^*) = 0$$

where $\mathbf{L}(\lambda) = \inf \{\mathbf{L}(\lambda, x) : x \in R^n\}$, λ_{\pm} and g_{\pm} denote the components of λ and g corresponding to indices $i \in I_{\pm}$, and

$$F_r = \nabla^2 f(x^*) + \sum_{i \in I_+} (\lambda_i^* \nabla^2 g_i(x^*) + 2r \nabla g_i(x^*)^T \nabla g_i(x^*)).$$

Hence, the Hessian (4.6) is singular when the number of elements in I_+ is greater than n .

Recall that the minimax problem corresponding to Y^N can be written as the inequality constrained problem

$$(4.7) \quad \begin{aligned} &\text{minimize } \rho \\ &\text{subject to } x \in X, \quad \rho \in R, \quad f(x, y) \leq \rho \quad \text{for every } y \in Y^N. \end{aligned}$$

Thus the y th component of g in (3.5) is identified with $f(x, y) - \rho$. Since the independent variables in (4.7) are x and ρ , the primal problem (4.7) is formulated in R^{n+1} when $X \subset R^n$ and the Hessian

$$\frac{\partial^2 L}{\partial \lambda^2}(\lambda^*)$$

is singular when the number of elements in I_+ is greater than $n+1$. Observe that the augmented Lagrangian is better conditioned than the ordinary Lagrangian, since the part of the Hessian corresponding to the second partial derivative with respect to λ_- is a multiple of the identity matrix which is perfectly conditioned. Nonetheless, as N grows, the Hessian can still become singular.

Now let us develop an algorithm to solve the minimax problem. Given $x \in X$, let $y_1(x), y_2(x), \dots$ denote the local maxima of $f(x, \cdot)$ on Y . Our algorithm for solving the minimax problem has two phases. In both phases, we utilize the inequality formulation (4.7). However, in *phase one*, Y^N is a fixed set $\{y_1, \dots, y_N\}$ contained in Y and N is "large." In *phase two*, Y^N has the form $\{y_1(x), \dots, y_N(x)\}$ and N is "small." If $f(x, \cdot)$ has a finite number of local maxima on Y , then the phase two problem

$$(4.8) \quad \begin{aligned} &\text{minimize } \rho \\ &\text{subject to } x \in X, \quad \rho \in R, \quad f(x, y_i(x)) \leq \rho \quad \text{for } i = 1, \dots, N \end{aligned}$$

is usually equivalent to (4.1) for N sufficiently large. Since (4.8) involves tracking the peaks $y_i(\cdot)$, solving (4.8) is more difficult than solving (4.7). Hence, phase two should only be activated when the algorithm applied to (4.8) converges rapidly. For many mathematical programming algorithms, rapid convergence only occurs in a *neighborhood* of an optimum. For this reason, it is more efficient to apply an unsophisticated algorithm to the ill conditioned problem (4.7) generating a starting guess for a fast algorithm that solves (4.8).

Let us now show in detail how an algorithm such as [18, Algorithm 5.2] or any other algorithm with similar structure can be used to solve either (4.7) or (4.8). Each iteration of Algorithm 5.2 has the following steps: A restoration step where the equality and binding inequality constraints are partially satisfied, a multiplier update where an improved approximation to the optimal dual multipliers is generated, an unconstrained minimization step where the augmented Lagrangian is minimized using (for example) several preconditioned conjugate gradient iterations, and an adjustment to the penalty when a minimizer of the augmented Lagrangian has essentially been computed. This algorithm monitors the convergence of the iterations to a Kuhn-Tucker point and typically, both the restoration step and the multiplier update are only activated in a neighborhood of an optimum. In other words, unless the iterations are in a neighborhood of an optimum, Algorithm 5.2 is essentially a preconditioned conjugate gradient method applied to an augmented Lagrangian. In [18] we show that Algorithm 5.2 is globally convergent while the iterations are locally quadratically convergent. When applying Algorithm 5.2 to either (4.7) or (4.8), the following four issues must be considered:

(1) *The initialization of phase one and phase two.* That is, given a guess for a solution to (4.1), what is the corresponding starting guess for the multipliers? Given an approximation to a solution to (4.7), what is the starting guess to (4.8)?

(2) *The addition and deletion of constraints.* After each iteration of an algorithm applied to either (4.7) or (4.8), we must delete "unnecessary" elements from Y^N and we must add "significant" elements to Y^N . The augmented Lagrangian will help to determine which elements to delete and which elements to add.

(3) *The elimination of ρ .* We introduced the parameter ρ to convert the minimax problem into an inequality constrained mathematical program. When applying a mathematical programming algorithm to either (4.7) or (4.8), we would like to eliminate the artificial variable ρ so that the iterations are expressed in terms of x and λ .

(4) *The computation of the gradient of the augmented Lagrangian.* When using the conjugate gradient method or any other gradient-based scheme to minimize the augmented Lagrangian, we need a formula for the gradient of the augmented Lagrangian with respect to x . Clarke's result [10, Thm. 2.1] can be used to compute this gradient.

To begin, let us consider the initialization of phase one. If x_1 is the starting guess in phase one, then in the absence of better information, let Y^N be the maximizers of $f(x_1, \cdot)$ on Y . In other words, $\eta \in Y^N$ if and only if

$$f(x, \eta) = \text{maximum}_{y \in Y} f(x, y).$$

In the absence of better information, the starting guess λ_1 for the multipliers corresponding to the constraint $f(x, y) \leq \rho$ is $\lambda_{1y} = 1/N$ for each $y \in Y^N$. The x starting guess for phase two is simply the final iteration x_k of phase one. To initialize the phase two multipliers, we collapse the components of the phase one multipliers around the nearest peak. That is, in phase one we generate a multiplier λ_k with support Y^N . Given an element y in Y^N , the index $\nu(y)$ of the nearest peak is

$$\nu(y) = \arg \min \{\|y - y_i(x_k)\| : i = 1, 2, \dots\}.$$

When more than one index achieves the minimum, let $\nu(y)$ be any one of them. Then the i th component of λ_1 , the starting guess for phase two, is the sum of the phase one multiplier components that correspond to elements of Y^N closest to $y_i(x_k)$:

$$\lambda_{1i} = \sum_{\substack{y \in Y^N \\ \nu(y)=i}} \lambda_{ky}.$$

Moreover, the starting set Y^N for phase two consists of those $y_i(\cdot)$ for which λ_{1i} is positive.

To reduce the computing time associated with algorithms to solve (4.7) or (4.8), we wish to keep N as small as possible. After each complete iteration of [18, Algorithm 5.2], we will drop those constraints that appear to be nonbinding and we will add constraints where the inequality $f(x, y) \leq \rho$ seems to be violated significantly. Let us now explain more precisely when to delete or add constraints. Given a finite set $S \subset Y$ and a multiplier λ with support in S , the augmented Lagrangian introduced in § 3 is

$$(4.9) \quad L(\lambda, S, x) = \text{minimum} \{L(\lambda, S, x, \rho) : \rho \in R\}$$

where

$$L(\lambda, S, x, \rho) = \rho + \sum_{y \in S_+} \{\lambda_y(f(x, y) - \rho) + r(f(x, y) - \rho)^2\} - \frac{1}{4r} \sum_{y \in S_-} \lambda_y^2.$$

As usual, the limits for the summations above are

$$(4.10) \quad \begin{aligned} S_+ &= \{y \in S : 2r(f(x, y) - \rho) + \lambda_y \geq 0\} \quad \text{and} \\ S_- &= \{y \in S : 2r(f(x, y) - \rho) + \lambda_y < 0\}. \end{aligned}$$

PROPOSITION 4.1. *For fixed r, λ, S , and x , suppose that $\rho = \rho^*$ attains the minimum in (4.9), and let S_+ be the corresponding set given in (4.10). Then we have*

$$L(\lambda, S, x) = L(\lambda_+, S_+, x) - \frac{1}{4r} \sum_{y \in S_-} \lambda_y^2$$

where λ_+ denotes the vector formed from λ by extracting those components λ_y corresponding to $y \in S_+$.

Proof. The identity

$$0 = \frac{d}{d\rho} L(\lambda, S, x, \rho^*) = \frac{d}{d\rho} L(\lambda_+, S_+, x, \rho^*)$$

implies that ρ^* also minimizes $L(\lambda_+, S_+, x, \cdot)$. \square

Since $L(\lambda, S, x)$ just differs from $L(\lambda_+, S_+, x)$ by a constant, Proposition 4.1 implies that, at least locally, the constraints $f(x, y) \leq \rho$ corresponding to $y \in S_-$ can be dropped. Consequently, our rule for deleting elements from Y^N can be stated:

CONSTRAINT DELETION

Let λ_k and x_k denote the approximations generated by one complete iteration of say [18, Algorithm 5.2]. Delete from Y^N those elements corresponding to $y \in Y_-^N$.

Now consider the addition of constraints. Again, the augmented Lagrangian helps us decide when the N in (4.8) must be increased. Suppose that η is not an element of S and $f(x, \eta) < \rho^*$ where $\rho = \rho^*$ attains the minimum in (4.9). Letting S_η denote $S \cup \{\eta\}$, we now show that $L(\lambda, S_\eta, x)$ is locally equal to $L(\lambda, S, x)$ if f is continuous and λ_η is zero. By the definition of the augmented Lagrangian, we have the inequality $L(\lambda, S_\eta, x) \geq L(\lambda, S, x)$ whenever $\lambda_\eta = 0$. Since $L(\lambda, S_\eta, x, \rho^*) = L(\lambda, S, x, \rho^*)$, it follows that $L(\lambda, S_\eta, x) = L(\lambda, S, x)$ whenever $\lambda_\eta = 0$. Since the inequality $f(x, \eta) < \rho^*$ is preserved for small perturbations in x when f is continuous, we conclude that $L(\lambda, S_\eta, x)$ is locally equal to $L(\lambda, S, x)$. Conversely, suppose that $f(x, \eta) > \rho^*$ and $\lambda_\eta = 0$. Since $L(\lambda, S_\eta, x, \rho) \geq L(\lambda, S, x, \rho)$ for every ρ when $\lambda_\eta = 0$ and since $L(\lambda, S_\eta, x, \rho^*) > L(\lambda, S, x, \rho^*)$, it follows from the uniqueness result Lemma 3.2 that $L(\lambda, S_\eta, x) > L(\lambda, S, x)$. To summarize, if $f(x, \eta) > \rho^*$, then $L(\lambda, S_\eta, x)$ is larger than $L(\lambda, S, x)$ and the gap between the value of the primal problem (4.1) and the value of the dual problem

$$\max_{\lambda, S} \min_{x \in X} L(\lambda, S, x)$$

may be reduced by inserting η into S . These observations lead us to the following rule for adding constraints in phase one:

CONSTRAINT ADDITION IN PHASE ONE

Let λ_k and x_k denote the approximations generated by one complete iteration of say [18, Algorithm 5.2] and let $\rho = \rho^*$ minimize $L(\lambda_k, Y^N, x_k, \rho)$ over ρ . Insert $y_i(x_k)$ into Y^N if $f(x_k, y_i(x_k)) > \rho^*$ and the distance between $y_i(x_k)$ and Y^N is greater than some fixed predetermined constant Δ .

Since phase one approximates the solution to the minimax problem, the local maximizer $y_i(x_k)$ appearing in the constraint addition step of phase one does not need to be computed very accurately. The positive parameter Δ introduced above prevents points in Y^N from clustering together. As the number of points in Y^N increases, the time to evaluate L increases and the Hessian of L becomes ill conditioned. Since it helps to keep the number of points in Y^N small, we exclude those local maxima which are already near elements of Y^N . In numerical experiments, the convergence speed is not very sensitive to the choice of Δ . In phase two, the elements of Y^N are local

maxima instead of fixed elements in Y . Hence, the analogous rule for adding constraints in phase two can be stated:

CONSTRAINT ADDITION IN PHASE TWO

Let λ_k and x_k denote the approximations generated by one complete iteration of say [18, Algorithm 5.2] and let $\rho = \rho^*$ minimize $L(\lambda_k, Y^N, x_k, \rho)$ over ρ . Insert $y_i(\cdot)$ into Y^N if $f(x_k, y_i(x_k)) > \rho^*$.

Up to here, we have explained how to initialize a mathematical programming algorithm to solve either (4.7) or (4.8) and we have explained how to add or delete constraints at the end of each iteration in the algorithm. Now let us consider the details of an iteration. In formulation (4.7) and (4.8), a parameter ρ is introduced and the number of independent variables is increased by one. For many algorithms, the artificial variable ρ can be eliminated and the iterations can be expressed in terms of x and λ . For notational convenience, we assume X is R^n . Let x_k be the k th approximation to a solution to the minimax problem and suppose that after deleting and adding constraints at the end of iteration k , we have $Y^N = \{y_1, \dots, y_N\}$. In [18, Algorithm 5.2], we estimate the multipliers corresponding to the constraints of (4.7) or (4.8) by computing the least squares solution λ to the system of equations

$$(4.11) \quad \sum_{i=1}^N \lambda_i = 1, \quad \sum_{i=1}^N \lambda_i \nabla f(x_k, y_i) = 0.$$

Computing the least squares solution to this system of $n+1$ equations in N unknowns is equivalent to computing the pseudoinverse of a $(n+1) \times N$ matrix. As an alternative to this procedure, we suggest the following: Solve the first equation in (4.11) for λ_1 in terms of λ_2 through λ_N and substitute into the second relation to obtain n equations in $N-1$ unknowns:

$$\sum_{i=2}^N \lambda_i (\nabla_x f(x_k, y_i) - \nabla_x f(x_k, y_1)) = -\nabla_x f(x_k, y_1).$$

The least squares solution to this system gives us an estimate for λ_2 through λ_N while λ_1 is determined from the relation $\lambda_1 = 1 - \lambda_2 - \lambda_3 - \dots - \lambda_N$.

The procedure outlined above to estimate the multipliers is quite effective in phase two. On the other hand, in phase one a simpler strategy involving the gradient approximation to the multipliers (see [2]) is often just as effective. Let λ_k be the k th approximation to the multipliers and let x_k be the corresponding approximation to a minimizer of the augmented Lagrangian $L(\lambda_k, Y^N, \cdot)$. Set $\lambda_{k+1,i} = \lambda_{ki} + 2r(f(x_k, y_{ki}) - \rho_k)$ for $i \in Y_+^N$ and set $\lambda_{k+1,i} = 0$ for $i \in Y_-^N$ if this rule generates a λ_{k+1} with the property that (λ_{k+1}, x_k) is a better approximation to a Kuhn-Tucker point for (4.7) than (λ_k, x_k) . Otherwise, set $\lambda_{k+1} = \lambda_k$. Here ρ_k denotes the minimizer in (4.9) corresponding to $\lambda = \lambda_k$, $S = Y^N$, and $x = x_k$. A technique for measuring the distance to a Kuhn-Tucker point is developed in [18].

In [18, Algorithm 5.2], the restoration step is essentially a Newton iteration applied to the system of N equations

$$(4.12) \quad f(x, y_i) = \rho \quad \text{for } i = 1 \text{ to } N$$

where the starting guess is x_k and the corresponding ρ_k generated the previous iteration. Since the N equations (4.12) are equivalent to the $N-1$ equations

$$(4.13) \quad f(x, y_i) - f(x, y_1) = 0 \quad \text{for } i = 2 \text{ to } N,$$

an alternative procedure is to apply one Newton iteration to the system (4.13). Observe that this Newton iteration involves computing the pseudoinverse of the same matrix used in the multiplier estimate.

In the minimization step of [18, Algorithm 5.2], we use a preconditioned conjugate gradient method to minimize $L(\lambda_k, Y^N, x)$ over x . (Here, λ_k denotes the multiplier associated with iteration k .) Near the optimum, the preconditioner is chosen to project the gradients into the null space of the binding constraints. Therefore, near the optimum, the preconditioner projects the gradients into the null space of the $(N-1) \times n$ matrix with rows

$$\nabla_x f(x_k, y_i) - \nabla_x f(x_k, y_1) \quad \text{for } i = 2 \text{ to } N.$$

Far from the optimum, the preconditioner is chosen to mitigate the ill conditioning due to penalty terms in the augmented Lagrangian. At the start of iteration $k+1$, the inequalities $f(x, y_i) \leq \rho$ are viewed as equalities and it follows from (4.9) that for x near x_k ,

$$L(\lambda_k, Y^N, x) = \sum_{i=1}^N \lambda_{ki} f(x, y_i) + r \sum_{i=1}^N f(x, y_i)^2 - \frac{r}{N} \left(\sum_{i=1}^N f(x, y_i) \right)^2.$$

The identity

$$\sum_{i=1}^N f(x, y_i)^2 - \frac{1}{N} \left(\sum_{i=1}^N f(x, y_i) \right)^2 = \sum_{i=1}^N \left(f(x, y_i) - \frac{1}{N} \sum_{j=1}^N f(x, y_j) \right)^2,$$

combined with the preconditioning theory developed in [18, § 4], tells us that a natural preconditioner for the minimax problem is the matrix $H = (I + B^T B)^{-1}$, where B is the $N \times n$ matrix with i th row

$$\nabla_x f(x_k, y_i) - \frac{1}{N} \sum_{j=1}^N \nabla_x f(x_k, y_j).$$

Observe that the rows of B are linearly dependent since their sum is zero. Let V denote the matrix $I - vv^T$, where

$$v = \frac{1}{\sqrt{\sqrt{N}(1+\sqrt{N})}} \begin{bmatrix} 1 + \sqrt{N} \\ 1 \\ \vdots \\ 1 \end{bmatrix}.$$

Since the first row of V is a multiple of $\mathbf{1}$, the first row of VB is zero. Let W be the matrix obtained by deleting the first row of VB . Since V is orthogonal, we have

$$B^T B = (VB)^T VB = W^T W.$$

Applying the Woodbury formula [15, p. 3], the preconditioner H can be written

$$H = (I + rW^T W)^{-1} = I - W^T (r^{-1}I + WW^T)^{-1} W.$$

When using any gradient technique to minimize $L(\lambda_k, Y^N, \cdot)$, we must compute the gradient of the augmented Lagrangian with respect of x . By [10, Thm. 2.1] this gradient can be expressed

$$\nabla_x L(\lambda, S, x) = \sum_{y \in S_+} (\lambda_y + 2r(f(x, y) - \rho^*)) \nabla_x f(x, y)$$

where ρ^* attains the minimum in (4.9). This formula for the gradient is also valid when the elements of S depend on x (as in (4.8)) provided these elements are local

extreme points of $f(x, \cdot)$ on Y and (for example) $\nabla_x f(x, y)$ is a continuous function of x and y .

Comparing our approach to the minimax problem to the approach of Murray and Overton [26], some similarities are that we both reformulate the minimax problem as a mathematical program with an extra unknown and we both estimate simultaneously the primal solution and the Lagrange multipliers. Some differences in our methods are the following: (i) In [26] Y is finite. (ii) We utilize an augmented Lagrangian while [26] considers the ordinary Lagrangian. (iii) Our strategy for adding and deleting constraints is different from [26]—our strategy ties in with the augmented Lagrangian. (iv) With our approach, nonlinear constraints contained in X can be incorporated in the augmented Lagrangian just as easily as the constraints $f(x, y) \leq \rho$.

5. Reducible minimax problems. So far we have viewed the minimax problem as an optimization problem with inequality constraints and we have applied a constrained optimization algorithm. Now let us develop an algorithm that is specially tailored to reducible minimax problems. That is, we assume that there exists a finite set $Y^* \subset Y$ and a x^* in X such that

$$\min_{x \in X} \max_{y \in Y^*} f(x, y) = \max_{y \in Y^*} f(x^*, y) = \max_{y \in Y} f(x^*, y) = \Phi(x^*),$$

and we search for the set Y^* . It is also assumed that there exists a real number r and a multiplier λ^* with support in Y^* such that $L(\lambda^*, Y^*) = \Phi(x^*)$ where $L(\cdot, \cdot)$ denotes the dual functional defined by

$$L(\lambda, S) = \inf \{L(\lambda, S, x) : x \in X\}.$$

Given an approximation $Y_k = \{y_{k1}, \dots, y_{kN}\}$ to Y^* and given an approximation λ_k to λ^* , the rules for computing Y_{k+1} and λ_{k+1} are the following:

PEAK CHASING ALGORITHM

(a) If x_k minimizes $L(\lambda_k, Y_k, x)$ over $x \in X$, then set $\lambda_{k+1,i} = \lambda_{ki} + 2r(f(x_k, y_{ki}) - \rho_k)$ for $i \in S_+$ and $\lambda_{k+1,i} = 0$ for $i \in S_-$ where ρ_k attains the minimum in (4.9) corresponding to $\lambda = \lambda_k$ and $S = Y_k$.

(b) If x_{k+1} minimizes $L(\lambda_{k+1}, Y_k, \cdot)$ over X and if $Z_k = \{z_1, \dots, z_N\}$ denotes a collection of local maxima for $f(x_{k+1}, \cdot)$ on Y where z_i is the closest local maximizer to y_{ki} , then we set $Y_{k+1} = \beta Y_k + (1 - \beta)Z_k$ where

$$\beta = \arg \max_{0 \leq \alpha \leq 1} L(\lambda_{k+1}, \alpha Y_k + (1 - \alpha)Z_k).$$

Step (a) is the usual gradient step for an augmented Lagrangian (see [2]). Since this algorithm is linearly convergent, the parameter β of step (b) can be imprecise. In practice, we find that the maximizer of the interpolating quadratic that agrees with $L(\lambda_{k+1}, \alpha Z_k + (1 - \alpha)Y_k)$ at $\alpha = 0$, at $\alpha = \frac{1}{2}$, and at $\alpha = 1$ works well. To show that the peak chasing algorithm is locally convergent, we verify that each iteration increases the value of the dual functional. That is, $L(\lambda_{k+1}, Y_{k+1}) \geq L(\lambda_k, Y_k)$ with equality only possible at $\lambda_k = \lambda^*$ and at $Y_k = Y^*$. In order to show that step (a) is an ascent step, let us first consider the equality constrained problem

$$(5.1) \quad \begin{aligned} &\text{minimize } f(x) \\ &\text{subject to } h(x) = 0, \quad x \in R^n \end{aligned}$$

where f is quadratic: $f(x) = x^T A x + a^T x$ and h is linear: $h(x) = Bx - b$. Here A is an $n \times n$ matrix, B is an $m \times n$ matrix, and a and b are vectors in R^n and R^m , respectively. The augmented Lagrangian corresponding to (5.1) is

$$L(\lambda, x) = f(x) + \lambda^T h(x) + r|h(x)|^2.$$

LEMMA 5.1. *Suppose that the rows of B are linearly independent and A is positive definite in the null space of B . Then there exist positive parameters α and s such that $A + rB^T B \geq \alpha I$ for every $r \geq s$. If $\lambda \in R^n$ and z minimizes $L(\lambda, \cdot)$ over R^n , then we have*

$$(5.2) \quad L(\lambda + 2rh(z), y) \geq L(\lambda, z) + \frac{1}{2}r|h(z)|^2 + \alpha|y - z|^2$$

for every $r \geq 3s$ and for every $y \in R^n$.

(If M_1 and M_2 are symmetric matrices of the same dimension, then the notation $M_1 > M_2$ means that $M_1 - M_2$ is positive definite.)

Proof. In [16, Lemma 2.6] we determine a parameter $s < \infty$ with the property that $A + rB^T B$ is positive definite for $r \geq s$. Let μ denote $\lambda + 2rh(z)$. Expanding the Lagrangian in a Taylor series, we have

$$(5.3) \quad L(\mu, y) = L(\mu, z) + \nabla_x L(\mu, z)(y - z) + \frac{1}{2}\nabla_x^2 L(\mu, \xi)(y - z)^2$$

where ξ lies on the line segment connecting y and z . The relation $\mu = \lambda + 2rh(z)$ implies that

$$(5.4) \quad L(\mu, z) = L(\lambda, z) + 2r|h(z)|^2$$

and

$$(5.5) \quad \nabla_x L(\mu, z)(y - z) = \nabla_x L(\lambda, z)(y - z) + 2rh(z)^T \nabla h(z)(y - z).$$

(Note that $\nabla_x L(\mu, z)$ is not equal to the gradient of the right side of (5.4) since μ is treated as a constant when computing $\nabla_x L(\mu, z)$.) If z minimizes $L(\lambda, \cdot)$, then $\nabla_x L(\lambda, z)$ is zero and by (5.5), we have

$$(5.6) \quad \nabla_x L(\mu, z)(y - z) = 2rh(z)^T \nabla h(z)(y - z).$$

By the definition of f and h , it follows that $\nabla h = B$ and $\nabla_x^2 L = 2(A + rB^T B)$. Combining (5.3), (5.4), and (5.6) gives us

$$L(\mu, y) = L(\lambda, z) + 2r|h(z)|^2 + 2rh(z)^T B(y - z) + (y - z)^T (A + rB^T B)(y - z).$$

Utilizing the inequality

$$ab \leq \frac{3}{4}a^2 + \frac{1}{3}b^2$$

where we identify a with $|h(z)|$ and b with $|B(y - z)|$ yields

$$(5.7) \quad L(\mu, y) \geq L(\lambda, z) + \frac{1}{2}r|h(z)|^2 + (y - z)^T (A + \frac{1}{3}rB^T B)(y - z).$$

Hence, (5.2) holds for $r \geq 3s$. \square

For a general f and h , the same argument employed in the proof of Lemma 5.1 can also be applied to a neighborhood of a local optimum. Removing the restriction that f is quadratic and h is linear, we have the following.

THEOREM 5.2. *Suppose that x^* and λ^* satisfy the hypotheses of Theorem 3.1. Then there exists a neighborhood N_x of x^* , a neighborhood N_λ of λ^* , and positive parameters α and s such that*

$$(5.8) \quad \nabla_x^2 l(\lambda, x) + 2r\nabla h(x)^T \nabla h(x) > \alpha I$$

whenever $r \geq s$, $\lambda \in N_\lambda$, and $x \in N_x$. Moreover, for s sufficiently large, a parameter c can be chosen so that $L(\lambda, \cdot)$ has a unique local minimizer $x(\lambda)$ inside N_x whenever $|\lambda - \lambda^*| \leq cr$ and $r \geq s$. And if λ and μ lie in N_λ , $y \in N_x$ with $|y - x^*| \leq c|\lambda - \lambda^*|/r$, and $r \geq s$, then we have

$$(5.9) \quad L(\mu, y) \geq L(\lambda, z) + \frac{1}{2}(r|h(z)|^2 + (\alpha - \delta|\lambda - \lambda^*|)|y - z|^2)$$

where $z = x(\lambda)$, $\mu = \lambda + 2rh(z)$, and δ is a constant that is independent of λ and y .

Proof. By [16, Lemma 2.6], $\nabla_x^2 L(\lambda^*, x^*)$ is positive definite for r sufficiently large, and by [16, Lemma 6.5], there exists a neighborhood of (λ^*, x^*) where (5.8) holds for α sufficiently small and r sufficiently large. The statement concerning the existence of a locally unique minimizer $x(\lambda)$ for $L(\lambda, \cdot)$ is established (for example) in [2]. To prove (5.9), we expand L in a Taylor series giving us the following analogue of (5.7):

$$(5.10) \quad L(\mu, y) \geq L(\lambda, z) + \frac{1}{2}r|h(z)|^2 + \frac{1}{2}(y - z)^T \nabla_x^2 L(\mu, \xi)(y - z) - \frac{2}{3}r|\nabla h(z)(y - z)|^2$$

where ξ lies between y and z . Utilizing the inequality

$$|x|^2 \leq (|x - y| + |y|)^2 \leq 5|x - y|^2 + \frac{5}{4}|y|^2$$

where x is identified with $\nabla h(z)(y - z)$ and y is identified with $\nabla h(\xi)(y - z)$, the last two terms in (5.10) satisfy the relation

$$(5.11) \quad \begin{aligned} & \frac{1}{2}(y - z)^T \nabla_x^2 L(\mu, \xi)(y - z) - \frac{2}{3}r|\nabla h(z)(y - z)|^2 \\ & \geq \frac{1}{2}(y - z)^T \nabla_x^2 L(\mu, \xi)(y - z) + \frac{1}{6}r|\nabla h(\xi)(y - z)|^2 \\ & \quad - \frac{10}{3}r|(\nabla h(\xi) - \nabla h(z))(y - z)|^2 \\ & \quad + r(y - z)^T \left(\sum_{i=1}^m h_i(\xi) \nabla^2 h_i(\xi) \right) (y - z). \end{aligned}$$

By (5.8), we have

$$(5.12) \quad \frac{1}{2}(y - z)^T \nabla_x^2 L(\mu, \xi)(y - z) + \frac{1}{6}r|\nabla h(\xi)(y - z)|^2 \geq \frac{1}{2}\alpha|y - z|^2$$

provided r is sufficiently large, $\mu \in N_\lambda$, and $\xi \in N_x$. By Theorem 3.1, $|z - x^*|$ is bound by a constant times $|\lambda - \lambda^*|/r$ and by assumption, $|y - x^*|$ is bound by a constant times $|\lambda - \lambda^*|/r$. Since $h(x^*)$ is zero, there exists a constant δ such that

$$(5.13) \quad \begin{aligned} & \frac{10}{3}r|(\nabla h(\xi) - \nabla h(z))(y - z)|^2 + r(y - z)^T \left(\sum_{i=1}^m h_i(\xi) \nabla^2 h_i(\xi) \right) (y - z) \\ & \leq \delta|\lambda - \lambda^*| |y - z|^2 \end{aligned}$$

for λ near λ^* . Combining (5.10)–(5.13), the proof is complete. \square

Theorem 5.2 implies that if λ is near λ^* , then for $\mu = \lambda + 2rh(x(\lambda))$, $L(\mu, x(\mu))$ is equal to $L(\lambda, x(\lambda))$ only if $h(x(\lambda)) = 0$. Since $x(\lambda)$ minimizes $L(\lambda, \cdot)$ over N_x , we conclude that if $L(\mu, x(\mu)) = L(\lambda, x(\lambda))$, then $x = x(\lambda)$ is a solution to the problem: minimize $f(x)$ subject to $x \in N_x$ and $h(x) = 0$. Therefore, $x(\lambda) = x^*$, the local minimizer corresponding to λ^* . Theorem 5.2 also applies to problems of the form

$$(5.14) \quad \begin{aligned} & \text{minimize } f(x) \\ & \text{subject to } h(x) = 0, \quad x \in X \end{aligned}$$

where X is a convex set. The proof of Theorem 5.2 in this more general setting involves an analogue of Theorem 3.1 that applied to (5.14). See Bertsekas [2] for the extension of Theorem 3.1 to problems with the constraint $x \in X$. Furthermore, referring to the

proof of Lemma 5.1, the constraint $x \in X$ alters the treatment of the term $\nabla_x \mathbf{L}(\lambda, z)(y - z)$. When X is R^n , $\nabla_x \mathbf{L}(\lambda, z)$ is zero and the term $\nabla_x \mathbf{L}(\lambda, z)(y - z)$ can be dropped. But for an arbitrary convex set, the corresponding relation is

$$\nabla_x \mathbf{L}(\lambda, z)(y - z) \geq 0.$$

This inequality has the right direction so that the term $\nabla_x \mathbf{L}(\lambda, z)(y - z)$ can still be dropped without affecting (5.7).

Although Theorem 5.2 is established for an equality constraint, it also applies to the inequality constraint $g(x) \leq 0$ provided λ_i^* is positive whenever $g_i(x^*)$ is zero. This follows from [31, Thm. 5.1] where Rockafellar proves that near λ^* , minimizing $\mathbf{L}(\lambda, \cdot)$ is equivalent to minimizing the augmented Lagrangian corresponding to an equality constraint $h(x) = 0$ —the components h_i of h are the components g_i of g for which $\lambda_i^* > 0$. Hence, by Theorem 5.2, step (a) of the peak chasing algorithm is an ascent step under appropriate assumptions. That is, $\mathbf{L}(\lambda_{k+1}, Y_k) \geq \mathbf{L}(\lambda_k, Y_k)$ with equality only possible when the x_k which locally minimizes $\mathbf{L}(\lambda_k, Y_k, \cdot)$ on X is a local minimizer for the problem

$$\underset{x \in X}{\text{minimize}} \quad \underset{y \in Y_k}{\text{maximum}} \quad f(x, y).$$

Now consider step (b) of the peak chasing algorithm. As noted above, in a neighborhood of an optimum, the augmented Lagrangian corresponding to an inequality constrained problem is the same as the augmented Lagrangian corresponding to an equality constrained problem. For this reason, we focus attention on the augmented Lagrangian

$$\mathbf{L}(\lambda, S, x) = \sum_{y \in S} \lambda_y f(x, y) + r \sum_{y \in S} f(x, y)^2 - \frac{r}{N} \left(\sum_{y \in S} f(x, y) \right)^2$$

which corresponds to the equality constrained problem

$$\text{minimize } \rho$$

$$\text{subject to } x \in X, \quad \rho \in R, \quad f(x, y) = \rho \quad \text{for every } y \in S.$$

And we establish the following property for the dual functional:

LEMMA 5.3. *Let λ be a fixed vector in R^N with nonnegative components, let $S_0 = \{y_1, \dots, y_N\}$ and let $S_1 = \{z_1, \dots, z_N\}$ be subsets of Y , and suppose that for $i = 0$ and for $i = 1$, x_i minimizes $\mathbf{L}(\lambda, S_i, x)$ over x in X . Then we have*

$$(5.15) \quad \mathbf{L}(\lambda, S_1) \geq \mathbf{L}(\lambda, S_0) + \sum_{i=1}^N (\lambda_i + 2r(f(x_1, y_i) - \rho_1))(f(x_1, z_i) - f(x_1, y_i))$$

where

$$\rho_1 = \frac{1}{N} \sum_{i=1}^N f(x_1, y_i).$$

Proof. Let $Q: R^N \rightarrow R$ be the quadratic defined by

$$Q(p) = \sum_{i=1}^N \lambda_i p_i + r p_i^2 - \frac{r}{N} \left(\sum_{i=1}^N p_i \right)^2.$$

The Hessian of Q is

$$\nabla^2 Q = 2r \left(I - \frac{1}{N} \mathbf{1} \mathbf{1}^T \right),$$

which is positive semidefinite by Gerschgorin's theorem. Hence, Q is a convex function which satisfies the standard inequality [29, p. 242]:

$$(5.16) \quad Q(p) - Q(q) \geq \nabla Q(q)(p - q) = \sum_{i=1}^N (\lambda_i + 2r(q_i - \rho))(p_i - q_i)$$

where

$$\rho = \frac{1}{N} \sum_{i=1}^N q_i.$$

Since x_0 minimizes $L(\lambda, S_0, x)$ over $x \in X$, it follows that $L(\lambda, S_0) \leq L(\lambda, S_0, x_1)$, or equivalently,

$$(5.17) \quad L(\lambda, S_1) - L(\lambda, S_0) \geq L(\lambda, S_1, x_1) - L(\lambda, S_0, x_1).$$

Applying (5.16) to the right side of (5.17) where $p_i = f(x_1, z_i)$ and $q_i = f(x_1, y_i)$ yields (5.15). \square

Lemma 5.3 can be used to show that under appropriate assumptions, step (b) of the peak chasing algorithm is an ascent step. Let S_α denote $\alpha S_1 + (1 - \alpha)S_0$. Suppose that for α between zero and one, the minimum of $L(\lambda, S_\alpha, x)$ over $x \in X$ is attained at a point labeled x_α which is a continuous function of α . Let λ^* maximize $L(\lambda, S_0)$ over λ . Typically the components of λ^* are positive. By Theorem 3.1, the vector μ_0 with components

$$\mu_{0i} = \lambda_i + 2r(f(x_0, y_i) - \rho_0), \quad \rho_0 = \frac{1}{N} \sum_{i=1}^N f(x_0, y_i)$$

satisfies the inequality $|\mu_0 - \lambda^*| \leq c|\lambda - \lambda^*|/r$ for some constant c . Hence, the components of μ_0 are positive for r sufficiently large. Letting μ_α be the vector defined by

$$\mu_{\alpha i} = \lambda_i + 2r(f(x_\alpha, y_i) - \rho_\alpha), \quad \rho_\alpha = \frac{1}{N} \sum_{i=1}^N f(x_\alpha, y_i),$$

it follows that the components of μ_α are positive for α sufficiently small. If z_i is a local maximizer of $f(x_0, \cdot)$, then we expect that $f(x_0, \cdot)$ is locally concave near z_i . Assuming that $f(z_\alpha, \cdot)$ is concave for α near zero on the line segment connecting y_i and z_i , we have

$$(5.18) \quad f(x_\alpha, \alpha z_i + (1 - \alpha)y_i) \geq \alpha f(x_\alpha, z_i) + (1 - \alpha)f(x_\alpha, y_i).$$

Combining (5.15) and (5.18) gives us

$$(5.19) \quad L(\lambda, S_\alpha) \geq L(\lambda, S_0) + \alpha \left\{ \sum_{i=1}^N \mu_{\alpha i} (f(x_\alpha, z_i) - f(x_\alpha, y_i)) \right\}.$$

Hence, for α sufficiently small, $L(\lambda, S_\alpha)$ is strictly larger than $L(\lambda, S_0)$ unless the y_i are equal to the z_i . Now let us state a more precise convergence result.

THEOREM 5.4. *We make the following assumptions:*

- I. X is R^n , Y is a convex, compact subset of a vector space, and $f(x, y)$ is a concave function of y for each fixed x . There exists x^* in X , a finite set $Y^* = \{y_1^*, \dots, y_N^*\}$ contained in Y , and a multiplier λ^* with support equal to Y^* such that

$$L(\lambda^*, Y^*) = \max_{y \in Y^*} f(x^*, y) = \max_{y \in Y} f(x^*, y).$$

II. The Hessian $\nabla_x^2 f(x, y)$ exists and depends continuously on x near x^* and on y near y_i^* for each i between 1 and N . Moreover, the mathematical program

minimize ρ

subject to $x \in R^n$, $f(x, y_i^*) - \rho = 0$ for $i = 1, \dots, N$

satisfies the assumptions of Theorem 3.1 at the optimum $x = x^*$ and $\rho = f(x, y_i^*)$ and if N is the neighborhood of x^* introduced in Theorem 3.1, then we have

$$L(\lambda, S) = \inf \{L(\lambda, S, x) : x \in N\}$$

for λ and S in some neighborhood W of (λ^*, Y^*) .

III. For $x \in N$ there exist local maxima $y_i(x)$ of $f(x, \cdot)$ on Y such that $y_i(x)$ approaches y_i^* as x approaches x^* for $i = 1, \dots, N$. Furthermore, $y_i(x)$ is the locally unique maximizer of $f(x, \cdot)$ for x near x^* and for x near x^* , we have

$$\max_{1 \leq i \leq N} f(x, y_i(x)) = \max_{y \in Y} f(x, y).$$

IV. $L(\lambda, S) < L(\lambda^*, Y^*)$ whenever $(\lambda, S) \in W$, $\lambda \neq \lambda^*$, and $S \neq Y^*$.

Under assumptions I-IV and for r large enough, the peak chasing algorithm converges to λ^* and Y^* starting from any point sufficiently close to λ^* and Y^* .

Proof. We just sketch the proof. For λ near λ^* and for S near Y^* , assumption II implies that there exists $x(\lambda, S)$ which minimizes $L(\lambda, S, x)$ over x in a neighborhood of x^* and $x(\lambda, S)$ depends continuously on λ and S . By Theorem 5.2, step (a) of the peak chasing algorithm is an ascent step for r sufficiently large. Since step (b) of the peak chasing algorithm does not decrease the value of the dual functional, assumption IV implies that if the iterations start near λ^* and Y^* , then the iterations remain near λ^* and Y^* . Since λ_k , Y_k , and x_k lie in compact sets, we can extract subsequences converging to limits λ_∞ , Y_∞ , and x_∞ , respectively. For convenience, these subsequences are also denoted λ_k , Y_k , and x_k . Since x_k minimizes $L(\lambda_k, Y_k, x)$ over $x \in X$, we conclude that x_∞ minimizes $L(\lambda_\infty, Y_\infty, x)$ over $x \in X$. Since $L(\lambda_k, Y_k)$ is bound above by $L(\lambda^*, Y^*)$, the difference $L(\lambda_{k+1}, Y_{k+1}) - L(\lambda_k, Y_k)$ approaches zero as k increases. Hence, Theorem 5.2 implies that $f(x_\infty, y_{\infty i}) - \rho_\infty$ is zero for each i . Also, it follows from (5.19) that the elements of Y_∞ are local maxima of $f(x_\infty, \cdot)$ on Y . Combining these relations, we conclude that

$$(5.20) \quad L(\lambda_\infty, Y_\infty) = \max_{y \in Y_\infty} f(x_\infty, y) = \max_{y \in Y} f(x_\infty, y).$$

The first equality in (5.20) implies through duality that x_∞ is the solution to the discrete minimax problem

minimize ρ

subject to $x \in R^n$, $f(x, y_{\infty i}) \leq \rho$ for $i = 1, \dots, N$.

And the second equality in (5.20) implies that x_∞ is a solution to the continuous minimax problem

$$\min_{x \in R^n} \max_{y \in Y} f(x, y).$$

By assumption IV, λ^* and Y^* are locally unique maxima of L . Consequently, $\lambda_\infty = \lambda^*$,

$Y_\infty = Y^*$, and $x_\infty = x^*$. It then follows from the ascent property that the original sequence (not just the extracted subsequence) converges to λ^* , Y^* , and x^* . \square

Acknowledgment. We wish to thank the referee for a suggestion that led to a shorter proof for Lemma 3.2 and a more direct treatment of the algorithm to minimize $L(\lambda, x, \rho)$ over ρ .

REFERENCES

- [1] K. J. ARROW AND R. M. SOLOW, *Gradient methods for constrained maxima, with weakened assumptions*, in Studies in Linear and Nonlinear Programming, K. Arrow, L. Hurwicz and H. Uzawa, eds., Stanford University Press, Stanford, CA, 1958.
- [2] D. P. BERTSEKAS, *Constrained Optimization and Lagrange Multiplier Methods*, Academic Press, New York, 1982.
- [3] R. W. CHANEY, *A method of centers algorithm for certain minimax problems*, Math. Programming, 22 (1982), pp. 202–226.
- [4] C. CHARALAMBOUS AND A. R. CONN, *An efficient method to solve the minimax problem directly*, SIAM J. Numer. Anal., 15 (1978), pp. 162–187.
- [5] A. CHARNES, W. W. COOPER AND K. KORTANEK, *Duality in semi-infinite programs and some works of Haar and Carathéodory*, Management Sci., 9 (1963), pp. 209–228.
- [6] ———, *On representations of semi-infinite programs which have no duality gaps*, Management Sci., 12 (1965), pp. 113–121.
- [7] ———, *On the theory of semi-infinite programming and a generalization of the Kuhn–Tucker saddle point theorem for arbitrary convex functions*, Naval Res. Logist. Quart., 16 (1969), pp. 41–51.
- [8] J. A. CHATELON, D. W. HEARN AND T. J. LOWE, *A subgradient algorithm for certain minimax and minisum problems*, Math. Programming, 15 (1978), pp. 130–145.
- [9] ———, *A subgradient algorithm for certain minimax and minisum problems—the constrained case*, this Journal, 20 (1982), pp. 455–469.
- [10] F. H. CLARKE, *Generalized gradients and applications*, Trans. Amer. Math. Soc., 205 (1975), pp. 247–262.
- [11] T. F. COLEMAN, *A note on ‘New Algorithms for constrained minimax optimization,’* Math. Programming, 15 (1978), pp. 239–242.
- [12] V. F. DEM’YANOV, *Algorithms for some minimax problems*, J. Comput. System Sci., 2 (1968), pp. 342–380.
- [13] V. F. DEM’YANOV AND V. N. MALOZEMOV, *Introduction to Minimax*, D. Louvish, transl., John Wiley, New York, 1974.
- [14] S. R. K. DUTTA AND M. VIDYASAGAR, *New algorithms for constrained minimax optimization*, Math. Programming, 13 (1977), pp. 140–155.
- [15] G. H. GOLUB AND C. F. VAN LOAN, *Matrix Computations*, Johns Hopkins University Press, Baltimore, MD, 1983.
- [16] W. W. HAGER, *Approximations to the multiplier method*, SIAM J. Numer. Anal., 22 (1985), pp. 16–46.
- [17] W. W. HAGER AND G. D. IANULESCU, *Dual approximations in optimal control*, this Journal, 22 (1984), pp. 423–465.
- [18] ———, *Dual techniques for constrained optimization*, J. Optim. Theory Appl., to appear.
- [19] W. W. HAGER AND R. ROSTAMIAN, *Optimal coatings, bang-bang controls, and gradient techniques*, in Optimal Control: Applications and Methods, to appear.
- [20] J. HALD AND K. MADSEN, *Combined LP and quasi-Newton methods for minimax optimization*, Math. Programming, 20 (1981), pp. 49–62.
- [21] S. P. HAN, *Variable metric methods for minimizing a class of nondifferentiable functions*, Math. Programming, 20 (1981), pp. 1–13.
- [22] M. R. HESTENES, *Multiplier and gradient methods*, J. Optim. Theory Appl., 4 (1969), pp. 303–320.
- [23] K. C. KIWIEL, *Methods of Descent for Nondifferentiable Optimization*, Lecture Notes in Mathematics, 1133, Springer-Verlag, New York, 1985.
- [24] R. KLESSIG AND E. POLAK, *A method of feasible directions using function approximations, with applications to min max problems*, J. Math. Anal. Appl., 41 (1973), pp. 583–602.
- [25] D. E. KNUTH, *The Art of Computer Programming, Volume III: Sorting and Searching*, Addison-Wesley, Reading, MA, 1973.
- [26] W. MURRAY AND M. L. OVERTON, *A projected Lagrangian algorithm for nonlinear minimax optimization*, SIAM J. Sci. Statist. Comput., 1 (1980), pp. 345–370.

- [27] B. T. POLYAK AND N. V. TRET'YAKOV, *The method of penalty estimates for conditional extremum problems*, Zh. Vychisl. Mat. i Mat. Fiz., 13 (1973), pp. 34–46 (translated in U.S.S.R. Comput. Math. and Math. Phys., 13 (1973), pp. 42–58).
- [28] M. J. D. POWELL, *A method for nonlienar constraints in minimization problems*, in Optimization, R. Fletcher, ed., Academic Press, New York, 1972.
- [29] R. T. ROCKAFELLAR, *Convex Analysis*, Princeton Univ. Press, Princeton, NJ, 1970.
- [30] ———, *The multiplier method of Hestenes and Powell applied to convex programming*, J. Optim. Theory Appl., 12 (1973), pp. 555–562.
- [31] ———, *A dual approach to solving nonlinear programming problems by unconstrained optimization*, Math. Programming, 5 (1973), pp. 354–373.
- [32] N. Z. SHOR, *Minimization Methods for Nondifferentiable Functions*, Springer-Verlag, Berlin, New York, 1985.
- [33] F. A. VALENTINE, *The problem of Lagrange with differential inequalities as added side conditions*, in Contributions to the Calculus of Variations, Univ. of Chicago Press, Chicago, 1937, pp. 407–448.
- [34] A. VARDI, *A new minimax algorithm*, Report no. 84–25, Institute for Computer Applications in Science and Engineering, Hampton, VA, 1984.

ON THE H^∞ -OPTIMAL SENSITIVITY PROBLEM FOR SYSTEMS WITH DELAYS*

CIPRIAN FOIAS†, ALLEN TANNENBAUM‡ AND GEORGE ZAMES§

Abstract. In this paper we extend some of the results of [IEEE Trans. Automat. Control, AC-31 (1986), pp. 763–766] to more general delay systems. In particular, we analyze the effect of the interaction of delays and nonminimum phase zeros on the H^∞ -optimal weighted sensitivity.

Key words. sensitivity minimization, delay, contraction, defect operator, distributed system

AMS(MOS) subject classifications. 93B35, 93C05

Notation and Terminology.

D = open unit disc

\bar{D} = closed unit disc

∂D = unit circle

H = open right half plane

\bar{H} = closed right half plane

$\hat{H} = \bar{H} \cup \{\infty\}$

$H^p(X)$ = the standard Hardy p -space ($1 \leq p \leq \infty$) on X where $X = D$ or H . See Duren [6] or Rudin [19] for details. We will also use some elementary facts about L^p -spaces and Hilbert spaces. Again see [6] or [19] for details.

$H^2(X) \ominus uH^2(X)$ = orthogonal complement of $uH^2(X)$ in $H^2(X)$ where $u \in H^\infty(X)$ is an inner function.

Let S denote an arbitrary Hilbert space with inner product $\langle \cdot, \cdot \rangle$. Then for $x, y \in S$, $x \otimes y$ denotes the operator defined by $(x \otimes y)w := \langle w, y \rangle x$ for $w \in S$.

On the unit circle ∂D we identify \bar{z} and $1/z$ in the usual way.

Finally we use all the standard notation from Hilbert space theory. See, e.g., [6], [19], [24].

Introduction. This paper is the sequel to [9]. We recall that in [9], the authors solved the weighted H^∞ -minimization problem for a plant consisting of a pure delay and arbitrary stable (with stable inverse) real rational proper weighting function. We saw that in contrast to the unweighted problem, which reduces to a simple classical Nevanlinna–Pick interpolation problem for a large class of distributed systems [7], [16], even for the simplest weighting function ($W(s) = 1/(as + 1)$, $a > 0$), the weighted problem reflects the distributed nature of systems with delays.

In this paper, we give a general procedure for computing the optimal weighted sensitivity for an arbitrary real rational stable (with stable inverse) weight, and for plants of the form $e^{-hs}P_0(s)$ where $P_0(s)$ is a proper real rational function with no poles or zeros on the $j\omega$ -axis.

In point of fact, we give a general procedure for solving the following kind of problem: Let $P(s)$ be a plant (perhaps distributed) and suppose that we have a

* Received by the editors December 20, 1985; accepted for publication (in revised form) March 24, 1986.

† Department of Mathematics, Indiana University, Bloomington, Indiana 47405.

‡ Department of Electrical Engineering, McGill University, Montréal, Québec, Canada H3A 2A7 and Department of Mathematics, Ben-Gurion University, Beer Sheva, Israel. Present address, Department of Electrical Engineering, University of Minnesota, Minneapolis, Minnesota 55455.

§ Department of Electrical Engineering, McGill University, Montréal, Québec, Canada H3A 2A7.

factorization $P(s) = P_1(s)P_2(s)$. Then for given weight, we can write down an expression for the H^∞ -optimal sensitivity of $P(s)$ in terms of data determined by $P_1(s)$ and $P_2(s)$. Moreover in this expression (see (3.2) below for a more precise statement) the data given by $P_1(s)$ is *decoupled* from the data given by $P_2(s)$. So for example when $P(s) = e^{-hs}P_0(s)$ as above, we can apply our procedure to $P_1(s) = e^{-hs}$ and $P_2(s) = P_0(s)$, a proper real rational function for which the optimal sensitivity problem is easy to solve.

Our methods are in a certain sense a generalization in the rational weighting case of the one-step extension technique of Adamjan, Arov and Krein [1], [2] and actually give new proofs to certain of their results (see Theorem 3.2, § 3.4 and Theorem 3.9 below for details). Basically what we have solved is an “ n -step” or even an “ ∞ -step” extension problem Theorem 3.2. Thus our techniques even give a new viewpoint to certain problems in Nevanlinna–Pick interpolation theory [14].

As in [9], our methods are heavily influenced by the results of Sarason [20] and Sz. Nagy and Foias [23], [24]. Consequently, we will be working in $H^2(D)$ where D is the unit disc. Moreover, the techniques we use have a strong complex-analytic flavor.

Finally in § 4, we will apply our procedure to the case

$$P(s) = e^{-hs} \left(\frac{s-b}{s+b} \right), \quad W(s) = \frac{1}{as+1},$$

$a, b, h > 0$. This will allow us to understand the coupling and effect of the three fundamental parameters a (the inverse of the bandwidth), b (the nonminimum phase zero), and h (the delay) on the optimal sensitivity. As expected for $b \rightarrow \infty$, our formula approaches that of [9] (see also § 1), and so our method here actually gives an alternative route to some of the results of [9].

1. Preliminaries. In this section we would like to briefly review some of the material from our paper [9], and set up some of the notation connected with the weighted sensitivity H^∞ -minimization problem posed by Zames [26]. We should note that independently David Flamm in his thesis [8] (done while at M.I.T.) has derived some results very similar to the ones that we will describe in this section. Israel Gohberg more recently discussed with the authors an approach to derive (1) below, similar to that of Flamm’s using the Hankel operator.

We begin by recalling the general weighted sensitivity H^∞ -minimization problem for SISO, LTI plants (see [11] for an excellent survey on all of this). We are given a SISO, LTI plant $P(s)$, and a stable (with stable inverse) proper real rational weight $W(s)$. Let $C(s)$ denote an internally stabilizing LTI controller for $P(s)$ in the feedback system of Fig. 1.

Then following [26], we define the *weighted sensitivity*:

$$S_W(s) := W(s)(1 + P(s)C(s))^{-1}.$$

The problem in which we are interested is in determining the existence of and computing

$$\inf \{ \|S_W(s)\|_\infty : C \text{ stabilizing} \}$$

where $\| \cdot \|_\infty$ denotes the H^∞ -norm in the right half plane H .

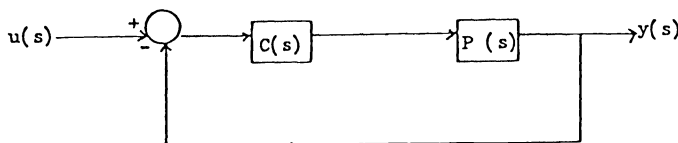


FIG. 1. Standard feedback configuration.

In the finite-dimensional case, this problem is discussed and solved in [13], [12], [15]. In our previous paper [9], we considered the case in which $P(s) = e^{-hs}$, and $W(s)$ a stable strictly proper real rational weighting function with stable inverse. Basically we showed that the problem of the computation of the optimal sensitivity could be reduced to computing the eigenvalues of a certain linear ordinary differential operator with constant coefficients of order $2n$ subject to $2n$ boundary conditions, where n = number of poles of $W(s)$. From the associated Wronskian determinant of the problem, we could then find the required minimal sensitivity (actually all of the singular values of the associated Hankel operator).

To see how this goes, let us briefly sketch the argument from [9]. (See [9] for all the rigorous details.) First of all using the results of [26], one can show that the computation of the optimal sensitivity amounts to finding:

$$\mu := \inf_{q \in H^\infty} \|W(s) - e^{-hs}q(s)\|_\infty.$$

(Throughout this section $H^2 := H^2(H)$, $H^\infty := H^\infty(H)$.) Let $\Pi: H^2 \rightarrow H^2 \ominus e^{-hs}H^2$ denote orthogonal projection. Moreover, we denote by M_W the operator $H^2 \rightarrow H^2$ defined by multiplication by W . Then by [1], [20], [23],

$$\mu = \|\Pi M_W|_{H^2 \ominus e^{-hs}H^2}\|.$$

Computing this norm is not difficult. Indeed we can show via the Fourier or Laplace transform (see [20]) that there exists an isometric isomorphism

$$\phi: H^2 \ominus e^{-hs}H^2 \xrightarrow{\sim} L^2[0, h].$$

Setting

$$\Gamma := \phi \circ (\Pi M_W|_{H^2 \ominus e^{-hs}H^2}) \circ \phi^{-1}$$

we are reduced to computing $\|\Gamma\|$. (Notice $\Gamma: L^2[0, h] \rightarrow L^2[0, h]$.) But again from [20] it follows that we can identify the operator “ $1/s$ ” on $H^2 \ominus e^{-hs}H^2$ with the Volterra operator

$$V: L^2[0, h] \rightarrow L^2[0, h]$$

$Vf(x) := \int_0^x f(t) dt$ via ϕ . The inverse operator (of course unbounded) of V is the derivative operator $Df = f'$ with domain consisting of

$$\{f \in L^2[0, h]: f' \in L^2[0, h], f(0) = 0\}$$

(i.e. the operator D corresponds to “ s ”).

Now to compute $\|\Gamma\|$, we need to compute the largest eigenvalue of $\Gamma^*\Gamma$ (since Γ is compact), or equivalently the smallest positive eigenvalue of $(\Gamma^*\Gamma)^{-1}$. To do this we clearly only need identify the adjoint D^* of D . But it is easy to compute (using integration by parts) that $D^* = -D$ with domain

$$\{f \in L^2[0, h]: f' \in L^2[0, h], f(h) = 0\}.$$

With these remarks one can derive the eigenvalue problem alluded to above [9].

In the particular case in which

$$W(s) = \frac{1}{as+1}, \quad a > 0,$$

we get $\Gamma \equiv (aD+1)^{-1}$ and one derives the eigenvalue problem of finding the largest positive ρ (it is straightforward to check $\rho < 1$) such that

$$(-a^2 D^2 + 1)f = \frac{1}{\rho^2} f, \quad f(h) = 0, \quad -af'(0) + f(0) = 0.$$

From the associated Wronskian, one is reduced to finding the largest $\rho \in (0, 1)$, say ρ_1 , that satisfies

$$(1) \quad \left(\sqrt{\frac{1}{\rho^2} - 1} \right) + \tan \left(\frac{h\sqrt{(1/\rho^2) - 1}}{a} \right) = 0.$$

Then ρ_1 is the required norm (and the first singular value of the associated Hankel operator).

Note that if $\rho_2 \in (0, 1)$, $\rho_2 < \rho_1$, is the next largest root of (1), then ρ_2 will be the second singular value of the associated Hankel, and so on. In other words we have an explicit procedure for computing all of the singular values of the associated Hankel from the Wronskian of a certain elementary eigenvalue problem. Moreover we can clearly even write down the Schmidt vectors using this procedure. (See [18] for the relevant definitions.)

In §§ 3 and 4 below, we will offer another procedure for computing the optimal sensitivity applicable to more general delay systems. Our new method only makes use of elementary properties of $H^2(D)$ and $H^\infty(D)$ and reduces the optimal sensitivity problem to an algebraic one. We will generalize (1) in § 4 to the case of a plant with a delay and a nonminimum phase zero.

2. Triangular operators. In this section we collect some standard facts about certain types of lower block triangular operators. Our basic references are [22], [23].

Let H_1, H_2 denote (complex) Hilbert spaces, and set $H := H_1 \oplus H_2$. Let $S: H \rightarrow H$ be a bounded linear operator such that H_2 is S -invariant subspace of H , i.e., $S|_{H_2}: H_2 \rightarrow H_2$. Then clearly we can write

$$S = \begin{bmatrix} S_1 & 0 \\ Y & S_2 \end{bmatrix}$$

where $S_1 := (S^*|_{H_1})^*$, $S_2 := S|_{H_2}$, and $Y: H_1 \rightarrow H_2$ is the *coupling operator*.

Next let $A: H \rightarrow H$ be an arbitrary contraction, i.e., $\|A\| \leq 1$. Then in the usual way [24] we can define the associated *defect operators* and *defect spaces*:

$$D_A := (I - A^*A)^{1/2}, \quad D_{A^*} := (I - AA^*)^{1/2},$$

$$\mathcal{D}_A := \overline{D_A H}, \quad \mathcal{D}_{A^*} := \overline{D_{A^*} H}.$$

We can now state one of the key results of [22].

THEOREM 2.1. *With the above notation, $\|S\| \leq \rho$ if and only if $\|S_i\| \leq \rho$ ($i = 1, 2$) and $Y = D_{S_2/\rho} L D_{S_1/\rho}$ for some $L: \mathcal{D}_{S_1/\rho} \rightarrow \mathcal{D}_{S_2/\rho}$ such that $\|L\| \leq \rho$. Moreover, if we set $\theta := \max\{\|S_1\|, \|S_2\|\}$, and assume $\rho > \theta$, then $\|S\| = \rho$ if and only if $\|L\| = \rho$.*

Proof. The first statement is Theorem 1 of [22]. The second statement is standard, but since we do not know a convenient reference, we will include the proof. By scaling we can assume $\rho = 1$. Therefore under the hypothesis that $1 > \theta$, we want to show $\|S\| = 1$ if and only if $\|L\| = 1$.

Suppose first $\|S\| = 1$. Then following [22, pp. 205–207], one can define an isometry $\sigma: \mathcal{D}_S \rightarrow \mathcal{D}_L \oplus \mathcal{D}_{S_2}$ such that

$$\sigma D_S h = \begin{bmatrix} D_L D_{S_1} & 0 \\ -S_2^* L D_{S_1} & D_{S_2} \end{bmatrix} \begin{bmatrix} h_1 \\ h_2 \end{bmatrix}.$$

Now $\|S\| = 1$ if and only if there exists a sequence

$$h^{(n)} = \begin{bmatrix} h_1^{(n)} \\ h_2^{(n)} \end{bmatrix},$$

$\|h^{(n)}\| = 1$ such that $D_S h^{(n)} \rightarrow 0$, which in turn is equivalent to

$$(*) \quad \|D_{S_2} h_2^{(n)} - S_2^* L D_{S_1} h_1^{(n)}\|^2 + \|D_L D_{S_1} h_1^{(n)}\|^2 \rightarrow 0.$$

We claim now that

$$\limsup_{n \rightarrow \infty} \|D_{S_1} h_1^{(n)}\| =: q > 0.$$

Indeed, suppose not. Then $\|h_1^{(n)}\| \rightarrow 0$ since D_{S_1} is invertible (S_1 by hypothesis is a strict contraction), and therefore by (*) $\|D_{S_2} h_2^{(n)}\| \rightarrow 0$, and so $\|h_2^{(n)}\| \rightarrow 0$ since D_{S_2} is invertible (S_2 is a strict contraction). But this contradicts our hypothesis that $\|h^{(n)}\| = 1$.

Choose a subsequence $\{h_1^{(n)}\}$ such that $\|D_{S_1} h_1^{(n)}\| > 0$, and $\|D_{S_1} h_1^{(n)}\| \rightarrow q$. Since by (*) $\|D_L D_{S_1} h_1^{(n)}\| \rightarrow 0$, we get that $\|D_L(D_{S_1} h_1^{(n)} / \|D_{S_1} h_1^{(n)}\|)\| \rightarrow 0$ which implies $\|L\| = 1$.

Conversely suppose $\|L\| = 1$ (and $1 > \theta$). By hypothesis $D_{S_1}^{-1}$ exists. Then we can choose a sequence $\{h_1^{(n)}\}$ such that $\|D_{S_1} h_1^{(n)}\| = 1$, and $\|D_L D_{S_1} h_1^{(n)}\| \rightarrow 0$. Set $h_2^{(n)} := D_{S_2}^{-1} S_2^* L D_{S_1} h_1^{(n)}$ (note $D_{S_2}^{-1}$ exists). Then clearly

$$\|D_{S_2} h_2^{(n)} - S_2^* L D_{S_1} h_1^{(n)}\|^2 + \|D_L D_{S_1} h_1^{(n)}\|^2 \rightarrow 0$$

and hence $D_S h^{(n)} \rightarrow 0$ where

$$h^{(n)} = \begin{bmatrix} h_1^{(n)} \\ h_2^{(n)} \end{bmatrix}.$$

To complete the proof therefore we need only show $\|h^{(n)}\| \geq M > 0$ for fixed positive constant M for all n . But clearly

$$\|h^{(n)}\| \geq \|h_1^{(n)}\| \geq \frac{1}{\|D_{S_1}\|}. \quad \square$$

Remark 2.2. For results related to (2.1) see [5] in which arbitrary block 2×2 matrices are considered.

So far we have been considering results about general contractions. In point of fact however, for our purposes the contractions we will need have a special form.

More precisely, let $m_1, m_2 \in H^\infty(D)$ be inner functions. Let $H_i := H^2 \ominus m_i H^2$ $i = 1, 2$ and set $H := H^2 \ominus m_1 m_2 H^2$ (where throughout this section $H^2 := H^2(D)$). We denote by T the *compression* (i.e. projection) of the unilateral shift on H^2 (defined by multiplication by z) to H . (Recall $T := \Pi M_z|_H$ where $M_z: H^2 \rightarrow H^2$ denotes multiplication by z and $\Pi: H^2 \rightarrow H$ is the orthogonal projection.)

Next we have that

$$\begin{aligned} H^2 \ominus m_1 m_2 H^2 &= (H^2 \ominus m_1 H^2) \oplus (m_1 H^2 \ominus m_1 m_2 H^2) \\ &\cong H_1 \oplus H_2. \end{aligned}$$

Note that by abuse of notation, the direct sum symbol in $H_1 \oplus m_1 H_2$ stands for “orthogonal direct sum,” while the direct sum symbol in $H_1 \oplus H_2$ stands for “external direct sum.” (See [19] for the relevant definitions.)

Moreover, we have the following.

LEMMA 2.3. $m_1 H^2 \ominus m_1 m_2 H^2$ is an invariant subspace of $H^2 \ominus m_1 m_2 H^2$ with respect to T .

Proof. Let $v \in m_1 H^2 \ominus m_1 m_2 H^2$. Set $m = m_1 m_2$. Clearly $\bar{m}v \perp H^2$. Let $v_{-1} = \langle \bar{m}v, \bar{z} \rangle$. Then it is easy to compute that

$$Tv = zv - v_{-1}m.$$

But then Tv is divisible by m_1 , i.e., $Tv \in m_1 H^2 \ominus m_1 m_2 H^2$. \square

Lemma 2.3 means that if we identify $H_2 \cong m_1 H^2 \ominus m_1 m_2 H^2$, then we can regard H_2 as an invariant subspace of H with respect to T . Thus with these identifications, we can write

$$T = \begin{bmatrix} T_1 & 0 \\ X & T_2 \end{bmatrix}$$

where the T_i are defined as above. Clearly T_1, T_2, T are contractions.

Now in this case it is well known [24] that $D_{T_1}, D_{T_2^*}$ are of rank 1. Indeed we can compute that

$$I - T_1^* T_1 = \mu_1 \otimes \mu_1, \quad I - T_2 T_2^* = \mu_{2*} \otimes \mu_{2*}$$

where

$$\mu_1 := \bar{z}(m_1(z) - m_1(0)); \quad \mu_{2*} := 1 - m_2(z) \overline{m_2(0)}$$

and where $(x \otimes y)w := \langle w, y \rangle x$.

For such T_1 and T_2 , it is easy to compute X .

PROPOSITION 2.4. $X = \mu_{2*} \otimes \mu_1$.

Proof. Since T is a contraction, by (2.1) we can write $X = D_{T_2^*} L D_{T_1}$ where $L: \mathcal{D}_{T_1} \rightarrow \mathcal{D}_{T_2^*}$ is a contraction between the corresponding defect spaces. Set

$$\hat{\mu}_1 := \frac{\mu_1}{\|\mu_1\|} \quad \text{and} \quad \hat{\mu}_{2*} := \frac{\mu_{2*}}{\|\mu_{2*}\|}$$

(note that $\|\mu_1\|^2 = 1 - |m_1(0)|^2$ and $\|\mu_{2*}\|^2 = 1 - |m_2(0)|^2$). Then $L: \mathcal{D}_{T_1} \rightarrow \mathcal{D}_{T_2^*}$ is such that $L\hat{\mu}_1 = \lambda\hat{\mu}_{2*}$ for some constant λ (since the defect spaces are one-dimensional). Hence, using the facts that $D_{T_1} = \|\mu_1\|(\hat{\mu}_1 \otimes \hat{\mu}_1)$, $D_{T_2^*} = \|\mu_{2*}\|(\hat{\mu}_{2*} \otimes \hat{\mu}_{2*})$, and $X = D_{T_2^*} L D_{T_1}$, we get that $X = \lambda \mu_{2*} \otimes \mu_1$. Note that since T is a contraction $|\lambda| \leq 1$. We have still not used the fact that T is the compression to $H^2 \ominus m_1 m_2 H^2$ of the shift. We do this now.

Indeed we can apply T to μ_1 . It is easy to compute that

$$\begin{aligned} T\mu_1 &= z\mu_1 - \overline{m_2(0)}(1 - |m_1(0)|^2)m_1 m_2 \\ &= z\mu_1 - (1 - |m_1(0)|^2)m_1 \\ &\quad + (1 - |m_1(0)|^2)m_1 + \overline{m_2(0)}(1 - |m_1(0)|^2)m_1 m_2. \end{aligned}$$

Under our isomorphisms,

$$\begin{aligned} H^2 \ominus m_1 m_2 H^2 &= (H^2 \ominus m_1 H^2) \oplus (m_1 H^2 \ominus m_1 m_2 H^2) \\ &\cong (H^2 \ominus m_1 H^2) \oplus (H^2 \ominus m_2 H^2), \end{aligned}$$

we can write

$$T\mu_1 = \begin{bmatrix} z\mu_1 - (1 - |m_1(0)|^2)m_1(z) \\ (1 - |m_1(0)|^2)(1 - \overline{m_2(0)}m_2(z)) \end{bmatrix}.$$

Finally it is easy to compute that

$$\begin{bmatrix} T_1 & 0 \\ \lambda\mu_{2*} \otimes \mu_1 & T_2 \end{bmatrix} \begin{bmatrix} \mu_1 \\ 0 \end{bmatrix} = \begin{bmatrix} z\mu_1 - (1 - |m_1(0)|^2)m_1(z) \\ \lambda(1 - |m_1(0)|^2)(1 - \overline{m_2(0)}m_2(z)) \end{bmatrix}.$$

Thus $\lambda = 1$ as required. \square

3. Weighted sensitivity minimization. In this section we explicitly solve the weighted sensitivity minimization problem for L^∞ -plants of the form $e^{-hs}P_0(s)$, where $P_0(s)$ is a real rational proper function with no poles or zeros on the $j\omega$ -axis. Actually our procedure does much more. Basically for given weight $W(s)$ (with the hypotheses discussed in § 1), we give a technique for solving the weighted H^∞ -minimization problem for a plant $P(s) = P_1(s)P_2(s)$ in terms of data determined independently by $P_1(s)$, and independently by $P_2(s)$. Our method only depends on one knowing the maximum of the optimal sensitivities of $P_1(s)$ and $P_2(s)$, and from this one can find the optimal sensitivity for $P(s)$.

As in [9], for simplicity we initially will take a weight of the form

$$W(s) = \frac{qs + r}{ms + n}$$

stable with stable inverse, and such that $\|W(s)\|_\infty \leq 1$. In § 3.8 we will explain how our method immediately applies to general real rational weights. Moreover we will assume that $P(s)$ is proper and stable with no zeros on the $j\omega$ -axis. Again in Remarks 3.10 we show how to extend our method to unstable plants. The example to keep in mind is $P(s) = e^{-hs}P_0(s)$ where $P_0(s)$ is a stable proper plant with no zeros on the $j\omega$ -axis. However, the technique we give applies much more generally.

Let $\phi: H \rightarrow D$ be a fixed conformal equivalence. Set

$$\hat{W}(z) = W(\phi^{-1}(z)), \quad \hat{P}(z) = P(\phi^{-1}(z)).$$

Let $\hat{P}_i(z)$ be the inner part of $\hat{P}(z)$. Then we assume $\hat{P}_i(z) = m_1(z)m_2(z)$ where the $m_i(z)$ are inner functions. As in § 2, set ($H^2 := H^2(D)$):

$$H := H^2 \ominus m_1 m_2 H^2,$$

$$H_i := H^2 \ominus m_i H^2, \quad i = 1, 2,$$

$$T := \text{compression of the unilateral shift on } H^2(D) \text{ to } H.$$

Then if we make the identifications

$$\begin{aligned} H &= H^2 \ominus m_1 m_2 H^2 \\ &= (H^2 \ominus m_1 H^2) \oplus (m_1 H^2 \ominus m_1 m_2 H^2) \\ &\cong H_1 \oplus H_2 \end{aligned}$$

we can regard H_2 as an invariant subspace of H with respect to T .

When $P(s) = e^{-hs}P_0(s)$ as above, we can take $m_1(z)$ to be the Blaschke product in D whose zeros consist of the images under ϕ of the nonminimum phase zeros of $P_0(s)$, and $m_2(z) = e^{-h\phi^{-1}(z)}$.

Then following the notation of § 1 and the constructions of [26], [13], [12] the problem of computing the optimal sensitivity

$$\inf_{C \text{ stabilizing}} \|W(1+PC)^{-1}\|_\infty$$

can be reduced to computing

$$\mu := \inf_{q \in H^\infty(D)} \|\hat{W}(z) - m_1(z)m_2(z)q(z)\|_\infty.$$

Remark 3.1. We should note that the existence of a $q(z)$, achieving the infimum μ for the given $m(z) = m_1(z)m_2(z)$ inner as above, only depends on the hypothesis that $\hat{W}(z) \in H^\infty(D)$. See [14], [20].

Now as in § 2, we have

$$T = \begin{bmatrix} T_1 & 0 \\ X & T_2 \end{bmatrix}$$

relative to the decomposition $H \cong H_1 \oplus H_2$. If we write

$$\hat{W}(z) = \frac{\alpha z + \beta}{\gamma z + \delta}$$

where $\Delta := \alpha\delta - \beta\gamma \neq 0$, then $\|\hat{W}(z)\|_\infty \leq 1$ since we assumed $\|W(s)\|_\infty \leq 1$. Moreover without loss of generality we may clearly assume $\|\hat{W}(z)\|_\infty = 1$. (Indeed, if necessary, we can always replace $\hat{W}(z)$ by $\hat{W}(z)/\|\hat{W}(z)\|_\infty$.) Thus

$$\|\hat{W}(T)\| \leq 1, \quad \|\hat{W}(T_1)\| \leq 1, \quad \|\hat{W}(T_2)\| \leq 1.$$

Moreover it is easy to compute that

$$\hat{W}(T) = \begin{bmatrix} \hat{W}(T_1) & 0 \\ \Delta(\gamma T_2 + \delta)^{-1}X(\gamma T_1 + \delta)^{-1} & \hat{W}(T_2) \end{bmatrix}.$$

Now it is well known (see [20], [23], [24]) that the infimum

$$\inf_{q \in H^\infty(D)} \|\hat{W}(z) - m_1(z)m_2(z)q(z)\|_\infty = \|\hat{W}(T)\|,$$

and what we will do now is give an explicit procedure for computing the latter norm in terms of data determined separately by the $\hat{W}(T_1)$ and $\hat{W}(T_2)$ parts of $\hat{W}(T)$. In effect we will decouple these in order to compute $\|\hat{W}(T)\|$. First note, however, that

$$\begin{aligned} \|\hat{W}(T)\| &\leq \|\hat{W}(z)\|_\infty = 1, \\ \|\hat{W}(T)\| &\geq \theta := \max \{\|\hat{W}(T_1)\|, \|\hat{W}(T_2)\|\} \end{aligned}$$

and so $\theta = 1$ implies that $\|\hat{W}(T)\| = 1$. Therefore we can clearly assume $\theta < 1$.

Using the defect operator notation of § 2 (as well as the functions μ_1 and μ_{2*}), define for $j = 1, 2$ and $\rho \in (0, 1]$ such that $\rho > \theta$

$$\begin{aligned} \mu_1^{(j)} &= D_{\hat{W}(T_1)/\rho}^{-j}(\bar{\gamma}T_1^* + \bar{\delta})^{-1}\mu_1, \\ \mu_{2*}^{(j)} &= D_{\hat{W}(T_2)^*/\rho}^{-j}(\gamma T_2 + \delta)^{-1}\mu_{2*}. \end{aligned}$$

We can now state (finally!) the following key result.

THEOREM 3.2. *With the above notation $\|\hat{W}(T)\| \leq \rho$ if and only if*

$$(2) \quad \langle (\bar{\gamma}T_2^* + \bar{\delta})^{-1}\mu_{2*}^{(2)}, \mu_{2*} \rangle \cdot \langle (\gamma T_1 + \delta)^{-1}\mu_1^{(2)}, \mu_1 \rangle \leq \rho^2 \Delta^{-2}.$$

Moreover $\|\hat{W}(T)\| = \rho$ if and only if (2) is an equality. (Note we are assuming $\rho \in (0, 1]$ is such that $\rho > \theta$.)

Proof by Theorem 2.1. $\|\hat{W}(T)\| \leq \rho$ if and only if

$$\frac{1}{\rho} \Delta(\gamma T_2 + \delta)^{-1} X(\gamma T_1 + \delta)^{-1} = D_{\hat{W}(T_2)^*/\rho} L_\rho D_{\hat{W}(T_1)/\rho}$$

where

$$L_\rho : \mathcal{D}_{\hat{W}(T_1)/\rho} \rightarrow \mathcal{D}_{\hat{W}(T_2)^*/\rho}$$

defines a contraction of the corresponding defect spaces. But then it is easy to compute that

$$L_\rho = \frac{1}{\rho} \Delta \mu_{2*}^{(1)} \otimes \mu_1^{(1)}.$$

Indeed this follows immediately from the definition of the $\mu_{2*}^{(1)}$ and $\mu_1^{(1)}$ once we show that

$$\mu_{2*}^{(1)} \otimes \mu_1(\gamma T_1 + \delta)^{-1} D_{\hat{W}(T_1)/\rho}^{-1} = \mu_{2*}^{(1)} \otimes D_{\hat{W}(T_1)/\rho}^{-1}(\bar{\gamma} T_1^* + \bar{\delta})^{-1} \mu_1.$$

But to see this just apply the first operator to an element ψ . We get

$$\begin{aligned} \mu_{2*}^{(1)} \otimes \mu_1(\gamma T_1 + \delta)^{-1} D_{\hat{W}(T_1)/\rho}^{-1} \psi &= \langle (\gamma T_1 + \delta)^{-1} D_{\hat{W}(T_1)/\rho}^{-1} \psi, \mu_1 \rangle \mu_{2*}^{(1)} \\ &= \langle \psi, D_{\hat{W}(T_1)/\rho}^{-1}(\bar{\gamma} T_1^* + \bar{\delta})^{-1} \mu_1 \rangle \mu_{2*}^{(1)} \end{aligned}$$

since $D_{\hat{W}(T_1)/\rho}^{-1}$ is self-adjoint.

Therefore

$$\|L_\rho\| \leq 1 \Leftrightarrow \|\mu_{2*}^{(1)}\| \|\mu_1^{(1)}\| \leq \rho \Delta^{-1}$$

\Leftrightarrow the inequality (2) holds.

Finally, under the assumption that $\rho > \theta$, by (2.1) $\|\hat{W}(T)\| = \rho$ if and only if $\|L_\rho\| = 1$ if and only if (2) is an equality. \square

Remarks 3.3. (i) In case $m_1(z) = (z - a)/(1 - \bar{a}z)$, $|a| < 1$, (2) is equivalent to certain inequalities derived by Adamjan, Arov and Krein [1], [2] in connection with the one-step extension problem. Hence what we have derived here is an expression for the norm of an “ n -step extension” (in case m_1 is a finite Blaschke product), or even an “ ∞ -step extension” (e.g., when m_1 is an infinite Blaschke product).

(ii) We will assume from now on that $\rho \geq \|\hat{W}(T)\|$, and $\rho > \theta$. Note that in our procedure below, we can compute $\|\hat{W}(T)\|$ explicitly once we know θ . Thus if we can find the optimal sensitivity for plants $P_1(s)$, $P_2(s)$ we can find it for $P(s) = P_1(s)P_2(s)$.

We now come to the crucial question of how to compute the inner products of (2). Again we can give an explicit procedure.

3.4. Computation of inner products. We will start with the computation of

$$\langle (\bar{\gamma} T_2^* + \bar{\delta})^{-1} \mu_{2*}^{(2)}, \mu_{2*} \rangle.$$

Set $\nu_* := (\bar{\gamma} T_2^* + \bar{\delta})^{-1} \mu_{2*}^{(2)}$. Since $\mu_{2*} = 1 - m_2(z) \overline{m_2(0)}$, and since $\nu_* \in H^2 \ominus m_2 H^2$, we have that $\langle \nu_*, \mu_{2*} \rangle = \nu_*(0)$. Thus we must show how to find $\nu_*(0)$. We give a simple algebraic procedure for doing this.

First note that

$$(\gamma T_2 + \delta)^{-1} \mu_{2*} = \left(1 - \frac{1}{\rho^2} \hat{W}(T_2) \hat{W}(T_2)^*\right) \mu_{2*}^{(2)}.$$

Therefore

$$(3) \quad \begin{aligned} \mu_{2*} &= \left[(\gamma T_2 + \delta)(\bar{\gamma} T_2^* + \bar{\delta}) - \frac{1}{\rho^2} (\alpha T_2 + \beta)(\bar{\alpha} T_2^* + \bar{\beta}) \right] \nu_* \\ &= (A + B T_2 + \bar{B} T_2^* + C T_2 T_2^*) \nu_* \end{aligned}$$

where

$$\begin{aligned} A &:= |\delta|^2 - \left(\frac{1}{\rho^2} \right) |\beta|^2, \\ B &:= \left(\gamma \bar{\delta} - \left(\frac{1}{\rho^2} \right) \alpha \bar{\beta} \right), \\ C &:= \left(|\gamma|^2 - \left(\frac{1}{\rho^2} \right) |\alpha|^2 \right). \end{aligned}$$

Now $(1 - T_2 T_2^*) \nu_* = \langle \nu_*, \mu_{2*} \rangle \mu_{2*} = \nu_*(0) \mu_{2*}$. Therefore, from (3) we see that

$$(4) \quad (1 + C \nu_*(0)) \mu_{2*} = (F + B T_2 + \bar{B} T_2^*) \nu_*$$

where $F := A + C$. But $\bar{m}_2 \nu_* \perp H^2$ so we can write $\bar{m}_2 \nu_* = \nu_{-1} \bar{z} + \nu_{-2} \bar{z}^2 + \dots$. Then

$$T_2^* \nu_* = \bar{z}(\nu_* - \nu_*(0)), \quad T_2 \nu_* = z \nu_* - m_2 \nu_{-1}.$$

Consequently, from (4) we see that

$$(5) \quad (1 + C \nu_*(0)) \mu_{2*} + \bar{B} \bar{z} \nu_*(0) + B m_2 \nu_{-1} = (F + B z + \bar{B} \bar{z}) \nu_*.$$

Finally, multiplying both sides of (4) by z and rearranging terms, we derive the following key relationship:

$$(6) \quad (C \mu_{2*} z + \bar{B}) \nu_*(0) + B m_2 z \nu_{-1} = (B z^2 + F z + \bar{B}) \nu_* - \mu_{2*} z.$$

(Note that even though this relationship has been derived on the boundary on D , since all the functions are in $H^2(D)$, they can be analytically continued to D .)

We are almost done! Indeed it is easy to see that the roots z_1, z_2 of $B z^2 + F z + \bar{B}$ are such that $|z_1 z_2| = 1$. If $B = \bar{B}$ is real (which always occurs in cases of interest in engineering) $z_1 z_2 = 1$. We can always assume $|z_1| \leq 1$. We have three cases.

CASE (i). $|F| > 2|B|$. Then $z_1 \in D, z_2 = 1/\bar{z}_1$. Now multiply (5) by \bar{m}_2 to get

$$(7) \quad (C \bar{m}_2 \mu_{2*} z + \bar{B} \bar{m}_2) \nu_*(0) + B z \nu_{-1} = (B z^2 + F z + \bar{B}) \bar{m}_2 \nu_* - \bar{m}_2 \mu_{2*} z.$$

Note that $\bar{m}_2 \mu_{2*}$ and $\bar{m}_2 \nu_*$ can be continued analytically in the complement of the unit disc and are 0 at ∞ . (On the boundary of D we identify \bar{z} and $1/z$.)

Then plugging z_1 into (6) and z_2 into (7) we get

$$(8) \quad (C \mu_{2*}(z_1) z_1 + \bar{B}) \nu_*(0) + B m_2(z_1) z_1 \nu_{-1} = -\mu_{2*}(z_1) z_1,$$

$$(9) \quad (C(\bar{m}_2 \mu_{2*})(z_2) \cdot z_2 + \bar{B} m_2(z_2)) \nu_*(0) + B z_2 \nu_{-1} = -(\bar{m}_2 \mu_{2*})(z_2) \cdot z_2.$$

Using the fact that $z_2 = 1/\bar{z}_1$, one can solve these equations for $\nu_*(0)$ and show

$$(10) \quad \frac{1}{|\nu_*(0)|} = \left| \frac{A - C}{2} + \frac{1}{2} \sqrt{F^2 - 4|B|^2} \frac{1 + |m_2(z_1)|^2}{1 - |m_2(z_1)|^2} \right|.$$

Note that the case in which $B=0$ is a limiting case of Case (i) in which $z_1=0$, $z_2=\infty$. When this occurs one can compute

$$(11) \quad \frac{1}{|\nu_*(0)|} = \left| \frac{A + C|m_2(0)|^2}{1 - |m_2(0)|^2} \right|.$$

Before stating Cases (ii) and (iii) we will need the following lemma.

LEMMA 3.5. *Let z_1 be such that $Bz_1^2 + Fz_1 + \bar{B} = 0$, and such that $|z_1| = 1$. Then $m_2(z)$ admits an analytic extension to a neighborhood of z_1 , and $|m_2(\zeta)| = 1$ for all ζ in an arc neighborhood of z_1 on the unit circle.*

Proof. First we claim $z_1 \notin \sigma(T_2)$ (where $\sigma(T_2)$ denotes the spectrum of the contraction T_2). Indeed if to the contrary $z_1 \in \sigma(T_2)$, then $\hat{W}(z_1) \in \sigma(\hat{W}(T_2))$. But by definition, since $Bz_1^2 + Fz_1 + \bar{B} = 0$, we have that $(1 - (|\hat{W}(z_1)|^2/\rho^2)) = 0$, that is $|\hat{W}(z_1)| = \rho$. But this would imply that $\|\hat{W}(T_2)\| \geq \rho$, which contradicts our assumption in Remark 3.3(ii) that $\|\hat{W}(T_2)\| < \rho$.

But since $z_1 \notin \sigma(T_2)$ we get the required result from [24, Chap. III, Thm. (5.1)]. \square

Remark 3.6. With the notation of (3.5), note that since $m_2(z)$ is analytic in a neighborhood of z_1 , μ_{2*} must be analytic in this neighborhood of z_1 , and ν_* can have at most a pole at z_1 . But since $\nu_* \in H^2(D)$, in point of fact ν_* must be analytic at z_1 as well. Moreover the derivatives of these functions will also be analytic in a neighborhood of z_1 , since the derivative of an analytic function is itself analytic.

We can now state Cases (ii) and (iii) (z_1 and z_2 are the roots of $Bz^2 + Fz + \bar{B}$).

CASE (ii). $|F| < 2|B|$ i.e. $|z_1| = |z_2| = 1$, $z_1 \neq z_2$. In this case plug the z_i $i = 1, 2$ into (6) to get two linear equations (one of which will be (7), and the other (7) with z_2 substituted for z_1) in the two unknowns $\nu_*(0)$, ν_{-1} and solve for $\nu_*(0)$. By (3.5) and (3.6) this is valid since the functions m_2 , μ_{2*} , ν_* are analytic in neighborhoods of z_1 and z_2 .

We can then compute that

$$(12) \quad \frac{1}{|\nu_*(0)|} = \left| \frac{A - C}{2} + \frac{j\sqrt{4|B|^2 - F^2}}{2} \cdot \frac{1 + m_2(z_1)\overline{m_2(z_2)}}{1 - m_2(z_1)\overline{m_2(z_2)}} \right|.$$

(When $m_2(z_1) = m_2(z_2)$, $z_1 \neq z_2$, it is easy to show that $\nu_*(0) = 0$.)

CASE (iii). $|F| = 2|B|$, i.e. $z_1 = z_2$. Then plug z_1 into (6), and z_2 into the derivative of (6). Once more by (3.5) and (3.6) this makes sense, and we can solve the two resulting equations in the two unknowns $\nu_*(0)$, ν_{-1} for $\nu_*(0)$.

Making the computation, we get that

$$(13) \quad \frac{1}{|\nu_*(0)|} = \left| -C + \varepsilon|B| - B \frac{m_2(z_1)}{m_2'(z_1)} \right|$$

where

$$\varepsilon = \begin{cases} 1 & \text{if } F > 0, \\ -1 & \text{if } F < 0, \end{cases}$$

for $m_2'(z_1) \neq 0$. When $m_2'(z_1) = 0$, it is easy to show that $\nu_*(0) = 0$.

In short from (6), using simple linear algebra, we can find $\nu_*(0)$, the value of the first inner product. Notice that Cases (i) and (ii) are generic, while Case (iii) is the nongeneric case in this situation.

Next we come to the computation of the second inner product of (2), namely

$$\langle (\gamma T_1 + \delta)^{-1} \mu_1^{(2)}, \mu_1 \rangle.$$

We will propose two methods for doing this. The first works for any inner function $m_1(z)$, and the second for a finite Blaschke product.

The first method is simply to imitate the procedure that we used previously in evaluating $\langle \nu_*, \mu_{2*} \rangle$. Indeed, set

$$\nu := (\gamma T_1 + \delta)^{-1} \mu_1^{(2)}.$$

Then we want to evaluate $\langle \nu, \mu_1 \rangle$. But

$$\begin{aligned} \langle \nu, \mu_1 \rangle &= \langle \nu, \bar{z}(m_1(z) - m_1(0)) \rangle \\ &= \langle \bar{m}_1 \nu, \bar{z} \rangle. \end{aligned}$$

Since $\bar{m}_1 \nu \perp H^2$, we may write $\bar{m}_1 \nu = \hat{\nu}_{-1} \bar{z} + (\text{higher order terms in } \bar{z})$ and so $\langle \nu, \mu_1 \rangle = \hat{\nu}_{-1}$. Playing the same game as above we end up with the following analogue of equation (6):

$$(14) \quad (C\mu_1 z + Bm_1 z) \hat{\nu}_{-1} + \bar{B}\nu(0) = (Bz^2 + Fz + \bar{B})\nu - \mu_1 z.$$

We again divide the analysis of (14) into the identical Cases (i), (ii) and (iii) depending upon the roots of $Bz^2 + Fz + \bar{B}$ from which we derive analogous formulae for $\hat{\nu}_{-1}$ (the required value of $\langle \nu, \mu_1 \rangle$) to those we found above for $\nu_*(0)$.

We should note that a deeper explanation of the analogy between (6) and (14) can be given via a beautiful result from [21]. In point of fact using this result it is possible to write down (14) immediately from (6) and the analogous formulae to those of (10)–(13) for $\hat{\nu}_{-1}$ just by inspection. However since these formulae may be derived by elementary linear algebra as above, we will leave it to the interested reader to consult [21].

The second method for finding $\hat{\nu}_{-1}$ works when $m_1(z)$ is a finite Blaschke product. In this case, $H^2 \ominus m_1 H^2$ is finite dimensional and it is easy to compute a basis for this space (see e.g. [21], [17]). Therefore the computation of the second inner product of (2) amounts to finite matrix operations once a suitable basis is chosen. For example, if

$$m_1(z) = \prod_{i=1}^n \left(\frac{z - a_i}{1 - \bar{a}_i z} \right)$$

with $a_i \neq a_j$ for $i \neq j$, then the elements

$$v_k := \frac{(1 - |a_k|^2)^{1/2}}{1 - \bar{a}_k z} \prod_{i=1}^{k-1} \left(\frac{z - a_i}{1 - \bar{a}_i z} \right)$$

for $k = 1, \dots, n$ form a unitary basis for $H^2 \ominus m_1 H^2$ relative to which all the relevant linear operators may be given a finite matrix form. For Blaschke products (in the unit disc) which have roots with multiplicities, it is again easy and standard to write down a similar unitary basis (see [20], [10], [17]).

We thus have an explicit procedure for computing the inner products (3.2). We now give an explicit algorithm for the computation of the optimal sensitivity.

3.7. Computation of optimal sensitivity. We will use the notation of (3.4). Note moreover that the computation of $\nu_*(0)$ and $\hat{\nu}_{-1}$ as functions of ρ divide into the identical Cases (i), (ii) and (iii) depending on the roots $Bz^2 + Fz + \bar{B}$.

To make the dependence of ρ explicit, let us set

$$\nu_1(\rho) := \nu_*(0), \quad \nu_2(\rho) := \hat{\nu}_{-1}.$$

Then (3.2) reads

$$(15) \quad \nu_1(\rho) \nu_2(\rho) \leq \rho^2 \Delta^{-2}.$$

Let us now recall some of our assumptions:

(a) $\hat{W}(z) = (\alpha z + \beta)/(\gamma z + \delta) \in H^\infty(D)$, $\hat{W}^{-1}(z)$ has no poles in D , the open unit disc. (This follows since we assumed $W(s) \in H^\infty(H)$ with stable inverse.) We should note that our methods immediately go through without the hypothesis that $\hat{W}^{-1}(z)$ has no poles in D , but we retain it since it will be easier to explain our algorithm this way.

(b) $\hat{W}(z)$ is normalized so that $\|\hat{W}(z)\|_\infty = 1$. (Again this can be done without loss of generality, by replacing if necessary $\hat{W}(z)$ by $\hat{W}(z)/\|\hat{W}(z)\|_\infty$. We make this normalization since it will be a bit easier to state our algorithm this way. Of course, one can easily write down a similar algorithm without such a normalization.)

(c) $\rho \in (0, 1]$ is such that $\rho \geq \|\hat{W}(T)\|$, and $\rho > \theta := \max\{\|\hat{W}(T_1)\|, \|\hat{W}(T_2)\|\}$. Note that for the algorithm to work we must know θ .

Here then is our algorithmic procedure for the computation of $\|\hat{W}(T)\|$ and hence the optimal sensitivity. We consider two cases.

(A) $|\beta/\alpha| = 1$. Then the algorithm is as follows:

- (i) We first consider Case (iii) of (3.4), i.e. $|F|^2 = 4|B|^2$. Regarding this as an equation in $\rho \in (0, 1]$, it is easy to see that the unique solution will be $\rho = 1$. (Just consider the locus

$$\{z: \rho^2 - |W(z)|^2 = 0\}$$

and notice that there exists $z_0 \in \partial D$ such that $\hat{W}(z_0) = 0$. See Fig. 2.)

We now check if $\rho = 1$ gives equality for (15) using the Case (iii) formulae of (3.4) ((13) and the analogous formula for $\nu_2(\rho)$). If we do get equality, then by Theorems 2.1 and 3.2, $\|\hat{W}(T)\| = 1$, and the algorithm terminates. If not, i.e. if we get strict inequality, we go to step (ii).

- (ii) If $\rho < 1$, then it is easy to check we are in Case (ii) of (3.4). (See Fig. 2.) Using the formulae we derived for Case (ii) ((12) and the analogous formula for $\nu_2(\rho)$), we check if there exists $\rho \in (0, 1)$ with $\rho > \theta$ which gives equality in (15). If there exists such a solution, say ρ_1 , then by Theorems 2.1 and 3.2 it is unique and $\|\hat{W}(T)\| = \rho_1$, i.e. the algorithm terminates. If not, i.e. if we get strict inequality for all $\rho \in (0, 1)$ with $\rho > \theta$, we go to step (iii).

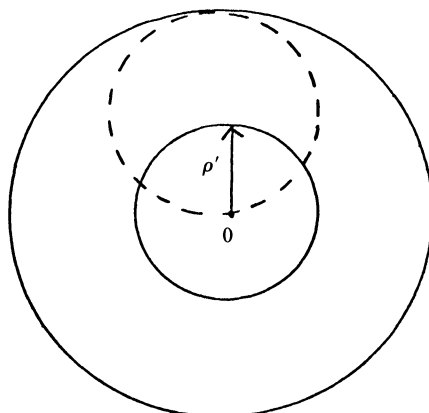


FIG. 2. Representation of the case $|\beta/\alpha| = 1$. Both solid circles are centered at 0, the larger being ∂D , the unit circle. The dashed circle represents the locus $\hat{W}(\partial D)$. Since $\|\hat{W}(z)\|_\infty = 1$, $\hat{W}^{-1}(z)$ has no poles in D , and $|\beta/\alpha| = 1$, $\hat{W}(\partial D)$ passes through the origin 0 and is tangent to ∂D . Note that the circle of radius ρ' intersects $\hat{W}(\partial D)$ in two points, i.e. for any $0 < \rho' < 1$ we are in Case (ii). When $\rho' = 1$, we are in Case (iii).

- (iii) If steps (ii) and (iii) fail to find the norm, then from our hypotheses and Theorems 2.1 and 3.2, we have

$$\|\hat{W}(T)\| = \theta = \max \{\|\hat{W}(T_1)\|, \|\hat{W}(T_2)\|\}$$

and once more we are done.

This completes the analysis of case (A).

- (B) $|\beta/\alpha| \neq 1$. Then the algorithmic procedure for finding $\|\hat{W}(T)\|$ is as follows:

- (i) As in (A), we first consider Case (iii) of (3.4), that is $|F|^2 = 4|B|^2$. Regarding this as an equation in $\rho \in (0, 1]$ and with the above hypotheses (a), (b), (c) one can easily show that we get precisely two solutions, namely $\rho = 1$, and a unique $0 < \rho_0 < 1$. (Again to see this, just consider the locus $\{\rho^2 - |\hat{W}(z)|^2 = 0\}$ and examine the cases $|\beta/\alpha| < 1$, $|\beta/\alpha| > 1$. See Figs. 3 and 4 below.)

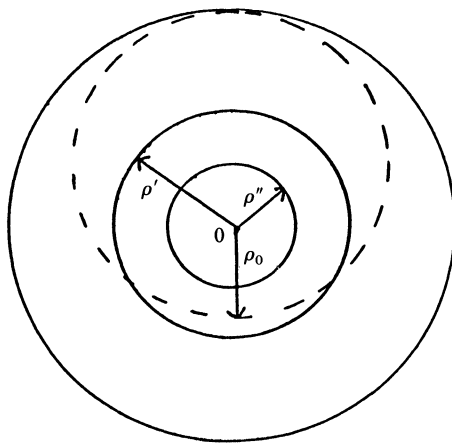


FIG. 3. Representation of the case $|\beta/\alpha| < 1$. All three solid circles are centered at 0, the largest being ∂D , the unit circle. The dashed circle represents the locus $\hat{W}(\partial D)$. Since $\|\hat{W}(z)\|_\infty = 1$, and $\hat{W}^{-1}(z)$ has no poles in D , $\hat{W}(\partial D)$ is tangent to ∂D . ρ_0 is the distance of 0 to the closest point on $\hat{W}(\partial D)$. Note that the circle of radius ρ' intersects $\hat{W}(\partial D)$ in two points, i.e. for $\rho_0 < \rho' < 1$ we are in Case (ii). For $0 < \rho' < \rho_0$ we are in Case (i). For $\rho' = 1$, or $\rho' = \rho_0$, we are in Case (iii).

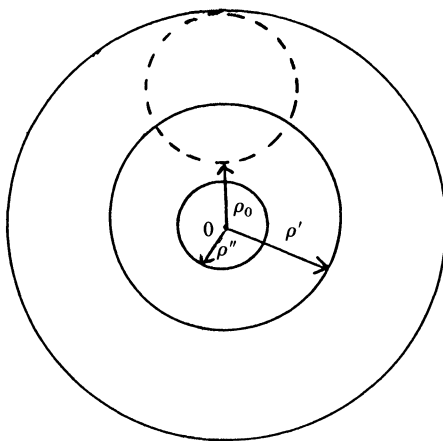


FIG. 4. Representation of the case $|\beta/\alpha| > 1$. Same explanation as for Fig. 3, except here the origin 0 lies to the exterior of $\hat{W}(\partial D)$ (which is represented by the dashed circle).

We now check if $\rho = 1$ gives equality for (15) using the Case (iii) formulae of (3.4) and if it does, then as before $\|\hat{W}(T)\| = 1$. If we get strictly inequality, we consider ρ_0 . If $\rho_0 > \theta$, and if it gives equality when substituted in (15) (using the Case (iii) formulae), then by Theorems 2.1 and 3.2, $\|\hat{W}(T)\| = \rho_0$. If not, we go to step (ii).

- (ii) If from step (i) we have failed to find the required norm, we consider now ρ such that $\rho > \theta$ and $\rho_0 < \rho < 1$. Then it is easy to check that we will be in Case (ii) of (3.4) (see Figs. 3 and 4). If we can find such a ρ , say ρ_2 , which gives equality in (15) (using the Case (ii) formulae of (3.4)), then by Theorems 2.1 and 3.2 ρ_2 will be unique and $\|\hat{W}(T)\| = \rho_2$, and so the algorithm terminates. If all such ρ with $\rho > \theta$ and $\rho_0 < \rho < 1$ give strict inequality, we go to step (iii).
- (iii) We consider ρ such that $\rho > \theta$ and $0 < \rho < \rho_0$. (Of course we need $\rho_0 > \theta$ in this step. If not, just go to step (iv).) Then one can easily check that we will be in Case (i) of (3.4) (see Figs. 3 and 4). If we can find such a ρ , say ρ_3 , which gives equality in (15) (using the Case (i) formulae of (3.4)), then by Theorems 2.1 and 3.2 ρ_3 will be unique and $\|\hat{W}(T)\| = \rho_3$, i.e., we are done. If all such ρ with $\rho > \theta$ and $0 < \rho < \rho_0$ give strict inequality, we go to step (iv).
- (iv) If in all three steps above we have failed to find the norm, then by Theorems 2.1 and 3.2 and the above hypotheses

$$\|\hat{W}(T)\| = \theta = \max \{ \|\hat{W}(T_1)\|, \|\hat{W}(T_2)\| \}$$

and once again the algorithm terminates.

In short, (3.7) gives an easily computable algorithm for finding $\|\hat{W}(T)\|$ once we know θ . Thus we have a technique for computing the H^∞ -optimal sensitivity for distributed systems like $e^{-hs}P_0(s)$, $P_0(s)$ rational stable, since we know the optimal sensitivities for e^{-hs} and $P_0(s)$ already. We now will discuss what occurs for more general weights.

3.8. General weights and one-step extensions. The above analysis was made for linear weights. Still keeping our assumptions on $P(s)$ (i.e. $P(s)$ is stable, proper, with no zeros on the $j\omega$ -axis), we would like to explicitly show how our methods carry over for a general real rational weight $W(s)$, $W(s) \in H^\infty$ with stable inverse, $\|W(s)\|_\infty \leq 1$. Using the above conformal equivalence $\phi: H \rightarrow D$ we set as before

$$\hat{W}(z) := W(\phi^{-1}(z))$$

and we write $\hat{W}(z) = p(z)/q(z)$ a ratio of relatively prime polynomials in z .

Then given as above that

$$T = \begin{bmatrix} T_1 & 0 \\ X & T_2 \end{bmatrix}$$

it is easy to compute that

$$\hat{W}(T) = \begin{bmatrix} \hat{W}(T_1) & 0 \\ q^{-1}(T_2)r(X)q^{-1}(T_1) & \hat{W}(T_2) \end{bmatrix}$$

where $r(X)$ has the form

$$(16) \quad r(X) = \sum_{0 \leq j, k \leq n-1} a_{jk} T_2^j \mu_{2*} \otimes T_1^{*k} \mu_1$$

for some constants a_{jk} , and where $n = \max \{ \text{degree } p(z), \text{degree } q(z) \}$.

Indeed (16) may be derived as follows: Set for $k \in \mathbb{N}$

$$X^{(k)} := \sum_{\substack{0 \leq i, j \leq k-1 \\ i+j=k-1}} T_2^i X T_1^j.$$

Given a polynomial

$$b(z) = \sum_{k=0}^s b_k z^k,$$

set

$$\hat{b}(X) = \sum_{k=1}^s b_k X^{(k)}.$$

Notice in $\hat{b}(X)$ we have dropped the constant term. Then by direct computation, one gets that

$$r(X) = q(T_2) \hat{p}(X) - p(T_2) \hat{q}(X).$$

Equation (16) now follows from the facts that $X = \mu_{2*} \otimes \mu_1$, and that

$$T_2^j X T_1^k = T_2^j \mu_{2*} \otimes T_1^{*k} \mu_1.$$

In short, $r(X)$ is a *finite rank operator*, and is composed of tensor products of the $T_2^j \mu_{2*}$ and $T_1^{*k} \mu_1$ all of which may be explicitly computed. Hence as in the linear weight case, the computation of $\|\hat{W}(T)\|$ may be reduced to an analogous (but of course messier) algebraic problem using the procedures discussed in Theorem 3.2 and (3.4).

In the most important (from a practical point of view) special case, in which $m_1(z)$ is a finite Blaschke product, we can even get simple closed form formulae as we did above. We will do this now.

Indeed first note that when $m_1(z)$ is a finite Blaschke product, we can in point of fact always reduce ourselves to the case in which $m_1(z) = (z - a)/(1 - \bar{a}z)$ for some $a \in D$. To see this let us suppose that

$$m_1(z) = \prod \left(\frac{z - a_i}{1 - \bar{a}_i z} \right).$$

Suppose moreover that we give a procedure for solving the optimal sensitivity problem for $((z - a_1)/(1 - \bar{a}_1 z))m_2(z)$ in terms of (decoupled) data determined by $m_2(z)$ and $(z - a_1)/(1 - \bar{a}_1 z)$ as we did in Theorem 3.2. Then we can take $\hat{m}_2(z) := ((z - a_1)/(1 - \bar{a}_1 z))m_2(z)$ as our new “ $m_2(z)$ ”, and $(z - a_2)/(1 - \bar{a}_2 z)$ as our new “ $m_1(z)$,” and solve the resulting problem for $((z - a_2)/(1 - \bar{a}_2 z))\hat{m}_2(z)$ in terms of $\hat{m}_2(z)$ and $(z - a_2)/(1 - \bar{a}_2 z)$, and so on. In other words, when $m_1(z)$ is a finite Blaschke product in order to solve the optimal sensitivity problem, it is enough to describe the solution to the problem when we add the zeros of $m_1(z)$ one at a time. This is, of course, the basic idea behind the classical recursive procedure of Nevanlinna–Pick interpolation [14], and the one-step extension procedure of Adamjan, Arov and Krein [1], [2].

Consequently, we will give an explicit solution now of the kind we gave for a linear weight, for a general real rational weight, $\hat{W}(z) = p(z)/q(z)$, $\|\hat{W}(z)\|_\infty \leq 1$, such that \hat{W}^{-1} has no poles in the unit disc, and an inner function $m(z) = m_1(z)m_2(z)$ where $m_1(z) = (z - a)/(1 - \bar{a}z)$, $a \in D$. Then with this notation, $\mu_1 = (1 - |a|^2)/(1 - \bar{a}z)$, and $T_1 \mu_1 = a \mu_1$.

From (16), we can write that

$$\begin{aligned} r(X) &= \sum_{0 \leq j, k \leq n-1} a_{jk} T_2^j \mu_{2*} \otimes \bar{a}^k \mu_1 \\ &= \sum_{j=0}^{n-1} b_j T_2^j \mu_{2*} \otimes \mu_1 \end{aligned}$$

(where $n = \max \{\text{degree } p(z), \text{degree } q(z)\}$), for some (explicitly computable) constants b_j .

Set

$$\mu_* = \sum_{j=0}^{n-1} b_j T_2^j \mu_{2*}.$$

Then

$$r(X) = \mu_* \otimes \mu_1.$$

Therefore we have

$$\hat{W}(T) = \begin{bmatrix} \hat{W}(T_1) & 0 \\ q(T_2)^{-1}(\mu_* \otimes \mu_1)q(T_1)^{-1} & \hat{W}(T_2) \end{bmatrix}.$$

We can now play precisely the same game that we did in the linear weight case. Once more without loss of generality we can assume that

$$\theta := \max \{\|\hat{W}(T_1)\|, \|\hat{W}(T_2)\|\} < 1.$$

Let $\rho \in (0, 1]$ and suppose $\rho > \theta$. Then we set for $j = 1, 2$

$$\begin{aligned} \mu_*^{(j)} &:= D_{\hat{W}(T_2)^*/\rho}^{-j}(q(T_2)^{-1})\mu_*, \\ \mu_1^{(j)} &:= D_{\hat{W}(a)/\rho}^{-j}(\overline{q(a)})^{-1}\mu_1 \end{aligned}$$

(since T_1 is multiplication by a).

Then the analogue of Theorem 3.2 in this case is the next theorem.

THEOREM 3.9. $\|\hat{W}(T)\| \leq \rho$ if and only if

$$(17) \quad \langle \bar{q}(T_2^*)^{-1}\mu_*^{(2)}, \mu_* \rangle \cdot \left\langle \frac{\mu_1^{(2)}}{q(a)}, \mu_1 \right\rangle \leq \rho^2.$$

Moreover $\|\hat{W}(T)\| = \rho$ if and only if equality holds in (17). (We are assuming $\rho > \theta$.)

Proof. As in Theorem 3.2, $\|\hat{W}(T)\| \leq \rho$ if and only if

$$\frac{1}{\rho} q(T_2)^{-1} r(X) q(T_1)^{-1} = D_{(1/\rho)\hat{W}(T_2)^*} L_\rho D_{(1/\rho)\hat{W}(T_1)}$$

for some contraction L_ρ . But it is easy to compute that

$$L_\rho = \frac{1}{\rho} (\mu_*^{(1)} \otimes \mu_2^{(1)}).$$

Therefore $\|\hat{W}(T)\| \leq 1$ if and only if we have the inequality (17). The second part of the theorem follows immediately from Theorem 2.1. \square

Remarks 3.10. (i) Clearly in this case the second inner product of (17) is trivial to compute. As for the first inner product, it is clear that one can use the same algebraic technique that we discussed in (3.4). Here from the roots of a polynomial of degree $2n$ one gets $2n$ linear equations in $2n$ unknowns from which one can solve for the

required value of the inner product, where $n := \max \{p(z), q(z)\}$. Depending upon the multiplicities of the roots and where they lie in relation to D , one can derive a procedure analogous to that of (3.7). We did this by hand for a simple quadratic weight ($W(s) = 1/(as + 1)^2$) and admittedly the computation becomes very messy. However, our procedure can certainly be programmed on computer for the kind of rational weights we have considered above.

(ii) Now we finally come to the case in which $P(s) \in L^\infty$ is not stable (but has no zeros on the $j\omega$ -axis). This poses no problem (at least theoretically). Indeed using the arguments of [13], [27] one may reduce the sensitivity minimization problem to the problem of computing

$$\inf_{q \in H^\infty} \|V - Bq\|_\infty$$

where $B(s) = \text{inner part of } P(s)$, $V \in H^\infty$.

Now from (3.1), with these hypotheses, the minimization problem will have a solution. If we then assume that the outer part of $P(s)$ is rational (of course we always consider rational weights that are in H^∞), V will be rational, and we can apply our techniques to the solution of the minimization problem. More explicitly, if $P(s) = P_1(s)P_2(s)$ and we could compute the minimal sensitivities of $P_1(s)$, $P_2(s)$, then we could use our preceding procedure in order to solve the problem for $P(s)$. This occurs for example when $P(s) = e^{-hs}P_0(s)$, $P_0(s)$ real rational and proper, $P_0 \in L^\infty$ with no zeros on the $j\omega$ -axis.

4. An explicit example. Given the general procedures of § 3, an illustrative non-trivial example is certainly called for. We will take

$$W(s) = \frac{1}{as + 1}, \quad a > 0$$

$$P(s) = e^{-hs} \left(\frac{s - b}{s + b} \right), \quad h, b > 0.$$

The minimum sensitivity in this case will allow us to understand the relationship among the quantities a , b , h .

So, let us plug these parameters into our machine and compute. First we choose $\phi: H \rightarrow D$ to be

$$z = \phi(s) := \frac{s - b}{s + b}.$$

Then

$$\hat{W}(z) = W(\phi^{-1}(z)) = \frac{1 - z}{(ab - 1)z + (ab + 1)},$$

$$e^{h\phi^{-1}(z)} = e^{hb((z+1)/(z-1))}$$

We now use the notation of (3.4). Note that $m_1(z) = z$, $m_2(z) = e^{hb((z+1)/(z-1))}$,

$$A = (ab + 1)^2 - \frac{1}{\rho^2}, \quad B = \bar{B} = ((ab)^2 - 1) + \frac{1}{\rho^2},$$

$$C = (ab - 1)^2 - \frac{1}{\rho^2}, \quad F = 2 \left((ab)^2 + 1 - \frac{1}{\rho^2} \right).$$

Then in this case, the two roots of the quadratic equation $Bz^2 + Fz + \bar{B}$ are

$$z_1 = \frac{-((ab)^2 + 1 - (1/\rho^2)) + 2abj\sqrt{(1/\rho^2) - 1}}{((ab)^2 - 1) + (1/\rho^2)}$$

and $z_2 = \bar{z}_1$. It is clear for our plant $P(s)$ that for $a, b, h > 0$ and finite, that the optimal sensitivity will always be strictly less than 1. Hence we can immediately remove Case (iii) of (3.4) from our considerations. (Note $|\beta/\alpha| = 1$ here. See (3.7) (A) above.)

Therefore since $|z_1| = |z_2| = 1$, we are in Case (ii) of the procedure (3.4) and (3.7). Then solving the corresponding linear equations (or using the formulae of (3.4)) we get

$$\nu_*(0) = \frac{\sin\left(\frac{h\sqrt{(1/\rho^2) - 1}}{a}\right)}{2ab \sin\left(\frac{h\sqrt{(1/\rho^2) - 1}}{a}\right) + 2ab\sqrt{(1/\rho^2) - 1} \cos\left(\frac{h\sqrt{(1/\rho^2) - 1}}{a}\right)}.$$

The second inner product is trivial to compute and turns out to be

$$\hat{\nu}_{-1} := \frac{\rho^2}{(\rho^2(ab+1)^2 - 1)}.$$

Next it is trivial to compute that $\Delta = -2ab$, and therefore from (2) we see

$$\nu_*(0)\hat{\nu}_{-1} \leq \frac{\rho^2}{4a^2b^2}.$$

Hence we get that

$$(18) \quad \frac{\sin\left(\frac{h\sqrt{(1/\rho^2) - 1}}{a}\right)}{\sin\left(\frac{h\sqrt{(1/\rho^2) - 1}}{a}\right) + \sqrt{(1/\rho^2) - 1} \cos\left(\frac{h\sqrt{(1/\rho^2) - 1}}{a}\right)} \leq \frac{\rho^2(ab+1)^2 - 1}{ab}.$$

Using our above notation set

$$\theta := \max \{ \|\hat{W}(T_1)\|, \|\hat{W}(T_2)\| \}$$

where T_1 is the compressed shift corresponding to $m_1(z) = z$, and T_2 is the compressed shift corresponding to $m_2(z) = e^{hb((z+1)/(z-1))}$. It is easy to compute that $\|\hat{W}(T_1)\| = 1/(ab+1)$, and $\|\hat{W}(T_2)\| = \rho_1$, the largest root of (1) (of § 1), $\rho_1 \in (0, 1)$.

Then if we algebraically manipulate (18) and invoke Theorems 2.1 and 3.2, (3.7) (A) we see that we are required to find ρ_{opt} , the unique root contained in $(\theta, 1)$ of the following equation (it is easy to check ρ_{opt} exists for $a, b, h \in (0, \infty)$):

$$(19) \quad \left(1 - \frac{2ab}{\rho^2(ab+1)^2 - 1}\right) \tan\left(\frac{h\sqrt{(1/\rho^2) - 1}}{a}\right) + \sqrt{\frac{1}{\rho^2} - 1} = 0.$$

By our above theory, $\|\hat{W}(T)\| = \rho_{\text{opt}}$. Equation (19) has a number of interesting properties a few of which we discuss here. For example, as $b \rightarrow \infty$, (19) approaches (1) of § 1; this just relates the a and the h . Hence in this sense (19) generalizes (1). As $b \rightarrow 0$, it is simple to check $\rho_{\text{opt}} \rightarrow 1$. In short, (19) gives the exact relationship among the fundamental parameters a, b, h in optimal sensitivity theory.

5. Conclusions. Once again we have seen the utility of the complex and functional-analytical methods of [21], [24] in dealing with systems with delays. In this paper we have solved (or at least given an implementable procedure to solve) the weighted H^∞ -minimization problem for an interesting class of delay systems. From our techniques, we have derived a precise picture of the interaction of a delay, nonminimum

phase zero, and given weight in an H^∞ -optimal sensitivity problem. Our work in a certain sense gives mathematically rigorous justification to results that one would hope to be true from just purely engineering considerations.

Finally, we have generalized some of the one-step extension results of Adamjan, Arov and Krein [1], [2], and perhaps given a new perspective to certain kinds of (generalized) interpolation problems. It should be interesting to try to push through the techniques we have given here for broader classes of distributed systems, for example those considered in [3], or even in [7].

REFERENCES

- [1] V. M. ADAMJAN, D. Z. AROV AND M. G. KREIN, *Analytic properties of Schmidt pairs for a Hankel operator and the generalized Schur-Takagi problem*, Math. USSR-Sb., 15 (1971), pp. 31-73.
- [2] ———, *Infinite block Hankel matrices and related extension problems*, Amer. Math. Soc. Transl., 111 (1978), pp. 133-156.
- [3] F. M. CALLIER AND C. A. DESOER, *An algebra of transfer functions for distributed time-invariant systems*, IEEE Trans. Circuits and Systems, CAS-25 (1978), pp. 651-662.
- [4] D. N. CLARK, *Concrete model theory for a class of operators*, J. Funct. Anal., 14 (1973), pp. 269-280.
- [5] C. DAVIS, W. M. KAHAN AND H. F. WEINBERGER, *Norm-preserving dilations and their applications to optimal error bounds*, SIAM J. Numer. Anal., 19 (1982), pp. 445-469.
- [6] P. L. DUREN, *Theory of H^p -Spaces*, Academic Press, New York, 1970.
- [7] A. FEINTUCH AND A. TANNENBAUM, *Gain optimization for distributed systems*, Systems Control Lett., 6 (1986), pp. 295-301.
- [8] D. FLAMM, M.I.T. thesis proposal, Department of Electrical Engineering, Massachusetts Institute of Technology, Cambridge, MA, 1985.
- [9] C. FOIAS, A. TANNENBAUM AND G. ZAMES, *Weighted sensitivity minimization for delay systems*, IEEE Trans. Automat. Control, AC-31 (1986), pp. 763-766.
- [10] B. A. FRANCIS, *Notes on H^∞ -optimal linear feedback systems*, Lecture Notes, Linköping Univ., 1983.
- [11] B. A. FRANCIS AND J. DOYLE, *Linear control theory with an H^∞ optimality criterion*, Systems Control Group report # 8501, Univ. Toronto, October 1985.
- [12] B. A. FRANCIS, J. W. HELTON AND G. ZAMES, *H^∞ -optimal feedback controllers for linear multivariable systems*, IEEE Trans. Automat. Control, AC-29 (1984), pp. 888-900.
- [13] B. A. FRANCIS AND G. ZAMES, *On H^∞ -optimal sensitivity theory for SISO feedback systems*, IEEE Trans. Automat. Control, AC-29 (1984), pp. 9-16.
- [14] J. B. GARNETT, *Bounded Analytical Functions*, Academic Press, New York, 1981.
- [15] K. GLOVER, *All optimal Hankel-norm approximations of linear multivariable systems and their L^∞ -error bounds*, Internat. J. Control, 39 (1984), pp. 1115-1193.
- [16] P. P. KHARGONEKAR AND K. POOLLA, *Robust stabilization of distributed systems*, Automatica, to appear.
- [17] M. G. KREIN AND A. A. NUDELMAN, *The Markov Moment Problem and External Problems*, Translations of Mathematical Monographs, 50, American Mathematical Society, Providence, RI, 1977.
- [18] S. C. POWER, *Hankel Operators on Hilbert Space*, Pitman Advanced Publishing Program, Boston, 1982.
- [19] W. RUDIN, *Functional Analysis*, McGraw-Hill, New York, 1973.
- [20] D. SARASON, *Generalized interpolation in H^∞* , Trans. Amer. Math. Soc. 127 (1967), pp. 179-203.
- [21] B. SZ. NAGY, *Unitary dilations of Hilbert space operators and related topics*, CBMS, Regional Conference Series in Mathematics 19, American Mathematical Society, Chap. 9, Par. 2, 1974.
- [22] B. SZ. NAGY AND C. FOIAS, *Forme triangulaire d'une contraction et factorisation de la fonction caractéristique*, Acta Sci. Math., 28 (1967), pp. 201-212.
- [23] ———, *Dilation des commutants*, C.R. Acad. Sci. Paris Sér. I Math., 266 (1968), pp. 493-495.
- [24] ———, *Harmonic Analysis of Operators on Hilbert Space*, North-Holland, Amsterdam, 1970.
- [25] A. TANNENBAUM, *Invariance and System Theory: Algebraic and Geometric Aspects*, Lecture Notes in Mathematics 845, Springer-Verlag, Berlin, New York, 1981.
- [26] G. ZAMES, *Feedback and optimal sensitivity: model reference transformations, seminorms, and approximate inverses*, IEEE Trans. Automat. Control, AC-23 (1981), pp. 301-320.
- [27] G. ZAMES AND B. FRANCIS, *Feedback, minimax sensitivity, and optimal robustness*, IEEE Trans. Automat. Control, AC-28 (1981), pp. 585-601.

SYNTHESIS OF OPTIMAL CONTROL FOR AN INFINITE DIMENSIONAL PERIODIC PROBLEM*

G. DA PRATO†

Abstract. We prove an existence and uniqueness result on periodic solutions of an infinite dimensional Riccati equation.

Key words. optimal control, periodic control, dynamic programming

AMS(MOS) subject classifications. 93C, 49B

1. Introduction. Consider the following optimal control problem: minimize

$$(1.1) \quad J(u) = \frac{1}{2} \int_0^\tau [\langle M(t)y(t), y(t) \rangle + \langle N(t)u(t), u(t) \rangle] dt$$

over all $u \in L^2(0, \tau; U)$ subject to

$$(1.2) \quad y'(t) = A(t)y(t) + B(t)u(t) + f(t), \quad y(0) = y(\tau).$$

Here $A(t)$ is a linear operator in a Hilbert space H , U is the Hilbert space of the controls, $M(t)$ is a linear operator in H , $N(t)$ is a linear operator in U , $B(t)$ is a linear operator from U into H and $f \in L^2(0, \tau, H)$. We give precise notations and assumptions in § 2. In § 3 we study existence and uniqueness of periodic solutions of the infinite dimensional Riccati equation

$$(1.3) \quad Q' + A^*Q + QA - QBB^*Q + M = 0$$

and in § 4 we prove that the optimal control for problem (1.1), (1.2) is a feedback control. We shall use an argument of dynamic programming, which follows closely [2] where a similar problem was studied in a finite dimensional space.

2. Notation and hypotheses. Let U and H be Hilbert spaces (scalar product $\langle \cdot, \cdot \rangle$). We shall denote by $L(H)$ the Banach algebra of all linear bounded operators in H . We set

$$(2.1) \quad \Sigma(H) = \{T \in L(H); T = T^*\}, \quad \Sigma^+(H) = \{T \in \Sigma(H); T \geq 0\}$$

where T^* represents the adjoint of T .

Given any interval $[a, b]$ we shall denote by $C_s([a, b]; L(H))$ the set of all the mappings $[a, b] \rightarrow L(H)$, $t \rightarrow T(t)$ such that $T(\cdot)x$ is continuous for any $x \in H$. If a and b are finite, then $C_s([a, b]; L(H))$, endowed with the norm

$$(2.2) \quad \|T\| = \sup \{\|T(t)\|; t \in [a, b]\},$$

is a Banach space (by the uniform boundedness theorem). We set moreover

$$(2.3) \quad C_s([a, b]; \Sigma(H)) = \{T \in C_s([a, b]; L(H)); T(t) \in \Sigma(H)\},$$

$$(2.4) \quad C_s([a, b]; \Sigma^+(H)) = \{T \in C_s([a, b]; L(H)); T(t) \in \Sigma^+(H)\}.$$

* Received by the editors November 12, 1984; accepted for publication (in revised form) April 14, 1986.

† Scuola Normale Superiore, 56100 Pisa, Italy.

$C_s([a, b]; L(U))$ and $C_s([a, b]; L(U, H))$ are defined analogously. Concerning the operators $A(t)$, $t \in \mathbb{R}$, we shall assume:

- (2.5) (i) $A(t) = A(t + \tau)$, $t \in \mathbb{R}$.
 (ii) There exists an evolution operator $U(t, s)$, $0 \leq s \leq t$ such that the initial value problem

$$z'(t) = A(t)z(t) + g(t), \quad z(0) = x$$

with $g \in L^2(0, \tau; H)$ and $x \in H$ has a unique mild solution z given by

$$z(t) = U(t, 0)x + \int_0^t U(t, s)g(s) ds.$$

- (iii) $A_n(t) = n^2(n - A(t))^{-1} - nI$ is defined for n sufficiently large. Moreover we have $z_n \rightarrow z$ in $C([0, \tau]; H)$, where z_n is the strict solution of the approximating problem

$$z'_n(t) = A_n(t)z_n(t) + g(t), \quad z_n(0) = x.$$

We shall denote by $U_n(t, s)$ the evolution operator relative to $A_n(t)$. We remark that (2.5) are fulfilled under the usual hypotheses of Tanabe and Kato-Tanabe (see for instance [3], [6], [8]).

Concerning M , N , B and f we shall assume:

- (2.6) (i) $f: \mathbb{R} \rightarrow H$ is τ -periodic and $f \in L^2(0, \tau; H)$,
 (ii) $B \in C_s(\mathbb{R}, L(U, H))$ and it is τ -periodic,
 (iii) $M \in C_s(\mathbb{R}; \Sigma^+(H))$ and it is τ -periodic,
 (iv) $N \in C_s(\mathbb{R}, \Sigma^+(U))$, it is τ -periodic and there exists $\varepsilon > 0$ such that $N(t) \geq \varepsilon I$, $t \leq 0$.

Finally, in order to solve uniquely problem (1.2), we need the following assumption:

- (2.7) 1 belongs to the resolvent set $\rho(U(\tau, 0))$ of $U(\tau, 0)$.

Under hypotheses (2.5)–(2.7) it is easy to prove that problem (1.2) has a unique mild solution y given by

$$(2.8) \quad y(t) = U(t, 0)(I - U(\tau, 0))^{-1} \int_0^\tau U(\tau, s)(f(s) + B(s)u(s)) ds \\ + \int_0^t U(t, s)(f(s) + B(s)u(s)) ds.$$

Returning now to the control problem (1.1), (1.2), we remark that the functional $J: L^2(0, \tau; U) \rightarrow \mathbb{R}$ has a unique minimum u^* (since it is a coercive quadratic form); u^* is called the *optimal control* and the corresponding solution of (1.2) the *optimal state*. Finally $J(u^*)$ is the *optimal cost*.

The optimality conditions are also easily derived. Namely if u is the optimal control and y the optimal state, we have:

$$(2.9) \quad \begin{aligned} y' &= Ay + Bu + f, & y(0) &= y(\tau), \\ p' &= -A^*p - My, & p(0) &= p(\tau), \\ u &= -N^{-1}B^*p. \end{aligned}$$

Concerning the synthesis problem we shall look for a linear operator Q such that

$$(2.10) \quad p = Qy + r.$$

As easily seen, Q and r must satisfy the equations

$$(2.11) \quad Q' + A^*Q + QA - QBN^{-1}B^*Q + M = 0,$$

$$(2.12) \quad r' + (A^* - QBN^{-1}B^*)r + Qf = 0$$

with the periodic conditions

$$(2.13) \quad Q(0) = Q(\tau), \quad r(0) = r(\tau).$$

The differential equations in (2.9), (2.12) are intended in the mild sense, whereas the precise meaning of a solution of (2.11) will be stated in the next section.

In § 4 we will prove that the optimal control u is given by the formula

$$(2.14) \quad u = -N^{-1}B^*(Qy + r)$$

where y (the optimal state) is the solution of the closed loop equation

$$(2.15) \quad y' = Ay - BN^{-1}B^*Qy - BN^{-1}B^*r + f$$

with the condition

$$(2.16) \quad y(0) = y(\tau).$$

We remark that if the following hypothesis holds:

$$(2.17) \quad 1 \text{ belongs to the resolvent sets of the evolution operators relative to } A - BN^{-1}B^*Q \text{ and } A^* - QBN^{-1}B^*,$$

then (2.12) and (2.15) have a unique τ -periodic solution.

3. Periodic solutions of the Riccati equation. We are here concerned with periodic solutions of the Riccati equation

$$(3.1) \quad Q' + A^*Q + QA - QBN^{-1}B^*Q + M = 0.$$

We first recall some result on the final value problem

$$(3.2) \quad Q' + A^*Q + QA - QBN^{-1}B^*Q + M = 0, \quad Q(\tau) = L \in \Sigma^+(H),$$

which we write in the following integral form:

$$(3.3) \quad \begin{aligned} Q(t)x &= U^*(\tau, t)LU(\tau, t)x \\ &- \int_t^\tau U^*(s, t)(Q(s)B(s)N^{-1}(s)B^*(s)Q(s) - M(s))U(s, t)x \, ds, \quad x \in H. \end{aligned}$$

Under suitable hypotheses (see Proposition 3.1 below) (3.3) has a unique solution $Q(t) = \Lambda(t, L)$.

We say that $Q \in C_s([0, \tau]; \Sigma^+(H))$ is a τ -periodic solution of (3.1) if it is a solution of (3.3) with $Q(\tau) = Q(0)$; this is equivalent to

$$(3.4) \quad Q(\tau) = \Lambda(0, Q(\tau)).$$

We shall consider also the approximating problem

$$(3.5) \quad Q'_n + A_n^*Q_n + Q_nA_n - Q_nBN^{-1}B^*Q_n + M = 0, \quad Q_n(\tau) = L$$

where $A_n(t) = n^2(n - A(t))^{-1} - nI$. Problem (3.5) has clearly a unique solution that we denote by $Q_n(t) = \Lambda_n(t, L)$.

PROPOSITION 3.1. Assume (2.5), (2.6) and let L belong to $\Sigma^+(H)$. Then

$$(3.6) \quad (i) \text{ There exists a unique solution } Q \text{ (resp. } Q_n \text{) of (3.3) (resp. (3.5)). Moreover } Q_n \rightarrow Q \text{ in } C_s([0, \tau]; \Sigma^+(H)).$$

(ii) If $L \leq \bar{L}$ we have:

$$\Lambda(t, L) \leq \Lambda(t; \bar{L}).$$

(iii) If $\{L_k\}$ is an increasing sequence in $\Sigma^+(H)$ that converges strongly to L , then $\Lambda(\cdot, L_k)$ converges to $\Lambda(\cdot, \bar{L})$ in $C_s([0, \tau]; \Sigma^+(H))$.

Proof. Statement (i) is essentially proved in [4] (see also [1, Thm. 1, p. 64]). The proof of (ii) is completely similar to that of [1, Lemma 16, p. 83]. Let us prove (iii). Setting $Q(t) = \Lambda(t, L)$, $Q_k(t) = \Lambda(t, L_k)$ we have

$$(3.7) \quad \begin{aligned} Q_k(t)x &= U^*(\tau, t)L_k U(\tau, t)x \\ &\quad - \int_t^\tau U^*(s, t)(Q_k(s)B(s)N^{-1}(s)B^*(s)Q_k(s) - M(s))U(s, t)x ds, \end{aligned} \quad x \in H.$$

By (3.6), $\{Q_k(t)\}$ is increasing for any t and $Q_k(t) \leq Q(t)$. It follows that there exists $\bar{Q}(t) \leq Q(t)$ such that $Q_k(t) \rightarrow \bar{Q}(t)x$ for any $x \in H$. By the dominated convergence theorem, taking the limit, as $k \rightarrow \infty$ in (3.7), we obtain

$$(3.8) \quad \begin{aligned} \bar{Q}(r)x &= U^*(\tau, t)L U(\tau, t)x \\ &\quad - \int_t^\tau U^*(s, t)(\bar{Q}(s)B(s)N^{-1}(s)B^*(s)\bar{Q}(s) - M(s))U(s, t)x ds, \end{aligned} \quad x \in H.$$

From (3.8) it follows that $\bar{Q} \in C_s([0, \tau]; \Sigma^+(H))$ so that, by uniqueness, we have $\bar{Q} = Q$. \square

In order to prove the existence of a periodic solution of (3.1), we need a stabilizability assumption:

(3.9) There exists a τ -periodic function $K \in C_s(\mathbb{R}; L(H, U))$ and two numbers, $\omega > 0$, $\mu > 0$ such that $\|U_{A-BK}(t, s)\| \leq \mu e^{-\omega(t-s)}$, $t > s$, where U_{A-BK} is the evolution operator relative to $A(t) - B(t)K(t)$, $t \in [0, \tau]$.

This hypothesis reduces to the usual one for the algebraic Riccati equation when A , B and M are time-independent (see [7]).

Remark 3.2. Hypothesis (3.9) is fulfilled if either $\|U(t, s)\| \leq a e^{-b(t-s)}$ with $b > 0$ or $B(t) \geq \sigma > 0$ and $a = 1$. \square

We are ready now to prove the following theorem:

THEOREM 3.3. Assume (2.5), (2.6) and (3.9). Then there exists a τ -periodic solution of (3.1).

Proof. We first recall a well-known identity (see for instance [1]). Let $u \in L^2(0, T; U)$, $T > 0$ and let y be the mild solution of the problem

$$(3.10) \quad y' = Ay + Bu, \quad y(0) = x, \quad x \in H.$$

Let W be the solution of the final value problem

$$(3.11) \quad W' + A^*W + WA - WB N^{-1} B^* W + M = 0, \quad W(T) = 0;$$

then we have:

$$(3.12) \quad \langle W(0)x, x \rangle + \int_0^T \|N^{-1/2} B^* W y + N^{1/2} u\|^2 ds = \int_0^T [\langle M y, y \rangle + \langle N u, u \rangle] ds.$$

We prove now the existence of a τ -periodic solution of (3.1). Set

$$(3.13) \quad S_0 = 0, \quad S_{n+1}(t) = \Lambda(t, S_n(0)), \quad n \in \mathbb{N}.$$

By (3.6) $\{S_n\}$ is increasing. For any $k \in \mathbb{N}$ we set

$$(3.14) \quad \begin{aligned} W_k(t) &= S_h(t - (k-h-1)\tau), \\ t &\in [(k-h-1)\tau, (k-h)\tau], \quad h = 1, \dots, k. \end{aligned}$$

As easily checked, W_k is a solution of the problem

$$(3.15) \quad \begin{aligned} W'_k + A^* W_k + W_k A - W_k B N^{-1} B^* W_k + M &= 0, \\ W_k(k\tau) &= 0, \quad 0 \leq t \leq k\tau. \end{aligned}$$

We now resort to (3.9) and (3.12) with u and y given by

$$(3.16) \quad u(t) = -K(t)U_{A-BK}(t, 0)x, \quad y(t) = U_{A-BK}(t, 0)x, \quad x \in H$$

and we get

$$(3.17) \quad \langle W_k(0)x, x \rangle \leq \frac{\mu^2}{2\eta} (\|M\| + \|N\| \|K\|) \|x\|^2,$$

which implies that the sequence $\{S_n(0)\}$ is bounded in $\Sigma^+(H)$. By a well-known result on the monotone sequences of linear operators it follows that there exists $\bar{S} \in \Sigma^+(H)$ such that $S_n(0)x \rightarrow \bar{S}x$ for any $x \in H$. Now, by Proposition 3.1(iii) and by (3.13) we have, as $n \rightarrow \infty$ $\bar{S} = \Lambda(0, \bar{S})$ so that $\Lambda(t, \bar{S})$ is the required periodic solution. \square

Remark 3.4. Theorem 3.3 generalizes a result in [9].

We consider now uniqueness and to this purpose we introduce a detectability assumption which reduces to the usual one for the algebraic Riccati equation (see [7]). We assume:

$$(3.18) \quad \text{There exists a } \tau\text{-periodic function } K_1 \in C_s(\mathbb{R}, L(H)) \text{ and two numbers } \omega_1 > 0, \mu_1 > 0 \text{ such that}$$

$$\|U_{A-K_1\sqrt{M}}(t, s)\| \leq \mu_1 e^{-\omega_1(t-s)}, \quad t \geq s$$

where $U_{A-K_1\sqrt{M}}$ is the evolution operator relative to

$$A(t) - K_1(t)\sqrt{M(t)}, \quad t \in [0, \tau].$$

We remark that (3.18) implies (2.17).

We first prove two lemmas as follows.

LEMMA 3.5. Assume (2.5), (2.6) and (3.18) and set $L = A - BN^{-1}B^*Q$ where Q is a τ -periodic solution of (3.1). Then there exists $c > 0$ such that

$$(3.19) \quad \int_s^\infty \|U_L(t, s)x\|^2 dt \leq c \|x\|^2.$$

Proof. Let Q be a τ -periodic solution of (3.1), fix $k \in \mathbb{N}$. Then we have

$$(3.20) \quad Q' + L^*Q + QL + QBN^{-1}B^*Q + M = 0, \quad Q(k\tau) = Q(0), \quad t \in [0, k\tau].$$

Let Q_n be the solution of the approximating problem

$$(3.21) \quad Q'_n + L_n^*Q_n + Q_nL_n + Q_nBN^{-1}B^*Q_n + M = 0, \quad Q_n(k\tau) = Q(0)$$

where $L_n = A_n - BN^{-1}B^*Q_n$. We remark that Q_n is not necessarily periodic. For any $x \in H$ we have

$$(3.22) \quad \begin{aligned} \frac{d}{dt} &< Q_n(t)U_{L_n}(t, s)x, U_{L_n}(t, s)x \rangle \\ &= -\|N^{1/2}BQ_nU_{L_n}(t, s)x\|^2 - \|\sqrt{M}U_{L_n}(t, s)x\|^2. \end{aligned}$$

By integrating in $[s, t]$ and letting n go to infinity we find

$$(3.23) \quad \begin{aligned} \langle Q(s)x, x \rangle &= \langle Q(k\tau)U_L(k\tau, s)x, U_L(k\tau, s)x \rangle \\ &+ \int_0^{k\tau} [\|N^{-1/2}Q(\sigma)U_L(\sigma, s)x\|^2 + \|\sqrt{M(\sigma)}U_L(\sigma, s)x\|^2] d\sigma. \end{aligned}$$

Then functions $N^{-1/2}QU_L(\cdot, s)x$ and $\sqrt{M}U_L(\cdot, s)x$ belong to $L^2(s, \infty; H)$. Let now Π be defined by

$$(3.24) \quad L = \Pi + (K_1\sqrt{M} - BN^{-1}B^*Q)$$

and remark that, by (2.18),

$$(3.25) \quad \|U_\Pi(t, s)\| \leq \mu_1 e^{-\omega_1(t-s)}.$$

By (3.24) it follows

$$(3.26) \quad \begin{aligned} U_L(t, s)x &= U_\Pi(t, s)x \\ &+ \int_s^t U_\Pi(t, \sigma)(K_1(\sigma)\sqrt{M(\sigma)} - B(\sigma)N^{-1}(\sigma)B^*(\sigma)Q(\sigma))U_L(\sigma, s)x d\sigma; \end{aligned}$$

now, by the Young inequality $U_L(\cdot, s)x$ belongs to $L^2(s, \infty; H)$ as required. \square

LEMMA 3.6. *Under the same hypotheses of Lemma 3.5 there exists a constant $c_1 > 0$ such that*

$$(3.27) \quad \|U_L(t, s)\| \leq \frac{c_1}{(t-s)}, \quad t \geq s.$$

Proof. Since L is τ -periodic, there exist $\mu_2 > 0$ and $\xi \in \mathbb{R}$ such that

$$(3.28) \quad \|U_L(t, s)\| \leq \mu_2 e^{\xi(t-s)}, \quad t > s.$$

For any $x \in H$ we have

$$\begin{aligned} \frac{1}{2\xi}(e^{2\xi(t-s)} - 1)\|U_L(t, s)x\| &= \int_s^t e^{2\xi(\sigma-s)}\|U_L(t, s)x\|^2 d\sigma \\ &\leq \int_s^t e^{2\xi(\sigma-s)}\|U_L(\sigma, s)x\|^2\|U_L(t, \sigma)\|^2 d\sigma \\ &\leq c\mu_2^2 e^{2\xi(t-s)}\|x\|^2 \end{aligned}$$

by (3.19); thus there exists $\gamma > 0$ such that

$$(3.29) \quad \|U_L(t, s)\| \leq \gamma, \quad t \geq s.$$

We have finally

$$\begin{aligned} (t-s)\|U_L(t, s)x\|^2 &= \int_s^t \|U_L(t, s)x\|^2 d\sigma \\ &\leq \int_s^t \|U_L(\sigma, s)x\|^2\|U_L(t, \sigma)\|^2 d\sigma \leq \gamma^2 c\|x\|^2 \end{aligned}$$

and the conclusion follows. \square

Remark 3.7. The above proof is inspired by the proof of the Datko theorem given in [7].

We are now ready to prove uniqueness.

THEOREM 3.8. *Assume (2.5), (2.6) and (3.18). Then (3.1) has at most one τ -periodic solution.*

Proof. Let Q, Q_1 be τ -periodic solutions of (3.1); set $R = Q - Q_1$. Then R verifies the equation

$$(3.30) \quad R' + L^*R + RL + RBN^{-1}B^*R = 0, \quad t \in [0, k\tau]$$

for any $k \in \mathbb{N}$. Let R_n be the solution of the final value problem

$$(3.31) \quad R'_n + L_n^*R_n + R_nL_n + R_nBN^{-1}B^*R_n = 0, \quad R_n(k\tau) = R(k\tau).$$

It follows that

$$(3.32) \quad \begin{aligned} & \frac{d}{dt} \langle R_n(t)U_{L_n}(t, s)x, U_{L_n}(t, s)x \rangle \\ &= -\|N^{-1/2}B^*(t)R_n(t)U_{L_n}(t, s)x\|^2 \leq 0, \quad x \in H, \end{aligned}$$

which implies

$$(3.33) \quad \langle R(0)U_L(k\tau, s)x, U_L(k\tau, s)x \rangle \leq \langle R(s)x, x \rangle.$$

Letting k go to infinity and using (3.27), we get $\langle R(s)x, x \rangle \geq 0$, that is, $Q(s) \geq Q_1(s)$; by interchanging Q and Q_1 we find $Q(s) \leq Q_1(s)$ and finally that $Q = Q_1$. \square

Remark 3.9. Stability. Assume the hypotheses of Theorems 3.3 and 3.8; let Q be the unique periodic solution of (3.1) and S a solution of the final value problem

$$S' + A^*S + SA - SBN^{-1}B^*S + M = 0, \quad S(0) = S_0 \in \Sigma^+(H), \quad -\infty < t \leq 0.$$

Setting $Z = Q - S$, $L = A - BN^{-1}B^*Q$, we have

$$Z' + L^*Z + ZL + ZBN^{-1}B^*Z = 0.$$

Thus, by (3.27), it follows that

$$\lim_{\|S_0\| \rightarrow 0} \|Q(t) - S(t)\| = 0 \quad \text{uniformly in } t$$

and the periodic solution Q is stable.

4. Dynamic programming. The Hamilton–Jacobi–Bellman equation corresponding to the control problem (1.1)–(1.2) is

$$(4.1) \quad \begin{aligned} & \psi_t(t, x) - \frac{1}{2} \|N(t)^{-1/2}B^*(t)\psi_x(t, x)\|^2 \\ & + \langle Ax + f(t), \psi_x(t, x) \rangle + \frac{1}{2} \langle M(t)x, x \rangle = 0. \end{aligned}$$

The following result is easily proved.

PROPOSITION 4.1. Assume (2.5)–(2.7), (2.17) and (3.9). Let Q be a τ -periodic solution of (3.1) and r the periodic solution of (2.12). Then the function

$$(4.2) \quad \psi(t, x) = \frac{1}{2} \langle Q(t)x, x \rangle + \langle r(t), x \rangle + s(t)$$

is a solution of (4.1) if and only if we have

$$(4.3) \quad s' - \frac{1}{2} \|N^{-1}B^*r\|^2 + \langle f(t), r(t) \rangle = 0.$$

LEMMA 4.2. Assume the hypotheses of Proposition 4.1. Let ψ be given by (4.2), $u \in L^2(0, \tau; U)$, y be defined by (1.2) and J by (1.1). Then the following identity holds:

$$(4.4) \quad \begin{aligned} J(u) &= \int_0^\tau \|N^{-1/2}B^*(Qy + r) + N^{1/2}u\|^2 dt \\ &+ \int_0^\tau [\langle f, r \rangle - \frac{1}{2} \|B^*r\|^2] dt. \end{aligned}$$

Proof. Let $Q_n(t) = \Lambda_n(t, Q(\tau))$, let r_n be the solution of the problem

$$(4.5) \quad r'_n + (A_n^* - Q_n B N^{-1} B^*) r_n + Q_n f = 0, \quad r_n(\tau) = r(\tau).$$

Let s_n be such that

$$(4.6) \quad s'_n - \frac{1}{2} \|N^{-1/2} B^* r_n\|^2 + \langle f, r_n \rangle = 0$$

and, finally, let y_n be the solution of the problem

$$(4.7) \quad y'_n = A_n y_n + B u + f, \quad y_n(0) = y(0).$$

Setting

$$(4.8) \quad \psi_n(t, y) = \frac{1}{2} \langle Q_n(t) y, y \rangle + \langle r_n(t), y \rangle + s_n(t)$$

we have

$$(4.9) \quad \frac{d}{dt} \psi_n(t, y_n) = \frac{1}{2} \|N^{-1/2} B^* (Q_n y_n + r_n)\|^2 - \frac{1}{2} [\langle M y_n + y_n \rangle + \langle N u, u \rangle].$$

Now the conclusion follows by integrating (4.9) in $[0, \tau]$ and by letting n go to infinity. \square

THEOREM 4.3. Assume (2.5)–(2.7), (2.17) and (3.9). Let Q be a τ -periodic solution of (3.1), let r be the corresponding τ -periodic solution of (2.12) and y the solution of the closed loop equation (2.15) with $y(0) = y(\tau)$. Then the optimal control u^* is given by

$$(4.10) \quad u^* = -N^{-1} B^* (Q y + r)$$

and the optimal cost results from

$$(4.11) \quad J(u^*) = \int_0^\tau \left[\langle f, r \rangle - \frac{1}{2} \|B^* r\|^2 \right] dt.$$

Proof. By (4.4) it follows that

$$(4.12) \quad J(u) \geq \int_0^\tau \left[\langle f, r \rangle - \frac{1}{2} \|B^* r\|^2 \right] dt = \Gamma;$$

now, if u is given by (4.10) we have $J(u^*) = \Gamma$ so that u is optimal. \square

Example 4.4. Let Ω be a bounded subset of R^n with smooth boundary $\partial\Omega$. Consider the following problem:

Minimize

$$(4.13) \quad J(u) = \frac{1}{2} \int_0^\tau dt \int_\Omega d\xi [|y(t, \xi)|^2 + |u(t, \xi)|^2]$$

over all $u \in L^2([0, \tau] \times \Omega)$

subject to

$$\frac{d}{dt} y(t, \xi) = \Delta_\xi y(t, \xi) - \phi(t) y(t, \xi) + u(t, \xi) + f(t, \xi),$$

$$(4.14) \quad y(t, \xi) = 0, \quad t \in [0, \tau], \quad \xi \in \partial\Omega,$$

$$y(0, \xi) = y(t, \xi)$$

where f and ϕ are continuous, τ -periodic in t and ϕ is nonnegative. Δ_ξ is the Laplace operator acting in the variable ξ .

Set $H = U = L^2(\Omega)$, $M(t) = N(t) = B = I$ and

$$(4.15) \quad A(t) = \Delta_\xi - \phi(t), \quad D(A(t)) = H^2(\Omega) \cap H_0^1(\Omega).$$

As easily seen, hypotheses (2.5) and (2.6) hold; moreover

$$(4.16) \quad U(t, s) = \exp \left(C(t-s) - \int_s^t \phi(\sigma) d\sigma \right)$$

where $Cy = \Delta_\xi y$ and $D(C) = D(A(t))$. By the maximum principle we have

$$(4.17) \quad \|U(t, s)\| \leq 1$$

so that $1 \in \rho(U(\tau, 0))$ and (2.7) is fulfilled. Moreover, (2.17) also holds because $A - BN^{-1}B^*Q = A - Q$ and Q is positive. Finally (3.9) holds by virtue of Remark 3.2. \square

REFERENCES

- [1] V. BARBU AND G. DA PRATO, *Hamilton-Jacobi Equations in Hilbert Spaces*, Pitman, London, 1983.
- [2] S. BITTANTI, A. LOCATELLI AND C. MAFFEZZONI, *Periodic optimization under small perturbations*, in Periodic Optimization, Vol. II, A. Marzollo, ed., Udine, Springer-Verlag, Berlin, New York, 1972, pp. 183-231.
- [3] R. F. CURTAIN AND A. J. PRITCHARD, *Infinite Dimensional Linear Systems*, Springer-Verlag, Berlin, New York, 1980.
- [4] G. DA PRATO, *Quelques résultats d'existence, unicité et régularité pour un problème de la théorie du contrôle*, J. Math. Pures Appl., 52 (1973), pp. 353-375.
- [5] R. E. KALMAN, P. L. FALB AND M. A. ARBIB, *Topics in Mathematical System Theory*, McGraw-Hill, New York, 1960.
- [6] A. PAZY, *Semigroups of Linear Operators and Applications to Partial Differential Equations*, Springer-Verlag, Berlin, New York, 1983.
- [7] A. J. PRITCHARD AND J. ZABCZYK, *Stability and stabilizability of infinite dimensional systems*, SIAM Rev., 23 (1981), pp. 25-52.
- [8] H. TANABE, *Equations of Evolution*, Pitman, London, San Francisco, 1979.
- [9] L. TARTAR, *Sur l'étude direct d'équations non linéaires intervenant en théorie du contrôle optimal*, J. Funct. Anal., 17 (1974), pp. 1-47.

CONTROLLABILITY OF SEMILINEAR CONTROL SYSTEMS DOMINATED BY THE LINEAR PART*

KOICHIRO NAITO†

Abstract. While various equivalent conditions for controllability have been obtained in the case of linear control systems, controllability problems of semilinear control systems usually require some complicated and limited assumptions. In this paper we show the approximate controllability of an abstract semilinear control system under the assumption, which has a simple form and can be easily checked in many examples.

Key words. semilinear control system, approximate controllability

AMS(MOS) subject classification. 93B

1. Introduction. In this paper we study the approximate controllability (a.c.) of the following semilinear control system.

$$(1.1) \quad \begin{aligned} \frac{dy}{dt} + Ay(t) &= F(y(t)) + (Bv)(t), & 0 < t < T, \\ y(0) &= 0. \end{aligned}$$

Let the state space X and the control space V be Hilbert spaces and the state function $y(t)$, $0 \leq t \leq T$, takes values in X and the control function $v(\cdot)$ is given in $L^2(0, T; V)$. Assume that the operator $-A$ generates a C_0 -semigroup $S(t)$ on X and B is a bounded linear mapping from $L^2(0, T; V)$ to $L^2(0, T; X)$. $F(\cdot)$ is a nonlinear operator on X . When $F \equiv 0$, system (1.1) is called the corresponding linear system, denoted by $(1.1)_1$.

The controllability theory for abstract linear control systems has been established; the necessary and sufficient conditions for the various types of controllability are well known (cf. [1], [7]). On the other hand, in the case of semilinear control systems, we need some complicated and restrictive conditions, which are of two classes, in the approximate controllability problem with dense reachable sets and in the local controllability problem with limited reachable sets. In either problem the restrictions are placed on the system components, such as range of B , continuity of F , regularity of $S(t)$ and control time T . For instance, in [11] Zhou showed a.c. of an abstract semilinear control system by assuming some inequality conditions, which are dependent on the properties of the system components, and also Quinn and Carmichael in [6] showed the diameters of the reachable sets by using the system constants determined restrictively.

Our purpose is to show a.c. of (1.1) under simple and fundamental assumptions on the system components. In particular, the range conditions of the operator B are most essential. We assume hypothesis (B1), which implies a.c. of $(1.1)_1$, and also, which is described as a geometrical relation in $L^2(0, T; X)$ between the range of the operator B and the subspace N^\perp related with the semigroup $S(t)$ (§ 2). On this space N^\perp , exactly, on an equivalent quotient space, we use the Schauder's degree theorem to show a.c. of (1.1). Furthermore, if the operator B is invertible (hypothesis (B2)), we can show an inclusive relation between the reachable sets of $(1.1)_1$ and those of (1.1) by determining the diameters of their control sets (§ 3). Because of its simple form, hypothesis (B1) can be easily checked in many examples (§ 4).

* Received by the editors April 29, 1985; accepted for publication (in revised form) April 10, 1986.

† Department of Information Sciences, Tokyo Institute of Technology, Oh-okayama, Meguro-ku, Tokyo, 152 Japan.

2. Notation and hypotheses. For system (1.1) we further assume

(F1) The nonlinear operator F on X is Lipschitz continuous; there exists a constant $K > 0$ such that

$$\|F(x_1) - F(x_2)\|_X \leq K \|x_1 - x_2\|_X, \quad x_1, x_2 \in X.$$

Then, we can consider a unique mild solution $y(t; v)$ for each $v \in L^2(0, T; V)$,

$$(2.1) \quad y(t; v) = \int_0^t S(t-s) \{F(y(s; v)) + (Bv)(s)\} ds, \quad 0 \leq t \leq T,$$

and we define the solution mapping W from $L^2(0, T; V)$ to $C(0, T; X)$ by

$$(Wv)(t) = y(t; v), \quad v \in L^2(0, T; V).$$

(See [5] for mild solutions of semilinear systems.) We also assume

(W) The solution mapping W is compact.

Remark 1. If A generates a compact semigroup, (W) is satisfied (cf. [4]). Recently, Seidman [8] shows the various sufficient conditions for the compactness of the solution mapping.

We introduce some definitions and notation. We define the linear operator \tilde{S} from $L^2(0, T; X)$ to X by

$$\tilde{S}p = \int_0^T S(T-s)p(s) ds \quad \text{for } p \in L^2(0, T; X).$$

Denote the kernel of the operator \tilde{S} by N , which is a closed subspace in $L^2(0, T; X)$, and its orthogonal space in $L^2(0, T; X)$ by N^\perp . Let P_J be the projection on $L^2(0, T; X)$ with the range N^\perp . Denote the range of the operator B by X_B and its closure by \bar{X}_B . We need the following hypothesis:

(B1) For each $p \in L^2(0, T; X)$ there exists a function $q \in \bar{X}_B$: $\tilde{S}p = \tilde{S}q$.

Remark 2. It is easily known that (B1) is equivalent to the following conditions: $L^2(0, T; X) = \bar{X}_B + N$ or $P_J \bar{X}_B = N^\perp$. We see later in Lemma 2 that (B1) implies a.c. of (1.1)₁. On the other hand, a.c. of (1.1)₁ is equivalent to the density condition $L^2(0, T; X) = \overline{X_B + N}$. Therefore, if the closedness of the product space is assumed (cf. Kato [2]), (B1) is equivalent to a.c. of (1.1)₁.

Under hypothesis (B1) we can define the mapping P from N^\perp to \bar{X}_B as follows: let $u_m \in N^\perp$ and define Pu_m as the unique minimum norm element u_X in $\{u_m + N\} \cap \bar{X}_B$:

$$\|Pu_m\| = \|u_X\| = \min \{\|u\| : u \in \{u_m + N\} \cap \bar{X}_B\}.$$

From hypothesis (B1) it follows that for each u_m in N^\perp $\{u_m + N\} \cap \bar{X}_B \neq \emptyset$. Thus the operator P is well defined.

LEMMA 1. *The operator P from N^\perp to \bar{X}_B is linear and continuous.*

Proof. Denote the range space of the mapping P by $R(P)$. Since $R(P)$ is identified with the quotient space $\bar{X}_B/(\bar{X}_B \cap N)$, $R(P)$ is a closed subspace in $L^2(0, T; X)$. Restrict the domain of the projection P_J to $R(P)$ and consider hypothesis (B1); then it is easily known that $P_J|_{R(P)}$ is an injective and surjective mapping from $R(P)$ to N^\perp . Since $(P_J|_{R(P)})^{-1} = P$, it follows from the open mapping theorem that P is linear and continuous.

In (2.1), consider the case in which $V = X$ and $B = I$. For each $u \in L^2(0, T; X)$, there exists a unique mild solution $y_u \in C(0, T; X)$ which satisfies

$$y_u(t) = \int_0^t S(t-s) \{F(y_u(s)) + u(s)\} ds.$$

So we can define the nonlinear operator \mathcal{F} on $L^2(0, T; X)$ by

$$(\mathcal{F}u)(t) = F(y_u(t)), \quad u \in L^2(0, T; X).$$

Since the solution mapping W is compact, hypothesis (F1) implies the compactness of the operator \mathcal{F} .

Let \tilde{Y} denote the quotient space; $\tilde{Y} = L^2(0, T; X)/N$ and define the norm of a coset $\tilde{y} = y + N$ in \tilde{Y} by $\|\tilde{y}\| = \inf \{\|y + f\| : f \in N\}$. Then we denote the isometric isomorphism from \tilde{Y} onto N^\perp by G . We will treat the controllability problem by using the degree theorem in \tilde{Y} . We define the operator $\tilde{\mathcal{F}}$ on \tilde{Y} by

$$\tilde{\mathcal{F}}\tilde{u} = \mathcal{F}(PG\tilde{u}) + N, \quad \tilde{u} \in \tilde{Y}.$$

From the previous argument we note that $\tilde{\mathcal{F}}$ is compact.

3. Approximate controllability. We denote the reachable set of system (1.1) by $K_T(F)$;

$$K_T(F) = \{y(T; v) : v \in L^2(0, T; V)\}$$

where $y(T; v)$ is the T -time value of the state function $y(t; v)$ which satisfies (2.1) with a given control function v in $L^2(0, T; V)$. If the reachable set is dense in X ; $\overline{K_T(F)} = X$, then the system is called approximately controllable. We also define the reachable set $K'_T(F)$ by

$$K'_T(F) = \{y(T; v) : v \in V_r\}$$

where V_r is an open ball in $L^2(0, T; V)$ with its radius $r > 0$. Similarly, as for the corresponding linear system (1.1)₁, we can define $K_T(0)$, $K'_T(0)$ and the approximate controllability.

For a.c. of (1.1)₁ we know the following.

LEMMA 2. Assume hypothesis (B1); then we have $\overline{K_T(0)} = X$.

Proof. Let $\xi \in D(A)$, then there exists a function $p \in C^1(0, T; X)$ such that

$$\xi = \int_0^T S(T-s)p(s) ds,$$

for instance, put $p(s) = (\xi + sA\xi)/T$. From (B1), there exists a function $q \in \bar{X}_B$ such that $p - q \in N$, that is,

$$\xi = \int_0^T S(T-s)p(s) ds = \int_0^T S(T-s)q(s) ds.$$

For every $\varepsilon > 0$ there exists a control function v_ε in $L^2(0, T; V)$ such that

$$\|Bv_\varepsilon - q\| < (MT)^{-1}\varepsilon$$

where M is a positive constant: $\|S(t)\| \leq M, 0 \leq t \leq T$. Put

$$\xi_\varepsilon = \int_0^T S(T-s)Bv_\varepsilon(s) ds;$$

then we have

$$\|\xi - \xi_\varepsilon\|_X \leq MT\|Bv_\varepsilon - q\| < \varepsilon.$$

Since ε is given arbitrarily, the density of the domain $D(A)$ in X implies a.c. of (1.1)₁.

To prove the controllability without the density condition for the range of the operator B and without any restrictions to the control time T , we need some additional

assumptions on the nonlinear function F . Here we consider the following case (cf. [3], [12]).

(F2) F is uniformly bounded; there exists a constant $M_F > 0$ such that

$$\|F(x)\|_X \leq M_F \quad \text{for all } x \in X.$$

Remark 3. As for the boundedness of the function \mathcal{F} on $L^2(0, T; X)$ and $\tilde{\mathcal{F}}$ on \tilde{Y} , it is easily known that

$$\|\tilde{\mathcal{F}}\tilde{u}\| \leq \|\mathcal{F}u\| \leq M_F\sqrt{T} \quad \text{for all } u \in L^2(0, T; X), \quad \tilde{u} \in \tilde{Y}.$$

THEOREM 1. *We assume the hypotheses (B1), (F1) and (F2); then we have*

$$K_T(0) \subset \overline{K_T(F)}.$$

Therefore, since $K_T(0)$ is dense in X , system (1.1) is approximately controllable.

Proof. We use the following open balls:

$$U_d = \{x \in L^2(0, T; X) : \|x\| < d\},$$

$$\tilde{U}_d = \{\tilde{x} \in \tilde{Y} : \tilde{x} = x + N, x \in U_d\},$$

$$\theta_d = \{\tilde{x} \in \tilde{Y} : \|\tilde{x}\|_{\tilde{Y}} < d\};$$

then we know that $\tilde{U}_d \subset \theta_d$. If $\eta \in K_T(0)$, then there exists a constant $r > 0$ and $v \in V_r$ such that $\eta \in K'_T(0)$ and

$$\eta = \int_0^T S(T-s)Bv(s) \, ds.$$

Put $z = Bv$ and $r_1 = \|B\|r$, then we have

$$\tilde{z} = z + N \in \tilde{U}_{r_1}.$$

Take a constant $R_1 > 0$ such that

$$R_1 > M_F\sqrt{T} + r_1.$$

We will apply the degree theorem on θ_{R_1} .

For the element $\tilde{z} \in \tilde{U}_{r_1}$, consider the equation

$$(3.1) \quad \tilde{z} = \lambda \tilde{\mathcal{F}}\tilde{u}_\lambda + \tilde{u}_\lambda, \quad 0 \leq \lambda \leq 1.$$

Since $\tilde{z} \in \theta_{R_1}$, (3.1) has a solution in θ_{R_1} when $\lambda = 0$. Let \tilde{u}_λ , $0 \leq \lambda \leq 1$ be a solution of (3.1); then we have

$$(3.2) \quad \begin{aligned} \|\tilde{u}_\lambda\| &\leq \|\tilde{z}\| + \|\tilde{\mathcal{F}}\tilde{u}_\lambda\| \\ &\leq r_1 + M_F\sqrt{T} < R_1. \end{aligned}$$

Thus, $\tilde{u}_\lambda \notin \partial\theta_{R_1}$, $0 \leq \lambda \leq 1$. It follows from the compactness of $\tilde{\mathcal{F}}$ that there exists a solution $\tilde{u} \in \theta_{R_1}$ which satisfies the equation

$$(3.3) \quad \tilde{z} = \tilde{\mathcal{F}}\tilde{u} + \tilde{u}.$$

Since

$$P(G\tilde{u}) \in P_J^{-1}(G\tilde{u}) \cap \bar{X}_B,$$

we have

$$P(G\tilde{u}) + N = \tilde{u}.$$

Put $u_B = P(G\tilde{u})$, then (3.3) is described by

$$\tilde{z} = \mathcal{F}(u_B) + u_B + N.$$

It follows that

$$\eta = \int_0^T S(T-s)z(s) ds = y_{u_B}(T) = \int_0^T S(T-s)\{F(y_{u_B}(s)) + u_B(s)\} ds.$$

Since the mapping P takes values in \bar{X}_B , for every $\varepsilon > 0$ there exists a function v_ε in $L^2(0, T; V)$ such that

$$\|PG\tilde{u} - Bv_\varepsilon\| < \varepsilon_0$$

where $\varepsilon_0 = \{M\sqrt{T} \exp(MKT)\}^{-1}\varepsilon$ and M is a positive constant such that $\|S(t)\| \leq M, 0 \leq t \leq T$. Consider the mild solution $y(t; v_\varepsilon) \in C(0, T; X)$;

$$y(t; v_\varepsilon) = \int_0^t S(t-s)\{F(y(s; v_\varepsilon)) + (Bv_\varepsilon)(s)\} ds.$$

Since the operator F is Lipschitz continuous, by using Hölder's inequality we have

$$\begin{aligned} \|y_{u_B}(t) - y(t; v_\varepsilon)\| &\leq \int_0^t \|S(t-s)\| \{\|F(y_{u_B}(s)) - F(y(s; v_\varepsilon))\| + \|u_B(s) - Bv_\varepsilon(s)\|\} ds \\ &\leq M\sqrt{T}\varepsilon_0 + KM \int_0^t \|y_{u_B}(s) - y(s; v_\varepsilon)\| ds, \quad 0 \leq t \leq T. \end{aligned}$$

By using Gronwall's lemma it follows that

$$\|y_{u_B}(t) - y(t; v_\varepsilon)\| \leq M\sqrt{T}\varepsilon_0 \exp(MKt), \quad 0 \leq t \leq T.$$

Thus for every $\varepsilon > 0$ there exists a control function v_ε in $L^2(0, T; V)$ such that

$$\|y_{u_B}(T) - y(T; v_\varepsilon)\| \leq \varepsilon.$$

It follows that

$$y_{u_B}(T) \in K_T(F) + S_\varepsilon$$

where $S_\varepsilon = \{x \in X : \|x\| \leq \varepsilon\}$. Since ε can be given arbitrarily, we have

$$\eta = y_{u_B}(T) \in \overline{K_T(F)},$$

which completes the proof.

Next we estimate the diameters of the admissible control sets, which implies the inclusive relation between the reachable set of system (1.1) and that of its corresponding linear system. We need the following hypothesis.

(B2) There exists a constant $k_B > 0$ such that

$$\|v\| \leq k_B \|Bv\|, \quad v \in L^2(0, T; V).$$

From the hypothesis (B2) it follows that X_B is a closed subspace in $L^2(0, T; X)$. We can see many examples which satisfy (B2) (cf. [12]).

THEOREM 2. *Under hypotheses (F1), (F2), (B1) and (B2), we have the following inclusive relation; for every $r > 0$, there exists $R > 0$ such that*

$$K_T^r(0) \subset K_T^R(F)$$

where R is a positive constant that satisfies

$$R > (r\|B\| + M_F\sqrt{T})\|P\|k_B.$$

Proof. We can use the same argument and the same notations in the proof of Theorem 1. Let $\eta \in K_T^r(0)$ and $z \in U_{r_1}$ such that

$$\eta = \int_0^T S(T-s)z(s) ds;$$

then we have $\tilde{z} = \tilde{\mathcal{F}}\tilde{u} + \tilde{u}$. By the definition of the operator P and hypothesis (B2) there exists a control function v in $L^2(0, T; V)$ such that $PG\tilde{u} = Bv$. Without the approximation argument in the proof of Theorem 1 we have

$$\eta = \int_0^T S(T-s)\{F(y(s; v)) + Bv(s)\} ds.$$

We also have

$$\begin{aligned} \|v\| &\leq k_B\|Bv\| = k_B\|PG\tilde{u}\| \\ &\leq k_B\|P\|\|\tilde{u}\| \\ &< k_B\|P\|(r\|B\| + M_F\sqrt{T}). \end{aligned}$$

Take a constant $R > k_B\|P\|(r\|B\| + M_F\sqrt{T})$; then we have

$$\eta \in K_T^R(F).$$

Remark 4. The hypotheses (F1) and (F2) can be weakened to the following hypothesis (F'):

$$\|F(x)\| \leq K'(1 + \|x\|^\alpha), \quad x \in X$$

where K' is a positive constant and $0 < \alpha < 1$. (See Seidman [9].) In fact, using an elementary inequality: $r^\alpha \leq 1 + r$ ($r \geq 0$) and Gronwall's lemma, we have

$$\|y_u\| \leq C'(1 + \|u\|), \quad u \in L^2(0, T; X)$$

and it follows that

$$\|\mathcal{F}u\| \leq C''(1 + \|u\|)^\alpha$$

where C' and C'' are positive constants dependent on K' , T , M . Dividing both sides of (2.3) by $\|\tilde{u}_\lambda\|^\alpha$ if $\|\tilde{u}_\lambda\|^\alpha \geq k$ for an arbitrarily given constant $k > 0$, we can obtain the similar estimation and use the same argument as that in the proof of Theorem 1.

4. Examples.

Example 1. In this section we consider the heat control system studied by Zhou [10]. Let $X = L^2(0, \pi)$ and $A = -d^2/dx^2$ with $D(A)$ consisting of all $y \in X$ with $d^2y/dx^2 \in X$ and $y(0) = y(\pi) = 0$. Put $\phi_n(x) = (2/\pi)^{1/2} \sin nx$, $0 \leq x \leq \pi$, $n = 1, 2, \dots$, then $\{\phi_n, n = 1, 2, \dots\}$ is an orthonormal base for X and ϕ_n is the eigenfunction corresponding to the eigenvalue $\lambda_n = -n^2$ of the operator $-A$, $n = 1, 2, \dots$. Define an infinite-dimensional space V by

$$V = \left\{ \sum_{n=2}^{\infty} u_n \phi_n, \text{ with } \sum_{n=2}^{\infty} u_n^2 < +\infty \right\}.$$

The norm in V is defined by $\|u\|_V = (\sum_{n=2}^{\infty} u_n^2)^{1/2}$. Define a continuous linear mapping from V to X as follows:

$$(4.1) \quad Bu = 2u_2\phi_1 + \sum_{n=2}^{\infty} u_n\phi_n \quad \text{for } u = \sum_{n=2}^{\infty} u_n\phi_n \in V.$$

Consider a control system governed by the semilinear heat equation

$$(4.2) \quad \begin{aligned} \frac{\partial y(t, x)}{\partial t} &= \frac{\partial^2 y(t, x)}{\partial x^2} + F(x, y(t, x)) + Bu(t, x), \quad 0 < t < T, \quad 0 < x < \pi, \\ y(t, 0) &= y(t, \pi) = 0, \quad 0 \leq t \leq T, \\ y(0, x) &= 0, \quad 0 \leq x \leq \pi. \end{aligned}$$

Now we can define the bounded linear operator \tilde{B} from $L^2(0, T; V)$ to $L^2(0, T; X)$ by $(\tilde{B}u)(t) = Bu(t)$, $u \in L^2(0, T; V)$. And the nonlinear operator F on X is assumed to satisfy hypotheses (F1) and (F2).

In [10] Zhou showed that this heat control system satisfied his inequality conditions and proved the approximate controllability of this system. Here we examine our condition (B1) and (B2) for this control system. By using Theorems 1 and 2 we show the approximate controllability without checking Zhou's inequality conditions.

It is well known that $-A$ generates a compact semigroup $S(t)$ and that

$$S(t)\phi_n = e^{-\lambda_n t}\phi_n, \quad n = 1, 2, \dots$$

Let $a \in N \subset L^2(0, T; X)$ and $a(s) = \sum_1^{\infty} a_n(s)\phi_n$, where $a_n(s) = (a(s), \phi_n)_X$. Since $S(t)a(s) = \sum_1^{\infty} a_n(s)e^{-\lambda_n t}\phi_n$, we have

$$\begin{aligned} \int_0^T S(T-s)a(s) ds &= \int_0^T \sum_1^{\infty} a_n(s) e^{-\lambda_n(T-s)} \phi_n ds \\ &= \sum_1^{\infty} \int_0^T a_n(s) e^{-\lambda_n(T-s)} ds \phi_n = 0. \end{aligned}$$

Thus we have

$$\int_0^T a_n(s) e^{-\lambda_n(T-s)} ds = 0, \quad n = 1, 2, \dots$$

Therefore, we know that $a \in N$ if and only if

$$a \in L^2(0, T; X) \quad \text{and} \quad \int_0^T a_n(s) e^{\lambda_n s} ds = 0, \quad n = 1, 2, \dots$$

We show the following claim which corresponds to hypothesis (B1): let $h \in L^2(0, T; X)$ and $h = \sum_1^{\infty} h_n(s)\phi_n$, then there exists a function u in $L^2(0, T; V)$ and a function a in N such that $h = a + \tilde{B}u$, that is, $h_1 = a_1 + 2u_2$, and $h_n = a_n + u_n$, $n = 2, 3, \dots$.

In Hilbert space $L^2(0, T)$ we easily know that

$$L^2(0, T) = \{e^s\}^{\perp} + \{e^{4s}\}^{\perp}.$$

So, for $h_1, h_2 \in L^2(0, T)$, there exists functions $a_1 \in \{e^s\}^{\perp}$, $a_2 \in \{e^{4s}\}^{\perp}$ which satisfy $h_1 - 2h_2 = a_1 - 2a_2$. Put $u_2 = h_2 - a_2$, then we have $h_1 = a_1 + 2u_2$ and $h_2 = a_2 + u_2$. And also, put $u_n = h_n$, $a_n = 0$, $n = 3, 4, \dots$, then we know that hypothesis (B1) is satisfied.

It is easily known that the operator \tilde{B} from $L^2(0, T; V)$ to $L^2(0, T; X)$ is one to one and the range of \tilde{B} is closed. It follows that the operator \tilde{B} satisfies hypothesis (B2).

Example 2. We introduce a simple example of the control operator B which satisfies hypothesis (B1). Consider the case $X = V$ and define the intercept operator $B_{(\alpha, T)}$, $0 < \alpha < T$, on $L^2(0, T; X)$ by

$$(B_{(\alpha, T)}v)(t) = \begin{cases} 0, & 0 \leq t < \alpha, \\ v(t), & \alpha \leq t \leq T, \end{cases} \quad v \in L^2(0, T; X).$$

We will show that for a given function $p \in L^2(0, T; X)$, there exists a control $v \in L^2(0, T; X)$: $\tilde{S}p = \tilde{S}B_{(\alpha, T)}v$. By using the semigroup property and the change of variation we obtain the following:

$$\begin{aligned} \int_0^T S(T-s)p(s) ds &= \int_0^\alpha S(T-s)p(s) ds + \int_\alpha^T S(T-s)p(s) ds \\ &= \int_\alpha^T S(T-s)S\left(s - \frac{\alpha}{T-\alpha}(s-\alpha)\right) \frac{\alpha}{T-\alpha} p\left(\frac{\alpha}{T-\alpha}(s-\alpha)\right) ds \\ &\quad + \int_\alpha^T S(T-s)p(s) ds. \end{aligned}$$

Thus, we can find a required control function $v(t)$:

$$v(t) = \begin{cases} 0, & 0 \leq t < \alpha, \\ p(t) + \frac{\alpha}{T-\alpha} S\left(t - \frac{\alpha}{T-\alpha}(t-\alpha)\right) p\left(\frac{\alpha}{T-\alpha}(t-\alpha)\right), & \alpha \leq t \leq T. \end{cases}$$

Furthermore, by using the similar calculations, we can see that if $-A$ generates a group $S(t)$, the intercept operator $B_{(0, \alpha)}$ and the impulsive operator $B_{(\alpha, \alpha+\varepsilon)}$, $0 < \varepsilon \ll 1$ also satisfy (B1).

Acknowledgments. The author wishes to express his deep appreciation to Professor T. I. Seidman for his many helpful suggestions and to Professor W. Takahashi for many stimulating conversations.

REFERENCES

- [1] A. V. BALAKRISHNAN, *Applied Functional Analysis*, Springer-Verlag, New York, Berlin, 1976.
- [2] T. KATO, *Perturbation Theory for Linear Operators*, Springer-Verlag, New York, Berlin, 1966.
- [3] K. MIRZA AND B. F. WOMACK, *On the controllability of a class of non-linear systems*, IEEE Trans. Automat. Control, AC-17 (1972), pp. 531-535.
- [4] K. NAITO, *Compactness of solution mappings for semilinear equations*, TIT Inf. Sci. Tech. Rep. A-98, 1985.
- [5] A. PAZY, *Semigroups of linear operators and applications to partial differential equations*, Springer-Verlag, New York, Berlin, 1983.
- [6] M. D. QUINN AND N. CARMICHAEL, *An approach to non-linear control problems using fixed-point methods, degree theory and pseudo-inverses*, Numer. Funct. Anal. Optim., 7 (1984-1985), pp. 197-219.
- [7] D. L. RUSSEL, *Controllability and stabilizability theory for linear partial differential equations: recent progress and open questions*, SIAM Rev., 20 (1978), pp. 639-739.
- [8] T. I. SEIDMAN, *Invariance under nonlinear perturbations for reachable and almost-reachable sets*, Proc. IFIP Symp. on Control Theory for P.D.E., to appear.
- [9] ———, *Invariance on the reachable sets under nonlinear perturbations*, this Journal, 25 (1987), to appear.
- [10] H. X. ZHOU, *A note on approximate controllability for semilinear one-dimensional heat equation*, Appl. Math. Optim., 8 (1982), pp. 275-285.
- [11] ———, *Approximate controllability for a class of semilinear abstract equations*, this Journal, 21 (1983), pp. 551-565.
- [12] ———, *Controllability properties of linear and semilinear abstract control systems*, this Journal, 22 (1984), pp. 405-422.

STABILIZABILITY AND STABLE-PROPER FACTORIZATIONS FOR LINEAR TIME-VARYING SYSTEMS*

KAMESHWAR POOLLA[†] AND PRAMOD KHARGONEKAR[‡]

Abstract. The objective of this paper is to study in detail stabilizability and stable-proper factorizations for linear *time-varying* discrete-time systems. Our main results are:

(i) If a linear-time-varying system can be stabilized by *dynamic* state feedback, then it can also be stabilized by *memoryless* state feedback.

(ii) A complete characterization of the *existence* of stable-proper factorizations for linear time-varying input/output operators. This characterization is nontrivial; there exist input/output operators that *do not* admit stable-proper factorizations.

Key words. time-varying systems, transfer-functions, stabilizability, skew-rings

AMS(MOS) subject classifications. 93D15, 93C50

1. Introduction. The aim of this paper is to study in detail stabilizability and stable-proper factorizations for linear *time-varying* discrete-time systems.

Many frequency-domain based synthesis methods used to design controllers for linear time-invariant plants begin with left- and/or right-stable proper coprime factorizations of the plant transfer-function matrix (see, for example, Zames and Francis [1983]). Given the power and utility of these representations, it seems natural and desirable to develop corresponding results for *time-varying* systems. In order to realize this objective we need to incorporate time-variance in a natural way into a transform type description of the system behavior. Attempts have been made to develop such a theory (e.g., the system function of Zadeh [1950]), but until recently, there has been no theory that has met with much success.

In § 2 of this paper, we establish some notation, and we outline the basic elements of a transfer-function approach to linear time-varying systems based on skew (noncommutative) rings developed by Kamen, Khargonekar and Poolla [1985]. For more detailed exposition, the reader is referred to the dissertation of Poolla [1984].

Following this, in § 3, we briefly review some well-known concepts related to the stability of linear time-varying systems. Then, in § 4, we introduce the key notion of *asycontrollability* of a linear-time-varying system Σ which is equivalent to being able to stabilize Σ via *dynamic* state feedback. Anderson and Moore [1981] have defined a notion of stabilizability for linear-time-varying systems, which is equivalent to being able to stabilize Σ via *memoryless* state feedback. One of the deepest results of this paper (Theorem 4.4) is the equivalence of asycontrollability and stabilizability. This result in particular implies that *dynamics* in state feedback buy nothing extra as far as the problem of stabilization is concerned.

Finally, in § 5 we introduce stable-proper factorizations for linear-time-varying systems. In glaring contrast to time-invariant systems, *not* all time-varying input/output maps admit stable-proper factorizations (see Example 5.8). Hence we first characterize the existence of stable-proper factorizations (see Theorem 5.2). Following this one can

* Received by the editors September 11, 1985; accepted for publication (in revised form) April 18, 1986. This work was supported in part by the National Science Foundation under grants ECS-82-00607 and ECS-84-00832 through the University of Florida. Part of this work was done while the authors were at the University of Florida, Gainesville, Florida 32611.

[†] Coordinated Science Laboratory, University of Illinois, Urbana, Illinois 61801.

[‡] Department of Electrical Engineering, University of Minnesota, Minneapolis, Minnesota 55455.

systematically employ the axiomatic theory of Desoer et al. [1980] to study feedback control problems, for instance by obtaining a complete parametrization of all controllers that internally stabilize a linear-time-varying plant. The results in this paper are a natural extension of our earlier work on a polynomial theory for linear-time-varying systems; the interested reader may consult Khargonekar and Poolla [1986].

2. The basic framework. With \mathbb{Z} = set of integers and \mathbb{R} = field of real numbers, let A denote \mathbb{R} -linear space of all functions from \mathbb{Z} into \mathbb{R} . With the operations of pointwise addition and pointwise multiplication, it is easy to see that A is a commutative ring with identity 1, where $1(k) = 1$ for all $k \in \mathbb{Z}$. Let σ denote the *right-shift operator* on A defined by

$$(\sigma\alpha)(k) = \alpha(k-1) \quad \text{all } k \in \mathbb{Z}.$$

Let A_+ denote the subring of A consisting of all functions with support bounded on the left; that is, for any $\alpha \in A_+$ there is an integer k_α (depending on α in general) such that $\alpha(k) = 0$ for all $k \leq k_\alpha$. Let $l^\infty(\mathbb{Z})$ denote the set of all bounded time-functions. Clearly, $l^\infty(\mathbb{Z})$ is a difference subring of A .

DEFINITION 2.1. Let m and p be positive integers. An m -input p -output linear time-varying causal input/output map f is an \mathbb{R} linear map

$$f: A_+^m \rightarrow A_+^p$$

such that $\{u(k) = 0, k \leq k_u\} \Rightarrow \{f(u)(k) = 0, k \leq k_u\}$.

It is well known that for any input/output map f as defined above, there exists a $p \times m$ matrix function $W_f(i, j)$ such that for any $u \in A_+^m$,

$$f(u)(i) = \sum_{j=-\infty}^i W_f(i, j)u(j).$$

The matrix function W_f is the *unit-pulse response function* associated with the input/output map f . Note that by causality, $W_f(i, j)$ is not defined for $i < j$.

Our next concept is the notion of a system.

DEFINITION 2.2. An m -input p -output n -dimensional linear time-varying system Σ over A is a quadruple $\Sigma = (F, G, H, J)$ of matrices over A where F is $n \times n$, G is $n \times m$, H is $p \times n$, and J is $p \times m$.

With a system $\Sigma = (F, G, H, J)$, we shall associate the dynamical equations

$$x(j+1) = F(j)x(j) + G(j)u(j),$$

$$y(j) = H(j)x(j) + J(j)u(j)$$

where $u(j)$, $x(j)$, $y(j)$ have the usual interpretations.

A system $\Sigma = (F, G, H, J)$ or the pair (F, G) over A is said to be *reachable in N steps* if for any $j \in \mathbb{Z}$ and any $x \in \mathbb{R}^n$ there exists an input sequence $u(j-N), u(n-N+1), \dots, u(j-1)$ which drives Σ from the zero state at time $j-N$ to the state x at time j . The dual notion of *observability in N steps* has the obvious system-theoretic interpretation.

Let \mathbb{R}_N denote the N -step reachability matrix

$$\mathbb{R}_N := [G|F(\sigma G)|\dots|F(\sigma F)\dots(\sigma^{N-2}F)(\sigma_{N-1}G)].$$

Weiss [1972] has shown that $\Sigma = (F, G, H, J)$ is reachable in N steps if and only if $\text{rank } \mathbb{R}_N(j) = n$, for all j in \mathbb{Z} . Dual criteria for observability also exist.

As is easy to see, the unit-pulse response function W_Σ associated with the system $\Sigma = (F, G, H, J)$ is given by

$$W_\Sigma(i, j) = \begin{cases} H(i)F(i-1)F(i-2) \cdots F(j+1)G(j), & i > j, \\ J(j), & i = j, \\ \text{not defined,} & i < j. \end{cases}$$

The input/output behavior of a system Σ is described by its input/output map f_Σ , where

$$f_\Sigma(u)(i) = \sum_{j=-\infty}^i W_\Sigma(i, j)u(j), \quad u \in A_+^m.$$

Given an input/output map $f: A_+^m \rightarrow A_+^p$, a *realization of f* is a system $\Sigma = (F, G, H, J)$ over A such that $f_\Sigma = f$. For results on realizability, we refer the reader to Weiss [1972].

The commutative rings of polynomials, power series and formal Laurent series all with coefficients in the reals \mathbb{R} , play a central role in the transfer-function theory of linear time-invariant systems. For *time-varying* systems the analogous objects are skew (noncommutative) rings with coefficients in the ring of time functions.

More precisely, with z equal to an indeterminate, let $A((z^{-1}))$ denote the set of all formal Laurent series of the form

$$\sum_{r=-N}^{\infty} z^{-r}\alpha_r, \quad \alpha_r \text{ in } A.$$

With the usual addition and multiplication defined by

$$z^r z^t = z^{r+t}, \quad \alpha z = z(\sigma\alpha), \quad \alpha \text{ in } A,$$

where $(\sigma\alpha)(k) = \alpha(k-1)$, $A((z^{-1}))$ is a noncommutative ring with identity, called the *skew ring of formal Laurent series* over A . There are two important subrings of $A((z^{-1}))$: The skew ring of polynomials $A[z]$ and the skew ring of formal power series $A[[z^{-1}]]$. These have the obvious definitions.

We now describe our transfer function approach to linear time-varying systems. All proofs are omitted; they can be found in Kamen, Khargonekar and Poolla [1985].

Again, let A_+ denote the subring of A consisting of all functions $\alpha: \mathbb{Z} \rightarrow \mathbb{R}$ with support bounded on the left. Let Δ denote the unit-impulse at the origin, i.e., $\Delta(k) = 1$ if $k = 0$, and $\Delta(k) = 0$ otherwise. Given any b in A_+ , the (generalized) *z-transform* of b written $B(z)$ is defined to be the skew Laurent series

$$B(z) = \sum_r z^{-r}b(r)\Delta.$$

Let f be an input/output map, and let W_f denote the unit-pulse response function associated with f . For each integer $r \geq 0$, define a $p \times m$ matrix W_r over A by

$$W_r(j) = W_f(r+j, j), \quad j \text{ in } \mathbb{Z}.$$

DEFINITION 2.3. The (formal) transfer-function matrix $\hat{W}_f(z)$ associated with the input/output map f is the $p \times m$ matrix over $A[[z^{-1}]]$ defined by

$$\hat{W}_f(z) = \sum_{r=0}^{\infty} z^{-r}W_r.$$

We would like to point out that $\hat{W}_f(z)$ is the same as the frequency-response function of Arveson [1975]. The input/output difference equation $y = f(u)$ can now be characterized in terms of the transfer-function matrix $\hat{W}_f(z)$ as follows.

PROPOSITION 2.4. *Let f be an input/output map. Let y be the output resulting from the input u in A_+^m . Let $Y(z)$ and $U(z)$ denote the (generalized) z -transforms of y and u . Then,*

$$(2.5) \quad Y(z) = \hat{W}_f(z)U(z).$$

Note the close resemblance of (2.5) to the time-invariant transfer function theory. This analogy to the time-invariant theory is further illustrated by the following proposition.

PROPOSITION 2.6. *Let $\Sigma = (F, G, H, J)$ be a linear time-varying system with input/output map f_Σ . Then the transfer-function matrix \hat{W}_Σ associated with f_Σ is given by*

$$(2.7) \quad \hat{W}_\Sigma(z) = H(zI - F)^{-1}G + J.$$

Despite the close resemblance these two results bear to the time-invariant theory, it must be emphasized that (2.5) and (2.7) are computed via the skew (noncommutative) multiplication defined earlier.

3. Stability. In this section we briefly review some well-known concepts dealing with the stability of linear time-varying systems, and we relate these concepts to the transfer-function theory based on skew-rings described in § 2. There is extensive literature available on the stability of linear time-varying systems (see, for example, Anderson and Moore [1981], Willems [1970], Cesari [1963], Freedman and Zames [1968], etc.).

For any vector x in \mathbb{R}^n let $\|x\|$ denote the Euclidean norm of x . For an $m \times m$ matrix M over \mathbb{R} , let $\|M\| := \sup_{x \neq 0} \|Mx\|/\|x\|$. For an $m \times m$ matrix N over A , define the norm of N to be $\|N\| := \sup_t \|N(t)\|$.

Let $f: A_+^m \rightarrow A_+^p$ be an input/output map. We shall say that f is *bounded input-bounded output (BIBO) stable* if for any bounded input sequence, i.e., for any u in $l^\infty(\mathbb{Z})_+$, the corresponding output sequence $y = f(u)$ is bounded. Let $\Sigma = (F, G, H, J)$ be any (fixed) realization of f . Consider the free behavior of the system Σ described by the vector difference equation

$$(3.1) \quad x(k+1) = F(k)x(k).$$

The system Σ is said to be *internally uniformly asymptotically stable*, henceforth *stable*, if for every number $\varepsilon > 0$, there exists a positive integer N_ε such that for any initial time t_0 in \mathbb{Z} and any initial state $x(t_0)$ with $\|x(t_0)\| \leq 1$, we have that $\|x(t_0+i)\| \leq \varepsilon$ for all $i \geq N_\varepsilon$. Here, $x(t_0+i)$ is the solution of (3.1) at time t_0+i starting from initial state $x(t_0)$.

Let $P(z)$ be an $n \times m$ matrix over the skew ring $A((z^{-1}))$, i.e., $P(z)$ is of the form

$$P(z) = \sum_{i=-N}^{\infty} z^{-i}P_i, \quad P_i \in A^{n \times m}.$$

DEFINITION 3.2. The matrix Laurent series $P(z)$ is said to be *stable* if $\lim_{i \rightarrow \infty} \|P_i\| \rightarrow 0$.

In terms of this notion, we can characterize internal stability as follows (see Green and Kamen [1984] and Kamen, Khargonekar and Poolla [1985, Prop. (4.4)]).

PROPOSITION 3.3. *Let $\Sigma = (F, G, H)$ be a linear time-varying system over A . Then, Σ is internally u.a.s. if and only if $(zI - F)^{-1}$ is a stable matrix power series.*

Note the similarity this result bears to the time-invariant theory. Let f be an input/output map, and let $\Sigma = (F, G, H, J)$ over A be any (fixed) realization of f . In contrast with the time-invariant case, internal stability of Σ does *not* in general imply

BIBO stability of f . However, for the class of *bounded* linear time-varying systems, we have the following result (the proof is omitted on account of its relative ease).

PROPOSITION 3.4. *Let f be an input/output map and let $\Sigma = (F, G, H, J)$ over $l^\infty(\mathbb{Z})$ be any (fixed) realization of f . Suppose Σ is internally u.a.s. Then f is BIBO stable.*

In subsequent sections, we shall deal only with linear time-varying systems with *bounded* coefficients, i.e., defined over the difference subring $l^\infty(\mathbb{Z}) \subset A$. We shall also require that any controller we design be over $l^\infty(\mathbb{Z})$. These are physically reasonable constraints since most time-varying plants that arise in practice have bounded time variation, and implementation of controllers with unbounded coefficients would be numerically ill-conditioned. (Technically, these constraints substantially complicate proofs and make results harder to obtain.)

4. Stabilizability and asycontrollability. In this section, we introduce the key notion of *asycontrollability*, which is closely related to being able to stabilize a linear time-varying system by *dynamic* state feedback. Anderson and Moore [1981] have defined the notion of *stabilizability*, which is equivalent to being able to stabilize a linear time-varying system by *nondynamic* (i.e., memoryless) state feedback. The central result in this paper, Theorem 4.4, shows the equivalence of stabilizability and asycontrollability. A striking conclusion of this theorem is that *dynamics* in state feedback buy nothing extra as far as the problem of stabilization is concerned.

We begin with the following key definition.

DEFINITION 4.1. A system $\Sigma = (F, G, H)$ over $l^\infty(\mathbb{Z})$ is said to be asycontrollable if for every $\varepsilon > 0$ there exists a real number \hat{M} and an integer N such that for any initial time t_0 in \mathbb{Z} and any initial state ξ in \mathbb{R}^n with $\|\xi\| = 1$, there exists an input sequence $u(t_0), u(t_0+1), \dots, u(t_0+N-1)$ which results in the state trajectory $\xi = x(t_0), x(t_0+1), \dots, x(t_0+N)$ and such that

$$\|x(t_0+N)\| < \varepsilon, \quad \|x(t_0+k)\|, \|u(t_0+k)\| < \hat{M} \quad \text{for all } k.$$

This technical definition has a precise system-theoretic interpretation in terms of stabilizing Σ via an open-loop control law. Roughly speaking, a system is asycontrollable if and only if it can be driven near zero “final” state asymptotically, using *uniformly* bounded input sequences along *uniformly* bounded state trajectories. The phrase asycontrollability (from *asymptotically controllable*) is borrowed from Khargonekar and Sontag [1982] for time-invariant systems over rings.

Remark 4.2. It can be shown (see Khargonekar and Poolla [1986]) that asycontrollability is equivalent to the existence of *stable* Laurent series Y_1 and Y_2 over $l^\infty(\mathbb{Z})((z^{-1}))$ such that $(zI - F)Y_1 + GY_2 = I$. In particular, this means that if Σ can be stabilized using a *dynamic* state-feedback controller, then Σ is asycontrollable.

Let $\Sigma = (F, G, H)$ be a linear time-varying system over $l^\infty(\mathbb{Z})$. Anderson and Moore [1981] have defined a notion of *stabilizability* which intuitively corresponds to requiring that unstable modes be controllable. The authors then show that this notion is equivalent to the existence of a stabilizing *memoryless* state feedback law $u(k) = -L(k)x(k)$. We shall take this to be the definition of stabilizability. In terms of our skew-ring framework, we phrase this as follows.

DEFINITION 4.3. Let $\Sigma = (F, G, H)$ over $l^\infty(\mathbb{Z})$ be a linear time-varying system. Then Σ is said to be stabilizable if there exists an $m \times n$ matrix L over $l^\infty(\mathbb{Z})$ such that $(zI - F + GL)^{-1}$ is a stable matrix power series.

One can similarly define the dual notion of *detectability*.

Let $\Sigma = (F, G, H)$ over $l^\infty(\mathbb{Z})$ be a stabilizable linear time-varying system. Let L be an $m \times n$ feedback matrix as in Definition 4.3. Notice that we can write

$$(zI - F)Y_1 + GY_2 = I,$$

where $Y_1 = (zI - F + GL)^{-1}$ and $Y_2 = LY_1$. Since Y_1 and Y_2 are *stable* Laurent series, it follows from Remark 4.2 that Σ is asycontrollable. Thus, stabilizability *implies* asycontrollability. The much more difficult converse is also true.

THEOREM 4.4. *Let $\Sigma = (F, G, H, J)$ be a linear time-varying system over $l^\infty(\mathbb{Z})$. Then, Σ is asycontrollable if and only if Σ is stabilizable.*

Proof. See Appendix A. We would like to remark that the essential technical difficulty in the proof is ensuring that the feedback matrix L is over $l^\infty(\mathbb{Z})$. \square

Recall (see Remark 4.2) that if a linear time-varying system Σ can be stabilized by a dynamic state-feedback controller, then Σ is asycontrollable. This observation of stabilizability, together with Definition 4.3, immediately offers the following surprising conclusion.

COROLLARY 4.5. *If a linear time-varying system Σ over $l^\infty(\mathbb{Z})$ can be stabilized using dynamic state-feedback, then Σ can also be stabilized using memoryless state-feedback.*

Remark 4.6. Indeed the above corollary is not (necessarily) *expected*, because there are classes of systems (for example, delay systems (see Kamen [1982, Ex. 3, p. 371])) for which it is *not* true. For time-invariant systems over a field, it is a well-known fact that dynamic state feedback is equivalent to memoryless state feedback as far as the problem of stabilization is concerned. The proof of this fact relies heavily on the Kalman canonical decomposition. No such decomposition exists for time-varying systems because the “dimension” of the reachable space could depend on time.

Since stabilizability and asycontrollability are equivalent notions, we shall henceforth only speak of stabilizability.

It is important to find “nice” necessary and sufficient tests for stabilizability. This problem appears to be quite formidable unless one specializes to particular classes (e.g., periodic) of time-varying systems. We do have, however, the following sufficient condition.

THEOREM 4.7. *Let $\Sigma = (F, G, H, J)$ over $l^\infty(\mathbb{Z})$ be a linear time-varying system. Suppose there exist integers N and K and a real number $\varepsilon > 0$ with the following property: For any integer t_0 , there exists some t_1 with $t_0 \leq t_1 \leq t_0 + K$ such that*

$$(4.8) \quad \det [R_N R'_N](t_1) \geq \varepsilon > 0$$

where R_N is the N step reachability matrix. Then, Σ is stabilizable.

Proof. Essentially condition (4.8) corresponds to the system being $l^\infty(\mathbb{Z})$ -reachable in N steps but *not* at all times.

Given any initial state ξ in \mathbb{R}^n and any initial time t_0 , we construct an open-loop stabilizing control law as follows: we apply zero control (i.e. $u(t) = 0$) for the time $t_0 \leq t < t_1$. Then, we drive the system to zero state in N steps. This can be done because from (4.8) Σ is reachable at time t_1 . Moreover, the input sequences applied are uniformly (in t_1) bounded in norm. Having brought the system to zero state we apply no further inputs. We can thus stabilize Σ by an open-loop control law. This implies that Σ is asycontrollable, which by Theorem 4.4 implies that Σ is stabilizable. \square

5. Stable-proper factorizations. Of late, the use of stable-proper factorizations for control system design has become increasingly popular. See, for example, Vidyasagar [1978], Desoer et al. [1980], Feintuch and Francis [1984], Khargonekar and Sontag [1982], etc. In this section we investigate in detail stable-proper factorizations for time-varying systems.

Define the set RP (for Rational and Proper) to be

$$RP := \{\phi \text{ in } l^\infty(\mathbb{Z})[[z^{-1}]]: \text{there exist } \alpha, \beta, \gamma, \delta \text{ such that } \phi = \alpha(zI - \beta)^{-1}\gamma + \delta\},$$

where α, β, γ , and δ are matrices of compatible dimensions over $l^\infty(\mathbb{Z})$. It is easy to

verify that RP forms a ring with the usual skew-multiplication and addition defined in $l^\infty(\mathbb{Z})[[z^{-1}]]$. Also, define the subring RP_s by

$$RP_s := \{\phi = \alpha(zI - \beta)^{-1}\gamma + \delta \text{ in } RP: (zI - \beta)^{-1} \text{ is stable}\}.$$

We shall call RP_s the *ring of stable, proper, rational functions* (the subscript s denotes stable). We shall abandon rigorous nomenclature and loosely refer to elements in RP_s or matrices over RP_s as being *stable-proper*.

Let $D(z)$ be an $n \times n$ matrix over RP . We shall say that $D(z)$ is *bicausal* if D has an inverse $D^{-1}(z)$ also over RP . Let f be an input/output map and let $\hat{W}_f(z)$ be its associated transfer-function matrix. Then, $\hat{W}_f(z)$ is said to admit a *right-Bezout stable-proper factorization* if there exist stable-proper matrices (i.e., over RP_s) N , D , X , and Y with D bicausal and such that

$$(5.1) \quad \hat{W}_f(z) = ND^{-1}, \quad XN + YD = I.$$

One can also similarly define *left-Bezout stable-proper factorization*. Such factorizations have been found to be extremely useful in tackling many control-theoretic problems (for example, H^∞ -optimal controller design, see Zames and Francis [1983], Francis and Zames [1984], Francis, Helton and Zames [1984]). We first characterize input/output maps whose transfer functions admit left- and/or right-Bezout stable-proper factorizations. We have the following central result.

THEOREM 5.2. *Let f be an input/output map and let $W_f(z)$ be its $p \times m$ associated transfer function matrix over $l^\infty(\mathbb{Z})[[z^{-1}]]$. Then, the following are equivalent:*

- (a) $\hat{W}_f(z)$ admits a right-Bezout stable-proper factorization;
- (b) $\hat{W}_f(z)$ admits a left-Bezout stable-proper factorization;
- (c) f admits a stabilizable and detectable realization.

Proof. (c) \rightarrow (a). Let $\Sigma = (F, G, H, J)$ be a stabilizable and detectable realization of f . Consequently, by definition, there exist matrices L and K over $l^\infty(\mathbb{Z})$ that $(zI - F + GL)^{-1}$ and $(zI - F + KH)^{-1}$ are stable matrix power series. Define the stable-proper (i.e., over RP_s) matrices N , D , X , and Y by

$$(5.3) \quad \begin{aligned} N &= (H - JL)(zI - F + GL)^{-1}G + J, & D &= I - L(zI - F + GL)^{-1}G, \\ X &= L(zI - F + KH)^{-1}K, & Y &= I + L(zI - F + KH)^{-1}(G - KJ). \end{aligned}$$

Clearly D is bicausal, and it can be mechanically verified that with these definitions

$$\hat{W}_f(z) := H(zI - F)^{-1}G + J = ND^{-1}, \quad XN + YD = I.$$

Thus, $\hat{W}_f(z)$ admits a right-Bezout stable-proper factorization. We would like to point out that formulae (5.3) have been available in the system theory literature for a number of years (see Khargonekar and Sontag [1982, pp. 635–636]; Nett, Jacobsen and Balas [1984]; and Vidyasagar [1985]). The difficulty, however, is in showing the *existence* of stabilizable and detectable realizations (i.e., (a) \rightarrow (c)) as this critically involves Theorem 4.4. We proceed to do this.

(a) \rightarrow (c). Now suppose that $\hat{W}_f(z)$ admits a right-Bezout stable-proper factorization, i.e., there exist stable-proper matrices N , D , X , and Y with D bicausal and such that $\hat{W}_f(z) = ND^{-1}$, $XN + YD = I$. From the definitions of a stable-proper function (i.e., the ring RP_s), it follows that we can write

$$N = N_1(zI - N_2)^{-1}N_3 + N_4, \quad D = D_1(zI - D_2)^{-1}D_3 + D_4$$

where $N_1, \dots, N_4, D_1, \dots, D_4$ are matrices of appropriate sizes over $l^\infty(\mathbb{Z})$, with $(zI - N_2)^{-1}$ and $(zI - D_2)^{-1}$ being *stable* matrix power series. Since D is bicausal,

without loss of generality, $D_4 = I$. Define the linear time-varying system (over $l^\infty(\mathbb{Z})$) $\Sigma = (F, G, H, J)$ by

$$(5.4) \quad \begin{aligned} F &= \begin{bmatrix} N_2 & -N_3 D_1 \\ 0 & D_2 - D_3 D_1 \end{bmatrix}, & G &= \begin{bmatrix} N_3 \\ D_3 \end{bmatrix}, \\ H &= [N_1 \quad -N_4 D_1], & J &= N_4. \end{aligned}$$

The system Σ is stabilizable because with $L = [0 \quad -D_1]$ (which is over $l^\infty(\mathbb{Z})$),

$$(zI - F + GL)^{-1} = \begin{bmatrix} (zI - N_2)^{-1} & 0 \\ 0 & (zI - D_2)^{-1} \end{bmatrix},$$

which is *stable*. We now show that Σ , defined by (5.4), is detectable.

Define the stable-proper matrix Q by

$$Q = \begin{bmatrix} (zI - N_2)^{-1} N_3 \\ (zI - D_2)^{-1} D_3 \end{bmatrix}.$$

It is easy to verify that $(zI - F)Q - GD = 0$ and $HQ + JD = N$. Combining these equations with $XN + YD = 1$, we can write

$$(5.5) \quad \begin{bmatrix} zI - F & G \\ XH & -(Y + XJ) \end{bmatrix} \begin{bmatrix} (zI - F + GL)^{-1} & Q \\ L(zI - F + GL)^{-1} & -D \end{bmatrix} = \phi\psi = \begin{bmatrix} I & O \\ V & I \end{bmatrix}$$

where V is some stable-proper matrix whose exact formula is not critical to our needs. It is clear from (5.5) that ϕ is *right-invertible* with ϕ_R^{-1} (the subscript R denotes right) a *stable-proper matrix*. We now show that ϕ is also left-invertible (and thus its inverse $\phi_R^{-1} = \phi_L^{-1} = \phi^{-1}$ is unique). Notice that

$$(5.6) \quad \phi = \begin{bmatrix} zI - F & O \\ O & I \end{bmatrix} \begin{bmatrix} I & (zI - F)^{-1} G \\ XH & -(Y + XJ) \end{bmatrix} := \phi_1 \phi_2.$$

The matrix ϕ_1 is clearly both left- and right-invertible. Also, $(Y + XJ)$ must be of the form $(Y + XJ) = I + \text{terms in } z^{-1}, z^{-2}, \dots$ since $XN + YD = 1$. Thus, the leading (z^0) coefficient ϕ_2 is invertible (over $l^\infty(\mathbb{Z})$). Consequently, ϕ_2 is both left- and right-invertible. Thus, from (5.6), ϕ must be both left- and right-invertible, proving our claim. Recalling that ϕ^{-1} is a stable-proper matrix, we can write

$$\phi^{-1}\phi = \begin{bmatrix} W_1 & W_2 \\ * & * \end{bmatrix} \begin{bmatrix} (zI - F) & G \\ XH & -(Y + XJ) \end{bmatrix}.$$

One component of the above equality is

$$W_1(zI - F) + (W_2 X)H = I.$$

Thus the dual of the pair (F, H) is asycontrollable, which by Theorem 4.4 implies that Σ is detectable. This completes the proof of (a) \rightarrow (c).

(b) \rightarrow (c) and (c) \rightarrow (b) follow from a dual argument. \square

Remark 5.7. In proving the above result, we have made critical use of Theorem 4.4, the fundamental result of the previous section. Recall that Theorem 4.4 states that asycontrollability is equivalent to stabilizability. There appears to be no way to circumvent Theorem 4.4 in studying the existence of stable-proper factorizations. We would further like to remark that (5.3) enables one to compute stable-proper factorizations once the stabilizing feedback matrices L and K are determined.

Theorem 5.2 essentially states that a time-varying system admits stable-proper factorizations if and only if it admits a stabilizable and detectable realization. This characterization is nontrivial because (in glaring contrast with time-invariant systems), *not all time-varying systems admit stabilizable realizations* (and therefore by Theorem 5.2, stable-proper factorizations). We illustrate this with the following example.

Example 5.8. Consider the linear time-varying system $\Sigma = (F, G, H)$ over $l^\infty(\mathbb{Z})$, where $F(k) = H(k) = 1$ for all k in \mathbb{Z} and $G = \Delta$ = the unit pulse concentrated at the origin. Let f_Σ be the input/output map associated with Σ . Suppose f_Σ admits a stabilizable realization $\bar{\Sigma} = (\bar{F}, \bar{G}, \bar{H})$. This would imply that for any initial state ξ at time $t = 1$, there exists an open loop control law that drives $\bar{\Sigma}$ from $\bar{x}(1) = \xi$ to zero final state (and therefore zero final output) asymptotically. However, from the unit-pulse response function of f_Σ , it is evident that application of inputs after $t = 1$ has no effect on the output. Thus, it must be that

$$(5.9) \quad \lim_{k \rightarrow \infty} \|\bar{H}(k)\bar{F}(k-1)\bar{F}(k-1) \cdots \bar{F}(1)\| = 0.$$

However,

$$W_\Sigma(z) := H(zI - F)^{-1}G = z^{-1}\Delta + z^{-2}\Delta + \cdots = \bar{H}(zI - \bar{F})^{-2}\bar{G}.$$

Consequently, for all integers $k \geq 1$

$$\bar{H}(k)\bar{F}(k-1) \cdots \bar{F}(1)\bar{G}(0) = 1.$$

This, together with the fact that \bar{G} is over $l^\infty(\mathbb{Z})$, renders (5.9) impossible. Therefore, f_Σ does *not* admit a stabilizable realization which implies (by Theorem 5.2) that $\hat{W}_{f_\Sigma}(z)$ does not admit a stable-proper factorization.

Armed with the powerful tool of stable-proper factorizations, we can systematically apply the axiomatic theory of Desoer et al. [1980] to study feedback control problems for linear time-varying plants. For instance we can readily obtain a complete parameterization of all stabilizing controllers as follows.

THEOREM 5.10. *Let $\Sigma = (F, G, H, J)$ be a stabilizable and detectable plant. Let*

$$\hat{W}_\Sigma(z) = ND^{-1} = D_1^{-1}N_1,$$

$$XN + YD = I = N_1X_1 + D_1Y_1$$

be any left- and right-stable-proper factorizations of W_Σ . Then, a controller Σ_c internally stabilizes Σ if and only if the controller transfer-function $\hat{W}_{\Sigma_c}(z)$ is of the form

$$(5.11) \quad \hat{W}_{\Sigma_c}(z) = (MN_1 + Y)^{-1}(-MD_1 + X)$$

for some matrix M over RP_s .

Appendix A: Proof of Theorem 4.4. We shall require several intermediate results before we are in a position to prove Theorem 4.4. We proceed to derive these:

Let $X = \{x_1, x_2, \dots, x_k\}$ be an ordered set of vectors in \mathbb{R}^n . Define $r(X) = r$ = the number of linearly independent vectors in X . The set X will be called *well ordered* if and only if the following conditions are satisfied:

- (a) $\{x_1, x_2, \dots, x_r\}$ are linearly independent;
- (b) For $i = r+1, \dots, k$, we can write

$$x_i = \sum_{j=1}^r \alpha_{i,j}x_j \quad \text{with } |\alpha_{i,j}| \leq 1;$$

- (c) On applying the Gram-Schmidt orthogonalization procedure to $\{x_1, x_2, \dots, x_r\}$ as

$$\begin{aligned}
x_1 &= b_1, \\
x_2 &= \alpha_{2,1}b_1 + b_2, \\
&\vdots \\
x_r &= \alpha_{r,1}b_1 + \alpha_{r,2}b_2 + \cdots + \alpha_{r,r-1}b_{r-1} + b_r
\end{aligned}$$

with $\{b_1, b_2, \dots, b_r\}$ orthogonal, we have

$$|\alpha_{i,j}| \leq 1, \quad \|b_1\| \geq \|b_2\| \geq \cdots \geq \|b_r\|.$$

We first have the following result (see Poolla [1984, Appendix C]):

PROPOSITION A.1. *Every ordered set $X = \{x_1, x_2, \dots, x_k\}$ of vectors in \mathbb{R}^n has a well-ordered rearrangement*

$$\hat{X} = \{\hat{x}_1, \hat{x}_2, \dots, \hat{x}_k\}.$$

We shall adopt the following notation through the remainder of this appendix: Let F and G be $n \times n$ and $n \times m$ matrices over $l^\infty(\mathbb{Z})$, with

$$\sup_t \|F(t)\|, \quad \sup_t \|G(t)\| < M.$$

If for some t in \mathbb{Z} an $m \times n$ matrix $L(t)$ is defined, then,

$$\tilde{F}(t) := F(t) + G(t)L(t).$$

Also define, for $t_1 > t_0$, the state-transition matrix

$$\phi(t_1, t_0) := F(t_1 - 1) \cdots F(t_0 + 1)F(t_0).$$

Let $N > 0$ be a (fixed) integer and let $\varepsilon > 0$ be a (fixed) real number. For a well-ordered set X , define

$$\lambda(X) := i \quad \text{such that} \quad \|b_i\| \geq \frac{\varepsilon}{M^N} > \|b_{i+1}\|.$$

Clearly, $\lambda \leq r \leq k$. We are now in a position to prove the key.

PROPOSITION A.2. *Let t be a (fixed) integer $0 \leq t \leq N$. Let $X_t = \{x_1(t), x_2(t), \dots, x_k(t)\}$ be a set of vectors in \mathbb{R}^n . Suppose that there exist $u_i(t)$, $i = 1, 2, \dots, k$ such that*

$$(A.3) \quad \|u_i(t)\| < \hat{M}, \quad \|\phi(N, t+1)x_i(t+1)\| < \varepsilon$$

where $x_i(t+1) := F(t)x_i(t) + G(t)u_i(t)$. Then there exists an $m \times n$ matrix $L(t)$ with

$$(A.4) \quad \|L(t)\| < \frac{n^2 \hat{M} M^N}{\varepsilon}$$

and such that

$$(A.5) \quad \|\phi(N, t+1)\tilde{F}(t)x_i(t)\| < \begin{cases} \varepsilon & \text{if } \lambda(X_t) = k, \\ \varepsilon n^4 & \text{otherwise.} \end{cases}$$

Proof. Without loss of generality, assume that X_t is well ordered. Let $b_i, \alpha_{i,j}$ be as in the definition of well-ordered sets. Let $\lambda = \lambda(X_t)$. Extend $\{x_1(t), x_2(t), \dots, x_\lambda(t), b_{\lambda+1}(t), \dots, b_k(t)\}$ orthogonally by $\{b_{k+1}(t), \dots, b_n(t)\}$ to form a basis for \mathbb{R}^n . Define $L(t)$ by

$$(A.6) \quad \begin{aligned} L(t): x_i(t) &\rightarrow u_i(t), & i &= 1, 2, \dots, \lambda, \\ L(t): b_i(t) &\rightarrow 0, & i &= \lambda + 1, \lambda + 2, \dots, n. \end{aligned}$$

We first prove (A.4). Notice that

$$(A.7) \quad \left\| \sum_{i=1}^{\lambda} \beta_i b(t) \right\|^2 = \sum_{i=1}^{\lambda} \beta_i^2 \|b_i(t)\|^2 \leq \frac{\varepsilon^2}{M^{2N}} \sum_{i=1}^{\lambda} \beta_i^2.$$

We now show inductively that for $i = 1, 2, \dots, \lambda$

$$(A.8) \quad \|L(t)b_i(t)\| < 2^{i-1}\hat{M} \leq 2^n\hat{M}.$$

For $i = 1$,

$$\|L(t)b_1(t)\| = \|L(t)x_1(t)\| = \|u_1(t)\| < \hat{M},$$

and the assertion is clearly true. Assume that the assertion holds for $i = 1, 2, \dots, s$. We then have

$$\begin{aligned} \|L(t)b_{s+1}(t)\| &\leq \|L(t)x_{s+1}(t)\| + \sum_{j=1}^s |\alpha_{s+1,j}| \cdot \|L(t)b_j(t)\| \\ &\leq \|u_{s+1}(t)\| + \sum_{j=1}^s |\alpha_{s+1,j}| 2^{j-1} \hat{M}. \end{aligned}$$

But $|\alpha_{s+1,j}| \leq 1$ and $\|u_{s+1}(t)\| \leq \hat{M}$ by (A.3). The above equation then becomes

$$\|L(t)b_{s+1}(t)\| \leq \hat{M} \left(1 + \sum_{j=1}^s 2^{j-1} \right) \leq \hat{M} 2^s,$$

completing the induction.

Combining (A.7) and (A.8) we see that

$$\begin{aligned} \|L(t)\|^2 &= \sup_{\beta_1 \cdots \beta_\lambda} \frac{\left\| \sum_{i=1}^{\lambda} \beta_i L(t)b_i(t) \right\|^2}{\left\| \sum_{i=1}^{\lambda} \beta_i b_i(t) \right\|^2} \leq \sup_{\beta_1 \cdots \beta_\lambda} \frac{(\sum_{i=1}^{\lambda} |\beta_i| 2^n \hat{M})^2}{\sum_{i=1}^{\lambda} \beta_i^2} \cdot \frac{M^{2N}}{\varepsilon^2} \\ &\leq \sup_{\beta_1 \cdots \beta_\lambda} \left(\frac{2^n \hat{M} M^N}{\varepsilon} \right)^2 \frac{(\sum_{i=1}^{\lambda} |\beta_i|)^2}{\sum_{i=1}^{\lambda} \beta_i^2} \leq \left(\frac{n 2^n \hat{M} M^N}{\varepsilon^2} \right), \end{aligned}$$

proving (A.4).

We now prove (A.5). Notice first that for $i = 1, 2, \dots, \lambda$

$$\tilde{F}(t)x_i(t) = F(t)x_i(t) + G(t)L(t)x_i(t) = x_i(t+1).$$

Therefore, from (A.3),

$$(A.9) \quad \|\phi(N, t+1)\tilde{F}(t)x_i(t)\| < \varepsilon \quad \text{for } i = 1, 2, \dots, \lambda.$$

Also, for $i = \lambda + 1, \dots, r$ from (A.7),

$$x_i(t) = \alpha_{i,1}b_1(t) + \alpha_{i,2}b_2(t) + \dots + \alpha_{i,\lambda}b_\lambda(t) + c(t)$$

where $\|c(t)\| < \varepsilon/M^N$. Using (c), we can rewrite the above equation as

$$x_i(t) = \gamma_{i,1}x_1(t) + \gamma_{i,2}x_2(t) + \dots + \gamma_{i,\lambda}x_\lambda(t) + c(t),$$

where $|\gamma_{i,j}| < n$. Consequently, for $i = \lambda + 1, \dots, r$,

$$(A.10) \quad \|\phi(N, t+1)\tilde{F}(t)x_i(t)\| < \lambda n \varepsilon + \varepsilon = (\lambda n + 1)\varepsilon.$$

For $i = r+1, r+2, \dots, k$, it follows from the definition of well-ordered sets that we can write

$$x_i(t) = \sum_{j=1}^r \alpha_{i,j} n_j(t), \quad |\alpha_{i,j}| \leq 1.$$

Therefore, for $i = r+1, \dots, k$,

$$(A.11) \quad \|\phi(N, t+1) \tilde{F}(t) x_i(t)\| < \lambda \varepsilon + (r - \lambda)(\lambda n + 1) \varepsilon < n^4 \varepsilon.$$

Summarizing (A.9)–(A.11), we see that for $i = 1, 2, \dots, k$,

$$\|\phi(N, t+1) \tilde{F}(t) n_i(t)\| < \begin{cases} \varepsilon & \text{if } \lambda = k, \\ n^4 \varepsilon & \text{otherwise.} \end{cases}$$

PROPOSITION A.12. *Suppose that there exist $u_i(t)$, $i = 1, 2, \dots, n$, $t = 0, 1, \dots, N-1$ with*

$$\|u_i(t)\| < \hat{M}$$

and such that

$$\|x_i(N)\| < \varepsilon$$

where $x_i(N)$ is defined recursively by $x_i(0) = e_i$ (the i th unit vector in \mathbb{R}^n), $x_i(t+1) = F(t)x_i(t) + G(t)u_i(t)$. Then, there exist $m \times n$ matrices $L(t)$, $t = 0, 1, \dots, N-1$ and a real number Δ with

$$\|L(t)\| < \Delta$$

and such that

$$\|\tilde{F}(N-1)\tilde{F}(N-2) \cdots \tilde{F}(1)\tilde{F}(0)\| < \varepsilon n^6.$$

Proof. We first set up some notation. For a set of vectors $X = \{x_1, x_2, \dots, x_k\}$ in \mathbb{R}^n , let $\hat{X} = \{\hat{x}_1, \hat{x}_2, \dots, \hat{x}_k\}$ be any well-ordered rearrangement of X . Recall that

$$\lambda = \lambda(X) := i \quad \text{such that } \|\hat{b}_i\| \geq \frac{\varepsilon}{M^N} > \|\hat{b}_{i+1}\|,$$

and that $k = k(X)$ = the number of vectors in X . Define a set of integers

$$I_X := \{j: x_j \in \{\hat{x}_1, \hat{x}_2, \dots, \hat{x}_\lambda\}\}.$$

For $t = 0, 1, 2, \dots, N-1$ define X_t and λ_t recursively by

$$\begin{aligned} \lambda_0 &= n, & X_0 &= \{x_1(0), x_2(0), \dots, x_2(0)\}, \\ X_{t+1} &= \{x_i(t+1): i \in I_{X_t}\}, & \lambda_{t+1} &= \lambda(X_{t+1}). \end{aligned}$$

From the above definition it is clear that

$$(A.13) \quad n = \lambda_0 = k_1 \geq \lambda_1 = k_2 \geq \lambda_2 \cdots \lambda_t = k_{t+1} \geq \lambda_{t+1} \cdots \geq \lambda_{N-1} \geq 0.$$

Define f_t , $t = 0, 1, \dots, N-1$ by

$$f_t := \begin{cases} 1 & \text{if } \lambda_t = k_t, \\ n^4 & \text{otherwise.} \end{cases}$$

It is clear from (A.13) that (and this is the *key step*)

$$(A.14) \quad f_{N-1} f_{N-2} \cdots f_0 \leq n^5.$$

We now apply Proposition A.2 to X_{N-1} to conclude that there exists an $m \times n$ matrix $L(N-1)$ with $\|(N-1)\| \leq CM^N$ where $C = n2^n(\hat{M}/\varepsilon)$ and such that

$$\|\tilde{F}(N-1)x_i(N-1)\| < \varepsilon f_{N-1}, \quad x_i(N-1) \in X_{N-1}.$$

Notice that

$$\|\tilde{F}(N-1)\| \leq \|F(N-1)\| + \|G(N-1)L(N-1)\| < M + M(CM^N) \leq CM^{N+2}.$$

We can again apply Proposition A.2 to X_{N-2} (with M being replaced by CM^{N+2} and ε being replaced by εf_{N-1}) to conclude that there exists an $m \times n$ matrix $L(N-2)$ with

$$\|L(N-2)\| < (CM)^{N^2}$$

and such that

$$\|\tilde{F}(N-1)\tilde{F}(N-2)x_i(N-2)\| < \varepsilon f_{N-1}f_{N-2}, \quad x_i(N-2) \in X_{N-2}.$$

Repeating this argument N times, we conclude that there exist $m \times n$ matrices $L(0), L(1), \dots, L(N-1)$ with

$$\|L(t)\| < (CM)^{N^N} =: \Delta$$

and such that for $i = 1, 2, \dots, n$,

$$\|\tilde{F}(N-1)\tilde{F}(N-2) \cdots \tilde{F}(0)e_i\| < \varepsilon f_{N-1}f_{N-2} \cdots f_0.$$

The above equation, along with (A.14), gives us

$$\|\tilde{F}(N-1)\tilde{F}(N-2) \cdots \tilde{F}(0)\| < n^6 \varepsilon,$$

completing the proof. \square

We are finally in a position to prove Theorem 4.4, which is restated below for convenience.

THEOREM 4.4. *Let $\Sigma = (F, G, H)$ be a linear time-varying system over $l^\infty(\mathbb{Z})$. Then Σ is asycontrollable if and only if Σ is stabilizable.*

Proof. We have already shown (see discussion preceding Theorem 4.4) that stabilizability implies asycontrollability. We now prove the converse.

Suppose Σ is asycontrollable. Choose in the Definition 4.1 of asycontrollability, $\varepsilon = \frac{1}{2}n^{-6}$. Then, by Proposition A.12, there exist matrices $L(t)$, $t = 0, 1, \dots, N-1$ such that

$$\|L(t)\| < \Delta, \quad \|\tilde{F}(N-1)\tilde{F}(N-2) \cdots \tilde{F}(1)\tilde{F}(0)\| < \frac{1}{2}$$

where $\tilde{F}(t) = F(t) + G(t)L(t)$. Recall that here N is the number of steps required to drive all states to energy $< \varepsilon$ as in Definition 4.1 of asycontrollability. Since all our arguments are independent of initial time, we can (again by Proposition A.12) find matrices $L(t)$ for all t in \mathbb{Z} such that $\|L(t)\| < 2\Delta n^6$ and

$$\|\tilde{F}(kN-1)\tilde{F}(kN-2) \cdots \tilde{F}(kN-N+1)\tilde{F}(kN-N)\| < \frac{1}{2}$$

for all k in \mathbb{Z} . Thus the matrix $L(t)$ viewed as being over A is bounded (i.e., in $l^\infty(\mathbb{Z})$), and $(zI - F - GL)^{-1}$ is a stable power series. This completes the proof. \square

REFERENCES

- B. D. O. ANDERSON AND J. B. MOORE (1981), *Detectability and stabilizability of time-varying discrete-time linear systems*, this Journal, 19, pp. 20-32.
 W. ARVESON (1975), *Interpolation problems in nest algebras*, J. Funct. Anal., 20, pp. 208-233.
 L. CESARI (1963), *Asymptotic Behavior and Stability Problems in Ordinary Differential Equations*, Academic Press, New York.

- A. FEINTUCH AND B. FRANCIS (1984), *Uniformly optimal control of linear time-varying systems*, Systems Control Lett., 5, pp. 67-71.
- C. A. DESOER, R. W. LIU, J. MURRAY AND R. SAEKS (1980), *Feedback system design: The fractional representation approach to analysis and synthesis*, IEEE Trans. Automat. Control, AC-25, pp. 399-412.
- B. FRANCIS, J. W. HELTON AND G. ZAMES (1984), *H^∞ -optimal feedback controllers for linear multivariable systems*, IEEE Trans. Automat. Control, AC-29, pp. 888-900.
- B. FRANCIS AND G. ZAMES (1984), *On H^∞ -optimal sensitivity theory for SISO feedback systems*, IEEE Trans. Automat. Control, AC-29, pp. 9-16.
- M. FREEDMAN AND G. ZAMES (1968), *Logarithmic variation criteria for the stability of systems with time-varying gains*, this Journal, 6, pp. 487-507.
- W. L. GREEN AND E. W. KAMEN, *On stability of linear difference equations with time-varying coefficients*, 1984, preprint.
- E. W. KAMEN (1982), *Linear systems with commensurate time delays: Stability and stabilization independent of delay*, IEEE Trans. Automat. Control, AC-27, pp. 367-375.
- E. W. KAMEN, P. P. KHARGONEKAR AND K. POOLLA (1985), *A transfer function approach to linear time-varying discrete-time systems*, this Journal, 23, pp. 550-565.
- P. P. KHARGONEKAR AND K. POOLLA (1986), *On polynomial matrix fraction representations for linear time-varying systems*, Linear Algebra Appl., 80, pp. 1-37.
- P. P. KHARGONEKAR AND E. D. SONTAG (1982), *On the relation between stable matrix fraction factorization and regulable realizations of linear systems over rings*, IEEE Trans. Automat. Control, AC-27, pp. 627-638.
- C. N. NETT, C. J. JACOBSEN AND M. J. BALAS (1984), *A connection between state space and doubly coprime fractional representations*, IEEE Trans. Automat. Control, AC-29, pp. 831-832.
- K. POOLLA (1984), *Linear-time-varying systems: Representations and control via transfer function matrices*, Ph.D. thesis, Univ. Florida, Gainesville.
- R. SAEKS AND J. MURRAY (1981), *Feedback system design: The tracking and disturbance rejection problems*, IEEE Trans. Automat. Control, AC-26, pp. 203-217.
- M. VIDYASAGAR (1978), *On the use of right-coprime factorizations in distributed feedback systems containing unstable subsystems*, IEEE Trans. Automat. Control, AC-29, pp. 916-921.
- (1985), *Control System Synthesis: A Factorization Approach*, MIT Press, Cambridge, MA.
- L. WEISS (1972), *Controllability, realization, and stability of discrete-time systems*, this Journal, 10, pp. 230-251.
- J. L. WILLEMS (1970), *Stability Theory of Dynamic Systems*, Nelson, London, 1970.
- L. A. ZADEH (1950), *Frequency analysis of variable networks*, Proc. IRE, 38, pp. 291-299.
- G. ZAMES AND B. FRANCIS (1983), *Feedback, minimax sensitivity, and optimal robustness*, IEEE Trans. Automat. Control, AC-28, pp. 585-601.

DIFFUSION FOR GLOBAL OPTIMIZATION IN \mathbb{R}^n *

TZUU-SHUH CHIANG†, CHII-RUEY HWANG† AND SHUENN-JYI SHEU†

Abstract. We seek a global minimum of $U: \mathbb{R}^n \rightarrow \mathbb{R}$. The solution to $(*) (d/dt)X(t) = -\nabla U(X(t))$ will find local minima. Using the idea of simulated annealing, we consider the diffusion process, $dX(t) = -\nabla U(X(t)) dt + \sigma(t) dW(t)$, $X(0) = x$, where $W(\cdot)$ is the n -dimensional standard Brownian motion and $\frac{1}{2}\sigma^2(t)$ is the annealing rate which decreases to zero as t goes to ∞ . Under suitable condition on $U(x)$, we prove that $X(t)$ converges weakly to a probability measure π if for large t , $\sigma^2(t) = c/\log t$ with $c > c_0$, where c_0 has a simple expression involving the action function of the dynamical system $(*)$, π concentrates on the global minima of U and is the weak limit of the Gibbs densities $\pi_t(x) \propto \exp(-2U(x)/\sigma^2(t))$.

The above result can also be formulated as follows: consider the Fokker-Planck equation (forward equation)

$$\frac{\partial}{\partial t} V(t, y) = \frac{1}{2} \sigma^2(t) \Delta V(t, y) + \nabla \cdot (V(t, y) \nabla U(y))$$

with $V(0, y) = \delta_x(y)$.

If $\sigma^2(t) = c/\log t$ for large t and $c > c_0$, then $V(t, y) \rightarrow \pi$ weakly.

Key words. diffusion, global optimization, simulated annealing, perturbed dynamical system, large deviation, action functional

AMS(MOS) subject classifications. GOH10, GOJ70

1. Introduction. For a fixed $U: \mathbb{R}^n \rightarrow [0, \infty)$, we give suitable conditions on U such that by choosing

$$\sigma^2(t) = \frac{c}{\log t} \quad \text{for large } t \text{ with } c > c_0 \quad \text{as } t \rightarrow \infty$$

$p(s, x, t, \cdot)$ converges weakly to a probability measure π concentrating on the global minima of U , $p(s, x, t, \cdot)$ is the transition probability of the diffusion process defined by

$$(1.1) \quad dZ(t) = -\nabla U(Z(t)) dt + \sigma(t) dW(t),$$

where $\frac{1}{2}\sigma^2(t)$ corresponding to the "temperature" is the annealing rate, $W(t)$ is a standard Brownian motion in \mathbb{R}^n . The probability π is the weak limit of the Gibbs density

$$(1.2) \quad \pi_t(x) \propto \exp\left(-\frac{2U(x)}{\sigma^2(t)}\right) \quad \text{as } t \rightarrow \infty.$$

The constant c_0 , which will be defined in § 2, has a simple expression involving the action function of the dynamical system

$$(1.3) \quad \frac{dY(t)}{dt} = -\nabla U(Y(t)).$$

The idea of our approach is as follows: Heuristically if we hold the temperature at time s for a *fairly large* amount of time, then $Z(t)$ defined by (1.1) and the fixed temperature process behaves almost the same at the end of that time interval. Hence, instead of (1.1) we may consider

$$(1.4) \quad \begin{aligned} dX(t) &= -\nabla U(X(t)) dt + \sigma(s) dW(t), \\ X(0) &= x. \end{aligned}$$

* Received by the editors October 23, 1985; accepted for publication (in revised form) April 16, 1986.

† Institute of Mathematics, Academia Sinica, Taipei, Taiwan.

Note that the weak limit π depends only on the local property of U near the minima [9]. If we modify U for large $|x|$, π remains unchanged. One may consider a modified version with $U(X) = |x|^4$ for large $|x|$. In this case $X(t)$ comes back from “infinity” to a fixed finite ball in a finite time which is independent of $\sigma(s)$. It is almost as in the compact situation. Some of the ideas used in [4], which dealt with a reflected version of (1.1), can be used again in here. Furthermore, results and ideas in [13], [14] are available when we consider (1.4).

Independently, Gidas and Kushner also consider (1.1) in their recent works [6], [11], respectively.

Our work was inspired by the “simulated annealing” [1], [10] which deals mainly with the discrete state space. A lot of research has been going on in this aspect, see e.g. [3], [5], [8].

The use of (1.1) as a global minimization algorithm is motivated by problems in imaging processing [4], [7] as well as in studying lattice gauge theory [12].

We think that the constant c_0 obtained here is not the best possible. One may argue heuristically as follows. For the fixed temperature process (1.4) with $\varepsilon = \sigma(s)$, Lemma 3 in § 3 describes a distance between p_t and π^ε . Let $L_\varepsilon = \frac{1}{2}\varepsilon^2\Delta - \nabla U \cdot \nabla$ and $\lambda_2(\varepsilon)$ denote the second eigenvalue of L_ε . Let $\|\cdot\|_{\pi^\varepsilon}$ denote the norm of $L^2(\pi^\varepsilon)$; then clearly

$$\|p_t^\varepsilon(x, f) - \pi^\varepsilon(f)\|_{\pi^\varepsilon} \leq \exp(t\lambda_2(\varepsilon))\|f\|_{\pi^\varepsilon}.$$

If $\lim_{\varepsilon \rightarrow 0} \varepsilon^2 \log(-\lambda_2(\varepsilon)) = -c_1$, then for $c > c_1$ such that $c > c_1 + a$ we have $-\lambda_2(\varepsilon) \geq \exp(-(c_1 + a)/\varepsilon^2)$ for small ε . For $\varepsilon^2 \approx c/\log t$

$$\|p_t^\varepsilon(x, f) - \pi^\varepsilon(f)\|_{\pi^\varepsilon} \leq \exp(-t^{1-((c_1+a)/c)}) \rightarrow 0 \quad \text{as } t \rightarrow \infty.$$

One would expect c_1 here is the critical constant.

Another heuristic approach is to consider the function

$$N(t) = \int \left| \frac{p(0, x, t, y)}{\pi_t(y)} - 1 \right|^2 \pi_t(y) dy, \quad t > 1,$$

which was discussed previously in [4]. If $N(t) \rightarrow 0$ as $t \rightarrow \infty$, then it is easy to see that $p(0, x, t, \cdot) \rightarrow \pi(\cdot)$ weakly. For simplicity, let us write $\sigma^2(t) = 2T(t)$ and heuristically one has

$$\begin{aligned} \frac{dN(t)}{dt} &= \left(\frac{d}{dt} \frac{1}{T(t)} \right) \int \frac{1}{\pi_t(y)} (U(y) - \pi_t(U)) p(0, x, t, y)^2 dy \\ &\quad - 2T(t) \int \left| \nabla_y \left(\frac{p(0, x, t, y)}{\pi_t(y)} \right) \right|^2 \pi_t(y) dy \\ &\leq \frac{c_2}{t} (N(t) + 1) - 2(-\lambda_2(\sigma(t))) N(t) \\ &= \frac{c_2}{t} + N(t) \left(\frac{c_2}{t} - 2t^{-(c_1+a)/c} \right) \end{aligned}$$

by

$$\int |f(y) - \pi_t(f)|^2 \pi_t(y) dy \leq \frac{T(t)}{-\lambda_2(\sigma(t))} \int |\nabla f(y)|^2 \pi_t(y) dy.$$

Then one can establish $N(t) \rightarrow 0$ from this differential inequality.

2. Statement of result. Let U be a twice continuously differentiable function from \mathbb{R}^n to $[0, \infty)$ such that the following assumptions hold:

- $$\min_{x \in \mathbb{R}^n} U(x) = 0,$$
- (A1) $U(x) \rightarrow \infty$ and $|\nabla U(x)| \rightarrow \infty$ as $|x| \rightarrow \infty$,
- $$\lim_{|x| \rightarrow \infty} |\nabla U(x)|^2 - \Delta U(x) > -\infty.$$
- For $0 < \varepsilon < 1$,
- (A2) $\pi^\varepsilon(x) := \frac{1}{c(\varepsilon)} \exp\left(-\frac{2U(x)}{\varepsilon^2}\right),$
- where $c(\varepsilon) = \int_{\mathbb{R}^n} \exp\left(-\frac{2U(x)}{\varepsilon^2}\right) dx < \infty.$
- (A3) π^ε has a unique weak limit π as $\varepsilon \downarrow 0$.

Clearly π concentrates on the global minima of U . The detailed discussion for the existence of π and its characterization in terms of the Hessian of U can be found in [9].

For simplicity we shall assume $\sigma^2(t) < 1$, $\sigma^2(t) = c/\log t$ for large t and the process $Z(t)$ starts at $Z(0) = x$.

Let S denote the set of all stationary points of U , i.e., $S = \{x | \nabla U(x) = 0\}$.

For any $\eta > 0$, $\xi > 0$, we define the following:

$$S(\eta) := \{x | d(x, S) < \eta\},$$

$$K(\eta) := \text{the set containing all the solutions of the dynamical system (1.3) with starting points in } S(\eta),$$

$$K(\eta, \xi) := \{x | d(x, K(\eta)) \leq \xi\},$$

$$I(t, x, y) := \inf_{\substack{\psi(0)=x \\ \psi(t)=y}} \frac{1}{2} \int_0^t |\dot{\psi}(s) + \nabla U(\psi(s))|^2 ds,$$

$$J(t, \eta, \xi) := \sup_{x, y \in K(\eta, \xi)} (I(t, x, y) - 2U(y)),$$

$$J(\eta, \xi) := \overline{\lim}_{t \rightarrow \infty} J(t, \eta, \xi),$$

$$c_0 := \frac{3}{2} \inf_{\eta} (\inf_{\xi} J(\eta, \xi)).$$

For a measure μ , $\mu(f) := \int f d\mu$.

THEOREM. Assume (A1), (A2) and (A3) and $c > c_0$; then for any bounded continuous function f

$$p(0, x, t, f) \rightarrow \pi(f) \quad \text{as } t \rightarrow \infty$$

and the convergence is uniform for x in a compact set. $p(s, x, t, \cdot)$ here is the transition probability of (1.1).

Remark 1. Without going into detail, we note that $J(\eta, \xi)$ is independent of η, ξ and

$$\begin{aligned} c_* &= J(\eta, \xi) = \sup_{x, y \in K(\eta, \xi)} (V(x, y) - 2U(y)) \\ &= \sup_{x, y \in S} (V(x, y) - 2U(y)), \end{aligned}$$

where $V(x, y) = \lim_{t \rightarrow \infty} I(t, x, y)$. This $V(x, y)$ is the same function used by Freidlin and Wentzell for describing the long time behavior of perturbed dynamical systems $dX(t) = -\nabla U(X(t)) dt + \varepsilon dw(t)$.

Remark 2. We suspect that $c > c_* = \frac{2}{3} c_0$ is enough for the result of the theorem to hold.

3. Proof of theorem. The proof of the main theorem is based on the following three lemmas.

LEMMA 1. $\lim_{t \rightarrow \infty} p(s, x, t, K(\eta, \xi)) = 1$. The convergence is uniform for x in a compact set.

LEMMA 2. Consider a family of processes defined by

$$\begin{aligned} (3.1) \quad dY(s, t) &= -\nabla U(Y(s, t)) dt + \sigma(s) dW(t), \\ Y(s, 0) &= y. \end{aligned}$$

Then for $h(s) \leq s^{2/3}$ and $h(s)$ increasing to ∞ ,

$$\lim_{s \rightarrow \infty} E_{0,y}(f(Y(s, h(s)))) - E_{s,y}(f(Z(\beta(s)))) = 0,$$

where $\beta(\cdot)$ is defined by

$$\int_s^{\beta(s)} \frac{\log s}{\log u} du = h(s).$$

And the convergence is uniform for y in a compact set.

LEMMA 3. Consider the following process

$$\begin{aligned} (3.2) \quad dX(t) &= -\nabla U(X(t)) + \varepsilon dW(t), \\ X(0) &= x. \end{aligned}$$

Then there exist $T_0 > 0 \ni \forall M > 0, \forall T > 2T_0, \forall \alpha > 0$

$$\overline{\lim}_{\varepsilon \rightarrow 0} |E_x^\varepsilon f(X(mT)) - \pi^\varepsilon(f)| \leq 4e^{-M} \|f\|,$$

where

$$m = M \exp\left(\frac{1}{\varepsilon^2}(J(t, \eta, \xi) + \alpha)\right), \quad t = T - 2T_0,$$

α is an arbitrary fixed positive constant. The convergence is uniform for x in a compact set.

Assuming the validity of these, we establish the theorem as follows: For a fixed $c > c_0$, there exists an $\alpha > 0$ such that for sufficiently large time t , sufficiently small η and ξ ,

$$(3.3) \quad c > \frac{3}{2}(J(t, \eta, \xi) + \alpha).$$

Choose a fixed large T such that (3.3) holds for time $T - 2T_0$, where T_0 is the constant in Lemma 3.

Choose $h(s)$ in Lemma 2 as

$$\begin{aligned} h(s) &= MT \exp \left(\frac{1}{\sigma^2(s)} (J(T - 2T_0, \eta, \xi) + \alpha) \right) \\ (3.4) \quad &= MT s^{(J(T - 2T_0, \eta, \xi) + \alpha)/c} \\ &< s^{2/3} \quad \text{for large } s. \end{aligned}$$

Note that h and β are strictly increasing functions and $s + h(s) \leq \beta(s) \leq s + 2h(s)$. Hence for $t \gg 1$, one can choose s such that $t = \beta(s)$. Clearly $s < t$ and $s \rightarrow \infty$.

$$\begin{aligned} p(0, x, t, f) - \pi_s(f) &= \int p(0, x, s, y) p(s, y, t, f) dy - \pi_s(f) \\ &= \int_{y \in K(\eta, \xi)} p(0, x, s, y) (p(s, y, t, f) - \pi_s(f)) dy \\ &\quad + \int_{y \notin K(\eta, \xi)} p(0, x, s, y) (p(s, y, t, f) - \pi_s(f)) dy. \end{aligned}$$

The second term is bounded by

$$2\|f\|(1 - p(0, x, s, K(\eta, \xi))),$$

which goes to zero uniformly over x in a compact set as $s \rightarrow \infty$ by Lemma 1. Note that $\pi_s(f) \rightarrow \pi(f)$.

By Lemma 2,

$$\begin{aligned} E_{0,y}(f(Y(s, h(s)))) - p(s, y, \beta(s), f) &\rightarrow 0, \\ E_{0,y}(f(Y(s, h(s)))) &= E_y^{\sigma(s)}(f(X(h(s)))) \\ &= E_y^{\sigma(s)}(f(X(mT))) \end{aligned}$$

by identifying $h(s)$ with mT and $\sigma(s)$ with ε .

Now by Lemma 3, we have the theorem.

4. Proof of Lemma 1. Let us first assume the validity of the following two lemmas.

LEMMA 4.1. For any compact set K in \mathbb{R}^n , the family of probability measures

$$\{p(s, x, t, \cdot) | s < t, x \in K\}$$

is tight.

LEMMA 4.2. For any compact set K , there exists T such that for any $t > T$, $Y(t) \in K(\eta)$, where

$$\frac{dY(t)}{dt} = -\nabla U(Y(t)), \quad Y(0) = y \in K.$$

The proof of Lemma 1 is as follows: By Lemma 4.1, for any $\delta > 0$ and for any given compact set J , there exists a compact set K such that

$$p(s, x, t, K) > 1 - \delta/2 \quad \text{for all } s < t, x \in J.$$

Choose T as in Lemma 4.2, then

$$\begin{aligned} p(s, x, t, K(\eta, \xi)) &= \int p(s, x, t - T, dy) p(t - T, y, t, K(\eta, \xi)) \\ &> \int_K p(s, x, t - T, dy) p(t - T, y, t, K(\eta, \xi)). \end{aligned}$$

It remains to show that there exists t_0 such that

$$p(t-T, y, t, K(\eta, \xi)) > 1 - \delta/2, \quad y \in K, \quad t > t_0.$$

Let $Y(\cdot)$ be the solution of (4.1) with $Y(t-T) = y$. Then by Lemma 4.2,

$$\begin{aligned} p(t-T, y, t, K(\eta, \xi)) &= E_{t-T, y} \{Z(t) \in K(\eta, \xi)\} \\ &= E_{t-T, y} \{|Z(t) - Y(t)| \leq \xi\} \\ &\quad + E_{t-T, y} \{|Z(t) - Y(t)| > \xi, Z(t) \in K(\eta, \xi)\} \\ &\geq E_{t-T, y} \{|Z(t) - Y(t)| \leq \xi\} \\ &\geq 1 - E_{t-T, y} \{\tau \leq t\}, \end{aligned}$$

where $\tau := \inf \{s > t-T, |Z(s) - Y(s)| > \xi\}$.

Now consider the process $Z(t)$ starting at $Z(t-T) = y$. Compare $Z(t)$ and $Y(t)$ up to τ . For $u \leq \tau$,

$$Z(u) - Y(u) = \int_{t-T}^u (-\nabla U(Z(s)) + \nabla U(Y(s))) ds + H(u),$$

where $H(u) = \int_{t-T}^u \sigma(s) dW(s)$. Note that for $t-T \leq s \leq \tau$, $Z(s)$ and $Y(s)$ are in a compact set in which U is Lipschitz with constant d , and we have

$$|Z(u) - Y(u)| \leq d \int_{t-T}^u |Z(s) - Y(s)| ds + |H(u)|.$$

By Gronwall inequality,

$$|Z(u) - Y(u)| \leq \exp(d(u - (t-T))) \sup_{t-T \leq s \leq u} |H(s)|.$$

For $\tau \leq t$,

$$\begin{aligned} \xi = |Z(\tau) - Y(\tau)| &\leq e^{dT} \sup_{t-T \leq s \leq t} |H(s)|, \\ p\{\tau \leq t\} &\leq p\left\{\sup_{t-T \leq s \leq t} |H(s)| \geq e^{-dT} \xi\right\} \\ &\leq 2n \exp\left\{\frac{-\xi^2 \log t}{2cnT} e^{-2dT}\right\} \\ &\leq \frac{\delta}{2} \quad \text{if } t \geq t_0 \text{ for a fixed large } t_0 \end{aligned}$$

[15, p. 87]. Hence,

$$p(t-T, y, t, K(\eta, \xi)) \geq 1 - \frac{\delta}{2}.$$

This completes the proof.

Proof of Lemma 4.1.

$$\begin{aligned} de^{U(Z(t))} e^{\lambda t} &= \left(\frac{\sigma^2(t)}{2} \Delta U(Z(t)) - \left(1 - \frac{\sigma^2(t)}{2}\right) |\nabla U(Z(t))|^2 + \lambda\right) e^{\lambda t} e^{U(Z(t))} dt \\ &\quad + e^{\lambda t} dM(t), \end{aligned}$$

where $M(t) = \int_0^t \sigma(s) \nabla U(Z(s)) e^{U(Z(s))} dW(s)$ is a local martingale.

For any $\lambda > 0$, there exists constant $A = A(\lambda) > 0$ such that

$$\begin{aligned} & \left(\frac{\sigma^2(t)}{2} \Delta U(z) - \left(1 - \frac{\sigma^2(t)}{2} \right) |\nabla U(z)|^2 + \lambda \right) e^{U(z)} \\ &= \left[\frac{\sigma^2(t)}{2} (\Delta U(z) - |\nabla U(z)|^2) - (1 - \sigma^2(t)) |\nabla U(z)|^2 + \lambda \right] e^{U(z)} \\ &\leq A \quad \forall t \text{ and } z \in \mathbb{R}^n, \end{aligned}$$

since for large $|z|$, the term in the bracket parentheses is negative for all t .

Let $\tau_m := \inf \{t; |Z(t)| > m\}$ and $\tau = \lim_{m \rightarrow \infty} \tau_m$ is the explosion time.

Then

$$E_{s,x} \{ e^{U(Z(t\wedge\tau_m))} e^{\lambda(t\wedge\tau_m)} \} \leq A E_{s,x} \left\{ \int_s^{t\wedge\tau_m} e^{\lambda u} du + e^{U(x)} e^{\lambda s} \right\}.$$

Let $m \rightarrow \infty$,

$$E_{s,x} \{ e^{U(Z(t\wedge\tau))} e^{\lambda(t\wedge\tau)} \} \leq \frac{A}{\lambda} (e^{\lambda t} - e^{\lambda s}) + e^{U(x)} e^{\lambda s}.$$

If $p\{\tau \leq \infty\} > 0$, then there exists t such that $E_{s,x} e^{U(Z(t\wedge\tau))} e^{\lambda(t\wedge\tau)} = \infty$. Hence we conclude that $p\{\tau = \infty\} = 1$.

Now we have

$$\begin{aligned} E_{s,x} e^{U(Z(t))} &\leq \frac{A}{\lambda} + e^{U(x)} e^{-\lambda(t-s)} \\ &\leq \frac{A}{\lambda} + e^{U(x)}. \end{aligned}$$

From this, it is easy to show that $\{p(s, x, t, \cdot), s < t, x \in K\}$ is tight.

Proof of Lemma 4.2.

$$(4.1) \quad U(Y(t)) - U(y) = - \int_0^t |\nabla U(Y(s))|^2 ds.$$

For $z \notin S(\eta)$, there exists $\nu > 0$ independent of z such that $U(z) > \nu$ and $|\nabla U(z)| > \nu$. Hence by (4.1) and the compactness of K , there exists T such that $Y(t) \in S(\eta)$ for some $t \leq T$. But by the definition of $K(\eta)$, once $Y(t) \in S(\eta) \subseteq K(\eta)$, then $Y(t') \in K(\eta)$ if $t' > t$. Therefore, $Y(t) \in K(\eta)$ if $t \geq T$.

5. Proof of Lemma 2. For simplicity, we shall write $b = -\nabla U$. Define $\beta(s, t)$ by

$$\int_s^{\beta(s,t)} \frac{\sigma^2(u)}{\sigma^2(s)} du = t.$$

Note that $\beta(s)$ defined in the statement is $\beta(s, h(s))$. For any fixed s , define $\tilde{Z}(s, t) = Z(\beta(s, t))$; then

$$\tilde{Z}(s, t) = x + \int_0^t b(\tilde{Z}(s, u)) \frac{\log \beta(s, u)}{\log s} du + \sigma(s) W(t).^1$$

¹ The Wiener process $W(t)$ may not be the same at each occurrence. This does not matter because we are only interested in the probability distributions.

Now compare $\tilde{Z}(s, \cdot)$ with $Y(s, \cdot)$,

$$Y(s, t) = x + \int_0^t b(Y(s, u)) du + \sigma(s) W(t).$$

Let us first consider $|b(x)| \leq M < \infty$. By the Girsanov theorem,

$$Ef(\tilde{Z}(s, t)) = E(f(Y(s, t)) \exp(A(t) - \frac{1}{2}B(t))),$$

where

$$\begin{aligned} A(t) &= \int_0^t b(Y(s, u)) \left(\frac{\log \beta(s, u)}{\log s} - 1 \right) \frac{1}{\sigma(s)} dW(u), \\ B(t) &= \int_0^t |b(Y(s, u))|^2 \left(\frac{\log \beta(s, u)}{\log s} - 1 \right)^2 \frac{1}{\sigma^2(s)} du, \\ Ef(\tilde{Z}(s, t)) &= Ef(Y(s, t)) + E \left\{ f(Y(s, t)) \left(\exp \left(A(t) - \frac{1}{2}B(t) \right) - 1 \right) \right\}. \end{aligned}$$

We shall show that the second term tends to zero for $t = h(s) \leq s^{2/3}$ as $s \rightarrow \infty$.

$$\begin{aligned} E(\exp(A(t) - \frac{1}{2}B(t)) - 1)^2 &= E(\exp(2A(t) - B(t)) - 1) \\ (5.1) \qquad \qquad \qquad &= E(\exp(2A(t) - 2B(t))(\exp B(t) - 1)), \end{aligned}$$

since $\exp(A(t) - \frac{1}{2}B(t))$ and $\exp(2A(t) - 2B(t))$ are martingales with expectation 1.

$$\begin{aligned} B(t) &= \int_0^t |b(Y(s, u))|^2 \left(\frac{\log \beta(s, u)}{\log s} - 1 \right)^2 \frac{1}{\sigma^2(s)} du \\ &\leq \frac{M^2}{c} \log s \int_0^t \left(\frac{\log \beta(s, u)}{\log s} - 1 \right)^2 du \\ &= \frac{M^2}{c} \log s \int_s^{\beta(s, t)} \left(\frac{\log u}{\log s} - 1 \right)^2 \frac{\log s}{\log u} du \\ &\leq \text{constant} \frac{1}{\log s} \int_s^{\beta(s, t)} \left(\frac{u}{s} - 1 \right)^2 du \\ &= \text{constant} \frac{1}{\log s} \frac{(\beta(s, t) - s)^3}{s^2} \\ &\leq \text{constant} \frac{1}{\log s} \rightarrow 0, \end{aligned}$$

since $s + 2t \geq \beta(s, t) \geq s + t$ and we choose $t = h(s) \leq s^{2/3}$. Then (5.1) is bounded by

$$\text{constant} \frac{1}{\log s} E(\exp(2A(t) - 2B(t))) = \text{constant} \frac{1}{\log s} \rightarrow 0.$$

Therefore for bounded $b(x)$, we have proved

$$(5.2) \qquad E_{s,x} f(Z(\beta(s))) - E_{0,x} f(Y(s, h(s))) \rightarrow 0.$$

Now let us prove the lemma for the general case. Let

$$\begin{aligned} \tau_r &= \inf \{t: U(Z(t)) > r\}, \\ \tau_r(s) &= \inf \{t: U(Y(s, t)) > r\}. \end{aligned}$$

Using the same argument as before by taking f an indicator function and noticing that b is bounded on the compact set $\{U(x) \leq r\}$, we can show that as $s \rightarrow \infty$,

$$(5.3) \quad E_{s,x}\{\tau_r > \beta(s)\} - E_{0,x}\{\tau_r(s) > h(s)\} \rightarrow 0.$$

If there exists r such that

$$(5.4) \quad E_{0,x}\{\tau_r(s) > h(s)\} \rightarrow 1 \text{ uniformly over } x \text{ in a compact set,}$$

then by combining (5.2) for bounded b and (5.3), one gets Lemma 2.

As for (5.4), it is an easy consequence of Lemma 6.4.

6. Proof of Lemma 3.

Super normal case. Let us first prove Lemma 3 for the following particular super normal case: there is a large fixed R_0 , such that

$$(6.1) \quad \begin{aligned} U(x) &= |x|^4 \quad \text{for } |x| > R_0; \quad \text{then} \\ |\nabla U(x)| &= 4|x|^3, \quad \Delta U(x) = (4n+8)|x|^2, \end{aligned}$$

and $K(\eta, \xi) \subseteq \{|x| < R_0\}$.

$$(6.2) \quad \begin{aligned} dX(t) &= -\nabla U(X(t)) dt + \varepsilon dW(t), \\ X(0) &= x. \end{aligned}$$

Let $\tau = \inf\{t \mid |X(t)| = 2R_0\}$.

CLAIM. There exists a constant c_1 such that for any $|x| > 2R_0$, for any $0 < \varepsilon < 1$, $E_x^\varepsilon(\tau) \leq c_1$.

Proof. For $|x| > 2R_0$, $\tau_0 := \inf\{t \mid |X(t)| = \frac{1}{2}|x|\}$, then

$$\begin{aligned} E_x^\varepsilon U(X(\tau_0)) - U(x) &= E_x^\varepsilon \int_0^{\tau_0} \left(-|\nabla U(X(s))|^2 + \frac{\varepsilon^2}{2} \Delta U(X(s)) \right) ds. \\ \left| \frac{1}{2}x \right|^4 - |x|^4 &= E_x^\varepsilon \int_0^{\tau_0} \left(-16|X(s)|^6 + \frac{\varepsilon^2}{2n+4} |X(s)|^2 \right) ds \\ &\leq -c_3 |x|^6 E_x^\varepsilon \tau_0. \end{aligned}$$

Therefore,

$$E_x^\varepsilon \tau_0 \leq c_4 |x|^{-2}.$$

Now let us define the following stopping times:

$$\begin{aligned} \tau_1 &= \inf\{t \mid |X(t)| = \frac{1}{2}|x|\}, \\ \tau_2 &= \inf\{t > \tau_1 \mid |X(t)| = \frac{1}{2}|X(\tau_1)|\}, \\ &\vdots \\ \tau_{i+1} &= \inf\{t > \tau_i \mid |X(t)| = \frac{1}{2}|X(\tau_i)|\}. \end{aligned}$$

Let m be a positive integer such that

$$2^m R_0 < |x| \leq 2^{m+1} R_0.$$

Then, $\tau \leq \tau_m$ and

$$\begin{aligned}
 E_x^\varepsilon(\tau) &\leq \sum_{k=2}^m E_x^\varepsilon(\tau_k - \tau_{k-1}) + E_x^\varepsilon(\tau_1) \\
 &= \sum_{k=2}^m E_x^\varepsilon E_{x(\tau_{k-1})}^\varepsilon(\tau_0) + E_x^\varepsilon(\tau_1) \\
 &\leq c_4 \sum_{k=2}^m E_x^\varepsilon |X(\tau_{k-1})|^{-2} + c_4 |x|^{-2} \\
 &\leq c_4 R_0^{-2} \sum_{k=1}^m (2^{m-k+1})^{-2} \leq \frac{1}{3} c_4 R_0^{-2} = c_1.
 \end{aligned}$$

CLAIM. For any $\delta > 0$ there exist T_0 and ε_0 such that

$$(6.3) \quad E_x^\varepsilon\{X(T_0) \in K(\eta, \xi)\} \geq 1 - \delta \quad \text{for all } x \in \mathbb{R}^n \text{ and } \varepsilon \leq \varepsilon_0.$$

Proof. First choose T_2 such that $c_1/T_2 < \delta/2$.

$$B(2R_0) = \{|x| \leq 2R_0\} \supset K(\eta, \xi).$$

T_1 is the time in Lemma 4.2 such that with initial point in $B(2R_0)$ the solution of the dynamic system will be contained in $K(\eta)$ after time T_1 . Now let $T_0 = T_1 + T_2$.

As in the proof of Lemma 1, we can choose an ε_0 such that

$$E_x^\varepsilon\{X(t) \in K(\eta, \xi)\} > 1 - (\delta/2) \quad \forall x \in B(2R_0), \quad \forall T_1 \leq t \leq T_0, \quad \forall \varepsilon \leq \varepsilon_0.$$

Now for any $x \in \mathbb{R}^n$,

$$\begin{aligned}
 E_x^\varepsilon\{X(T_0) \in K(\eta, \xi)\} &\geq E_x^\varepsilon\{E_{X(\tau)}^\varepsilon\{X(T_0 - \tau) \in K(\eta, \xi)\}, \tau \leq T_2\} \\
 &\quad (\text{for } \tau \leq T_2, T_1 \leq T_0 - \tau \leq T_0, \text{ and } X(\tau) \in B(2R_0)) \\
 &\geq \left(1 - \frac{\delta}{2}\right) E_x^\varepsilon\{\tau \leq T_2\} \\
 &\geq \left(1 - \frac{\delta}{2}\right) \left(1 - \frac{c_1}{T_2}\right) > 1 - \delta.
 \end{aligned}$$

LEMMA 6.1. Let $p_t^\varepsilon(x, y)$ denote the transition density of (6.2) and define

$$q_t^\varepsilon(x, y) \pi^\varepsilon(y) = p_t^\varepsilon(x, y).$$

Then for any x_0, y_0 in \mathbb{R}^n , $\varepsilon \leq \varepsilon_0$, $t > 0$,

$$q_{t+2T_0}^\varepsilon(x_0, y_0) \geq \inf_{x, y \in K(\eta, \xi)} q_t^\varepsilon(x, y) (1 - \delta)^2,$$

the relation between δ , T_0 , ε_0 is the same as in (6.3). And one may take any fixed δ , say $\delta = \frac{1}{2}$.

Proof. For $\varepsilon < 1$, by a similar argument as in Lemma 4.1, $X(t)$ has no explosion. By the Girsanov theorem it is obvious that $X(t)$ has transition densities.

Since the infinitesimal generator $(\varepsilon^2/2)\Delta - \nabla U \cdot \nabla$ is self-adjoint in the weighted space $L^2(\mathbb{R}^n, \pi^\varepsilon)$, it is not hard to show that

$$(6.4) \quad q_t^\varepsilon(x, y) = q_t^\varepsilon(y, x),$$

$$\begin{aligned}
 q_{t+2T_0}^\varepsilon(x_0, y_0) &= \int p_{T_0}^\varepsilon(x_0, x) p_t^\varepsilon(x, y) q_{T_0}^\varepsilon(y, y_0) dx dy \\
 &\geq \int_{x, y \in K(\eta, \xi)} p_{T_0}^\varepsilon(x_0, x) q_t^\varepsilon(x, y) \pi^\varepsilon(y) q_{T_0}^\varepsilon(y, y_0) dx dy \\
 &\geq \inf_{x, y \in K(\eta, \xi)} q_t^\varepsilon(x, y) p_{T_0}^\varepsilon(x_0, K(\eta, \xi)) \\
 &\quad \cdot \int_{K(\eta, \xi)} q_{T_0}^\varepsilon(y, y_0) \pi^\varepsilon(y) dy \quad (\text{by 6.4}) \\
 &= \inf_{x, y \in K(\eta, \xi)} q_t^\varepsilon(x, y) p_{T_0}^\varepsilon(x_0, K(\eta, \xi)) p_{T_0}^\varepsilon(y_0, K(\eta, \xi)) \\
 &\geq (1 - \delta)^2 \inf_{x, y \in K(\eta, \xi)} q_t^\varepsilon(x, y), \quad \varepsilon \leq \varepsilon_0.
 \end{aligned}$$

This completes the proof.

LEMMA 6.2 (Sheu [13, Cor. 2.5]).

$$\lim_{\varepsilon \rightarrow 0} \varepsilon^2 \log p_t^\varepsilon(x, y) \geq -I(t, x, y)$$

uniformly for x, y in a compact set.

COROLLARY 6.1. For any $t > 0, \alpha > 0$, there is $\varepsilon_0 > 0$ such that for $\varepsilon \leq \varepsilon_0, x_0, y_0 \in \mathbb{R}^n$

$$q_{t+2T_0}^\varepsilon(x_0, y_0) \geq \exp\left(-\frac{1}{\varepsilon^2}(J(t, \eta, \xi) + \alpha)\right).$$

LEMMA 6.3 (Super normal case.) For a fixed $t > 0$, let $T = t + 2T_0$. Then $\forall \alpha > 0, \forall M > 0$ there is $\varepsilon_0 > 0$ such that for $\varepsilon \leq \varepsilon_0$

$$|p_{mT}^\varepsilon(x, f) - \pi^\varepsilon(f)| < 4\|f\|\exp(-M),$$

where

$$m = M \exp\left(\frac{1}{\varepsilon^2}(J(t, \eta, \xi) + \alpha)\right).$$

Proof. Let $\beta = \exp(-1/\varepsilon^2(J(t, \eta, \xi) + \alpha))$.

$$\begin{aligned}
 &p_{mT}^\varepsilon(x_1, f) - p_{mT}^\varepsilon(x_2, f) \\
 &= \int p_T^\varepsilon(x_1, z) p_{(m-1)T}^\varepsilon(z, f) dz - \int p_T^\varepsilon(x_2, z) p_{(m-1)T}^\varepsilon(z, f) dz \\
 &= \int q_T(x_1, z) \pi^\varepsilon(z) p_{(m-1)T}^\varepsilon(z, f) dz \\
 &\quad - \int q_T(x_2, z) \pi^\varepsilon(z) p_{(m-1)T}^\varepsilon(z, f) dz \\
 &= \int (q_T(x_1, z) - \beta) \pi^\varepsilon(z) p_{(m-1)T}^\varepsilon(z, f) dz \\
 &\quad - \int (q_T(x_2, z) - \beta) \pi^\varepsilon(z) p_{(m-1)T}^\varepsilon(z, f) dz \\
 &\leq (1 - \beta) (\max_z p_{(m-1)T}^\varepsilon(z, f) - \min_x p_{(m-1)T}^\varepsilon(x, f)) \\
 &= (1 - \beta) \sup_{x_1, x_2 \in \mathbb{R}^n} |p_{(m-1)T}^\varepsilon(x_1, f) - p_{(m-1)T}^\varepsilon(x_2, f)|.
 \end{aligned}$$

By induction,

$$\sup_{x_1, x_2 \in \mathbb{R}^n} |p_{mT}^\varepsilon(x_1, f) - p_{mT}^\varepsilon(x_2, f)| \leq 2\|f\|(1-\beta)^{[m]}.$$

Since π^ε is the invariant measure of $p_t^\varepsilon(x, y)$ [16, p. 243],

$$\begin{aligned} |\pi^\varepsilon(f) - p_{mT}^\varepsilon(x, f)| &\leq \left| \int \pi^\varepsilon(z) (p_{mT}^\varepsilon(z, f) - p_{mT}^\varepsilon(x, f)) dz \right| \\ &\leq 2(1-\beta)^{[m]}\|f\|. \end{aligned}$$

General case. In order to compare the general case with the super normal case, we need the following lemma.

LEMMA 6.4. Let $B(r) = \{x | U(x) \leq r\}$ and $\tau_r = \inf \{t | X(t) \notin B(r)\}$. Then there exists $c(r)$ for large r

$$(i) \quad c(r) \rightarrow \infty \quad \text{as } r \rightarrow \infty,$$

$$(ii) \quad \lim p_x^\varepsilon \left\{ \tau_r > \exp \left(\frac{1}{\varepsilon^2} c(r) \right) \right\} = 1 \quad \text{uniformly for } x \in K(\eta, \xi) \subseteq B(r).$$

Suppose that Lemma 6.4 holds. Choose r large enough such that

$$c(r) > J(t, \eta, \xi) + 1, \quad K(\eta, \xi) \subset B(r).$$

Let \hat{U} satisfy (6.1) for $R_0 > r$ and $\hat{U} = U$ on $B(r)$. Let $\hat{\pi}^\varepsilon$ denote the modified version.

$$\begin{aligned} |p_{mT}^\varepsilon(x, f) - \pi^\varepsilon(f)| &\leq |p_{mT}^\varepsilon(x, f) - \hat{p}_{mT}^\varepsilon(x, f)| \\ &\quad + |\hat{p}_{mT}^\varepsilon(x, f) - \hat{\pi}^\varepsilon(f)| + |\hat{\pi}^\varepsilon(f) - \pi^\varepsilon(f)|. \end{aligned}$$

The second term goes to zero by Lemma 6.3. Since $\hat{\pi}^\varepsilon$ and π^ε have the same weak limit, the third term also tends to zero.

$$|p_{mT}^\varepsilon(x, f) - \hat{p}_{mT}^\varepsilon(x, f)| \leq 2\|f\| E_x^\varepsilon \{ \tau_r \leq mT \} \rightarrow 0 \quad \text{as } \varepsilon \rightarrow 0$$

by Lemma 6.4.

Proof of Lemma 6.4. Choose r_0 such that $K(\eta, \xi) \subseteq B(r_0) =: \Omega_1$ and $\Omega_2 := B(r_0 + 1) \subseteq B(r) =: \Omega_3$. Define

$$\begin{aligned} \sigma_1 &= \inf \{t | X(t) \in \Omega_1\}, \\ \theta_1 &= \inf \{t > \sigma_1 | X(t) \notin \Omega_2\}, \\ &\vdots \\ \sigma_m &= \inf \{t > \theta_{m-1} | X(t) \in \Omega_1\}, \\ \theta_m &= \inf \{t > \sigma_m | X(t) \notin \Omega_2\}. \end{aligned}$$

If one can prove that before exit from Ω_3 , the path spends a lot of time jumping between Ω_1 and Ω_2 , then τ_r will have a good lower estimate.

Let $U(x) = r_0 + 1$ and Q_x^ε denote the measure of the zero drift process, then

$$\begin{aligned} p_x^\varepsilon \{ \tau_r < \sigma_1 \} &= Q_x^\varepsilon \left\{ \tau_r < \sigma_1, \exp \left(\frac{1}{\varepsilon^2} \int_0^{\tau_r} (-\nabla U(X(s)) \cdot dX(s) \right. \right. \\ &\quad \left. \left. - \frac{1}{2\varepsilon^2} \int_0^{\tau_r} |\nabla U(X(s))|^2 ds \right) \right\} \\ &= Q_x^\varepsilon \left\{ \tau_r < \sigma_1, \exp \left(-\frac{1}{\varepsilon^2} \{ U(X(\tau_r)) - U(x) \} \right. \right. \\ &\quad \left. \left. - \frac{1}{2\varepsilon^2} \int_0^{\tau_r} (|\nabla U(X(s))|^2 - \varepsilon^2 \Delta U(X(s))) ds \right) \right\}. \end{aligned}$$

For $s \leq \tau_r < \sigma_1$, $U(X(s)) > r_0$, then there exist $M_1 > 0$ and $M_2 > 0$ such that

$$\begin{aligned} & |\nabla U(X(s))|^2 - \varepsilon^2 \Delta U(X(s)) \\ &= \varepsilon^2 (|\nabla U(X(s))|^2 - \Delta U(X(s))) + (1 - \varepsilon^2) |\nabla U(X(s))|^2 \\ &\geq -\varepsilon^2 M_1 + (1 - \varepsilon^2) M_2 > 0 \quad \text{for small } \varepsilon. \end{aligned}$$

Hence,

$$\begin{aligned} p_x^\varepsilon\{\tau_r < \sigma_1\} &\leq Q_x^\varepsilon\left\{\tau_r < \sigma_1, \exp\left(-\frac{1}{\varepsilon^2}(r - r_0 - 1)\right)\right\} \\ &\leq \exp\left(-\frac{1}{\varepsilon^2}(r - r_0 - 1)\right), \\ p_x^\varepsilon(\tau_r < \sigma_m) &= \sum_{k=1}^m p_x^\varepsilon(\tau_r < \sigma_k, \tau_r \geq \sigma_{k-1}) \\ &= \sum_{k=1}^m p_x^\varepsilon\left\{E_{X(\sigma_{k-1})}\{\tau_r < \sigma_1\}, \tau_r \geq \sigma_{k-1}\right\} \\ &\leq m \exp\left(-\frac{1}{\varepsilon^2}(r - r_0 - 1)\right). \end{aligned}$$

Now we shall show that σ_1 is not too small. Let

$$T^* = \inf_{U(x)=r_0+1} \inf\{t \mid Y(0) = x, Y(t) \in \Omega_1, Y(s) \text{ satisfies (1.3) for } 0 \leq s \leq t\}.$$

Let

$$0 < \delta_0 < d \left(\Omega_1, \left\{ Y(T^*/2) \mid Y(0) = x, U(x) = r_0 + 1, \right. \right. \\ \left. \left. Y(s) \text{ satisfies (1.3) for } 0 \leq s \leq T^*/2 \right\} \right).$$

Let $T_0 \leq T^*/2$; then by a similar method as in the end of the proof of Lemma 1,

$$\begin{aligned} p_x^\varepsilon\{\sigma_1 < T_0\} &\leq p\{\tau < T_0\} \leq (2n) \exp(-e^{-2dT_0}\delta_0^2/(2nT_0\varepsilon^2)) \\ &\leq 2n \exp(-e^{-2dT_0}\delta/(T_0\varepsilon^2)) \end{aligned}$$

(n is the dimension and $\delta = \delta_0^2/2n$, $\tau = \inf\{t \mid |X(t) - Y(t)| > \delta_0\}$, d is the corresponding Lipschitz constant of ∇U in a compact set).

Then, it is obvious that

$$\begin{aligned} p_x^\varepsilon\{\sigma_m < mT_0\} &\leq 2nm \exp\left(-e^{-2dT_0}\frac{\delta}{T_0\varepsilon^2}\right), \\ p_x^\varepsilon\{\tau_r < mT_0\} &\leq p_x^\varepsilon\{\tau_r < \sigma_m\} + p_x^\varepsilon\{\sigma_m < mT_0\} \\ &\leq m \exp\left(-\frac{1}{\varepsilon^2}(r - r_0 - 1)\right) + 2nm \exp\left(-\frac{1}{\varepsilon^2}\frac{e^{-2dT_0}\delta}{T_0}\right). \end{aligned}$$

Choose T_0 such that

$$\frac{e^{-2dT_0}\delta}{T_0} > (r - r_0 - 1).$$

And choose $m-1 = [\exp(1/\varepsilon^2(r-r_0-1-v))]$, where v is an arbitrary fixed small positive number

$$\begin{aligned} p_x^\varepsilon \left\{ \tau_r \geq \exp \left(\frac{1}{\varepsilon^2} (r-r_0-1-v) \right) T_0 \right\} \\ > 1 - (2n+3) \exp \left(-\frac{v}{\varepsilon^2} \right) \rightarrow 1 \quad \text{as } \varepsilon \rightarrow 0. \end{aligned}$$

Hence we may choose $c(r) = r-r_0-1-v$ for any fixed $v > 0$.

For $x \in K(\eta, \xi)$,

$$\begin{aligned} p_x^\varepsilon \{ \tau_r < \exp(c(r)/\varepsilon^2) \} &= p_x^\varepsilon \{ E_{X(\theta)} \{ \tau_r < \exp(c(r)/\varepsilon^2) \}, \theta < \exp(c(r)/\varepsilon^2) \} \\ &\rightarrow 0 \quad \text{as } \varepsilon \rightarrow 0, \end{aligned}$$

where $\theta = \inf \{ t | U(X(t)) = r_0 + 1 \}$.

7. Appendix.

(1) Properties of $I(t, x, y)$ can be found in [2] and [14].

(2) $c_0 < \infty$ is obvious. In fact if we consider

$$\begin{aligned} \phi(u) &= x + u(z-x) \quad \text{for } 0 \leq u \leq 1, \\ &= z \quad \text{for } 1 \leq u \leq t-1, \\ &= z + (u-(t-1))(y-z) \quad \text{for } t-1 \leq u \leq t, \end{aligned}$$

where z is a stationary point, then it is easy to see why $c_0 < \infty$.

(3) Suppose that the set S of all stationary points has only finitely many, say l , connected components. We also assume that points in each component can be connected by smooth curves. Choose ε small enough such that $S(\varepsilon)$ has l disjoint components S_1, \dots, S_l . For x, y belong to the same S_i ,

$$I(t, x, y) = O(1/t) + O(\varepsilon).$$

Indeed, connect x to a stationary point v in S_i with $|x-v| < 2\varepsilon$ by straight line for time interval $[0, 1]$. Connect y to a stationary point w in S_i with $|y-w| < 2\varepsilon$ by straight line for time interval $[t-1, t]$. Connect v, w by a fixed smooth curve ψ with all the curve stationary points. Rescale the parameter of the curve to the interval $[1, t-1]$ and denote it by ϕ , then

$$\begin{aligned} \int_1^{t-1} |\dot{\phi}(u) + \nabla U(\phi(u))|^2 du &= \int_1^{t-1} |\dot{\phi}(u)|^2 du \\ &= \frac{1}{t-2} \int_0^1 |\dot{\psi}(u)|^2 du. \end{aligned}$$

If $\phi(t-u)$, $0 \leq u \leq t$, is a solution, then

$$\int_0^t |\dot{\phi}(u) + \nabla U(\phi(u))|^2 du = 0.$$

For any starting point x , and end point y , since U is a Lyapunov function for the dynamical system (1.4), within a fixed finite time x will reach some S_i via a solution of (1.4) and y some S_j via a solution. If ϕ is a solution, then

$$\frac{1}{4} \int_0^t |\dot{\phi}(u) + \nabla U(\phi(u))|^2 du = U(\phi(t)) - U(\phi(0)).$$

(4) A good upper bound for $I(t, x, y)$ is a curve $\phi(0) = x$, $\phi(t) = y$ and ϕ spends most of its time at a stationary point. From (3), we only have to count the contribution from connecting different components. Consider $\{S_i\}_{i=1, \dots, l}$ as nodes of a graph, and define S_i and S_j as neighboring nodes if there is a trajectory of (1.4) connecting S_i and S_j . Suppose S_1, S_2, \dots, S_m , $m \leq l$, are in the same connected component, and assume there exist points $x_1, \bar{x}_2, x_2, \bar{x}_3, x_3, \dots, \bar{x}_{m-1}, x_{m-1}, \bar{x}_m$ in S_1, \dots, S_m , respectively, such that a trajectory connects x_i to \bar{x}_{i+1} (see Fig. 1). If $U(\bar{x}_{i+1}) > U(x_i)$, then the contribution

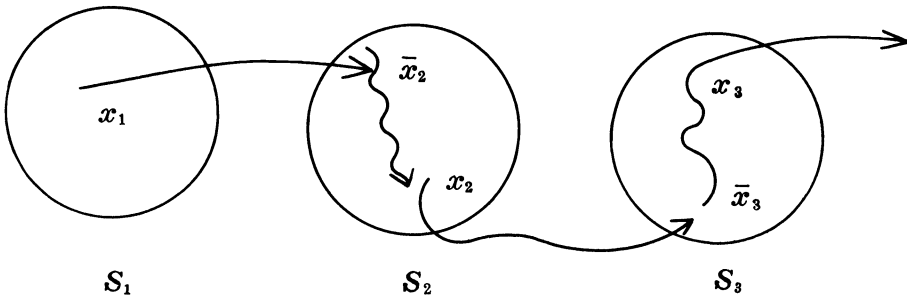


FIG. 1

is $2(U(\bar{x}_{i+1}) - U(x_i))$. Otherwise, it is a free ride. Of course, there are other paths to connect S_1, \dots, S_m . Note that $U(x_j)$, $U(\bar{x}_j)$ and $U(z_j)$ are almost of the same value, where z_i is a stationary point in S_j . Two consecutive increasing trajectories contribute

$$2(U(\bar{x}_{i+2}) - U(x_{i+1}) + U(\bar{x}_{i+1}) - U(x_i)) \sim 2(U(\bar{z}_{i+2}) - U(z_i)).$$

Hence, $2([m/2] \max_{1 \leq j \leq m} U(z_j))$ is a bound.

(5) Suppose the graph $\{S_j\}_{1 \leq j \leq l}$ has, say, G_1, \dots, G_k components. Let K be any bounded connected set containing all S_j 's. For any $x \in K$ either x in some G_i or there exists a unique G_i such that x is connected to G_i by a trajectory. Now we have partitioned K into K_1, \dots, K_k disjoint sets. Define K_i, K_j are neighbors if $d(K_i, K_j) = 0$. If we regard $\{K_j\}_{j=1, \dots, k}$ as nodes of a graph, then we can show it is a connected graph since K is connected. In other words we can connect G_i to G_j via trajectories of (1.4) and at most $k-1$ line segments of arbitrary small length. The same argument as in 4 yields that the contribution to connect different components is less than

$$2\left(\left[\frac{k}{2}\right] \max_{1 \leq i \leq l} U(z_i)\right).$$

(6) One may use $3([l/2] + [k/2]) \max_{1 \leq i \leq l} U(z_i)$ as a rough bound for c_0 .

(7) We give some examples (see Figs. 2-5) to calculate c_* for the one-dimensional case.

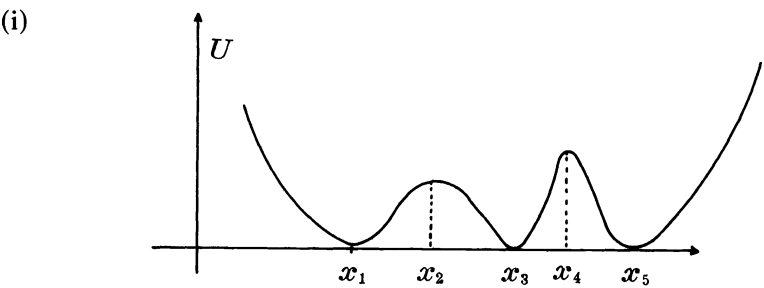


FIG. 2. $c_* = (U(x_2) - U(x_1)) + (U(x_4) - U(x_3)) = (U(x_2) - U(x_3)) + (U(x_4) - U(x_5))$.

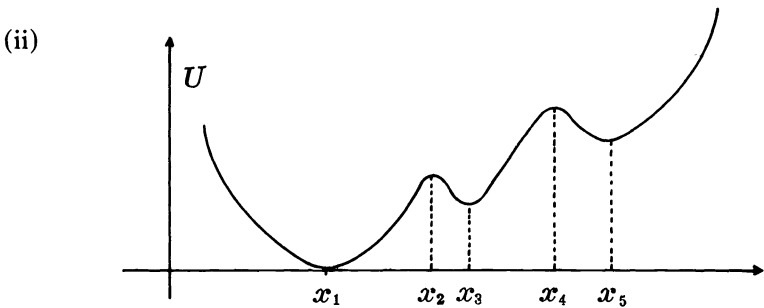


FIG. 3. $c_* = (U(x_2) - U(x_1)) + (U(x_4) - U(x_3)) - U(x_5) = (U(x_4) - U(x_5)) + (U(x_2) - U(x_3))$, ($U(x_1) = 0$).

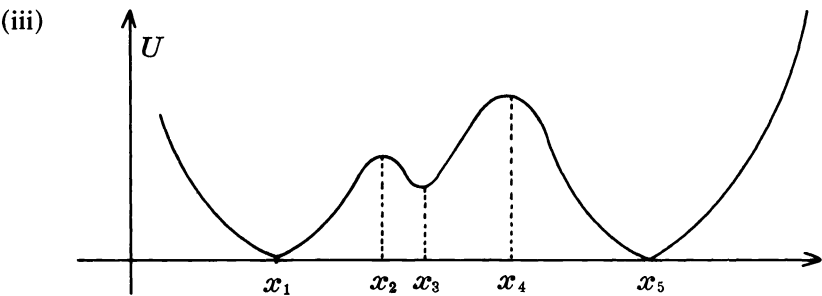


FIG. 4. $c_* = (U(x_2) - U(x_1)) + (U(x_4) - U(x_3)) = (U(x_4) - U(x_5)) + (U(x_2) - U(x_3))$.

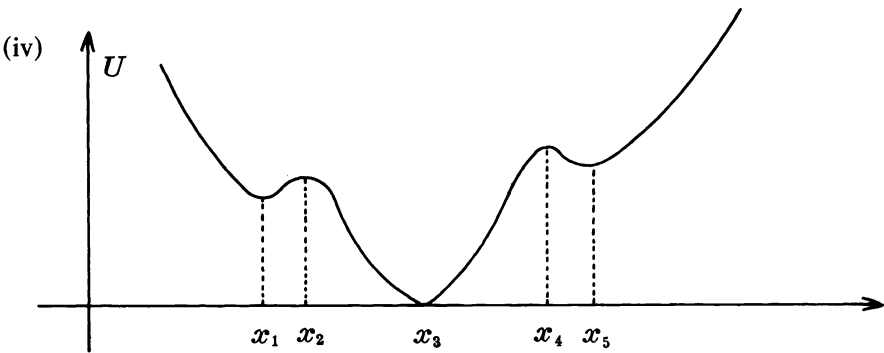


FIG. 5. $c_* = (U(x_2) - U(x_1)) + (U(x_4) - U(x_3)) - U(x_5) = (U(x_4) - U(x_5)) + (U(x_2) - U(x_3)) - U(x_1)$.

Acknowledgment. The authors would like to thank Professor Daniel Stroock for very helpful suggestions and discussions.

REFERENCES

- [1] V. ČERNÝ, *A thermodynamical approach to the travelling salesman problem: an efficient simulation algorithm*, preprint, Inst. of Physics and Biophysics, Comenius Univ., Bratislava, 1982.
- [2] M. I. FREIDLIN AND A. D. WENTZELL, *Random Perturbations of Dynamical Systems*, Springer-Verlag, Berlin, New York, 1984.
- [3] S. GEMAN AND D. D. GEMAN, *Stochastic relaxation, Gibbs distribution, and the Bayesian restoration of images*, IEEE Trans. Pattern Anal. and Machine Intelligence, 6 (1984), pp. 721–741.
- [4] S. GEMAN AND C.-R. HWANG, *Diffusion for global optimization*, this Journal, to appear.
- [5] B. GIDAS, *Non-Stationary Markov chains and convergence of the annealing algorithm*, J. Statist. Phys., 39 (1985), pp. 73–131.
- [6] ———, *Global minimization via the Langevin equation*, in preparation.
- [7] U. GRENANDER, *Tutorial in Pattern Theory*, Brown Univ., Providence, RI, 1983.
- [8] B. HAJEK, *Cooling schedules for optimal annealing*, Dept. of Electrical & Computer Engineering and the Coordinated Science Lab., Univ. of Illinois, Champaign-Urbana, IL, 1985.
- [9] C.-R. HWANG, *Laplace's method revisited: weak convergence of probability measures*, Ann. Probab., 8 (1980), pp. 1177–1182.
- [10] S. KIRKPATRICK, C. D. GELATT, JR. AND M. P. VECCHI, *Optimization by simulated annealing*, Science, 220 (1983), pp. 621–680.
- [11] H. J. KUSHNER, *Asymptotic global behavior for stochastic approximations and diffusion with slowly decreasing noise effects: global minimization via Monte Carlo*, preprint, Div. of Applied Mathematics, Brown Univ., Providence, RI, 1985.
- [12] G. PARISI, *Prolegomena to any further computer evaluation of the QCD mass spectrum*, in Progress in Gauge Field Theory, Cargese, 1983.
- [13] S.-J. SHEU, *Asymptotic behavior of transition density of diffusion Markov process with small diffusion*, Stochastics, 13 (1984), pp. 131–163.
- [14] ———, *Asymptotic behavior of invariant density of diffusion Markov process with small diffusion*, SIAM J. Math. Anal., 12 (1985), pp. 451–460.
- [15] D. W. STROOCK AND S. R. S. VARADHAN, *Multidimensional Diffusion Processes*, Springer-Verlag, Berlin, New York, 1979.
- [16] S. R. S. VARADHAN, *Lectures on Diffusion Problems and Partial Differential Equations*, Tata Inst., Bombay, 1980.

SPECTRAL FACTORIZATION AND NEVANLINNA-PICK INTERPOLATION*

TRYPHON T. GEORGIOU† AND PRAMOD P. KHARGONEKAR‡

Abstract. We develop a spectral factorization algorithm based on linear fractional transformations and on the Nevanlinna-Pick interpolation theory. The algorithm is recursive and depends on a choice of points $(z_k, k = 1, 2, \dots)$ inside the unit disk. Under a mild condition on the distribution of the z_k 's, the convergence of the algorithm is established. The algorithm is flexible and convergence can be influenced by the selection of z_k 's.

Key words. spectral factorization, interpolation theory, positive-real functions, Nevanlinna-Pick interpolation

AMS(MOS) subject classifications. 93C05, 30E05, 30D50

1. Introduction. Interpolation theory for complex analytic functions has a long history in mathematics and engineering. The origin of the subject can be traced back at least to the work of Caratheodory, Schur, Nevanlinna and Pick (see [7]) and continues with the recent works of Adamjan, Arov and Krein [1], Sarason [24], Sz. Nagy and Foias [20], and Ball and Helton [3] which have extended the theory to a general operator theoretic setting. In engineering, interpolation theory has been used in a variety of problem areas. Passive circuit synthesis, optimal control, stability theory, representation and prediction theory for stochastic processes, and control theory are some of the engineering disciplines where interpolation theory of complex analytic functions has played a significant role. For these, see for example [10], [11], [18], [19], [25], [26] and the references therein.

In the present work we use interpolation theoretic ideas and, in particular, linear fractional transformations to develop a general scheme for spectral factorization. Spectral factorization is a key problem in a variety of engineering fields, and has been investigated extensively. In particular, see [2], [5], [6], [11], [12], [15], [21], [23]. The approach we have taken leads to a connection with ideas from interpolation theory and in particular to the use of linear fractional transformations. *Our main contribution in this paper is a new and versatile theoretical algorithm for spectral factorization.* However, numerical properties of this algorithm are not addressed here and will be pursued elsewhere.

We denote by \mathbb{U} the unit ball in H^∞ ; i.e., $\mathbb{U} := \{f(z) \text{ analytic in } \mathbb{D} \text{ such that } |f(z)| \leq 1 \text{ for all } z \in \mathbb{D}\}$, where \mathbb{D} denotes the *open unit disc* in the complex plane. The classical Nevanlinna-Pick interpolation problem requires finding a function f in \mathbb{U} that satisfies the following interpolation conditions:

$$f^{(m)}(z_k) = w_{m,k}, \quad m = 0, 1, \dots, N_k, \quad \text{for } k = 1, 2, \dots, N.$$

* Received by the editors December 9, 1985; accepted for publication (in revised form) April 28, 1986. This work was supported in part by the National Science Foundation under grant ECS-8451519, and in part by grants from Honeywell, 3M, and the MEIS Center at the University of Minnesota, Minneapolis, Minnesota.

† Department of Electrical and Computer Engineering, Iowa State University, Ames, Iowa 50011.

‡ Department of Electrical Engineering and Center for Control Science and Dynamical Systems, University of Minnesota, Minneapolis, Minnesota 55455.

(The case $N = 1$ is known as Caratheodory–Schur interpolation.) The *Nevanlinna–Pick recursion* allows the solution, i.e., existence and characterization of all solutions, of this problem by iteratively reducing it to an equivalent one with fewer interpolation constraints. The general form of the solutions is then presented in terms of a linear fractional transformation

$$(1.1) \quad f(z) = [A(z) + B(z)h(z)] \times [C(z) + D(z)h(z)]^{-1},$$

where A, B, C and D are functions depending on the data of the problem and $h(z)$ is an arbitrary function in \mathbb{U} that parametrizes the set of solutions.

With every function f in \mathbb{U} , such that

$$\ln [1 - |f(e^{i\theta})|^2] \text{ is in } L^1 := L^1[-\pi, \pi],$$

there is associated a unique outer function (see [22])

$$g_f(z) := \exp \left\{ (4\pi)^{-1} \int_{-\pi}^{\pi} [e^{i\theta} + z][e^{i\theta} - z]^{-1} \ln [1 - |f(e^{i\theta})|^2] d\theta \right\}$$

such that $|g_f(e^{i\theta})|^2 = 1 - |f(e^{i\theta})|^2$ a.e. on $[-\pi, \pi]$. The function g_f is known as the **canonical spectral factor** of f and plays an important role in several problem areas such as representation of stochastic processes, optimal control, network synthesis, etc. Under fairly general conditions g_f can in fact be defined as a meromorphic function on the whole complex plane (see [8]). This is certainly true for the important case where f is also a rational function, and this is precisely the case we consider in the present paper.

If g_f and g_h denote the spectral factors of f and h respectively that are related as in (1.1), then it can be shown that g_f and g_h under certain conditions have the same zeros. (The general problem of describing invariants of the action of the semigroup of linear fractional transformations has been considered by Helton [17].) Utilizing the invariance of the so-called spectral zeros (or transmission zeros) under linear fractional transformations, we developed a spectral factorization algorithm along the lines of Caratheodory–Schur interpolation [14], [15]. The present work extends our earlier results to the Nevanlinna–Pick setting and gives rise to a general spectral factorization algorithm.

2. The Nevanlinna–Pick recursion. We begin with the following well-known lemma, which is simply an invariant formulation of Schwarz’s lemma. This has provided an important tool in the theory of interpolation with complex analytic functions and was utilized in a masterful way in that context by Nevanlinna (see Garnett [13]).

LEMMA 2.1. *Let f_1 be in \mathbb{U} , and consider a sequence of points $(z_k \in \mathbb{D}, k = 1, 2, \dots)$ and a sequence of parameters $(c_k \in \mathbb{D}, k = 1, 2, \dots)$. Define*

$$(2.2a) \quad w_k := f_k(z_k),$$

$$(2.2b) \quad \tilde{f}_{k+1} := \frac{1 - \bar{z}_k z}{z - z_k} \frac{f_k - w_k}{1 - \bar{w}_k f_k},$$

$$(2.2c) \quad f_{k+1} := \frac{\tilde{f}_{k+1} - c_k}{1 - \bar{c}_k \tilde{f}_{k+1}},$$

for $k = 1, 2, \dots$. Then, $f_k, k = 2, 3, \dots$, is a sequence of \mathbb{U} -functions. In case $|w_n| = 1$ for a value $k = n$, then the above sequence terminates to a function $f_n \equiv w_n$. This last case

occurs only if f_1 is a finite Blaschke product; i.e., f_1 is of the form

$$f_1(z) = e^{i\varphi} \prod_{k=1}^n \frac{(z - \xi_k)}{(1 - \bar{\xi}_k z)},$$

and consequently has modulus equal to one on the boundary of the disk.

The case where f_1 is a finite Blaschke product is of no interest to us because in this case the spectral factor of f_1 is the zero function, and the spectral factorization problem becomes trivial. Hence, in the sequel, even when not explicitly stated, we tacitly assume that this is not the case.

The sequence of the parameters c_k in the Nevanlinna recursion can be taken to be arbitrary constants in \mathbb{D} . However, we follow a standard and convenient normalization (see [4], [8]) described in the lemma below.

LEMMA 2.3. *Let $f_1 \in \mathbb{U}$ (but not a finite Blaschke product), and let $f_1(0) = 0$. Given a sequence of points $(z_k \in \mathbb{D}, k = 1, 2, \dots)$, and letting*

$$(2.4) \quad c_k := \tilde{f}_{k+1}(0) = \begin{cases} \frac{w_k}{z_k} & \text{whenever } z_k \neq 0, \\ \lim_{z \rightarrow 0} \frac{f_k(z)}{z} & \text{whenever } z_k = 0, \end{cases}$$

the Nevanlinna recursion (2.2a-c) produces a sequence $f_k, k = 2, 3, \dots$, of \mathbb{U} -functions that satisfy $f_k(0) = 0$, for all k .

The Nevanlinna recursion in Lemma 2.1 can be used to provide a constructive approach to the Nevanlinna-Pick problem (see Garnett [13, p. 166]). It can also be used to generate, from a known function f_1 in \mathbb{U} , the associated sequence of the so-called Schur parameters/reflection coefficients w_k . This is the way we apply the Nevanlinna recursion. In fact, our objective in the next section is to study the limiting behavior of the “by-product” f_k as k tends to ∞ .

3. Some convergence results. Let f_1 (different from a finite Blaschke product) be in \mathbb{U} and assume that $f_1(0) = 0$. This causes no loss of generality from our standpoint because the functions f in \mathbb{U} , and zf which is also in \mathbb{U} , have the same spectral factor. The assumption $f_1(0) = 0$ simplifies the computations required in the sequel.

Compute now the sequence $f_k, k = 2, 3, \dots$, of \mathbb{U} -functions from f_1 and the sequence of points $(z_k \in \mathbb{D}, k = 1, 2, \dots)$ via the Nevanlinna recursion (2.2) and (2.4). Recall that the choice (2.4) for the constants c_k readily implies that $f_k(0) = 0$ for $k = 2, 3, \dots$. This is very convenient as we will see shortly.

Using (2.2b) and (2.2c) it easily follows that

$$\frac{(1 - |w_k|^2)(1 - |f_k|^2)}{|1 - \bar{w}_k f_k|^2} = \frac{(1 - |c_k|^2)(1 - |f_{k+1}|^2)}{|1 + \bar{c}_k f_{k+1}|^2} \quad \text{for } z \text{ on } \mathbb{T},$$

and $k = 1, 2, \dots, n-1$. Applying $\ln(\cdot)$ to both sides of the above equation, we obtain that

$$(3.1) \quad \begin{aligned} & \ln(1 - |w_k|^2) + \ln(1 - |f_k|^2) - \ln|1 - \bar{w}_k f_k|^2 \\ & = \ln(1 - |c_k|^2) + \ln(1 - |f_{k+1}|^2) - \ln|1 + \bar{c}_k f_{k+1}|^2 \end{aligned}$$

for $z \in \mathbb{T}$. Note that $|\bar{w}_k f_k| < 1$ in \mathbb{D} . Hence $1 - \bar{w}_k f_k$ is an analytic function with no roots in \mathbb{D} . Therefore, (see Rudin [22, Thm. 13.12]) $u(z) := \ln|1 - \bar{w}_k f_k|^2$ is a harmonic function

in \mathbb{D} . Also, since $f_k(0) = 0$, it follows that $u(0) = 0$. Consequently, using the well-known mean value property of harmonic functions (see [22])

$$\int_{-\pi}^{\pi} \ln |1 - \bar{w}_k f_k(e^{i\theta})|^2 d\theta = u(0) = 0.$$

A similar argument applies to $\ln |1 + \bar{c}_k f_{k+1}|^2$ and yields

$$\int_{-\pi}^{\pi} \ln |1 + \bar{c}_k f_{k+1}(e^{i\theta})|^2 d\theta = 0.$$

Now, from (3.1), we integrate over the interval $[-\pi, \pi]$ and exponentiate both sides to obtain that

$$\begin{aligned} \exp \left\{ (2\pi)^{-1} \int_{-\pi}^{\pi} \ln (1 - |f_k(e^{i\theta})|^2) d\theta \right\} \\ = \left(\frac{1 - |c_k|^2}{1 - |w_k|^2} \right) \exp \left\{ (2\pi)^{-1} \int_{-\pi}^{\pi} \ln (1 - |f_{k+1}|^2) d\theta \right\} \end{aligned}$$

for all k . Finally by induction we conclude that

$$\begin{aligned} \exp \left\{ (2\pi)^{-1} \int_{-\pi}^{\pi} \ln (1 - |f_1(e^{i\theta})|^2) d\theta \right\} \\ (3.2) \quad = \left(\prod_{k=1}^n \frac{1 - |c_k|^2}{1 - |w_k|^2} \right) \exp \left\{ (2\pi)^{-1} \int_{-\pi}^{\pi} \ln (1 - |f_{k+1}|^2) d\theta \right\}. \end{aligned}$$

We are interested in the case where f_1 is a rational function in \mathbb{U} but different from a finite Blaschke product. In this case the above integrals are different from zero and we can obtain the following theorem.

THEOREM 3.3. *Let f_1 be a rational function in \mathbb{U} such that $\ln (1 - |f_1(e^{i\theta})|^2)$ is in L^1 (hence not a Blaschke product), and $f_1(0) = 0$. Let $(z_k, k = 1, 2, \dots)$ be a sequence of points in \mathbb{D} satisfying the property*

$$(3.4) \quad \sum_{k=1}^{\infty} (1 - |z_k|) = \infty,$$

and obtain the corresponding sequence of w_k 's and c_k 's from (2.2) and (2.4). Then

$$\exp \left\{ (2\pi)^{-1} \int_{-\pi}^{\pi} \ln (1 - |f_1(e^{i\theta})|^2) d\theta \right\} = \lim_{n \rightarrow \infty} \prod_{k=1}^n \left(\frac{1 - |c_k|^2}{1 - |w_k|^2} \right).$$

The proof of the above theorem is based on certain classical facts in function theory and some results obtained by Dewilde and Dym [8], [9] and Bulteel and Dewilde [4], and is given in § 6. Below we give an immediate corollary of Theorem 3.3 and relation (3.2).

COROLLARY 3.5. *Under the conditions of Theorem 3.3,*

$$\lim_{k \rightarrow \infty} \exp \left\{ (2\pi)^{-1} \int_{-\pi}^{\pi} \ln (1 - |f_{k+1}|^2) d\theta \right\} = 1$$

and

$$\lim_{k \rightarrow \infty} f_{k+1}(z) = 0 \text{ a.e. on } \mathbb{T}.$$

4. Invariance of spectral zeros–spectral factorization. Let f be a rational \mathbb{U} -function (but not a finite Blaschke product) and let it be represented as the ratio of two coprime polynomials $a(z)/b(z)$. Since f is in \mathbb{U} ,

$$1 - |f(z)|^2 = \frac{|b(z)|^2 - |a(z)|^2}{|b(z)|^2} \geq 0 \quad \text{on } \mathbb{T},$$

and it admits a factorization

$$(4.1) \quad 1 - |f(z)|^2 = \frac{|\mu\eta(z)|^2}{|b(z)|^2} \quad \text{for } z = e^{i\theta},$$

where $\eta(z)$ is a polynomial that *can be taken to have no root inside \mathbb{D} with $\eta(0) = 1$, and μ a positive constant*. Under this normalization $\eta(z)$ is uniquely defined by f and will be called the **spectral numerator of f** . The canonical spectral factor of f is then given by

$$g(z) = \mu \frac{\eta(z)}{b(z)},$$

and is defined on the whole complex plane. Moreover, (4.1) extends to an equality of meromorphic functions

$$(4.2) \quad 1 - f(z)f(z^{-1}) = \mu^2 \frac{\eta(z)}{b(z)} \frac{\bar{\eta}(z^{-1})}{\bar{b}(z^{-1})}$$

valid throughout the complex plane.

Let now f_1 be a rational \mathbb{U} -function and $f_k, k = 2, 3, \dots$, be the sequence of \mathbb{U} -functions obtained from f_1 and from a sequence of points $(z_k, k = 1, 2, \dots)$ via the Nevanlinna recursion. In view of (2.2), it is clear that f_{k+1} is also a rational function. Thus, the Nevanlinna recursion produces a sequence of rational functions $f_k, k = 2, 3, \dots$.

We now express the recurrence formulas (2.2, 2.4) in terms of fractional representations a_k/b_k for the functions $f_k, k = 1, 2, \dots$. First

$$(4.3) \quad w_k := \frac{a_k(z_k)}{b_k(z_k)}.$$

By solving (2.2) for f_{k+1} in terms of f_k we obtain (see also Garnett [13, p. 167])

$$(4.4) \quad f_{k+1} = \frac{\alpha_k - \gamma_k f_k}{-\beta_k + \delta_k f_k},$$

where

$$(4.5) \quad \begin{aligned} \alpha_k(z) &= w_k(1 - \bar{z}_k z) + c_k(z - z_k), \\ \beta_k(z) &= \bar{c}_k w_k(1 - \bar{z}_k z) + (z - z_k), \\ \gamma_k(z) &= (1 - \bar{z}_k z) + c_k \bar{w}_k(z - z_k), \\ \delta_k(z) &= \bar{c}_k(1 - \bar{z}_k z) + \bar{w}_k(z - z_k) \end{aligned}$$

and we use (2.4), which becomes

$$(4.6) \quad c_k = \begin{cases} \frac{w_k}{z_k} & \text{when } z_k \neq 0, \\ \lim_{z \rightarrow 0} \frac{a_k(z)}{zb_k(z)} & \text{when } z_k = 0. \end{cases}$$

Starting from a coprime fraction for $f_1 = a_1/b_1$, i.e., a_1 and b_1 are polynomials with no common factor, define a sequence of pairs of functions (that will turn out to be polynomials) (a_k, b_k) , for $k = 2, 3, \dots$, via the following:

$$(4.7) \quad \begin{bmatrix} a_{kf1} \\ b_{k+1} \end{bmatrix} = \frac{1}{(z - z_k)(1 - |c_k|^2)} \begin{bmatrix} \gamma_k & -\alpha_k \\ -\delta_k & \beta_k \end{bmatrix} \begin{bmatrix} a_k \\ b_k \end{bmatrix}$$

for $k = 2, 3, \dots$. We can now state the following proposition:

PROPOSITION 4.8. *Let f_1 be a rational \mathbb{U} -function that is not a finite Blaschke product, a_1/b_1 be a polynomial coprime fraction for f_1 , and generate the sequence of (a_k, b_k) and of f_k via (4.3)–(4.7). Then the following hold.*

$$(4.8a) \quad f_k = a_k/b_k \quad \text{for all } k.$$

$$(4.8b) \quad (a_k, b_k) \quad \text{for } k = 2, 3, \dots, \text{ are polynomials in } z.$$

$$(4.8c) \quad \text{The maximum degree of the polynomials } (a_k, b_k) \text{ never exceeds the maximum degree of the polynomials } (a_1, b_1).$$

$$(4.8d) \quad \text{If } f_1(0) = 0, \text{ and the polynomials } a_1, b_1 \text{ have been normalized to satisfy } a_1(0) = 0, \text{ and } b_1(0) = 1, \text{ then}$$

$$a_k(0) = 0 \quad \text{and} \quad b_k(0) = 1 \quad \text{for all } k.$$

Proof.

$$(4.8a) \quad \text{It follows immediately by comparison of (4.4) with (4.7).}$$

$$(4.8b) \quad \text{For } z = z_k, \text{ the expressions } (\gamma_k a_k - \alpha_k b_k) \text{ and } (\delta_k a_k - \beta_k b_k) \text{ become equal to } [(1 - \bar{z}_k z_k) a_k(z_k) - w_k(1 - \bar{z}_k z_k) b_k(z_k)], \text{ and } [\bar{c}_k(1 - \bar{z}_k z_k) a_k(z_k) - \bar{c}_k w_k(1 - \bar{z}_k z_k) b_k(z_k)], \text{ respectively. But } a_k(z_k) = w_k b(z_k). \text{ Hence, both expressions become equal to zero. Therefore, } (\gamma_k a_k - \alpha_k b_k) \text{ and } (\delta_k a_k - \beta_k b_k) \text{ are polynomial expressions divisible by } (z - z_k). \text{ From (4.7) we now conclude that } (a_k, b_k), \text{ for } k = 2, 3, \dots \text{ are polynomials in } z.$$

$$(4.8c) \quad \text{The polynomials } \alpha, \beta, \gamma \text{ and } \delta, \text{ have degree equal to one. Hence the maximum degree of } \{(\gamma_k a_k - \alpha_k b_k)/(z - z_k), (\delta_k a_k - \beta_k b_k)/(z - z_k)\} \text{ does not exceed the maximum degree of } (a_k, b_k).$$

$$(4.8d) \quad \text{It follows by straightforward computation. } \square$$

However, $a_k(z)$ and $b_k(z)$ might have a common factor. The determinant of the transformation matrix in (4.7) is computed directly and is given below

$$(4.9) \quad \gamma_k \beta_k - \alpha_k \delta_k = (1 - |c_k|^2)(1 - |w_k|^2)(z - z_k)(1 - \bar{z}_k z)$$

for $k = 1, 2, \dots$. (Note that $(z - z_k)$ cannot be a factor of a_{k+1} or b_{k+1} since it has been divided out in (4.7).) Thus, the only possible common factor of a_{k+1} and b_{k+1} , in addition to common factors of a_k and b_k , is $(1 - \bar{z}_k z)$. In fact a_{k+1} and b_{k+1} will have $(1 - \bar{z}_k z)$ as a common factor precisely when it is also a factor in $\eta_k(z)$. (Then, this becomes a common factor of every pair (a_{k+l}, b_{k+l}) for $l = 1, 2, \dots$.) The following theorem addresses exactly this point.

THEOREM 4.10. *Let f_1 be a rational \mathbb{U} -function (different from a finite Blaschke product) and a_k/b_k be a coprime polynomial fraction description of f_1 satisfying (4.8d). Let (a_k, b_k) , $k = 2, 3, \dots$, be obtained from (4.3), (4.5)–(4.7) and a sequence of points $(z_k$ in \mathbb{D} , for $k = 1, 2, \dots$). Define d_k to be the greatest common divisor of (a_k, b_k)*

normalized by $d_k(0) = 1$, and let η_k denote the spectral numerator of the \mathbb{U} -function $f_k := a_k/b_k$. Then the following hold:

(i) If for $k = n$, $(1 - \bar{z}_n z)$ is a factor of $\eta_n(z)$, then

$$d_{n+1} = (1 - \bar{z}_n z) d_n, \text{ whereas}$$

$$\eta_n = (1 - \bar{z}_n z) \eta_{n+1};$$

(ii) If for $k = n$, $(1 - \bar{z}_n z)$ is not a factor of $\eta_n(z)$, then

$$d_{n+1} = d_n, \text{ and } \eta_n = \eta_{n+1}.$$

Proof. We begin by recalling first a certain well-known property of the transformation matrix

$$M_k(z) := \frac{1}{(z - z_k)(1 - |c_k|^2)} \begin{bmatrix} \gamma_k(z) & -\alpha_k(z) \\ -\delta_k(z) & \beta_k(z) \end{bmatrix}$$

used in (4.7). Define by

$$M_k(z)_* := M_k^*(z^{-1}),$$

where $(\)^*$ denotes complex conjugation of the coefficients and transposition of the matrix. Then,

$$(4.11) \quad M_k(z)_* \begin{bmatrix} -1 & 0 \\ 0 & 1 \end{bmatrix} M_k(z) = \frac{(1 - |w_k|^2)}{(1 - |c_k|^2)} \begin{bmatrix} -1 & 0 \\ 0 & 1 \end{bmatrix}.$$

This property is known as *J-unitarity*; e.g., see [9]. (It follows directly by algebraic manipulations and the use of (4.5) and (4.9).)

We now compute

$$\begin{aligned} & \bar{b}_{k+1}(z^{-1})b_{k+1}(z) - \bar{a}_{k+1}(z^{-1})a_{k+1}(z) \\ &= [\bar{a}_{k+1}(z^{-1})\bar{b}_{k+1}(z^{-1})] \begin{bmatrix} -1 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} a_{k+1}(z) \\ b_{k+1}(z) \end{bmatrix}, \end{aligned}$$

which because of (4.11)

$$\begin{aligned} &= \frac{(1 - |w_k|^2)}{(1 - |c_k|^2)} [\bar{a}_k(z^{-1})\bar{b}_k(z^{-1})] \begin{bmatrix} -1 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} a_k(z^{-1}) \\ b_k(z^{-1}) \end{bmatrix} \\ &= \frac{(1 - |w_k|^2)}{(1 - |c_k|^2)} (\bar{b}_k(z^{-1})b_k(z) - \bar{a}_k(z^{-1})a_k(z)). \end{aligned}$$

This last equality implies that

$$(4.12) \quad |d_{k+1}(z)\eta_{k+1}(z)|^2 = |d_k(z)\eta_k(z)|^2, \quad z = e^{i\theta},$$

where μ_k is a nonzero constant that can be taken to be positive. Also note that, as it was argued before on the basis of (4.9) and (4.7),

$$(4.13) \quad d_{k+1}(z) \text{ is either equal to } d_k(z) \text{ or equal to } (1 - \bar{z}_k z)d_k(z)$$

for all values of k . Consequently, $d_k(z)$ has no roots in \mathbb{D} . Since the same applies to $\eta_k(z)$, we now conclude from (4.12) that

$$(4.14) \quad d_{k+1}(z)\eta_{k+1}(z) = d_k(z)\eta_k(z) \quad \text{for } k = 1, 2, \dots$$

Now, if $(1 - \bar{z}_k z)$ is not a factor of $\eta_k(z)$, then we conclude from (4.13) and (4.14) that

$$d_{k+1}(z) = d_k(z) \quad \text{and} \quad \eta_{k+1}(z) = \eta_k(z).$$

If $(1 - \bar{z}_k z)$ is a factor of $\eta_k(z)$, we only need to consider the case where $z_k \neq 0$. In this case

$$\eta_k(\bar{z}_k^{-1})\bar{\eta}_k(\bar{z}_k) = 0,$$

which implies that

$$(4.15) \quad 1 - f_k(\bar{z}_k^{-1})\bar{f}_k(\bar{z}_k) = 0.$$

But $\bar{f}_k(\bar{z}_k) = \overline{a_k(z_k)/b_k(z_k)} = \bar{w}_k$; therefore (4.15) implies that

$$(4.16) \quad b_k(\bar{z}_k^{-1}) = \bar{w}_k a_k(\bar{z}_k^{-1}).$$

From (4.5) we now have that

$$\begin{aligned} a_{k+1}(\bar{z}_k^{-1}) &= \frac{1}{(\bar{z}_k^{-1} - z_k)(1 - |c_k|^2)} [\gamma_k(\bar{z}_k^{-1})a_k(\bar{z}_k^{-1}) - \alpha_k(\bar{z}_k^{-1})b_k(\bar{z}_k^{-1})] \\ &= \frac{1}{(\bar{z}_k^{-1} - z_k)(1 - |c_k|^2)} [c_k \bar{w}_k(\bar{z}_k^{-1} - z_k)a_k(\bar{z}_k^{-1}) - c_k(\bar{z}_k^{-1} - z_k)b_k(\bar{z}_k^{-1})]. \end{aligned}$$

Because of (4.16), the above expression is equal to zero, hence

$$a_{k+1}(\bar{z}_k^{-1}) = 0,$$

and in fact $(1 - \bar{z}_k z)$ is a factor of $a_{k+1}(z)$. Clearly, even if $(1 - \bar{z}_k z)$ is already a factor of $d_k(z)$, the above can be used to show that $a_{k+1}(z)$ is divisible by $(1 - \bar{z}_k z)d_k(z)$. In a similar way we conclude that $(1 - \bar{z}_k z)d_k(z)$ divides $b_{k+1}(z)$. Therefore,

$$d_{k+1}(z) = (1 - \bar{z}_k z)d_k(z),$$

and from (4.14) we deduce that

$$\eta_{k+1}(z)(1 - \bar{z}_k z) = \eta_k(z).$$

This concludes the proof. \square

An immediate consequence of the above is that *if for no point in the sequence $(z_k, k = 1, 2, \dots)$ we have $(1 - \bar{z}_k z)$ as a factor of $\eta_1(z)$, then*

$$\eta_k(z) = \eta_1(z) \quad \text{for } k = 1, 2, \dots,$$

and also $d_k(z) \equiv 1$ for all values of k . Alternatively, *if all roots of $\eta_1(z)$, including multiplicities, have inverse complex conjugate values belonging to $(z_k, k = 1, 2, \dots)$, then after a finite number of steps we will have that*

$$d_n(z) = d_{n+l}(z) = \eta_1(z) \quad \text{for } l = 1, 2, \dots,$$

while $\eta_{n+l}(z) = 1$.

However, regardless of how the points $z_k, k = 1, 2, \dots$, are chosen (provided a mild condition on their distribution is met), the polynomials $a_k(z)$ and $b_k(z)$ as $k \rightarrow \infty$, tend to the zero polynomial and $\eta_1(z)$ respectively. This is the content of the next theorem.

THEOREM 4.17. *Let $f_1(z)$ be a rational \mathbb{U} -function (with $f_1(0) = 0$), $\eta_1(z)$ be the associated spectral numerator, and $f_1(z) = a_1(z)/b_1(z)$ a representation of f_1 as the ratio of two coprime polynomials satisfying $a_1(0) = 0$ and $b_1(0) = 1$. Let $z_k, k = 1, 2, \dots$, be a sequence of points in \mathbb{D} satisfying*

$$(3.4) \quad \sum_{k=1}^{\infty} (1 - |z_k|) = \infty,$$

and $(a_k, b_k), k = 2, 3, \dots$, be obtained from (4.3), (4.5)–(4.7). Then as $k \rightarrow \infty$,

$$b_k(z) \rightarrow \eta_1(z), \quad a_k(z) \rightarrow 0$$

coefficientwise.

Proof. Let

$$(4.18) \quad \frac{a_k(z)}{b_k(z)} = \rho_1 z + \rho_2 z^2 + \cdots + \rho_m z^m + \cdots$$

be a Taylor series for $f_k(z)$ around the origin. Since $f_k(z)$ is a rational function in \mathbb{U} , it is analytic in \mathbb{D} and also continuous on the boundary (because of the rationality). Corollary 3.5 now implies that $f_k(z)$ tends uniformly to zero on compact subsets of \mathbb{D} and in particular that $\rho_m \rightarrow 0$, for all m , as $k \rightarrow \infty$.

From Proposition 4.8, the polynomial $b_k(z)$ satisfies $b_k(0) = 1$ for all values of k and has no root in \mathbb{D} (because f_k is in \mathbb{U} and $d_k(z)$ has no root in \mathbb{D} as discussed earlier). Therefore, the coefficients of b_k are all bounded by one. Also $\rho_m \rightarrow 0$, uniformly in m for $m = 1, \dots, l$, when $k \rightarrow \infty$, and l being any finite integer. We conclude that $a_k(z) \rightarrow 0$ as a polynomial; i.e., its coefficients tend to zero.

From the proof of Theorem 4.10 we now have that

$$\begin{aligned} |b_k(z)|^2 - |a_k(z)|^2 &= |\mu_k|^2 |d_k(z) \eta_k(z)|^2 \\ &= |\mu_k \eta_1(z)|^2 \quad \text{on } \mathbb{T}. \end{aligned}$$

But $a_k(z) \rightarrow 0$, whereas $b_k(z)$ and $\eta_1(z)$ are polynomials that have no root in \mathbb{D} and have value one at the origin. Therefore

$$b_k(z) \rightarrow \eta_1(z),$$

and $\mu_k \rightarrow 1$, when $k \rightarrow \infty$. \square

Theorem 4.17 provides a general recursive scheme in the form of relations (4.3), (4.5)–(4.7), for obtaining the spectral factor of a rational \mathbb{U} -function $f_1(z) = a_1(z)/b_1(z)$ as summarized below:

1. Select a sequence of points $(z_k$ in \mathbb{D} : $k = 1, 2, \dots$) satisfying (3.4).

Iterate step 2 for $k = 1, 2, 3, \dots$:

2. Given $(a_k(z), b_k(z))$ compute $(a_{k+1}(z), b_{k+1}(z))$ using (4.7), and λ_{k+1} using

$$\lambda_{k+1} = \frac{1 - |c_k|^2}{1 - |w_k|^2} \lambda_k,$$

with $\lambda_1 = 1$, and the parameters w_k and c_k obtained from (4.3) and (4.6) respectively.

3. Then as $k \rightarrow \infty$, $b_k(z)$ approaches the spectral numerator of $f_1(z)$ and $\lambda_k b_k(z)/b_1(z)$ approaches the canonical spectral factor of $f_1(z)$.

The choice of the sequence $(z_k, k = 1, 2, \dots)$ is arbitrary provided they do not converge too fast towards the boundary; i.e., condition (3.4) of the theorem is met. However, the choice of this sequence influences the speed of the convergence $b_k(z) \rightarrow \eta_1(z)$. But the convergence itself is guaranteed by the theorem. It appears that the choice of the z_k 's in the vicinity of the roots of η_1 results in a relatively fast convergence. This may potentially be useful when η_1 has roots on or very near the boundary of \mathbb{D} . However, a thorough analysis of the numerical properties and speed of convergence will be pursued elsewhere.

5. Proof of Theorem 3.3. Define $\mathbb{C} := \{F(z) \text{ analytic and with positive real part in } \mathbb{D}\}$. (\mathbb{C} for Caratheodory; also the class of positive real functions.) With any function f in \mathbb{U} we associate the function

$$(5.1) \quad F(z) = \frac{1 - f(z)}{1 + f(z)} \quad \text{for } z \in \mathbb{D}.$$

It is well known that F is a \mathbb{C} -function. (In fact (5.1) sets up a bijective correspondence between \mathbb{U} and \mathbb{C} .) Also, the condition $f(0) = 0$, which appeared earlier, translates into $F(0) = 1$.

The real part of $F(z)$ is defined almost everywhere on \mathbb{T} and is equal to

$$(5.2) \quad \tau(\theta) := \operatorname{Re} \{F(e^{i\theta})\} = \frac{1 - |f(e^{i\theta})|^2}{|1 + f(e^{i\theta})|^2} \quad \text{a.e. on } \mathbb{T}.$$

The function $1 + f$ has no root in \mathbb{D} . Hence, $\ln |1 + f|$ is a harmonic function in \mathbb{D} . Provided $f(0) = 0$, $\ln |1 + f(z)|$ has value at the origin equal to zero. Therefore, the integral of $\ln |1 + f(e^{i\theta})|$ in the interval $[-\pi, \pi]$ is equal to zero. Then, by applying $\ln(\cdot)$ to both sides in (5.2) and integrating we derive that

$$(5.3) \quad \int_{-\pi}^{\pi} \ln [\tau(\theta)] d\theta = \int_{-\pi}^{\pi} \ln [1 - |f(e^{i\theta})|^2] d\theta.$$

The above integral plays a key role in an approximation problem for analytic functions (Szegő's Theorem—see Grenander and Szegő [16]):

$$\begin{aligned} & \inf \left\{ (2\pi)^{-1} \int_{-\pi}^{\pi} |p(e^{i\theta})|^2 \tau(\theta) d\theta : p(z) \text{ polynomial with } p(0) = 1 \right\} \\ &= \exp \left\{ (2\pi)^{-1} \int_{-\pi}^{\pi} \ln [\tau(\theta)] d\theta \right\}. \end{aligned}$$

(The left-hand side of the above equality can be seen as the error of approximating 1 with polynomials vanishing at the origin, in $L^2[\tau(\theta)d\theta]$.) In general *the infimum is attained for a function $p_0(z)$ in H^2* —the subspace of L^2 functions with analytic continuation inside \mathbb{D} . (This in general is not a polynomial and turns out to be a scalar multiple of the inverse of the canonical spectral factor corresponding to $\tau(\theta)$.) Hence,

$$(5.4) \quad (2\pi)^{-1} \int_{-\pi}^{\pi} |p_0(e^{i\theta})|^2 \tau(\theta) d\theta = \exp \left\{ (2\pi)^{-1} \int_{-\pi}^{\pi} \ln [\tau(\theta)] d\theta \right\},$$

where $p_0(0) = 1$.

Let now $(z_k, k = 1, 2, \dots)$ be a sequence of points in \mathbb{D} . Define $K_n := (zB_n(z)H^2)^\perp$, where $B_n(z)$ denotes the Blaschke product corresponding to the first n “interpolation” points

$$B_n(z) := \prod_{k=1}^n \frac{(z - z_k)}{(1 - \bar{z}_k z)},$$

and “ $^\perp$ ” denotes the “orthogonal complement of.” K_n is a finite dimensional linear space. Let $\tau(\theta)$ be the real part of a \mathbb{C} -function $F(z)$ for $z = e^{i\theta}$. Then $\tau(\theta)$ is defined a.e. in $[-\pi, \pi]$ and it is a nonnegative valued function. K_n can be endowed with an inner product defined by

$$\langle f, g \rangle_\tau := (2\pi)^{-1} \int_{-\pi}^{\pi} f(e^{i\theta}) \overline{g(e^{i\theta})} \tau(\theta) d\theta,$$

and then we will use the obvious notation $\|p(z)\|_{\tau(\theta)}^2$. Bultheel and Dewilde [4, Cor. 1] and Dewilde and Dym [8, Lemma 4.3] have considered approximation with functions in K_n and have shown that

$$(5.5) \quad \inf \{ \|p(z)\|_{\tau(\theta)}^2 : p(z) \text{ in } K_n \text{ and } p(0) = 1 \} = \prod_{k=1}^n \left(\frac{1 - |c_k|^2}{1 - |w_k|^2} \right).$$

Clearly,

$$\begin{aligned} \inf \{ \|p(z)\|_{\tau(\theta)}^2 : p(z) \text{ in } K_n \text{ and } p(0) = 1 \} &\geq \|p_0(z)\|_{\tau(\theta)}^2 \\ &= \exp \left\{ (2\pi)^{-1} \int_{-\pi}^{\pi} \ln [\tau(\theta)] d\theta \right\} = \exp \left\{ (2\pi)^{-1} \int_{-\pi}^{\pi} \ln [1 - |f(e^{i\theta})|^2] d\theta \right\}. \end{aligned}$$

In order to prove the theorem, we need to establish that the above holds with equality. To show this it suffices to show that

$$K := \bigcup_{n=1}^{\infty} K_n \text{ is dense in } H^2$$

with respect to $\langle \cdot, \cdot \rangle_{\tau(\theta)}$.

We now briefly indicate that it is sufficient to show that the aforementioned space is dense in H^2 with respect to the standard norm. Since f is a rational function, it can be readily shown that $\tau(\theta) = |g(e^{i\theta})|^2$ where $g(z)$ is a rational function with no poles on the unit circle (since $g(z)$ is an outer function). This implies that $\tau(\theta)$ is bounded from above for all $\theta \in [-\pi, \pi]$. Consequently, convergence in the standard norm implies convergence in $\|\cdot\|_{\tau(\theta)}^2$ and this establishes our claim.

Now we shall use a classical result of Blaschke, which states that

$$(5.6) \quad \sum_{k=1}^{\infty} (1 - |z_k|) = \infty,$$

holds if and only if $B_n(z)$ tends to zero at every point in \mathbb{D} as $n \rightarrow \infty$. Any function q in H^2 that is orthogonal to K belongs to $zB_n H^2$, for all n . Therefore, q must be the zero function. Hence the closure of K is in fact the whole of H^2 . Therefore (5.6) implies that

$$(5.7) \quad \liminf_{n \rightarrow \infty} \{ \|p(z)\|_{\tau(\theta)}^2 : p \in K_n \text{ and } p(0) = 1 \} = \|p_0(z)\|_{\tau(\theta)}^2$$

and consequently that

$$\lim_{n \rightarrow \infty} \prod_{k=1}^n \left(\frac{1 - |c_k|^2}{1 - |w_k|^2} \right) = \exp \left\{ (2\pi)^{-1} \int_{-\pi}^{\pi} \ln [1 - |f(e^{i\theta})|^2] d\theta \right\}.$$

This completes the proof of the theorem. \square

6. Remarks on spectral factorization of \mathbb{C} -functions. So far we have considered spectral factorization of rational \mathbb{U} -functions. In many cases one is given a \mathbb{C} -function $F(z)$ instead. So let

$$F(z) = \frac{\pi(z)}{\chi(z)} \quad \text{be in } \mathbb{C},$$

where $\pi(z)$ and $\chi(z)$ are polynomials in z . Then

$$\operatorname{Re} \{ F(e^{i\theta}) \} = \frac{\pi(z)\bar{\chi}(z^{-1}) + \chi(z)\bar{\pi}(z^{-1})}{\chi(z)\bar{\chi}(z^{-1})} \geq 0 \quad \text{for } z = e^{i\theta},$$

and assumes a factorization

$$\operatorname{Re} \{ F(e^{i\theta}) \} = \frac{|\kappa\eta(z)|^2}{|\chi(z)|^2} \quad \text{for } z = e^{i\theta}, \quad \theta \in [-\pi, \pi],$$

where κ is a positive constant and $\eta(z)$ can be assumed to have no root in \mathbb{D} and to have value equal to one at the origin. Then $\kappa\eta(z)/\chi(z)$ is called the *canonical spectral factor* of $F(z)$, and $\eta(z)$ will be said to be the *spectral numerator* of $F(z)$.

With no loss in generality we may assume that $F(0) = 1$. From (6.1) we can obtain an associated function

$$f(z) = \frac{1 - F(z)}{1 + F(z)} = \frac{a(z)}{b(z)}$$

of class \mathbb{U} . Then $a(z) := \chi(z) - \pi(z)$ and $b(z) := \chi(z) + \pi(z)$ are also polynomials. Let $\eta(z)$ be the spectral numerator of $f(z)$, and

$$g_f(z) = \frac{\kappa\eta(z)}{b(z)}$$

be the canonical spectral factor of f . Then, from (6.1)–(6.2) we obtain that

$$\operatorname{Re}\{F(e^{i\theta})\} = \frac{|\kappa\eta(z)|^2}{|b(z) + a(z)|^2} = \frac{|\kappa\eta(z)|^2}{|\chi(z)|^2} \quad \text{for } z = e^{i\theta}, \quad \theta \in [-\pi, \pi],$$

and the spectral factor of $F(z)$ is $\kappa\eta(z)/\chi(z)$. Thus, the spectral numerator of both $F(z)$ and $f(z)$ is the same. Therefore, when looking for the spectral factor of F , we may consider the corresponding \mathbb{U} -function $f(z)$. Then take

$$a(z) = \chi(z) - \pi(z) \quad \text{and} \quad b(z) = \chi(z) + \pi(z),$$

and apply the algorithm of Theorem 4.17 given by (4.3), (4.5)–(4.7) and an appropriate choice of points $(z_k \in \mathbb{D}, k = 1, 2, \dots)$ to obtain the spectral numerator $\eta(z)$. The algorithm expressed by (4.3), (4.5)–(4.7) can be written directly in terms of polynomial fractions of \mathbb{C} -functions. However, this offers no advantage over (4.3), (4.5)–(4.7), which seem to be simpler and thus preferable.

REFERENCES

- [1] V. M. ADAMJAN, D. Z. AROV AND M. G. KREIN, *Infinite Hankel block matrices and related extension problems*, AMS Transl., (2)111 (1978), pp. 133–156.
- [2] B. D. O. ANDERSON, K. L. HITZ AND N. D. DIEM, *Recursive algorithm for spectral factorization*, IEEE Trans. Circuits and Systems, CAS-21 (1974), pp. 742–750.
- [3] J. A. BALL AND J. W. HELTON, *A Beurling–Lax theorem for the Lie group $U(m, n)$ which contains most of classical interpolation theory*, J. Operator Theory, 9 (1983), pp. 107–142.
- [4] A. BULTHEEL AND P. DEWILDE, *Orthogonal functions related to the Nevanlinna–Pick problem*, Proc. Internat. Conf. on the Mathematical Theory of Networks and Systems, 1981, pp. 207–211.
- [5] F. M. CALLIER, *On polynomial matrix spectral factorization by symmetric extraction*, IEEE Trans. Automat. Control, AC-30 (1985), pp. 453–464.
- [6] P. DELSARTE, Y. GENIN AND Y. KAMP, *A simple algorithm for spectral factorization*, IEEE Trans. Circuits and Systems, CAS-25 (1978), pp. 943–946.
- [7] ———, *On the role of the Nevanlinna–Pick problem in circuit and system theory*, Circuit Theory Appl., 9 (1981), pp. 177–187.
- [8] P. DEWILDE AND H. DYM, *Schur recursions, error formulas and convergence of rational estimators for stationary stochastic sequences*, IEEE Trans. Inform. Theory, IT-27(4) (1981), pp. 446–455.
- [9] ———, *Lossless chain scattering matrices and optimum linear prediction: the vector case*, Circuit Theory Appl., 9 (1981), pp. 135–675.
- [10] P. DEWILDE, A. VIEIRA AND T. KAILATH, *On a generalized Szegő–Levinson realization algorithm for optimal linear predictors based on a network synthesis approach*, IEEE Trans. Circuits and Systems, CAS-25 (1978), pp. 663–675.
- [11] P. FAURRE, M. CLERGET AND F. GERMAIN, *Opérateurs rationnels positifs: application à l’hyperstabilité et aux processus aléatoires*, Dunod, Paris, 1978.
- [12] B. FRIEDLANDER, *A lattice algorithm for factoring the spectrum of a moving average process*, Proc. Conf. on Information Sciences and Systems, Princeton, NJ, 1982, pp. 5–9.
- [13] J. B. GARNETT, *Bounded Analytic Functions*, Academic Press, New York, 1981.
- [14] T. T. GEORGIU AND P. P. KHARGONEKAR, *On the partial realization problem for covariance sequences*, Proc. Conf. on Information Sciences and Systems, Princeton, NJ, 1982, p. 181.

- [15] ———, *Linear fractional transformations and spectral factorization*, IEEE Trans. Automat. Control, AC-31 (1986), pp. 345–347.
- [16] U. GRENANDER AND G. SZEGÖ, *Toeplitz Forms and their Applications*, Chelsea, 2nd edition, New York, 1985.
- [17] J. W. HELTON, *Orbit structure of the Mobius transformation semigroup acting on H^∞ (broadband matching)*, in Topics in Functional Analysis, Adv. in Math. Suppl. Stud., 3, 1978, pp. 129–197.
- [18] T. KAILATH, *A view of three decades of linear filtering theory*, IEEE Trans. Inform. Theory, IT-20 (1974), pp. 146–181.
- [19] P. P. KHARGONEKAR AND A. R. TANNENBAUM, *Noneuclidean metrics and the robust stabilization of systems with parameter uncertainty*, IEEE Trans. Automat. Control, AC-30 (1985), pp. 1005–1013.
- [20] B. SZ. NAGY AND C. FOIAŞ, *Harmonic Analysis of Operators on Hilbert Space*, North-Holland, Amsterdam, 1970.
- [21] J. RISSANEN AND T. KAILATH, *Partial realization of random systems*, Automatica, 8 (1972), pp. 389–396.
- [22] W. RUDIN, *Real and Complex Analysis*, McGraw-Hill, New York, 1974.
- [23] R. SAEKS, *The factorization problem—A survey*, Proc. IEEE, 64 (1976), pp. 90–95.
- [24] D. SARASON, *Generalized interpolation in H^∞* , Trans. AMS, 127 (1967), pp. 179–203.
- [25] D. C. YOULA AND M. SAITO, *Interpolation with positive-real functions*, J. Franklin Inst., 284 (1970), pp. 77–1108.
- [26] G. ZAMES AND B. A. FRANCIS, *Feedback, minimax sensitivity, and optimal robustness*, IEEE Trans. Automat. Control, AC-28 (1983), pp. 585–601.

AVERAGING AND DETERMINISTIC OPTIMAL CONTROL*

F. CHAPLAIS†

Abstract. Averaging is often used in ordinary differential equations when dealing with fast periodic phenomena. It is shown here that it can be used efficiently in optimal control. As the period tends to zero, a limit or “averaged” problem is defined. The open loop optimal control of the limit problem induces a cost which is optimal up to the second order when evaluated through the original dynamics. The definition of the averaged problem is then generalized to the nonperiodic case. It is shown that the Bellman function of the original “fast” problem tends uniformly on any compact set to that of the averaged problem.

Key words. averaging, optimal control, time scales, perturbations in control, Hamilton–Jacobi equations

AMS(MOS) subject classifications. 34C29, 41A60, 4900, 49C20

Introduction: A perturbation approach. Averaging can be seen as part of the perturbation theory of differential equations. Consider

$$(1) \quad \frac{dx}{dt} = f^\varepsilon(x, t), \quad x(0) = x_0, \quad x(t) \in \mathbf{R}^n, \quad t \in [0, T].$$

Regular perturbations correspond to the situation when f^ε has a limit in $C^0(\mathbf{R}^n \times [0, T], \mathbf{R}^n)$ as ε tends to zero. Singular perturbations [6] can be seen as f^ε having a limit in $L^2(\mathbf{R}^n \times [0, T], \mathbf{R}^n)$. Roughly speaking, averaging is the case when f^ε has a limit in $L^2(\mathbf{R}^n \times [0, T], \mathbf{R}^n)$ in the weak topology. As an example, consider

$$(2) \quad \frac{dx}{dt} = f\left(x, t, \frac{t}{\varepsilon}\right)$$

where f is periodic in the last variable. Define $f^\varepsilon(x, t) = f(x, t, t/\varepsilon)$; clearly f^ε has no limit either pointwise or in L^2 ; yet f^ε tends to f^0 defined by $f^0(x, t) = 1/\omega \int_0^\omega f(x, t, \theta) d\theta$, weakly in L^2 .

It is well known [1] that the solution of (2) can be approximated by the solution of

$$(3) \quad \frac{dy}{dt} = f^0(y, t), \quad y(0) = x_0$$

with an error of order $\varepsilon\omega$, provided that f be regular enough. Solving (3) instead of (2) is known as “averaging”.

What is good for differential equations is often good for optimal control. Regular and singular perturbations have been extensively studied ([2], [6]). As far as averaging is concerned, it has been studied in the context of partial differential equations [3] or stochastic optimal control [5]. We present here some approximation results in deterministic optimal control.

1. The periodic case.

1.1. The averaged problem. Let

$$f: \mathbf{R}^n \times \mathbf{R}^p \times [0, T] \times \mathbf{R} \rightarrow \mathbf{R}^n, \\ (x, u, t, \theta) \rightarrow f(x, u, t, \theta),$$

* Received by the editors February 3, 1986; accepted for publication (in revised form) May 5, 1986.

† Centre d'Automatique et Informatique de l'Ecole Nationale Supérieure des Mines de Paris, 35, rue Saint-Honoré, 77305 Fontainebleau Cedex, France.

$$L: \mathbf{R}^n \times \mathbf{R}^p \times [0, T] \times \mathbf{R} \rightarrow \mathbf{R},$$

$$(x, u, t, \theta) \rightarrow L(x, u, t, \theta),$$

with f and L periodic in θ with a period ω independent of x, u and t (regularity will be considered in § 1.2). Let U^{ad} be a constraint domain and $W^{ad} = \{u \in L^2([0, T], \mathbf{R}^p), u(t) \in U^{ad} \text{ for almost every } t\}$ the set of admissible controls. We define the problem (P^ε) as:

$$(4) \quad \left. \begin{aligned} \frac{dx}{dt} &= f\left(x(t), u(t), t, \frac{t}{\varepsilon}\right), \quad x(0) = x_0, \\ \text{Minimize } \int_0^T L\left(x(t), u(t), t, \frac{t}{\varepsilon}\right) dt &\text{ in } W^{ad}, \end{aligned} \right\} (P^\varepsilon).$$

We define the associated problem \bar{P} as follows:

$$(5) \quad \left. \begin{aligned} V^{ad} &= \{v \in L^2([0, T] \times [0, \omega], \mathbf{R}^p), v(t, \theta) \in U^{ad} \text{ a.e. } (t, \theta)\}, \\ \frac{dy}{dt} &= \frac{1}{\omega} \int_0^\omega f(y(t), v(t, \theta), t, \theta) d\theta, \quad y(0) = x_0, \\ \text{Minimize } \int_0^T \frac{dt}{\omega} \int_0^\omega L(y(t), v(t, \theta), t, \theta) d\theta &\text{ in } V^{ad}, \end{aligned} \right\} (\bar{P}).$$

Problem (4) can be seen as the perturbation of (5) by a fast oscillating input of null average. Notice that if f is Lipschitz in x and measurable in u, t, θ , then, for v in V^{ad} , the average \bar{f} of f as defined in (5) is Lipschitz in x and measurable in t . Hence, both (4) and (5) have a unique solution over $[0, T]$.

Remark 1. At first sight, it would seem reasonable to consider in (5) the averages of f and L in θ *independently* of u , that is, u being a constant vector in \mathbf{R}^p and not a function in $L^2([0, \omega], \mathbf{R}^p)$. This amounts to restricting V^{ad} to controls which are constant over $[0, \omega]$ at time t . As we shall see, this may lead to a severe loss of optimality when ε tends to 0. In short, it is necessary to have a feedback on the fast time θ . Consider for instance the following problem: $n = p = 1$, $U^{ad} = \mathbf{R}$, $f(x, u, t, \theta) = -x + u \sin(\theta)$ and $L(x, u, t, \theta) = x^2 + (u^2/6)$. If u is independent of θ , the average of f is equal to $-x$, which is itself independent of u . Within this class of functions, the optimal cost for (P^ε) is asymptotically equal to $\bar{J}(0)$, if $\bar{J}(u)$ denotes the cost of u in (\bar{P}) . $\bar{J}(0)$ is equal to $x_0^2(1 - e^{-2T})/2$.

Now use our definition of the averaged problem to compute a better control. Let q be the solution of the Riccati equation of the averaged problem: $dq/dt = 3q^2 + 2q - 1$, $q(T) = 0$; then $q(0) = (1 - e^{-4T})/(3 + e^{-4T})$. Define z by $dz/dt = -z(1 + 3q)$, $z(0) = x(0)$; z is the optimal trajectory for the averaged problem. Now use the open-loop control $u^\varepsilon(t) = -6q(t)z(t) \sin(t/\varepsilon)$ in problem (P^ε) . Then it is easy to see that x^ε driven by u^ε in (4) is asymptotic to z . Hence the cost of u^ε in (P^ε) is asymptotic to $\int_0^T z^2(t) \times (1 + 3q^2(t)) dt$. Thanks to the Riccati equation, the latter quantity is equal to $q(0)x^2$, which is smaller than $\bar{J}(0)$; thus for ε small enough, u^ε is better than any "slow" control.

Notice that the same phenomenon can be observed in stochastic optimal control, where it is well known that open-loop controls alone are not enough to ensure optimality. In both cases this is due to the presence of averages w.r.t. the events or fast time in the problem (the same can be said of the "ordinary," "slow" time, of course).

Remark 2. Even though the controls of problem (\bar{P}) are in an infinite dimensional state, the minimum principle and dynamic programming apply; the important fact is that the state is finite-dimensional. One also checks that the Bellman function is the viscosity solution of the Hamilton-Jacobi-Bellman equation [7].

Let $\bar{f}(x, v, t)$ denote the average of f and \bar{L} the average of L . The adjoint state equations for the problem (\bar{P}) are:

$$\frac{dp}{dt} = -\frac{\partial \bar{f}^T}{\partial x} p - \frac{\partial \bar{L}}{\partial x}, \quad p(T) = 0$$

and if v^* is an optimal control for the problem (\bar{P}) , the minimum principle says that $v^*(t, \cdot)$ minimizes the Hamiltonian of (\bar{P}) , that is, $p(t)^T \bar{f}(x, v, t) + \bar{L}(x, v, t)$, with respect to v in V^{ad} . This is equivalent to: $v^*(t, \theta)$ minimizes $p(t)^T f(x, v, t, \theta) + L(x, v, t, \theta)$ with respect to v in U^{ad} , this for almost every θ . We see that the control appears naturally as a feedback on the fast time (cf. Remark 1).

Remark 3. (\bar{P}) is better conditioned numerically than (P^ε) , since (P^ε) involves a time grid of order $\delta t/\varepsilon$, while (\bar{P}) involves only a grid of order δt , at least in the simulation part. A time grid in $\delta t/\varepsilon$ is still needed to compute the averages and to minimize the Hamiltonian.

1.2. An approximation theorem. The averaged problem is used here to compute a near optimal control for the problem (P^ε) , with an error of order ε^2 . The proof and assumptions are close to those used in [2].

Assumptions.

- (H1) $U^{ad} = \mathbf{R}^p$; no constraint.
- (H2) L does not depend on θ ; f and L are C^0 , of class C^2 in x , and u ; the first- and second-order derivatives of f are bounded, and Lipschitz; the second-order derivatives of L are bounded, and Lipschitz.
- (H3) (\bar{P}) has a solution, with optimal control u_0 , trajectory y and adjoint state q .
- (H4) Let $H(p, x, u, t, \theta) = p^T f(x, u, t, \theta) + L(x, u, t, \theta)$. There exists $\beta > 0$ such that, for all (x, u, t, θ) , $\partial^2 H / \partial v^2$ is greater than β Identity at point $(q(t), x, u, t, \theta)$, in the sense of the cone of positive semidefinite matrices.
- (H5) One has

$$\left[\frac{\partial^2 H}{\partial x^2} - \frac{\partial^2 H}{\partial x \partial v} \left(\frac{\partial^2 H}{\partial v^2} \right)^{-1} \frac{\partial^2 H}{\partial v \partial x} \right] \geq 0 \quad \text{at } (q(t), x, u, t, \theta) \text{ for all } (x, u, t, \theta).$$

- (H6) $f(y(t), u_0(t, \theta), t, \theta)$ and $\frac{\partial H}{\partial x}(q(t), y(t), u_0(t, \theta), t, \theta)$ are C^1 in t with a Lipschitz derivative.

Assumptions (H4) and (H5) ensure sufficient regularity of the control with respect to the cost. Assumption (H6) is specific to averaging: if it did not hold, there would not be a true separation of time scales.

THEOREM 1. *Let f and L meet Assumptions (H1)–(H6). Define u^ε by $u^\varepsilon(t) = u_0(t, t/\varepsilon)$. Then u^ε is near optimal for the problem (P^ε) with an error on the cost of order ε^2 .*

Sketch of proof. (A detailed proof can be found in [4].) We will proceed in two stages. First, we will exhibit a lower bound of the cost for any control u . We will then show that the control u^ε induces a cost which approximates this lower bound with an error of order ε^2 . It is then easy to conclude that u^ε is near optimal for the problem (P^ε) .

Remark 4. Unless otherwise stated, all partial derivatives will be taken at time t , with x being $y(t)$, u being $u_0(t, t/\varepsilon) = u^\varepsilon(t)$, θ being t/ε and the adjoint state being $q(t)$. We will make use of the following conventions: if $g(\sigma, \theta)$ is a periodic function

of θ , we will denote by $\bar{g}(\sigma)$ or $\bar{g}(\sigma, \cdot)$ or $\text{av}(g(\sigma, \cdot))$ the average of g in θ . We then define the operator Π on periodic functions g , by $\Pi(g(\sigma, \cdot))$ being the only primitive in θ of $g - \bar{g}$ with a null average. Π plays a key role in all developments of integrals or solutions of differential equations involving a periodicity in fast time. Finally, the superscript T denotes transposition.

LEMMA 1. Let $x_2(t, \theta) = \Pi(f(y(t), u_0(t, \cdot), t, \cdot)(\theta))$. There exists $\varepsilon_0 > 0$, $k \geq 0$, such that, for any control u and any $\varepsilon \in]0, \varepsilon_0[$, x being the trajectory driven by u in (4), the following estimation holds:

$$\begin{aligned} \int_0^T L(x(t), u(t), t) dt &\geq \int_0^T L\left(y(t), u_0\left(t, \frac{t}{\varepsilon}\right), t\right) dt \\ &\quad + \varepsilon \int_0^T \text{av} \left[\frac{\partial H}{\partial x}(y(t), q(t), u_0(t, \cdot), t, \cdot) x_2(t, \cdot) \right] dt \\ &\quad - \varepsilon q(0)^T x_2(0, 0) - k\varepsilon^2. \end{aligned}$$

Proof of Lemma 1. We will denote by \tilde{x} and \tilde{u} the following errors:

$$\tilde{x}(t) = x(t) - y(t) - \varepsilon x_2\left(t, \frac{t}{\varepsilon}\right), \quad \tilde{u}(t) = u(t) - u_0\left(t, \frac{t}{\varepsilon}\right).$$

It should be noticed that, if x_1 is defined by

$$(6) \quad \frac{dx_1}{dt} = \frac{\partial \tilde{f}}{\partial x} x_1 + \frac{\partial \tilde{f}}{\partial x} x_2, \quad x_1(0) + x_2(0, 0) = 0$$

and if $\tilde{u} = 0$ (that is, x is the trajectory driven by u^ε), then $y + \varepsilon x_1 + \varepsilon x_2(t, t/\varepsilon)$ is a uniform approximation of x with an ε^2 error.

We will use developments of functions with integral remains. To this end, we will use the following notation: for λ in $[0, 1]$, μ in $[0, 1]$, $F_t(\lambda, \mu)$ will denote the Hessian symmetric operator associated with the second-order derivatives of f in x and u , at point $(y(t), \lambda \mu(\varepsilon x_2(t, t/\varepsilon) + \tilde{x}(t)), u_0(t, t/\varepsilon) + \lambda \mu \tilde{u}(t), t, t/\varepsilon)$. $L_t(\lambda, \mu)$ will denote the analogue for L and $H_t(\lambda, \mu) = q^T(t) F_t(\lambda, \mu) + L_t(\lambda, \mu)$ will denote the Hessian of H at the same point.

We will also make use of the following linear quadratic oscillatory “tangent” problem:

$$(7) \quad \left. \begin{aligned} \frac{dz}{dt} &= \frac{\partial f}{\partial x} \left(z + x_2\left(t, \frac{t}{\varepsilon}\right) \right) + \frac{\partial f}{\partial u} v, \quad z(0) + x_2(0, 0) = 0, \\ \text{Min}_v \int_0^T &\left[\frac{1}{2} (z^T, v^T) H_t(0, 0) (z^T, v^T)^T + p_2^T \left(t, \frac{t}{\varepsilon} \right) \left(\frac{\partial f}{\partial x} z + \frac{\partial f}{\partial u} v \right) \right] dt, \end{aligned} \right\} (TP^\varepsilon)$$

where $p_2 = -\Pi(\partial H^T / \partial x)$. p_2 is the analogue of x_2 for the adjoint state.

Thanks to (H4) and (H5), (TP^ε) has a unique solution with trajectory y_1 , optimal control v_1 and adjoint state q_1 . Moreover, $\|y_1\|_\infty$ and $\|v_1\|_\infty$ are bounded when ε ranges within a neighbourhood of zero. Finally estimates will bear on the quantity

$$z_\varepsilon^2 = \int_0^1 \lambda d\lambda \int_0^1 d\mu \|Z_\varepsilon(\lambda, \mu)\|_{L^2}^2$$

where

$$\begin{aligned} Z_\varepsilon(\lambda, \mu)(t) &= v(t) + \left[\left(\frac{\partial^2 H}{\partial v^2} v \right)^{-1} \frac{\partial^2 H}{\partial v \partial x} \right] \left(r(t) + \varepsilon x_2\left(t, \frac{t}{\varepsilon}\right) \right), \\ r &= \tilde{x} - \varepsilon y_1, \quad v = \tilde{u} - \varepsilon v_1, \end{aligned}$$

the derivatives being taken at the same interpolation points as for $F_t(\lambda, \mu)$, $L_t(\lambda, \mu)$ and $H_t(\lambda, \mu)$.

At last, \approx will denote an approximation with an ε^2 error.

LEMMA 1.1.

$$\begin{aligned} \int_0^T L(x, u, t) dt &\approx \int_0^T L(y, u^\varepsilon, t) dt + \varepsilon \int_0^T \frac{\partial \bar{H}}{\partial x} x_2 dt - \varepsilon q(0)^T x_2(0, 0) \\ &\quad + \int_0^T \left[\frac{\partial H}{\partial x} - \frac{\partial \bar{H}}{\partial x} \right] \tilde{x} dt \\ &\quad + \int_0^T dt \int_0^1 \lambda d\lambda \int_0^1 d\mu \left(\tilde{x}^T + \varepsilon x_2^T \left(t, \frac{t}{\varepsilon} \right), \tilde{u}^T \right) H_t(\lambda, \mu) \\ &\quad \times \begin{pmatrix} \tilde{x} + \varepsilon x_2(t, t/\varepsilon) \\ \tilde{u} \end{pmatrix}. \end{aligned}$$

Proof of Lemma 1.1. The cost is expanded at the second order using an integral remain. Since there is no constraint, $q^T(\partial f/\partial u) + (\partial L/\partial u) = 0$; the Hamiltonian appears after a classical integration by parts. We then neglect all integrals depending on fast time when of order larger than or equal to ε^2 .

LEMMA 1.2. *There exists $k \geq 0$ such that*

$$\|r\|_\infty^2 \leq k(\varepsilon^2 + z_\varepsilon^2) \quad \text{and} \quad \|v\|_{L^2}^2 \leq k(\varepsilon^2 + z_\varepsilon^2).$$

Proof of Lemma 1.2.

$$\begin{aligned} \frac{dr}{dt} &= \frac{dx}{dt} - \frac{dy}{dt} - \varepsilon \frac{d}{dt} \left(x_2 \left(t, \frac{t}{\varepsilon} \right) \right) - \varepsilon \frac{dy_1}{dt} \\ (8) \quad &= f \left(r + y + \varepsilon x_2 + \varepsilon y_1, v + u_0 + \varepsilon v_1, t, \frac{t}{\varepsilon} \right) - f \left(y, u_0 \left(t, \frac{t}{\varepsilon} \right), t, \frac{t}{\varepsilon} \right) - \varepsilon \frac{\partial f}{\partial x} (y_1 + x_2) \\ &\quad - \varepsilon \frac{\partial f}{\partial u} v_1 - \varepsilon \Pi \left[\frac{\partial}{\partial t} (f(y, u_0(t, \cdot), t, \cdot)) \right] \left(\frac{t}{\varepsilon} \right), \end{aligned}$$

since $\partial/\partial\theta[\Pi(f)] = f - \bar{f}$. The second expression is equal to

$$\begin{aligned} &f \left(r + y + \varepsilon x_2 + \varepsilon y_1, v + u_0 + \varepsilon v_1, t, \frac{t}{\varepsilon} \right) - f \left(y + \varepsilon x_2 + \varepsilon y_1, u_0 + \varepsilon v_1, t, \frac{t}{\varepsilon} \right) \\ &\quad + \varepsilon^2 \int_0^1 \lambda d\lambda \int_0^1 d\mu (x_2^T + y_1^T, v_1^T) F_t(\lambda, \mu) \begin{pmatrix} x_2 + y_1 \\ v_1 \end{pmatrix} - \varepsilon \Pi \left[\frac{\partial}{\partial t} (f) \right] \left(\frac{t}{\varepsilon} \right). \end{aligned}$$

v then appears as a difference in the controls and is replaced by

$$Z_\varepsilon(\lambda, \mu) - \left[\left(\frac{\partial^2 H}{\partial v^2} \right)^{-1} \frac{\partial^2 H}{\partial v \partial x} \right] (r + \varepsilon x_2).$$

Using the Gronwall lemma, and neglecting the integrals of fast periodic functions at the second order, we get the estimate on $\|r\|_\infty$. The estimation on $\|v\|_{L^2}$ is obtained through the definition of Z_ε .

LEMMA 1.3. *r can be approximated uniformly with an error of order ε^2 by r_1 , with*

$$\begin{aligned} (9) \quad \frac{dr_1}{dt} &= f \left(r_1 + y + \varepsilon x_2 + \varepsilon y_1, v + u_0 + \varepsilon v_1, t, \frac{t}{\varepsilon} \right) - f \left(y + \varepsilon x_2 + \varepsilon y_1, u_0 + \varepsilon v_1, t, \frac{t}{\varepsilon} \right), \\ r_1(0) &= 0. \end{aligned}$$

In particular, one has $\|r_1\|_\infty^2 \leq 2k(\varepsilon^2 + z_\varepsilon^2)$.

Proof of Lemma 1.3. Note that (9) is close to (8). We then get the result by using the Gronwall lemma. We will now use r_1 rather than r .

LEMMA 1.4. *There exists $k \geq 0$ such that, for $\varepsilon \in]0, 1]$:*

$$\int_0^T \left(\frac{\partial H}{\partial x} - \frac{\partial \bar{H}}{\partial x} \right) \tilde{x} dt \geq \varepsilon \int_0^T p_2^T \left(t, \frac{t}{\varepsilon} \right) \left[\frac{\partial f}{\partial x} r_1 + \frac{\partial f}{\partial u} v \right] dt - k\varepsilon^2 - k\varepsilon z_\varepsilon^2.$$

Proof of Lemma 1.4.

$$-\varepsilon \frac{d}{dt} \left(p_2^T \left(t, \frac{t}{\varepsilon} \right) \right) = \varepsilon \Pi \left[\frac{\partial}{\partial t} \left(\frac{\partial H}{\partial x} (q(t), y(t), u_0(t, \cdot), t, \cdot) \right) \right] \left(\frac{t}{\varepsilon} \right) + \frac{\partial H}{\partial x} - \frac{\partial \bar{H}}{\partial x}$$

and $\tilde{x} \approx r_1 + \varepsilon y_1$. Hence, integrating by parts, and neglecting second order terms,

$$\begin{aligned} \int_0^T \left(\frac{\partial H}{\partial x} - \frac{\partial \bar{H}}{\partial x} \right) \tilde{x} dt &\approx -\varepsilon \int_0^T \Pi \left[\frac{\partial}{\partial t} \left(\frac{\partial H}{\partial x} (q, y, u_0(t, \cdot), t, \cdot) \right) \right] \left(\frac{t}{\varepsilon} \right) r_1 dt \\ &\quad + \varepsilon \int_0^T p_2^T \left(t, \frac{t}{\varepsilon} \right) \left[f \left(x, u, t, \frac{t}{\varepsilon} \right) - f \left(x - r_1, u - v, t, \frac{t}{\varepsilon} \right) \right] dt. \end{aligned}$$

But

$$\begin{aligned} &\varepsilon \int_0^T p_2^T \left(t, \frac{t}{\varepsilon} \right) \left[f \left(x, u, t, \frac{t}{\varepsilon} \right) - f \left(x - r_1, u - v, t, \frac{t}{\varepsilon} \right) \right] dt \\ &\quad \geq \varepsilon \int_0^T p_2^T \left(t, \frac{t}{\varepsilon} \right) \left[\frac{\partial f}{\partial x} r_1 + \frac{\partial f}{\partial u} v \right] dt - k\varepsilon \int_0^T \left| p_2^T \left(t, \frac{t}{\varepsilon} \right) \right| (|r_1|^2 + |v|^2) dt \\ &\quad \geq \varepsilon \int_0^T p_2^T \left(t, \frac{t}{\varepsilon} \right) \left[\frac{\partial f}{\partial x} r_1 + \frac{\partial f}{\partial u} v \right] dt - k\varepsilon (\varepsilon^2 + z_\varepsilon^2) \end{aligned}$$

by Lemma 1.2. On the other hand,

$$\begin{aligned} &\varepsilon \int_0^T \Pi \left[\frac{\partial}{\partial t} \left(\frac{\partial H}{\partial x} (q, y, u_0(t, \cdot), t, \cdot) \right) \right] \left(\frac{t}{\varepsilon} \right) r_1 dt \\ &\quad = \varepsilon \left\{ \int_0^T \Pi \left[\frac{\partial}{\partial t} \left(\frac{\partial H}{\partial x} (q, y, u_0(t, \cdot), t, \cdot) \right) \right] \left(\frac{t}{\varepsilon} \right) dt \right\} r_1(T) \\ &\quad \quad - \varepsilon \int_0^T \left\{ \int_0^t \Pi \left[\frac{\partial}{\partial t} \left(\frac{\partial H}{\partial x} (q, y, u_0(s, \cdot), s, \cdot) \right) \right] \left(\frac{s}{\varepsilon} \right) ds \right\} \frac{dr_1}{dt} dt \\ &\quad \geq -k\varepsilon^2 \left(\|r_1\|_\infty + \left\| \frac{dr_1}{dt} \right\|_{L^1} \right), \end{aligned}$$

by averaging estimations. But, from (9) and Lemmas 1.2 and 1.3, one sees that there exists $k > 0$ such that $\|dr_1/dt\|_{L^1} \leq k(1 + z_\varepsilon^2)$; this completes the proof.

We are now going to study the second order term with $H_t(\lambda, \mu)$ in Lemma 1.1.

LEMMA 1.5. *There exists $k > 0$ such that:*

$$\begin{aligned} &\int_0^1 dt \int_0^1 \lambda d\lambda \int_0^1 d\mu (\tilde{x}^T + \varepsilon x_2^T, \tilde{u}^T) H_t(\lambda, \mu) \begin{pmatrix} \tilde{x} + \varepsilon x_2 \\ v \end{pmatrix} \\ &\quad \geq 2\varepsilon \int_0^1 dt \int_0^1 \lambda d\lambda \int_0^1 d\mu (y_1 + v_1^T) H_t(\lambda, \mu) \begin{pmatrix} r + \varepsilon x_2 \\ v \end{pmatrix} - k\varepsilon^2 + \beta z_\varepsilon^2. \end{aligned}$$

Proof of Lemma 1.5. By substituting $Z_\varepsilon - [(\partial^2 H / \partial v^2)^{-1} \partial H / \partial v \partial x](r + \varepsilon x_2)$ to v in the integral and using Assumptions (H4) and (H5), one shows that the expression on the left-hand side is greater than $\int_0^1 dt \int_0^1 \lambda d\lambda \int_0^1 d\mu \beta |Z_\varepsilon(\lambda, \mu)(t)|^2$, that is, βz_ε^2 .

This will be the only positive term in z_ε^2 ; the others will be of the form $-k\varepsilon z_\varepsilon^2$.

Note that Lemma 1.4 already displays one part of the cost to be minimized in problem (TP^ε) . The other part will appear by replacing $H_t(\lambda, \mu)$ by $H_t(0, 0)$ in the estimates of Lemma 1.5.

LEMMA 1.6. *There exists $k \geq 0$ such that:*

$$\begin{aligned} 2\varepsilon \int_0^T dt \int_0^1 \lambda d\lambda \int_0^1 d\mu (y_1^T, v_1^T) H_t(\lambda, \mu) \begin{pmatrix} r + \varepsilon x_2 \\ v \end{pmatrix} \\ \geq \varepsilon \int_0^T (y_1^T, v_1^T) H_t(0, 0) \begin{pmatrix} r + \varepsilon x_2 \\ v \end{pmatrix} dt - k\varepsilon^2 - k\varepsilon z_\varepsilon^2. \end{aligned}$$

Proof of Lemma 1.6. Since the second derivatives of f and L are Lipschitz,

$$\|H_t(\lambda, \mu) - H_t(0, 0)\| \leq K\lambda\mu(|x - y|^2 + |u - u_0|^2)^{1/2} = K\lambda\mu(|r + \varepsilon x_2 + \varepsilon y_1|^2 + |v - \varepsilon v_1|^2)^{1/2}.$$

As x_2, y_1 and v_1 are bounded, there exists $k > 0$ such that:

$$\begin{aligned} \varepsilon \left| \int_0^T dt \int_0^1 \lambda d\lambda \int_0^1 d\mu (y_1 + v_1^T) [H_t(\lambda, \mu) - H_t(0, 0)] \begin{pmatrix} r + \varepsilon x_2 \\ v \end{pmatrix} \right| \\ \leq k\varepsilon(\varepsilon^2 + \|r\|_\infty^2 + \|v\|_{L^2}^2). \end{aligned}$$

We get the result from Lemma 2.

LEMMA 1.7. *There exists $k \geq 0$ such that:*

$$\begin{aligned} \int_0^T L(x, u, t) dt &\geq \int_0^T L(x, u^\varepsilon, t) dt + \varepsilon \int_0^T \overline{\frac{\partial H}{\partial x}} x_2 dt - \varepsilon q(0)^T x_2(0, 0) \\ &\quad + \varepsilon \int_0^T p_2^T \left(t, \frac{t}{\varepsilon} \right) \left[\frac{\partial f}{\partial x} r_1 + \frac{\partial f}{\partial u} v \right] dt \\ &\quad + \varepsilon \int_0^T (y_1^T, v_1^T) H_t(0, 0) \begin{pmatrix} r + \varepsilon x_2 \\ v \end{pmatrix} dt + z_\varepsilon^2(\beta - k\varepsilon) - k\varepsilon^2. \end{aligned}$$

Proof of Lemma 1.7. Combine the results of Lemmas 1.1, 1.4, 1.5 and 1.6.

We are going now to complete the estimate by using the problem (TP^ε) .

LEMMA 1.8. *There exists $k \geq 0$ such that:*

$$\varepsilon \int_0^T p_2^T \left(t, \frac{t}{\varepsilon} \right) \left[\frac{\partial f}{\partial x} r_1 + \frac{\partial f}{\partial v} v \right] dt + \varepsilon \int_0^T (y_1^T, v_1^T) H_t(0, 0) \begin{pmatrix} r + \varepsilon x_2 \\ v \end{pmatrix} dt \geq -k\varepsilon(\varepsilon^2 + z_\varepsilon^2).$$

Proof of Lemma 1.8. Transformations using the adjoint equations of (TP^ε) and the explicit value of v_1 as a feedback yield the following estimate for the expression on the left-hand side:

$$\varepsilon \int_0^T q_1^T \left[f \left(x, u, t, \frac{t}{\varepsilon} \right) - f \left(x - r_1, u - v, t, \frac{t}{\varepsilon} \right) - \frac{\partial f}{\partial x} r_1 - \frac{\partial f}{\partial u} v \right] dt$$

which is clearly of second order in $(\|r_1\|^2 + \|v\|_{L^2}^2)^{1/2}$; from Lemmas 1.2 and 1.3 we get the result.

From Lemmas 1.8 and 1.7 the estimation proposed in Lemma 1 is proved for ε sufficiently small to make βz_ε^2 dominant against $-k\varepsilon$. This completes the proof of Lemma 1.

LEMMA 2. *We are going to estimate the cost induced by U^ε .*

Let now $u = u^\varepsilon$, that is $\tilde{u} = 0$. Then:

$$\int_0^T L(x, u^\varepsilon, t) dt \approx \int_0^T L(y, u^\varepsilon, t) dt - \varepsilon q^T(0)x_2(0, 0) + \varepsilon \int_0^t \overline{\frac{\partial H}{\partial x}} x_2 dt.$$

Proof of Lemma 2. We can use Lemma 1.1 to get a first estimate:

$$\begin{aligned} \int_0^T L(x, u^\varepsilon, t) dt &\approx \int_0^T L(x, u^\varepsilon, t) dt - \varepsilon q^T(0)x_2(0, 0) + \varepsilon \int_0^t \overline{\frac{\partial H}{\partial x}} x_2 dt \\ &\quad + \int_0^T \left(\frac{\partial H}{\partial x} - \frac{\partial \bar{H}}{\partial x} \right) \tilde{x} dt \\ &\quad + \int_0^T dt \int_0^1 \lambda d\lambda \int_0^1 d\mu (\tilde{x}^T + \varepsilon x_2^T) \frac{\partial^2 H}{\partial x^2}(\lambda, \mu) (\tilde{x} + \varepsilon x_2) \end{aligned}$$

where $\partial^2 H / \partial x^2$ is computed at the same point as for $H_t(\lambda, \mu)$. Since \tilde{x} is of order one in ε , the last integral is of order two. Proceeding as in Lemma 1.4, we also show that the integral before that one is of order two. This completes the proof of Lemma 2. Theorem 1 follows from Lemmas 1 and 2.

COROLLARY 1. *If u is a "better" control than u^ε for the problem (P^ε) , then:*

$$\|x - y\|_\infty \leq k\varepsilon \left(1 + \frac{1}{\sqrt{\beta}} \right) \quad \text{and} \quad \|u - u_0\|_{L^2} \leq k\varepsilon \left(1 + \frac{1}{\sqrt{\beta}} \right).$$

Proof of Corollary 1. Denote by $J^\varepsilon(u)$ the cost in problem (P^ε) . For $\varepsilon \in]0, 1]$, one has $J^\varepsilon(u) \geq J^\varepsilon(u^\varepsilon) - k\varepsilon^2 + (\beta - k\varepsilon)z_\varepsilon^2$, with $k \geq 0$. If u is better than u^ε , then $(\beta - k\varepsilon)z_\varepsilon^2 \leq k\varepsilon^2$. Take $\varepsilon < \beta/2k$; then $z_\varepsilon^2 \leq 2k/\beta\varepsilon^2$. The result follows from Lemma 1.2.

2. The nonperiodic case.

2.1. An ergodic theorem on O.D.E. Let

$$f: \begin{cases} \mathbf{R}^n \times [0, T] \times \mathbf{R}_+ \rightarrow \mathbf{R}^n \\ (x, t, \theta) \rightarrow f(x, t, \theta) \end{cases}$$

meeting the following assumptions:

(H7) f is Lipschitz in (x, t) with Lipschitz constant Λ , and integrable in θ .

(H8) f has an average \bar{f} in the sense that, for any bound B , one has:

$$\sup_{\substack{|x| \leq B \\ t \in [0, T] \\ \hat{t} \geq 0}} \left| \frac{1}{\tau} \int_{\hat{t}}^{\hat{t}+\tau} f(x, t, \theta) d\theta - \bar{f}(x, t) \right| \xrightarrow{\tau \rightarrow +\infty} 0.$$

As before, (H7) ensures that one has a true separation of time scales.

Let x^ε and y be defined by:

$$(10) \quad \frac{dx^\varepsilon}{dt} = f\left(x^\varepsilon, t, \frac{t}{\varepsilon}\right), \quad x^\varepsilon(0) = x_0, \quad t \in [0, T],$$

$$(11) \quad \frac{dy}{dt} = \bar{f}(y, t), \quad y(0) = x_0, \quad t \in [0, T].$$

Then

$$\sup_{t \in [0, T]} |x^\varepsilon(t) - y(t)| \xrightarrow{\varepsilon \rightarrow 0} 0.$$

Proof. At a fast time scale, the slow time t can be considered as constant, as well as $x^\varepsilon(t)$ and $y(t)$. Hence the dynamics f can be approximated by \bar{f} ; integration yields the result.

More precisely, let $D \in \mathbf{N}^*$ and, for $t \in [0, T]$, let $t_k = kt/D$. Then:

$$\begin{aligned} |x^\varepsilon(t) - y(t)| &\leq \Lambda \int_0^t |x^\varepsilon(s) - y(s)| ds + \sum_{k=0}^{D-1} \int_{t_k}^{t_{k+1}} \left| f\left(y(s), s, \frac{s}{\varepsilon}\right) - f\left(y(t_k), t_k, \frac{s}{\varepsilon}\right) \right| ds \\ &\quad + \sum_{k=0}^{D-1} \int_{t_k}^{t_{k+1}} |\bar{f}(y(s), s) - \bar{f}(y(t_k), t_k)| ds \\ &\quad + \sum_{k=0}^{D-1} \int_{t_k}^{t_{k+1}} \left| f\left(y(t_k), t_k, \frac{s}{\varepsilon}\right) - \bar{f}(y(t_k), t_k) \right| ds \\ &\leq \Lambda \int_0^t |x^\varepsilon(s) - y(s)| ds + \Lambda_1 \frac{t^2}{D} + t \sup_{\substack{|x| \leq \|y\|_\infty \\ t \in [0, T] \\ \hat{t} \geq 0}} \left| \frac{D\varepsilon}{t} \int_{\hat{t}}^{\hat{t}+\tau} f(x, t, \theta) d\theta \right. \\ &\quad \left. - \bar{f}(x, t) \right| \xrightarrow{\tau \rightarrow +\infty} 0. \end{aligned}$$

Choose $D = D(\varepsilon)$ such that $D(\varepsilon) \rightarrow_{\varepsilon \rightarrow 0} \infty$ and $\varepsilon D(\varepsilon) \rightarrow_{\varepsilon \rightarrow 0} 0$; use of the Gronwall lemma yields the result.

2.2. The averaged problem. Section 2.1 can be viewed as a generalization of averaging techniques to the nonperiodic case. We are now going to use it to define the averaged problem in the nonperiodic case.

Let f and L be as in § 1.1, except that now they need not be periodic in θ , and define problem (P^ε) accordingly. The important point in the definition of the averaged problem is that of the set W^{ad} of admissible controls. W^{ad} will be the set of functions u from $[0, T] \times \mathbf{R}_+$ to \mathbf{R}^p , with values in U^{ad} for almost every (t, θ) , and such that $f(x, u(t, \theta), t, \theta)$ and $L(x, u(t, \theta), t, \theta)$ have an average in θ for every (x, t) .

Note that W^{ad} may be empty. However, we will show that, if averaging can reasonably be expected to be used (i.e., the minimized Hamiltonian has an average), then W^{ad} is nonempty (see § 2.3).

For u in W^{ad} we define the averaged problem:

$$\left. \begin{aligned} \frac{dy}{dt} &= \bar{f}(y, u(t, \cdot), t, \cdot), \quad y(0) = x_0, \quad t \in [0, T], \\ \text{Minimize } \int_0^T \bar{L}(y, u(t, \cdot), t, \cdot) dt, \end{aligned} \right\} (\bar{P}).$$

If f and L are periodic, we find the same definition as in § 1.1.

2.3. The Hamiltonian of the averaged problem. In this section, we will omit the mention x and t . All assumptions made will be supposed to hold for every x and t .

Define the pseudo-Hamiltonian $h(p, u, \theta)$ by $h(p, u, \theta) = p^T f(u, \theta) + L(u, \theta)$ and let $H(p, \theta) = \min_{u \in U^{ad}} h(p, u, \theta)$ when it exists.

We are going to show that if H has an average for any p , then its average is the minimum of the Hamiltonian of the averaged problem. We will use the following assumptions:

- (H9) For any bounded part B of \mathbf{R}^p , f and L are bounded and uniformly continuous on $B \times \mathbf{R}_+$ (i.e., f and L are in $\text{BUC}(B, \mathbf{R}_+)$).
- (H10) For any (p, θ) , $h(p, u, \theta)$ has a minimum on U^{ad} . Moreover, for any bounded domain B in \mathbf{R}^p there exists a compact set K in U^{ad} such that the minimum can be reached in K for any (p, θ) in $B \times \mathbf{R}_+$.
- (H11) H has an average \bar{H} , i.e.,

$$\frac{1}{\tau} \int_0^\tau H(p, \theta) d\theta \xrightarrow{\tau \rightarrow +\infty} \bar{H}(p) \quad \text{for all } p \text{ in } \mathbf{R}^n.$$

THEOREM 2. *Let f and L meeting (H9), (H10) and (H11). Then W^{ad} is nonempty and $\bar{H}(p) = \text{Min}_{v \in W^{ad}} p^T \bar{f}(v) + \bar{L}(v)$.*

Proof. Denote $p^T \bar{f}(v) + \bar{L}(v)$ by $\bar{h}(p, v)$ for v in W^{ad} and let $E = \{p \in \mathbf{R}^n \mid \exists v \in W^{ad}, \bar{H}(p) = \bar{h}(p, v)\}$. As obviously $\bar{h}(p, v) \geq \bar{H}(p)$ for any v in W^{ad} , Theorem 2 is equivalent to $E = \mathbf{R}^n$. We are going to show that E is closed and that $\mathbf{R}^n - E$ is of null measure (Lebesgue).

(i) E is closed.

If $E = \emptyset$, this is true. If $E \neq \emptyset$, let p_n be a sequence in E , converging in \mathbf{R}^n , with $\bar{H}(p_n) = \bar{h}(p_n, v_n)$. Thanks to (H9) and (H10), $\bar{f}(v_n)$ and $\bar{L}(v_n)$ are bounded; let \bar{f} and \bar{L} be two cluster points, and w_n a subsequence such that $\bar{f}(w_n) \rightarrow_{n \rightarrow \infty} \bar{f}$ and $\bar{L}(w_n) \rightarrow_{n \rightarrow \infty} \bar{L}$. We are going to exhibit a control v such that $p^T \bar{f} + \bar{L} = p^T \bar{f}(v) + \bar{L}(v)$.

Let τ_n be an increasing sequence in \mathbf{R}_+ such that $\tau_n \geq n!$ and such that, for $\tau \geq \tau_n$:

$$\left| \frac{1}{\tau} \int_0^\tau f(w_n(\theta), \theta) d\theta - \bar{f}(w_n) \right| < \frac{1}{n}.$$

τ_n exists since w_n is in W^{ad} . Define v by $v(\theta) = w_n(\theta)$ for θ in $[\tau_n, \tau_{n+1}[$; it is then easy to check that v is in W^{ad} and that $p^T \bar{f} + \bar{L} = \bar{h}(p, v)$.

(ii) $\mathbf{R}^n - E$ is of null measure.

\bar{H} is locally Lipschitz, thanks to Assumptions (H9) and (H10), and thus, if F denotes the set of differentiability points of \bar{H} , $\mathbf{R}^n - F$ is of null measure. We are going to show that $F \subset E$.

Notice that, thanks to (H9) and (H10), there exists a measurable function u from $\mathbf{R}^n \times \mathbf{R}_+$ to U^{ad} such that $H(p, \theta) = h(p, u(p, \theta), \theta)$. Let ρ be a small positive number, p in F and q a direction in \mathbf{R}^n . Then:

$$\frac{H(p + \rho q) - H(p, \theta)}{\rho} \leq q^T f(p, u(p, \theta), \theta) \leq \frac{H(p, \theta) - H(p - \rho q, \theta)}{\rho}.$$

Let l be a cluster point of $1/\tau \int_0^\tau q^T f(p, u(p, \theta), \theta) d\theta$ as $\tau \rightarrow +\infty$; l exists thanks to (H9) and (H10). Then let $\tau \rightarrow \infty$ first, then $\rho \rightarrow 0$ in the above inequalities. We conclude that $l = \partial \bar{H} / \partial p$ in the direction q .

As this is true for any cluster point and any direction q , we have

$$\lim_{\tau \rightarrow \infty} \frac{1}{\tau} \int_0^\tau f(p, u(p, \theta), \theta) d\theta = \frac{\partial \bar{H}^T}{\partial p}.$$

Since f and H have an average, L has an average and $u(p, \theta)$ is in W^{ad} , with $\bar{H}(p) = p^T \bar{f}(u) + \bar{L}(u)$.

Remark 5. The concavity of H in p is essential. Let $H(p, \theta) = \sin(p, \theta)$. $\bar{H} = 0$, yet $\partial H / \partial p$ has no average.

2.4. A limit theorem on the Bellman function. Assumption (H11) has given sense to the averaged problem by ensuring that W^{ad} is nonempty. It also yields a limit theorem on the Bellman function. We will consider the latter as the unique viscosity solution of the Hamilton–Jacobi–Bellman equation (see [7]).

Assumptions. Let

$$H: \begin{cases} \mathbf{R}^n \times \mathbf{R}^n \times [0, T] \times \mathbf{R}_+ \rightarrow \mathbf{R}, \\ (p, x, t, \theta) \rightarrow H(p, x, t, \theta). \end{cases}$$

H may represent, for instance, the minimized Hamiltonian of § 2.3. The assumptions are the following:

$$(H12) \quad H \in \text{BUC}(B \times \mathbf{R}^n \times [0, T] \times \mathbf{R}_+) \text{ for any bounded part } B \text{ of } \mathbf{R}^n, \text{ and } |H(p, x, t, \theta) - H(p, y, t, \theta)| \leq C(1 + |p|)(|x - y|).$$

$$(H13) \quad H \text{ has an average } \bar{H} \text{ such that, for any } p, x, t,$$

$$\sup_{\hat{t} \geq 0} \left| \frac{1}{\tau} \int_{\hat{t}}^{\hat{t} + \tau} H(p, x, t, \theta) d\theta - \bar{H}(p, x, t) \right| \xrightarrow{\tau \rightarrow +\infty} 0.$$

THEOREM 3. *Let V be the viscosity solution of:*

$$(12) \quad \frac{\partial V}{\partial t} + \bar{H}\left(\frac{\partial V}{\partial x}, x, t\right) = 0, \quad V(x, T) \equiv 0, \quad x \in \mathbf{R}^n, \quad t \in [0, T].$$

Let V^ε be the viscosity solution of:

$$(13) \quad \frac{\partial V^\varepsilon}{\partial t} + H\left(\frac{\partial V^\varepsilon}{\partial x}, x, t, \frac{t}{\varepsilon}\right) = 0, \quad V^\varepsilon(x, T) \equiv 0, \quad x \in \mathbf{R}^n, \quad t \in [0, T].$$

Then V^ε converges to V as ε tends to zero, uniformly on any compact subset of $\mathbf{R}^n \times [0, T]$.

Proof. We will make use of the following notation:

$$L^p_{\text{unif}}(\mathbf{R}^n \times [0, T]) = \left\{ \phi \in L^p_{\text{loc}}(\mathbf{R}^n \times [0, T], \mathbf{R}), \sup_{y \in \mathbf{R}^n} \int_{t \in [0, T]} \int_{|x-y| < 1} |\phi(x, t)|^p dx dt < \infty \right\}.$$

$W^{2,1,p}_{\text{unif}}(\mathbf{R}^n \times [0, T])$ is the set of functions ϕ in $L^p_{\text{unif}}(\mathbf{R}^n \times [0, T])$ such that $\partial \phi / \partial t$, $\partial \phi / \partial x$ and $\partial^2 \phi / \partial x^2$ exist and are in L^p_{unif} . $W^{2,1,p}_{\text{unif}}$ is the analogue of $W^{2,1,p}$ except that L^p is replaced by L^p_{unif} . We can use the norm on $W^{2,1,p}_{\text{unif}}$ defined by the sup of $W^{2,1,p}$ norms over all $\bar{B}(y, 1) \times [0, T]$, where $\bar{B}(y, 1)$ is the closed ball of center y and radius 1.

We will denote, for $\alpha > 0$, by V^ε_α the unique solution in $\cap_{p \geq 0} W^{2,1,p}_{\text{unif}}$ of

$$(14) \quad \frac{\partial V^\varepsilon_\alpha}{\partial t} + \alpha \Delta V^\varepsilon_\alpha + H\left(\frac{\partial V^\varepsilon_\alpha}{\partial x}, x, t, \frac{t}{\varepsilon}\right) = 0, \quad V^\varepsilon_\alpha(x, T) \equiv 0$$

and V_α the analogue for

$$(15) \quad \frac{\partial V_\alpha}{\partial t} + \alpha \Delta V_\alpha + \bar{H}\left(\frac{\partial V_\alpha}{\partial x}, x, t\right) = 0, \quad V_\alpha(x, T) \equiv 0.$$

It is well known [7] that

$$\|V^\varepsilon_\alpha - V^\varepsilon\|_\infty \leq k\sqrt{\alpha} \quad \text{and} \quad \|V_\alpha - V\|_\infty \leq k_2\sqrt{\alpha}.$$

Moreover, k does not depend on the behaviour of H in time, that is, in particular k does not depend on ε . Hence, the theorem will be proved if, α being fixed, V_α^ε converges to V_α for ε in $]0, 1]$. Estimates on the derivatives of V_α^ε ensure that the V_α^ε are bounded in (and thus form a weakly relatively compact subset of) $W_{\text{unif}}^{2,1,p}$ for any $p > 0$. Thus, the $(V_\alpha^\varepsilon, \partial V_\alpha^\varepsilon / \partial x)$ are relatively compact in $C^0(K \times [0, T], \mathbf{R})$, K being any compact subset of \mathbf{R}^n . Let $(\bar{V}_\alpha, \partial \bar{V}_\alpha / \partial x)$ a cluster point and $(V_{\alpha^n}^\varepsilon, \partial V_{\alpha^n}^\varepsilon / \partial x)$ a sequence converging to $(\bar{V}_\alpha, \partial \bar{V}_\alpha / \partial x)$ uniformly on any compact, with $V_{\alpha^n}^\varepsilon \rightarrow_{n \rightarrow \infty} \bar{V}_\alpha$ in $W_{\text{unif}}^{2,1,p}$ weakly.

Proving the theorem thus amounts to showing that \bar{V}_α is a solution of (15) in the weak sense. Let ϕ a function in $C^\infty(\mathbf{R}^n \times [0, T])$ with a compact domain.

$$\begin{aligned} \int_{t_k}^{t_{k+1}} dt \int_{\mathbf{R}^n} dx \left[\frac{\partial \bar{V}_\alpha}{\partial t} + \alpha \Delta \bar{V}_\alpha + \bar{H} \left(\frac{\partial \bar{V}_\alpha}{\partial x}, x, t \right) \right] \phi(x, t) \\ = \int_{t_k}^{t_{k+1}} dt \int_{\mathbf{R}^n} dx \left[\frac{\partial \bar{V}_\alpha}{\partial t} - \frac{\partial V_\alpha^\varepsilon}{\partial t} + \alpha \Delta \bar{V}_\alpha - \alpha \Delta V_\alpha^\varepsilon \right] \phi(x, t) \\ + \int_{t_k}^{t_{k+1}} dt \int_{\mathbf{R}^n} dx \left[H \left(\frac{\partial \bar{V}_\alpha}{\partial x}, x, t, \frac{t}{\varepsilon} \right) - H \left(\frac{\partial V_\alpha^\varepsilon}{\partial x}, x, t, \frac{t}{\varepsilon} \right) \right] \phi(x, t) \\ + \int_{t_k}^{t_{k+1}} dt \int_{\mathbf{R}^n} dx \left[\bar{H} \left(\frac{\partial \bar{V}_\alpha}{\partial x}, x, t \right) - H \left(\frac{\partial \bar{V}_\alpha}{\partial x}, x, t, \frac{t}{\varepsilon} \right) \right] \phi(x, t). \end{aligned}$$

As ε tends to zero, the first expression has limit zero by weak convergence in $W_{\text{unif}}^{2,1,2}$. The second one tends to zero thanks to the Lebesgue dominated convergence theorem. A discretization scheme similar to that used in § 2.1 ensures that the third one has limit zero, thanks to Assumption (H13).

Remark 6. It can be proved [4] that, if H is locally Lipschitz in p and if the convergence in Assumption (H13) is uniform for x in \mathbf{R}^n , then the convergence of V^ε to V is uniform on $\mathbf{R}^n \times [0, T]$.

One may wonder when (H12) is true with $H = p^T f + L$. It is true if, for instance, f and L are BUC and Lipschitz in x . However, this can be extended to the case where f is uniformly continuous, Lipschitz in x with $|f| \leq k(1 + |x| + |u|)$; and L is uniformly continuous, locally Lipschitz in x , $|\partial L / \partial x| \leq k(1 + |x| + |u|)$ and $|L| \leq k(1 + |x|^2 + |u|^2)$, $L \geq k_1(-1 + |u|^2)$.

This includes the linear quadratic case; actually, we might call this the “sublinear quadratic case.” With a suitable truncation of f and L on the phase domain, we may keep V^ε and V unchanged on a portion of $\mathbf{R}^n \times [0, T]$, while retrieving Assumption (H12). The convergence is still uniform on any compact.

3. Perspectives. We have proven here, at least in the periodic case, that averaging can be used as an efficient tool in deterministic optimal control. It is efficient for two reasons.

First, the averaged problem (\bar{P}) is easier to solve numerically than the original problem (P^ε) , since a “fast” time grid is no longer needed in the simulation part. Gains should also be expected on the state space grid, since it is often related to the former one; it is an important point if one thinks, for instance, of dynamic programming.

Second, and thanks to Theorem 1, the solution of (\bar{P}) is known to be near optimal for (P^ε) ; hence, we do not lose much by solving the averaged problem instead of the original one.

We have shown that use of the averaged problem can also be expected to be efficient in the nonperiodic case, as the optimal cost of (P^ε) is close to that of (\bar{P}) when ε is small (Theorem 3).

However, practical problems are seldom under the form (P^ε) , be it in the periodic or nonperiodic case. Most of the time, there is, for instance, no explicit separation of the time scales under the form (t, θ) , with θ ranging from 0 to $+\infty$; in particular, periodicity or averaging assumptions cannot be checked directly. Nevertheless, the results presented here provide an important theoretical background for developments of both theoretical and practical interest.

From the theoretical point of view, it is reasonable to expect problem (TP^ε) in Lemma 1 to provide further expansions of the cost (J^ε) , as similar methods have already been used with success in the case of regular and singular perturbations [2]. A complete expansion of the Bellman function in the linear quadratic periodic case has already been obtained [4]. In particular, the terms of order higher than two are in the form $V(x, t, t/\varepsilon, (T - t/\varepsilon))$, with periodicity in both the forward and backward fast times. This is probably related to the existence of the terms x_2 and p_2 in the expansion of the primal and dual trajectories. The same phenomenon exists in singular perturbations with boundary layers instead of phase terms.

We have seen that x_2 and p_2 are defined through the operator Π . In fact, Π appears in any expansion of an integral with an integrand periodic in fast time. A generalization of Π would be welcome if we hope to find some results equivalent to Theorem 1 in the nonperiodic case.

At last, links should be developed with singular perturbations. From the practical point of view, we have seen that the assumptions in Theorems 1 and 3 cannot be checked directly. It should be noticed (especially in the nonperiodic case) that the question is not so much that the assumptions might not hold, as it is rather to immerse the optimization problem in the "right" family of problems (P^ε) . Moreover, the ideas are sufficiently simple and general to be used in heuristics. One can think, for instance, of separating the time scales through the use of "moving averages." Heuristics can also be developed to generalize the operator Π to the nonperiodic case and use it to improve performances. We are going to experiment numerically on these ideas.

Conversely, averaging has been often used empirically by engineers in practical problems. The results and notions presented here may provide them some guidelines in further applications.

We have discussed *how* averaging could be used practically. We shall discuss now *when* it could be used. Heuristically, averaging can be expected to yield good results and performances when the dynamics of a system depend on a fast "erratic" exogenous phenomenon. A good example is given by weather disturbances.

However, these phenomena are often modeled by stochastic processes. As we mentioned before, both approaches are similar in the sense that they make use of averages; their theoretical background is, however, quite different. Moreover, averaging has also been used in stochastic control ([5], for instance). In all cases, the original data consists often of a finite number of physical measures, in no way probabilistic or two-time scaled in nature. Thus, neither approach is justified a priori. Therefore, it should prove quite interesting to experiment all methods on various sets of data. We plan to conduct such experiments.

REFERENCES

- [1] V. ARNOLD, *Chapitres supplémentaires de la théorie des équations différentielles ordinaires*, French translation from the Russian, Mir, Moscow, 1980.
- [2] A. BENSOUSSAN, *Singular Perturbations in Systems and Control*, Mark Ardem, ed., Springer-Verlag, Berlin-Heidelberg-New York, 1983, pp. 169-185.

- [3] A. BENSOUSSAN, J. L. LIONS AND G. PAPANICOLAOU, *Asymptotic Analysis for Periodic Structures*, North-Holland, Amsterdam, 1978.
- [4] F. CHAPLAIS, *Averaging et contrôle optimal déterministe*, Thèse de Docteur-Ingénieur, Ecole nationale Supérieure des Mines de Paris, Paris, 1984.
- [5] F. DELBECQUE AND J. P. QUADRAT, *Contribution of Stochastic Control Singular Perturbation Averaging and Team Theory to an Example of Large-Scale Systems: Management of Hydropower Production*, IEEE Trans. Automat. Control, AC-23, April 1978, pp. 209–221.
- [6] P. V. KOKOTOVIC, R. E. O'MALLEY, JR. AND P. SANNUTI, *Singular perturbations and order reduction in control theory: An overview*, Automatica, 12 (1976), pp. 123–132.
- [7] P. L. LIONS, *Generalized Solutions of Hamilton–Jacobi Equations*, Pitman, Boston, 1982.

LINEAR-QUADRATIC PROGRAMMING AND OPTIMAL CONTROL*

R.T. ROCKAFELLAR†

Abstract. A generalized approach is taken to linear and quadratic programming in which dual as well as primal variables may be subjected to bounds, and constraints may be represented through penalties. Corresponding problem models in optimal control related to continuous-time programming are then set up and theorems on duality and the existence of solutions are derived. Optimality conditions are obtained in the form of a global saddle point property which decomposes into an instantaneous saddle point condition on the primal and dual control vectors at each time, along with an endpoint condition.

Key words. Linear-quadratic programming, dual control problems, intertemporal programming, continuous-time programming, penalty representation of constraints

AMS(MOS) subject classifications. 49B10, 90C20, 90C05

1. Introduction. In finite-dimensional optimization a great importance is attached to problems of linear and quadratic programming. Such problems serve as mathematical models for a large number of applications. They are relatively easy to work with and possess duality properties that yield valuable insights and are the basis for many special algorithms. They are useful in methods of solving more general problems, for instance, in connection with sequential approximation or direction-finding subroutines. For such purposes they can be extended beyond traditional formulations to admit piecewise linear-quadratic objectives and penalty representations of constraints, although this possibility has not yet fully been utilized.

In optimal control there has not been a comparable emphasis on a "linear-quadratic" class of problems. The linear-quadratic regulator problem fits the picture to some degree but is virtually unconstrained. The continuous-time linear programming problems first introduced as "bottleneck" problems by Bellman [1] include certain types of control problems with constraints on states and controls (possibly mixed), but they carry no provision for quadratic terms in the objective and are very narrow in their treatment of initial and terminal conditions. Continuous-time linear programming problems do enjoy a strong duality theory, thanks to efforts of Tyndall [2], [3], Levinson [4], Grinold [5], [6], Schecter [7], Reiland [8], Meidan and Perold [9], and others. Continuous-time nonlinear programming has also been investigated, chiefly for duality; cf. Hanson [10], Hanson and Mond [11], Grinold [12], Farr and Hanson [13], Reiland and Hanson [14], Reiland [15]. This nonlinear literature covers certain classes of optimal control problems with quadratic terms, subject to the same limitations on the treatment of initial and terminal states. However, the quadratic case has not been worked out to take advantage of its special nature, and, in any case, the results are based on a Lagrange multiplier approach that does not yield even in finite dimensions a duality theory as broad and flexible as may currently be needed.

*Received by the editors December 16, 1985; accepted for publication (in revised form) June 24, 1986. This work was supported in part by a grant from the National Science Foundation at the University of Washington, Seattle.

† Department of Mathematics, University of Washington, Seattle, Washington 98195.

Our goal in this paper is to develop a theory of linear-quadratic programming-type problems specifically adapted to the optimal control setting and capable eventually of being used in new computational schemes, as well as directly.

Some of the motivation comes from mathematical modeling. Linear-quadratic models do not appear to have been used so far to their full potential. An obstacle may lie in the format in which finite-dimensional problems in linear programming and quadratic programming are ordinarily presented. In this format it is hard to deal with piecewise linear or piecewise quadratic functions, such as often are important in penalty representations, except by reformulations that disrupt the fundamental relationships, especially duality.

An alternative approach in finite dimensions, which we have followed recently in work on algorithms in stochastic programming [16], [17], [18], is to give primacy to an underlying saddle point problem (minimax problem). Thus we think of finite-dimensional linear-quadratic programming in a more general sense than usual as corresponding to finding a saddle point of a convex-concave quadratic (or linear) function on a product of polyhedral convex sets. Any such saddle point problem generates a primal problem of minimization and a dual problem of maximization. The classical case of linear and quadratic programming duality is the one where the polyhedral convex sets are orthants.

The problems in the general case could be reduced individually to the classical case, but by working directly in the broader format one gains several advantages. The most significant is the perception that bounds can reasonably be introduced for dual variables as well as primal variables, and moreover that this amounts to passing from exact representations of certain constraints to penalty representations.

We begin in §2 and §3 by explaining this unconventional approach to finite-dimensional linear-quadratic programming and the kinds of problem forms it handles. A particular aim is the elucidation of circumstances under which a model involving bounds on both primal and dual variables is appropriate, at least for computation. Then in §4 and §5 we introduce corresponding problems in optimal control, of a sort we call *intertemporal linear-quadratic programming*. The main results are obtained in §6. They consist of theorems on existence, duality, and the characterization of optimal controls. They are tied to an infinite-dimensional saddle point representation in terms of a convex-concave quadratic functional on a product of generalized polyhedral sets.

Our problems in optimal control have dynamics that are essentially linear, although “polyhedral differential inclusions” are also encompassed by the formulation. The expression of the objective and constraints involves, in general, terms that may be piecewise linear-quadratic. To clarify the nature of such terms in this introduction would take us too far. A brief description of one of the basic *linear* models covered by our theory is feasible, however, and may help to put the approach and results in perspective.

Over a fixed time interval $[t_0, t_1]$ we consider a dynamical system

$$(1.1) \quad \dot{x}(t) = A(t)x(t) + B(t)u(t) + b(t), \quad x(t_0) = B_e u_e + b_e,$$

where $x(t) \in \mathbb{R}^n$ is the state, $u(t) \in \mathbb{R}^k$ is the instantaneous control and $u_e \in \mathbb{R}^{k_e}$ is an additional vector to be chosen, an “endpoint control.” The incorporation of such a vector u_e may seem odd relative to the customary patterns in control theory, but it greatly aids in dualizing various conditions. Of course u_e could be trivialized by taking the dimension k_e to be 0 (then $x(t_0) = b_e$ in (1.1)). Another case to note is the one of a free initial point: $B_e = I$, $b_e = 0$ (then $x(t_0) = u_e$ in (1.1)). The subscript e will consistently be used in our notation for elements connected with endpoints.

For the basic linear case in question, the problem we associate with the system (1.1) takes the form

$$\begin{aligned}
 & \text{minimize} && \int_{t_0}^{t_1} [p(t) \cdot u(t) - c(t) \cdot x(t)] dt + [p_e \cdot u_e - c_e \cdot x(t_1)] \\
 (\mathcal{P}_1) \quad & \text{subject to} && (1.1) \text{ with} \quad C(t)x(t) + D(t)u(t) \geq q(t), \quad u(t) \geq 0, \\
 & && C_e x(t_1) + D_e u_e \geq q_e, \quad u_e \geq 0.
 \end{aligned}$$

Discussion of the exact technical assumptions is postponed until §4. Observe, however, that the formulation allows for constraints only on the controls (rows of $C(t)$ consisting of 0's), constraints only on the states (rows of $D(t)$ consisting of 0's), and mixed constraints. The endpoint conditions allow for any system of finitely many linear equations or inequalities to be imposed on the pair $x(t_0), x(t_1)$ (as explained in detail in Examples 5.1 and 5.2 in §5).

In dualizing (\mathcal{P}_1) we pass to the dynamical system

$$(1.2) \quad -\dot{y}(t) = A^*(t)y(t) + C^*(t)v(t) + c(t), \quad y(t_1) = C_e^* v_e + c_e,$$

where $y(t) \in \mathbb{R}^n$ is the state, $v(t) \in \mathbb{R}^\ell$ is the instantaneous control, and $v_e \in \mathbb{R}^{\ell_e}$ is the endpoint control; the asterisk $*$ denotes the transpose of a matrix. The dual problem over the system (1.2) is

$$\begin{aligned}
 & \text{maximize} && \int_{t_0}^{t_1} [q(t) \cdot v(t) - b(t) \cdot y(t)] dt + [q_e \cdot v_e - b_e \cdot y(t_0)] \\
 (\mathcal{Q}_1) \quad & \text{subject to} && (1.2) \text{ with} \quad B_e^*(t)y(t) + D^*(t)v(t) \leq p(t), \quad v(t) \geq 0, \\
 & && B_e y(t_0) + D_e^* v_e \leq p_e, \quad v_e \geq 0.
 \end{aligned}$$

Although (\mathcal{P}_1) and (\mathcal{Q}_1) have been written with inequality constraints only, there is no difficulty about extending the formulation to include equations in the manner familiar in linear programming. Thus, for example, the condition $C(t)x(t) + D(t)u(t) \geq q(t)$ in (\mathcal{P}_1) can be converted to $C(t)x(t) + D(t)u(t) = q(t)$ by dropping the condition $v(t) \geq 0$ in (\mathcal{Q}_1) .

In contrast to (\mathcal{P}_1) and (\mathcal{Q}_1) the continuous-time linear programming problems mentioned earlier take the primal form

$$\begin{aligned}
 & \text{minimize} && \int_{t_0}^{t_1} p(t) \cdot u(t) dt \\
 & \text{subject to} && \int_{t_0}^t K(t, \tau) u(\tau) d\tau + D(t)u(t) \geq q(t), \quad u(t) \geq 0,
 \end{aligned}$$

and the dual form

$$\begin{aligned}
 & \text{maximize} && \int_{t_0}^{t_1} q(t) \cdot v(t) dt \\
 & \text{subject to} && \int_t^{t_1} K^*(\tau, t) v(\tau) d\tau + D^*(t)v(t) \leq p(t), \quad v(t) \geq 0,
 \end{aligned}$$

where the matrix $K(t, \tau)$ is some "kernel" with transpose $K^*(t, \tau)$. These are not necessarily problems of optimal control but become so in choosing

$$K(t, \tau) = C(t)A(t)A(\tau)^{-1}B(\tau)$$

with $A(t)$ the fundamental matrix corresponding to the differential equation (1.1) (i.e., $A(t)x_0$ is the unique solution to $\dot{x}(t) = A(t)x(t)$, $x(t_0) = x_0$), and setting

$$x(t) = A(t) \int_{t_0}^t A(\tau)^{-1} B(\tau) u(\tau) d\tau, \quad y(t) = A^*(t)^{-1} \int_t^{t_0} A^*(\tau) C^*(\tau) v(\tau) d\tau.$$

Then one gets the case of (P_1) and (Q_1) where $b(t) = 0$, $c(t) = 0$, and all the e terms trivialize: the primal has $x(t_0) = 0$ but $x(t_1)$ free, whereas the dual has $y(t_1) = 0$ but $y(t_0)$ free.

In the work that has been done on special computational methods in continuous-time linear programming, e.g. Perold [19], [20], Anstreicher [21], attention has typically been limited further to the case where the kernel K is a *constant* matrix. In optimal control this corresponds not merely to having $A(t)$, $B(t)$ and $C(t)$ constant, but $A(t) \equiv 0$, a severe restriction.

Because of these distinctions and the desirability of being able to treat discrete-time analogues under the same heading, we shall refer to (P_1) and (Q_1) as problems of "intertemporal linear programming" (in continuous time) rather than "continuous-time linear programming."

The possibility of mixed constraints on states and controls is important in accommodating many applications of an economic nature, involving planned activities with cumulative effects. But it also puts problems like (P_1) and (Q_1) beyond the range of the Pontryagin maximum principle. Mixed constraints can be readily handled, however, in the versions of optimal control and variational calculus that have been developed over the years in the conceptual framework of convex analysis and, more recently, nonsmooth analysis in the sense of Clarke [22].

The theory of convex problems of Bolza type, developed by the author in [23]–[29], is specifically applicable to problems (P_1) , (Q_1) and their quadratic programming counterparts after a transformation which expresses everything through the trajectories x and y , as outlined in [30]. By this route it would be possible, with a degree of technical elaboration, to derive sharp duality theorems that characterize solutions and the circumstances in which they exist. Full justice to constraints involving states would, however, require us in the context of such duality to pass beyond the formulation of our primal and dual problems in terms of control *functions* u and v to one in which "impulse controls" may occur. An extension along those lines is indeed appropriate, and for the basic linear programming case in (P_1) , (Q_1) , it has been carried out by Murray [31] under a somewhat different choice of endpoint expressions.

For the present purpose we are able to postpone working with such an extension. We follow a different path and sidestep the difficulties posed by state constraints by appealing instead to alternative problem formulations where the constraints may be enforced by linear or piecewise linear-quadratic penalty expressions. We argue that as a practical matter of mathematical modeling and computation this is an often reasonable tactic which can be served by a much simpler theory where solutions always exist and strong duality always holds. The supporting results in finite-dimensional linear-quadratic programming provided in §§2 and 3 are critical in understanding this.

The saddle point representation furnished in §6 for the duality between our two infinite-dimensional problems of intertemporal linear-quadratic programming is of a kind not previously seen in optimal control. Moreover the representation has a separate decomposition property in each argument that may open the way to new saddle point techniques for computation such as extensions of the finite generation method

devised by R.J.-B. Wets and the author in a similar setting in stochastic programming [17]. Decomposition of the intertemporal saddle point condition leads to a characterization of optimality in terms of a “instantaneous” saddle point condition satisfied at each time t and an “endpoint” saddle point condition. This is a sort of “minimax principle” which has some precedent in continuous-time linear programming (Grinold [5, p. 46]) and the theory of Bolza problems (Rockafellar [23, Thm. 6]) but is new in this context of optimal control.

2. Linear-quadratic programming in finite dimensions. The infinite-dimensional control problems that are the subject of this paper, and our approach to them, will better be understood after a brief treatment of the formulation and duality properties of finite-dimensional linear-quadratic programming problems in the generalized sense. Such a treatment will also introduce facts and concepts that will be needed in later sections.

A simple foundation for almost all kinds of duality theory in optimization starts with a function $J(u, v)$ on a product set $U \times V$, where J is real-valued or possibly extended-real-valued. Regardless of the nature of J and the sets U and V (as long as the latter are nonempty), there is an associated *primal* problem

$$(\mathcal{P}_0) \quad \text{minimize } f(u) \text{ over } U \quad \text{where } f(u) = \sup_{v \in V} J(u, v),$$

and a *dual* problem

$$(\mathcal{Q}_0) \quad \text{maximize } g(v) \text{ over } V \quad \text{where } g(v) = \inf_{u \in U} J(u, v).$$

The relationship between these problems is tied to the *saddle point*, or *minimax* problem for J on $U \times V$, a saddle point being by definition a pair $(\bar{u}, \bar{v}) \in U \times V$ such that

$$(2.1) \quad J(u, \bar{v}) \geq J(\bar{u}, \bar{v}) \geq J(\bar{u}, v) \quad \text{for all } u \in U, v \in V.$$

The following facts are well known (cf. [32, Thm. 2], for example).

PROPOSITION 2.1. *It is always true that $\inf(\mathcal{P}_0) \geq \sup(\mathcal{Q}_0)$. Furthermore a pair (\bar{u}, \bar{v}) is a saddlepoint of J on $U \times V$ if and only if \bar{u} solves (\mathcal{P}_0) , \bar{v} solves (\mathcal{Q}_0) , and $\min(\mathcal{P}_0) = \max(\mathcal{Q}_0)$.*

Here we use the notation that $\inf(\mathcal{P}_0)$ is the optimal value in (\mathcal{P}_0) , namely the infimum of f over U . We allow ourselves to write $\min(\mathcal{P}_0)$ in place of $\inf(\mathcal{P}_0)$ if the infimum is actually attained at some \bar{u} . Similarly for $\sup(\mathcal{Q}_0)$, $\max(\mathcal{Q}_0)$.

By finite-dimensional (*piecewise*) *linear-quadratic programming* in the general sense we shall mean the case of problems (\mathcal{P}_0) and (\mathcal{Q}_0) where U is a nonempty convex polyhedron in a space \mathbb{R}^k , V is a nonempty convex polyhedron in space \mathbb{R}^ℓ , and J is a convex-concave function of the form

$$(2.2) \quad J(u, v) = p \cdot u + v \cdot q + \frac{1}{2}u \cdot Pu - \frac{1}{2}v \cdot Qv - v \cdot Du,$$

where $p \in \mathbb{R}^k$, $q \in \mathbb{R}^\ell$, $P \in \mathbb{R}^{k \times k}$, $Q \in \mathbb{R}^{\ell \times \ell}$ and $D \in \mathbb{R}^{\ell \times k}$, with P and Q symmetric and positive *semidefinite*. When $P = 0$ and $Q = 0$, we speak of (*piecewise*) *linear programming* in the general sense. This includes classical linear programming, of course (cf. Example 3.1 below).

In the linear-quadratic programming case the objective functions in (\mathcal{P}_0) and (\mathcal{Q}_0) take the form

$$(2.3) \quad f(u) = p \cdot u + \frac{1}{2}u \cdot Pu + \rho_{V, Q}(q - Du),$$

$$(2.4) \quad g(v) = q \cdot v - \frac{1}{2}v \cdot Qv - \rho_{U,P}(D^*v - p),$$

where

$$(2.5) \quad \rho_{V,Q}(s) = \sup_{v \in V} \{s \cdot v - \frac{1}{2}v \cdot Qv\},$$

$$(2.6) \quad \rho_{U,P}(r) = \sup_{u \in U} \{r \cdot u - \frac{1}{2}u \cdot Pu\}.$$

When $P = 0$ and $Q = 0$, the functions $\rho_{V,Q}$ and $\rho_{U,P}$ reduce to the *support functions*

$$(2.7) \quad \sigma_V(s) = \sup_{v \in V} s \cdot v, \quad \sigma_U(r) = \sup_{u \in U} r \cdot u.$$

The specific nature of these various expressions will be explored in the examples in §3. The central fact is that strong duality always holds for such problems.

THEOREM 2.2. *In the case where (P_0) and (Q_0) are finite-dimensional linear-quadratic programming problems in the general sense just described, one has*

$$\infty > \min(P_0) = \max(Q_0) > -\infty,$$

unless the optimal values $\inf(P_0)$ and $\sup(Q_0)$ are both infinite. In particular, any finite-dimensional linear-quadratic programming problem with finite optimal value has an optimal solution.

Theorem 2.2 can easily be derived from known results about quadratic programming in the standard sense, specifically the duality theorem of Dorn [33] and Cottle [34] and the existence criterion of Frank and Wolfe [35]. We have given the argument in full in [17, Thm. 2].

Incidentally, the suprema in (2.5) and (2.6) must be attained also, when finite. Indeed, these formulas give the optimal values in certain quadratic programming problems and are covered by the result just cited.

The sense in which the terminology "linear-quadratic programming in the general sense" is appropriate for the problems in Theorem 2.2 is elucidated by our next result.

PROPOSITION 2.3. *The function $\rho_{V,Q}$ is lower semicontinuous, convex, and piecewise linear-quadratic: its effective domain*

$$(2.8) \quad L = \{s \in \mathbb{R}^\ell \mid \rho_{V,Q}(s) < \infty\}$$

is a nonempty convex polyhedron that can be decomposed into finitely many polyhedral convex sets, on each of which $\rho_{V,Q}$ is quadratic (or linear).

The same holds of course for $\rho_{U,P}$ and its effective domain

$$(2.9) \quad K = \{r \in \mathbb{R}^k \mid \rho_{U,P}(r) < \infty\}.$$

Proof. Define

$$(2.10) \quad \begin{aligned} \varphi(v) &= \begin{cases} \frac{1}{2}v \cdot Qv & \text{when } v \in V, \\ \infty & \text{when } v \notin V \end{cases} \\ &= j_Q(v) + \delta_V(v), \end{aligned}$$

where j_Q is the quadratic convex function corresponding to the positive definite form Q , and δ_V is the indicator of the convex polyhedron V :

$$(2.11) \quad \delta_V(v) = \begin{cases} 0 & \text{when } v \in V, \\ \infty & \text{when } v \notin V. \end{cases}$$

Clearly φ is convex, and its conjugate

$$(2.12) \quad \varphi^*(s) = \sup_{v \in \mathbb{R}^\ell} \{s \cdot v - \varphi(s)\}$$

is given by

$$(2.13) \quad \varphi^*(s) = \rho_{V,Q}(s).$$

The latter is therefore lower semicontinuous and convex in s , and its effective domain L is a nonempty convex set (these properties being true for the conjugate of any proper convex function [36, §12]).

For each $s \in L$, the supremum in (2.12) (equivalently (2.5)) must actually be attained, as noted above. On the other hand we know from convex analysis [36, Thm. 23.5] that the supremum in (2.12) is attained at v if and only if $v \in \partial\varphi^*(s)$, which is equivalent to $s \in \partial\varphi(v)$. Thus L coincides with the effective domain of the subdifferential multifunction $\partial\varphi^*$, which is also the range of $\partial\varphi$. We shall use this fact to demonstrate that L is polyhedral and has the decomposition claimed.

Because $\varphi = j_Q + \delta_V$ and j_Q is finite everywhere on \mathbb{R}^ℓ , we have by [36, Thm. 23.8] that

$$(2.14) \quad \partial\varphi(v) = \partial j_Q(v) + \partial\delta_V(v) = Qv + N_V(v),$$

where $N_V(v)$ is the normal cone to V at v [36, p. 215]. This normal cone is polyhedral, because V is polyhedral, and it depends only on the face of V to which v belongs. There are only finitely many faces of V , so it follows from (2.14) that $\partial\varphi$ is a polyhedral multifunction in the sense of Robinson [37], namely its graph in $\mathbb{R}^\ell \times \mathbb{R}^\ell$ is the union of finitely many polyhedral convex sets (one for each face of V). The same is then true for the multifunction $\partial\varphi^* = \partial\varphi^{-1}$, whose domain, already identified with L , must therefore be the projection of the union of finitely many polyhedral convex sets. We may conclude that the convex set L is actually polyhedral and can be decomposed into finitely many polyhedral convex sets L_i , over each of which the graph of $\partial\varphi^*$ is a polyhedral convex set. In the case of such a subset L_i having $\text{int } L_i \neq \emptyset$, $\partial\varphi^*$ must by this reduce to a single-valued affine mapping on $\text{int } L_i$, inasmuch as $\partial\varphi^*$ is single-valued almost everywhere on $\text{int } L$ (a fact true of the subdifferential of any proper convex function on the interior of its effective domain [36, Thm. 24.5]). Therefore φ^* is quadratic (or linear) on $\text{int } L_i$ by the lower semicontinuity of φ^* . For L_i with $\text{int } L_i = \emptyset$, a slightly more general argument based on relative interiors of convex sets leads to the same conclusion. Thus the function $\varphi^* = \rho_{V,Q}$ is piecewise linear-quadratic as claimed. \square

The terminology “linear programming in the general sense” in the case where $P = 0$ and $Q = 0$ is justified similarly. The functions $\rho_{V,Q}$ and $\rho_{U,P}$ reduce then to the support functions σ_V and σ_U in (2.7), which are polyhedral convex (piecewise linear) because U and V are polyhedral [36, Cor. 19.2.1].

Because $\rho_{V,Q}$ and $\rho_{U,P}$ can take ∞ as a value in some cases, the linear-quadratic programming problems (P_0) and (Q_0) may have implicit constraints. Thus in minimizing the function f given by (2.3) we are really interested only in the choices of u that satisfy

$$(2.15) \quad q - Du \in L \quad \text{as well as } u \in U.$$

Likewise in maximizing the function g in (2.4) we focus on v satisfying

$$(2.16) \quad D^*v - p \in K \quad \text{as well as } v \in V.$$

The polyhedral convexity of L and K in Proposition 2.3 together with that of U and V means that these constraint systems can be represented in principle by finitely many linear equations and inequalities.

A closer analysis of the sets L and K reveals additional structure that will be of use to us. Here we denote the *null space* of Q by

$$\text{nl } Q = \{w \in \mathbb{R}^\ell \mid Qw = 0\}$$

and the *recession cone* [36, §8] of V by

$$\text{rc } V = \{w \in \mathbb{R}^\ell \mid v + \lambda w \in V, \forall \lambda \geq 0\} \quad \text{for } v \in V.$$

The latter is the same regardless of the choice of $v \in V$. It is a polyhedral convex cone (always containing 0), because V is a polyhedral convex set [36, Thm. 19.5]. Indeed, if $V = \{v \mid Mv \leq m\}$, one has $\text{rc } V = \{w \mid Mw \leq 0\}$. We denote the *polar* of a cone G as usual by

$$(2.17) \quad G^\circ = \{z \mid z \cdot w \leq 0, \forall w \in G\}.$$

PROPOSITION 2.4. *The effective domains L and K in Proposition 2.3 are the polar cones*

$$(2.18) \quad L = [\text{rc } V \cap \text{nl } Q]^\circ \quad \text{and} \quad K = [\text{rc } U \cap \text{nl } P]^\circ.$$

Thus

$$(2.19) \quad L = \mathbb{R}^\ell \iff [\text{the only } w \in \text{rc } V \text{ with } Qw = 0 \text{ is } w = 0],$$

$$(2.20) \quad K = \mathbb{R}^k \iff [\text{the only } z \in \text{rc } U \text{ with } Pz = 0 \text{ is } z = 0].$$

In particular $L = \mathbb{R}^\ell$ if V is bounded or if Q is positive definite, whereas $K = \mathbb{R}^k$ if U is bounded or if P is positive definite.

Proof. Let φ be given again by (2.10), so that $\varphi^* = \rho_{V,Q}$ as in (2.13). Since $L = \text{dom } \varphi^*$ and L is closed, we have by [36, Thm. 13.3] that the indicator δ_L is conjugate to the recession function

$$(\text{rc } \varphi)(w) = \lim_{\lambda \rightarrow \infty} \varphi(v + \lambda w)/\lambda,$$

where $v \in \text{dom } \varphi = V$ (the limit being independent of the particular choice of v [36, Thm. 8.5]). The limit works out to

$$(\text{rc } \varphi)(w) = \begin{cases} 0 & \text{if } w \in \text{rc } V \text{ and } Qw = 0, \\ \infty & \text{otherwise.} \end{cases}$$

Thus $\text{rc } \varphi = \delta_G$ for $G = \text{rc } V \cap \text{nl } Q$. The indicators δ_G and δ_L being conjugate to each other, we conclude that G and L are cones polar to each other [36, §14]. \square

An important question of mathematical modeling and computation in applications both finite and infinite-dimensional is whether a problem (\mathcal{P}_0) , associated with a certain choice of J, U , and V , can reasonably be replaced by a more amenable problem $(\hat{\mathcal{P}}_0)$ obtained in substituting for U and V a pair of smaller sets \hat{U} and \hat{V} , e.g. bounded sets. The theorem we state next provides the answers for finite-dimensional linear-quadratic programming, although its full import will not be clear until the end of §3. It will be the basis for an infinite-dimensional generalization at the end of §6.

THEOREM 2.5. *Let (\mathcal{P}_0) and (\mathcal{Q}_0) be a pair of finite-dimensional linear-quadratic programming problems in the general sense. Consider also an auxiliary pair of such problems $(\hat{\mathcal{P}}_0)$ and $(\hat{\mathcal{Q}}_0)$ which corresponds to the same function J but subsets $\hat{U} \subset U$ and $\hat{V} \subset V$.*

(a) If \bar{u} and \bar{v} are solutions to (\mathcal{P}_0) and (\mathcal{Q}_0) such that actually $\bar{u} \in \hat{U}$ and $\bar{v} \in \hat{V}$, then \bar{u} and \bar{v} are also solutions to $(\hat{\mathcal{P}}_0)$ and $(\hat{\mathcal{Q}}_0)$.

(b) Conversely, if \bar{u} and \bar{v} are solutions to $(\hat{\mathcal{P}}_0)$ and $(\hat{\mathcal{Q}}_0)$, and if U coincides with \hat{U} around \bar{u} (i.e. $U \cap N = \hat{U} \cap N$ for some neighborhood N of \bar{u}) and V coincides with \hat{V} around \bar{v} , then \bar{u} and \bar{v} are actually solutions to (\mathcal{P}_0) and (\mathcal{Q}_0) .

Proof. From Proposition 2.1 and Theorem 2.2 we know that \bar{u} and \bar{v} solve (\mathcal{P}_0) and (\mathcal{Q}_0) if and only if (\bar{u}, \bar{v}) is a saddle point of J relative to $U \times V$. Likewise, \bar{u} and \bar{v} solve $(\hat{\mathcal{P}}_0)$ and $(\hat{\mathcal{Q}}_0)$ if and only if (\bar{u}, \bar{v}) is a saddle point of J relative to $\hat{U} \times \hat{V}$. The former trivially implies the latter when $\hat{U} = U$ and $\hat{V} = V$, and this establishes (a). Under the assumptions in (b), (\bar{u}, \bar{v}) is a saddle point relative to certain neighborhoods of \bar{u} in U and \bar{v} in V , i.e. it is a *local* saddle point relative to $U \times V$. But any local saddle point must be a global saddle point by the convexity-concavity of J . \square

3. Basic models in linear-quadratic programming. The nature of the ρ functions appearing in the finite-dimensional linear-quadratic programming problems in §2 is revealed more clearly in the examples that follow. These examples illustrate various possibilities in formulation that one needs to appreciate in order to see the broad scope of the optimal control problems which will be introduced in §4.

Example 3.1. (Classical linear programming.) Let $P = 0$, $Q = 0$, $U = \mathbb{R}_+^k$, $V = \mathbb{R}_+^\ell$. Then

$$(3.1) \quad \rho_{V,Q}(s) = \sigma_{\mathbb{R}_+^\ell}(s) = \begin{cases} 0 & \text{if } s \leq 0, \\ \infty & \text{if } s \not\leq 0, \end{cases}$$

$$(3.2) \quad \rho_{U,P}(r) = \sigma_{\mathbb{R}_+^k}(r) = \begin{cases} 0 & \text{if } r \leq 0, \\ \infty & \text{if } r \not\leq 0. \end{cases}$$

It follows that in (\mathcal{P}_0) we

$$\begin{aligned} & \text{minimize} && p \cdot u \\ & \text{subject to} && Du \geq q, \quad u \geq 0, \end{aligned}$$

whereas in (\mathcal{Q}_0) we

$$\begin{aligned} & \text{maximize} && q \cdot v \\ & \text{subject to} && D^*v \leq p, \quad v \geq 0. \end{aligned}$$

Note the role of ∞ in (3.1) and (3.2) in representing constraints in these problems as discussed in connection with the sets L and K in Proposition 2.3.

Versions of linear programming that involve equality constraints or variables not restricted to be nonnegative correspond to other choices of U and V as polyhedral convex cones.

Example 3.2. (Standard quadratic programming.) Let $Q = 0$ (but $P \neq 0$) and take $U = \mathbb{R}^k$, $V = \mathbb{R}_+^\ell$. Then (2.8) holds, and in (\mathcal{P}_0) we

$$\begin{aligned} & \text{minimize} && p \cdot u + \frac{1}{2} u \cdot Pu \\ & \text{subject to} && Du \geq q. \end{aligned}$$

This is quadratic programming in the traditional sense. To see what the dual is we must determine

$$(3.3) \quad \rho_{\mathbb{R}^k, P}(r) = \sup_{u \in \mathbb{R}^k} \{r \cdot u - \frac{1}{2} u \cdot Pu\}.$$

If P is positive definite, we easily calculate the supremum to be $\frac{1}{2}r \cdot P^{-1}r$, so that in (\mathcal{Q}_0) we

$$\begin{aligned} & \text{maximize} && q \cdot v - \frac{1}{2}[D^*v - p] \cdot P^{-1}[D^*v - p] \\ & \text{subject to} && v \geq 0. \end{aligned}$$

If P is only positive semidefinite, the dualization is more subtle and is facilitated by an algebraic normalization. First we can decompose $U = \mathbb{R}^k$ into $U_1 \times U_2$, where $U_1 = \{u \mid Pu = 0\}$ and $U_2 = U_1^\perp$. Then by a change of coordinates if necessary we can actually suppose that $U_1 \times U_2 = \mathbb{R}^{k_1} \times \mathbb{R}^{k_2}$ for some $k_1 + k_2 = k$, so that $P = \text{diag}(P_1, 0)$ for an positive definite matrix $P_1 \in \mathbb{R}^{k_1 \times k_1}$. Writing $u = (u_1, u_2)$ with $u_1 \in \mathbb{R}^{k_1}$, $u_2 \in \mathbb{R}^{k_2}$, and correspondingly $r = (r_1, r_2)$ in (3.4), we calculate

$$\begin{aligned} \rho_{U,P}(r_1, r_2) &= \sup_{u_1, u_2} \{r_1 \cdot u_1 + r_2 \cdot u_2 - \frac{1}{2}u_1 \cdot P_1 u_1\} \\ (3.4) \qquad &= \begin{cases} \frac{1}{2}r_1 \cdot P_1^{-1}r_1 & \text{if } r_2 = 0, \\ \infty & \text{if } r_2 \neq 0. \end{cases} \end{aligned}$$

Also writing $p = (p_1, p_2)$ and $D = (D_1, D_2)$, we see that in (\mathcal{P}_0) we

$$\begin{aligned} & \text{minimize} && p_1 \cdot u_1 + p_2 \cdot u_2 + \frac{1}{2}u_1 \cdot P_1 u_1 \\ & \text{subject to} && D_1 u_1 + D_2 u_2 \geq q \end{aligned}$$

whereas in (\mathcal{Q}_0) we

$$\begin{aligned} & \text{maximize} && q \cdot v - \frac{1}{2}[D_1^*v - p_1] \cdot P_1^{-1}[D_1^*v - p_1] \\ & \text{subject to} && D_2^*v = p_2, \quad v \geq 0. \end{aligned}$$

Mixed systems of equality and inequality constraints can be handled by choosing $V = \mathbb{R}_+^{\ell_1} \times \mathbb{R}^{\ell_2}$ for some $\ell_1 + \ell_2 = \ell$.

With further algebra transformations it is possible actually to normalize the study of quadratic programming to the case where the matrix P is always *diagonal*. All one has to do is provide a factorization

$$(3.5) \qquad P = M^*M \quad \text{with} \quad M \in \mathbb{R}^{m \times k} \text{ for some dimension } m.$$

Then the problem (\mathcal{P}_0) at the beginning of this example can be written as:

$$\begin{aligned} & \text{minimize} && p \cdot u + 0 \cdot u' + \frac{1}{2}u' \cdot u' \quad \text{over all } (u, u') \in \mathbb{R}^k \times \mathbb{R}^m \\ & \text{satisfying} && Du + 0u' \geq q, \quad Mu - Iu' = 0. \end{aligned}$$

This can be identified as a quadratic programming problem which can be written in terms of the enlarged vector (u, u') in the same format as the original (\mathcal{P}_0) , but with mixed equality and inequality constraints and a diagonalized quadratic form (actually with diagonal entries that are 0 for the components of u and 1 for the components of u').

Incidentally, some quadratic programming models can be set up more easily by taking advantage of the matrix Q instead of P . For example, the problem

$$\begin{aligned} & \text{minimize} && \frac{1}{2}|Du - q|^2 \\ & \text{subject to} && u \geq 0, \end{aligned}$$

where $|\cdot|$ is the Euclidean norm, can be regarded as the case of (\mathcal{P}_0) where $p = 0$, $P = 0$, $Q = I$, $U = \mathbb{R}_+^k$, $V = \mathbb{R}^\ell$, inasmuch as

$$(3.6) \qquad \rho_{\mathbb{R}^\ell, I}(s) = \frac{1}{2}|s|^2.$$

The corresponding dual problem (\mathcal{Q}_0) is:

$$\begin{aligned} & \text{maximize} && q \cdot v - \frac{1}{2}|v|^2 \\ & \text{subject to} && D^*v \leq 0. \end{aligned}$$

Example 3.3. (Basic piecewise linear programming.) Suppose $P = 0$, $Q = 0$. Let U be any convex polyhedron in \mathbf{R}^k (expressible by some system of linear constraints which, for now, does not need to be specified), and let V be the unit simplex in \mathbf{R}^ℓ :

$$(3.7) \quad V = \{v \in \mathbf{R}_+^\ell \mid v \cdot \mathbf{1} = 1\} \quad \text{where } \mathbf{1} = (1, 1, \dots, 1).$$

Then

$$(3.8) \quad \rho_{V,Q}(r) = \sigma_V(r) = \max_{i=1, \dots, \ell} r_i \quad \text{for } r = (r_1, \dots, r_\ell).$$

It follows that in (\mathcal{P}_0) we

$$\text{minimize} \quad p \cdot u + \max_{i=1, \dots, \ell} \{q_i - d_i \cdot u\} \quad \text{over } u \in U,$$

where q_i is the i th component of q and d_i the i th row of D . The “max” expression in the objective in (\mathcal{P}_0) is the pointwise maximum of a finite collection of affine functions of u and represents a general piecewise linear (i.e. polyhedral convex) function of u in the sense of [36, §19]. In the corresponding dual problem (\mathcal{Q}_0) we

$$\begin{aligned} & \text{maximize} && q \cdot v - \sigma_U(p - D^*v) \\ & \text{subject to} && v \geq 0, \quad v \cdot \mathbf{1} = 1, \end{aligned}$$

where σ_U is the support function of U as in (2.7).

The constraint structure represented so far by the set U can be handled more directly under a different choice of notation. Still with $P = 0$ and $Q = 0$, simply take $U = \mathbf{R}^k$ but

$$V = \left\{ v \in \mathbf{R}_+^\ell \mid \sum_{i=1}^m v_i = 1 \right\} \quad \text{for an index } m \text{ satisfying } 1 < m < \ell,$$

where v_i is the i th component of v . This time

$$(3.9) \quad \rho_{V,Q}(r) = \sigma_V(r) = \begin{cases} \max_{i=1, \dots, m} r_i & \text{if } r_{m+1} \geq 0, \dots, r_\ell \geq 0, \\ \infty & \text{otherwise.} \end{cases}$$

Then in (\mathcal{P}_0) we

$$\begin{aligned} & \text{minimize} && p \cdot u + \max_{i=1, \dots, m} \{q_i - d_i \cdot u\} \\ & \text{subject to} && d_i \cdot u \geq q_i \quad \text{for } i = m+1, \dots, \ell, \end{aligned}$$

whereas in (\mathcal{Q}_0) we

$$\begin{aligned} & \text{maximize} && \sum_{i=1}^{\ell} q_i v_i \\ & \text{subject to} && v_i \geq 0 \quad \text{for } i = 1, \dots, \ell, \quad \sum_{i=1}^m v_i = 1, \quad \sum_{i=1}^{\ell} v_i d_i = p. \end{aligned}$$

Example 3.4. (Bounded linear programming.) The linear programming problems in Example 3.1 are stated in terms of unbounded variables, but in practice this may not always be wise or convenient. Many linear programming codes ask the user to specify

both upper and lower bounds for the vector u in the primal problem, say $\hat{u}^- \leq u \leq \hat{u}^+$. The effects on duality, however, are not widely appreciated. In fact there is reason impose upper and lower bounds on the dual variables too, say $\hat{v}^- \leq v \leq \hat{v}^+$. What this corresponds to is a representation of constraints in terms of *linear penalties*, like those in the currently popular ℓ_1 penalty function approach to nonlinear programming (cf. Fletcher [38]).

To be specific, suppose $P = 0$, $Q = 0$, and let U and V be vectorial intervals ("boxes") defined by upper and lower bounds:

$$U = [\hat{u}^-, \hat{u}^+], \quad V = [\hat{v}^-, \hat{v}^+].$$

Adopting the notation

$$(3.10) \quad [s]_+ = \max\{0, s\}, \quad [s]_- = \min\{0, s\}$$

in the vectorial sense, where the max is taken component by component (so that $s = [s]_+ + [s]_-$), we get

$$(3.11) \quad \rho_{V,Q}(s) = \sigma_V(s) = \max_{\hat{v}^- \leq v \leq \hat{v}^+} v \cdot s = \hat{v}^+ \cdot [s]_+ + \hat{v}^- \cdot [s]_-,$$

$$(3.12) \quad \rho_{U,P}(r) = \sigma_U(r) = \max_{\hat{u}^- \leq u \leq \hat{u}^+} u \cdot r = \hat{u}^+ \cdot [r]_+ + \hat{u}^- \cdot [r]_-.$$

It follows that in (\mathcal{P}_0) we

$$\begin{aligned} &\text{minimize} && p \cdot u + \hat{v}^+ \cdot [q - Du]_+ + \hat{v}^- \cdot [q - Du]_- \\ &\text{subject to} && \hat{u}^- \leq u \leq \hat{u}^+, \end{aligned}$$

whereas in (\mathcal{Q}_0) we

$$\begin{aligned} &\text{maximize} && q \cdot v + \hat{u}^+ \cdot [D^*v - p]_+ + \hat{u}^- \cdot [D^*v - p]_- \\ &\text{subject to} && \hat{v}^- \leq v \leq \hat{v}^+. \end{aligned}$$

Observe that these problems have piecewise linear objectives of a special kind. The optimal values are always finite, so optimal solutions always exist (Theorem 2.2).

Bounded linear programming in this sense may be a more natural vehicle in some applications than standard linear programming. Furthermore, problems in such a format can be solved directly, without reformulating them in the traditional way. Versions of the simplex method developed by Fourer [39] and the author [40, Chap. 11] can be used instead, for example.

Example 3.5. (Bounded quadratic programming.) This is an extension of the preceding example to allow for quadratic terms. Let

$$U = [\hat{u}^-, \hat{u}^+] \quad \text{and} \quad V = [\hat{v}^-, \hat{v}^+]$$

again, and take

$$P = \text{diag} [\beta_1, \dots, \beta_k], \quad Q = \text{diag} [\gamma_1, \dots, \gamma_\ell],$$

where $\beta_j \geq 0$, $\gamma_i \geq 0$. (The assumption of a diagonal form for P and Q does not entail the loss of generality that might be imagined; cf. Example 3.2.) The calculation of the ρ functions (2.5) and (2.6) decomposes into one-dimensional calculations of the form

$$(3.13) \quad \max_{\alpha \in [\alpha^-, \alpha^+]} \left\{ \tau \alpha - \frac{1}{2} \lambda \alpha^2 \right\}$$

for various intervals $[\alpha^-, \alpha^+]$ and constants $\lambda \geq 0$. The maximum value in (3.13) is a function of $\tau \in \mathbb{R}$ that depends on the parameters α^-, α^+ and λ , and it is given by

$$(3.14) \quad \theta(\tau; \alpha^-, \alpha^+, \lambda) = \begin{cases} (2\tau - \lambda\alpha^+)(\alpha^+/2) & \text{when } \tau \geq \lambda\alpha^+, \\ (1/2\lambda)\tau^2 & \text{when } \lambda\alpha^- < \tau < \lambda\alpha^+, \\ (2\tau - \lambda\alpha^-)(\alpha^-/2) & \text{when } \tau \leq \lambda\alpha^-. \end{cases}$$

Despite its formula, this function of τ has a simple form and a natural meaning. In the case where $\lambda = 0$, it vanishes at $\tau = 0$, is linear with slope α_+ for $\tau > 0$ and linear with slope α_- for $\tau < 0$. In the case where $\lambda > 0$, it has a similar structure but with a quadratic interpolation instead of a “corner.” Indeed, it is the unique *smooth* function whose values are given by $\alpha^+\tau + \text{const.}$ for τ sufficiently high, by $\alpha^-\tau + \text{const.}$ for τ sufficiently low, and by $(1/2\lambda)\tau^2$ on the interval between.

With this notation, and denoting the components of p, q , and D by p_j, q_i, d_{ij} , and so forth, we can express the primal and dual problems as follows. In (\mathcal{P}_0) we

$$\begin{aligned} &\text{minimize} && \sum_{j=1}^k [p_j u_j + \tfrac{1}{2} \beta_j u_j^2] + \sum_{i=1}^{\ell} \theta \left(q_i - \sum_{j=1}^k d_{ij} u_j; \hat{v}_i^-, \hat{v}_i^+, \gamma_i \right) \\ &\text{subject to} && \hat{u}_j^- \leq u_j \leq \hat{u}_j^+ \quad \text{for } j = 1, \dots, k, \end{aligned}$$

whereas in (\mathcal{Q}_0) we

$$\begin{aligned} &\text{maximize} && \sum_{i=1}^{\ell} [q_i v_i - \tfrac{1}{2} \gamma_i v_i^2] - \sum_{j=1}^k \theta \left(\sum_{i=1}^{\ell} v_i d_{ij} - p_j; \hat{u}_j^-, \hat{u}_j^+, \beta_j \right) \\ &\text{subject to} && \hat{v}_i^- \leq v_i \leq \hat{v}_i^+ \quad \text{for } i = 1, \dots, \ell. \end{aligned}$$

When $\beta_j = 0, \gamma_i = 0$, these problems reduce to the bounded linear case in Example 3.4. They are useful in modeling situations where constraints are not necessarily sharp, as in stochastic programming (see Rockafellar and Wets [18] and King et al. [41]). Thus for instance if $[\hat{v}_i^-, \hat{v}_i^+] = [0, \alpha_i^+]$ the corresponding θ term in (\mathcal{P}_0) imposes no penalty if the putative constraint $\sum_{j=1}^k d_{ij} u_j \geq q_i$ is satisfied, a slight penalty at a marginal cost that grows linearly (at the rate $1/\gamma_i$) from 0 as this constraint begins to be violated, and eventually for large violations a penalty with constant marginal cost α_i^+ .

Of course it is also possible to get versions of these problems in which the penalty expressions do not eventually become linear but stay quadratic for arbitrarily large violations. These correspond to limiting cases of $\theta(\tau; \alpha^-, \alpha^+, \lambda)$ where $\alpha^- = -\infty$ or $\alpha^+ = \infty$, or both. They can be obtained by taking U and V not to be “boxes” but orthants or products of orthants and subspaces, as in Example 2.3. \square

In understanding the relationship between penalty models such as Examples 3.4 and 3.5 and the more traditional models without penalties, such as Examples 3.1 and 3.2, the facts in Theorem 2.5 are essential. As an illustration of the way Theorem 2.5 can be employed, let us look again at the standard linear programming problems in Example 3.1. Suppose we know that an optimal solution \bar{u} to (\mathcal{P}_0) will exist within certain upper bounds, say $\bar{u} \leq \hat{u}$, and also that a dual optimal solution \bar{v} to (\mathcal{Q}_0) will exist within certain upper bounds, say $\bar{v} \leq \hat{v}$. Then according to Theorem 2.5(a), \bar{u} and \bar{v} can be found by solving, instead of the given problems, the bounded linear programming problems in Example 3.4 with

$$(3.15) \quad U = [\hat{u}^-, \hat{u}^+] = [0, \hat{u}], \quad V = [\hat{v}^-, \hat{v}^+] = [0, \hat{v}].$$

The idea here that dual bounds can be given along with primal bounds is not so far-fetched as it might seem. The components of a dual optimal solution \bar{v} often have interpretation as marginal prices, or as rates of change with respect to certain perturbations of constraints. Economic limitations or experience may dictate appropriate bounds. Anyway, there is no great harm in going ahead with solving the bounded versions of the problems in terms of *estimated* bounds \hat{u} and \hat{v} . If solutions \bar{u} and \bar{v} are obtained for which the upper bounds are not tight, then \bar{u} and \bar{v} actually solve the original problems, according to Theorem 2.5(b). If the upper bounds are tight in some components, they can be loosened and the procedure repeated.

4. Intertemporal linear-quadratic programming. The general problems of optimal control that are the main object of our study can now be formulated. The time interval $[t_0, t_1]$ is fixed. The primal problem is:

(P) Over the dynamical system

$$\dot{x}(t) = A(t)x(t) + B(t)u(t) + b(t) \quad \text{a.e.}, \quad x(t_0) = B_e u_e + b_e,$$

with control space

$$\mathcal{U} = \{(u, u_e) \mid u \in \mathcal{L}^1, u(t) \in U(t) \text{ a.e.}, u_e \in U_e\}$$

minimize the functional

$$\begin{aligned} \mathcal{F}(u, u_e) = & \int_{t_0}^{t_1} [p(t) \cdot u(t) + \frac{1}{2}u(t) \cdot P(t)u(t) - c(t) \cdot x(t)]dt + [p_e \cdot u_e + \frac{1}{2}u_e \cdot P_e u_e - c_e \cdot x(t_1)] \\ & + \int_{t_0}^{t_1} \rho_{V(t), Q(t)}(q(t) - C(t)x(t) - D(t)u(t)) + \rho_{V_e, Q_e}(q_e - C_e x(t_1) - D_e u_e). \end{aligned}$$

The dual problem is:

(Q) Over the dynamical system

$$-\dot{y}(t) = A^*(t)y(t) + C^*(t)v(t) + c(t) \quad \text{a.e.}, \quad y(t_1) = C_e^* v_e + c_e,$$

with control space

$$\mathcal{V} = \{(v, v_e) \mid v \in \mathcal{L}^1, v(t) \in V(t) \text{ a.e.}, v_e \in V_e\},$$

maximize the functional

$$\begin{aligned} \mathcal{G}(v, v_e) = & \int_{t_0}^{t_1} [q(t) \cdot v(t) - \frac{1}{2}v(t) \cdot Q(t)v(t) - b(t) \cdot y(t)]dt + [q_e \cdot v_e - \frac{1}{2}v_e \cdot Q_e v_e - b_e \cdot y(t_0)] \\ & - \int_{t_0}^{t_1} \rho_{U(t), P(t)}(B^*(t)y(t) + D^*(t)v(t) - p(t))dt - \rho_{V_e, P_e}(B_e^* y(t_0) + D_e^* v_e - p_e). \end{aligned}$$

Here

$$u(t) \in \mathbf{R}^k, \quad u_e \in \mathbf{R}^{k_e}, \quad x(t) \in \mathbf{R}^n, \quad v(t) \in \mathbf{R}^\ell, \quad v_e \in \mathbf{R}^{\ell_e}, \quad y(t) \in \mathbf{R}^n,$$

and dimensions of the other elements are determined accordingly. The matrices $P(t)$, P_e , $Q(t)$ and Q_e are assumed to be symmetric and positive *semidefinite* (possibly 0). The sets $U(t) = \mathbf{R}^k$, $U_e = \mathbf{R}^{k_e}$, $V(t) \subset \mathbf{R}^\ell$ and $V_e \subset \mathbf{R}^{\ell_e}$ are assumed to be polyhedral

convex. The ρ terms are defined by (2.5) and (2.6). In general they are piecewise linear-quadratic convex functions that may take on the value ∞ ; cf. Proposition 2.3. Various cases based in part on the finite-dimensional models in §3 will be viewed in §5. First we must clarify our technical foundations.

All the data elements in problems (\mathcal{P}) and (\mathcal{Q}) , namely

$$A(t), B(t), C(t), D(t), b(t), c(t), P(t), Q(t), p(t), q(t), U(t), V(t),$$

are assumed to depend *continuously* on t . For the sets $U(t)$ and $V(t)$ this means continuity with respect to the usual notions of convergence of subsets of Euclidean space that are not necessarily bounded; see Salinetti and Wets [42] for an exposition of the convex case. Thus the multifunctions $t \mapsto U(t)$ and $t \mapsto V(t)$ should be lower semicontinuous and of closed graph. Lower semicontinuity of $t \mapsto U(t)$ implies that the multifunction $t \mapsto \text{int } U(t)$ is of open graph; indeed, by virtue of the convexity of $U(t)$, lower semicontinuity is equivalent to the latter property if $\text{int } U(t) \neq \emptyset$ for all $t \in [t_0, t_1]$ (Rockafellar [43, p. 458]). A special case of continuous dependence, of course, is the one where $U(t)$ and $V(t)$ are constant with respect to t .

Under these assumptions the dynamical systems in (\mathcal{P}) and (\mathcal{Q}) are well defined with respect to the control spaces \mathcal{U} and \mathcal{V} . They determine unique absolutely continuous functions x and y from $[t_0, t_1]$ to \mathbb{R}^n .

In showing that the integrals in the objective functionals in (\mathcal{P}) and (\mathcal{Q}) are well defined too, we shall make use of the following.

PROPOSITION 4.1. *The expression $\rho_{V(t), Q(t)}(s)$ is lower semicontinuous jointly in t and s , in fact continuous relative to $\{(t, s) \mid s \in \text{int } L(t)\}$, where*

$$(4.1) \quad L(t) = \{s \in \mathbb{R}^\ell \mid \rho_{V(t), Q(t)}(s) < \infty\}.$$

Moreover $L(t)$ depends lower semicontinuously on t .

The same holds for the expression $\rho_{U(t), P(t)}(r)$ and the effective domain

$$(4.2) \quad K(t) = \{r \in \mathbb{R}^k \mid \rho_{U(t), P(t)}(r) < \infty\}.$$

Proof. Our argument is based on showing that the function $\rho_{V(t), Q(t)}$ depends *epicontinuously* on t , i.e. its epigraph set

$$(4.3) \quad E(t) = \{(s, \alpha) \in \mathbb{R}^\ell \times \mathbb{R} \mid \rho_{V(t), Q(t)}(s) \leq \alpha\},$$

which is convex, depends continuously on t . Epicontinuity corresponds to a notion of function convergence first considered by Wijsman [44] and subsequently developed by others; see Wets [45]. It yields all the properties claimed. Indeed, if the multifunction $t \mapsto E(t)$ is continuous, then by definition it is lower semicontinuous and of closed graph. The closed graph property is equivalent to the lower semicontinuity of the function

$$(4.4) \quad (t, s) \mapsto \rho_{V(t), Q(t)}(s).$$

The lower semicontinuity of $t \mapsto E(t)$ implies from its definition the lower semicontinuity of the domain multifunction $t \mapsto L(t)$, since $L(t)$ is the projection in \mathbb{R}^ℓ of the epigraph (4.3). (Recall from Proposition 2.3 that $L(t)$ is a *closed* convex set, since it is polyhedral.) The multifunction $t \mapsto \text{int } L(t)$ is then of open graph, as cited above, i.e. the set $\{(t, s) \mid s \in \text{int } L(t)\}$ is open in the space $[t_0, t_1] \times \mathbb{R}^\ell$. The upper semicontinuity of $\rho_{V(t), Q(t)}(s)$ on this open set follows then from the lower semicontinuity of $t \mapsto E(t)$ again and the corresponding openness of $\{(t, s, \alpha) \mid (s, \alpha) \in \text{int } E(t)\}$, and with the lower semicontinuity noted earlier for (4.4) one gets continuity.

To prove that $\rho_{V(t), Q(t)}$ depends epicontinuously on t , we resort to the notation of Proposition 2.3, where now, however, everything depends on t . We identify $\rho_{V(t), Q(t)}$ with the conjugate φ_t^* of the convex function $\varphi_t = j_{Q(t)} + \delta_{V(t)}$, where $\delta_{V(t)}$ is the indicator of $V(t)$ and

$$(4.5) \quad j_{Q(t)}(v) = \frac{1}{2}v \cdot Q(t)v.$$

Trivially $\delta_{V(t)}$ depends epicontinuously on t , since its epigraph is just $V(t) \times \mathbb{R}_+$. Furthermore the convex function $j_{Q(t)}$ is finite everywhere on \mathbb{R}^ℓ , and its values depend continuously on t because $Q(t)$ depends continuously on t . This implies by Wets [45, p. 392] that $j_{Q(t)}$ depends epicontinuously on t and by McLinden and Bergstrom [46, Thm. 6] that the sum $\varphi_t = j_{Q(t)} + \delta_{V(t)}$ depends epicontinuously on t . The operation of passing to the conjugate of a convex function is known to preserve epicontinuity (Wijsman [44]), so we may conclude that the function $\rho_{V(t), Q(t)} = \varphi_t^*$ does depend epicontinuously on t , as claimed. \square

THEOREM 4.2. *In problem (P) the control space \mathcal{U} is a nonempty closed convex subset of $\mathcal{L}^1([t_0, t_1], \mathbb{R}^k) \times \mathbb{R}^{k_e}$, and the objective functional \mathcal{F} is well defined, lower semicontinuous and convex, with values that are finite or ∞ .*

Likewise, in problem (Q) the control space \mathcal{V} is a nonempty closed convex subset of $\mathcal{L}^1([t_0, t_1], \mathbb{R}^\ell) \times \mathbb{R}^{\ell_e}$, and the objective functional \mathcal{G} is well defined, upper semicontinuous and concave, with values that are finite or $-\infty$.

Proof. Only the first half has to be argued; the second half is parallel. The convexity and closedness of \mathcal{U} is obvious from the convexity and closedness of the sets $U(t)$ and U_e . The nonemptiness of \mathcal{U} comes from the nonemptiness of $U(t)$ and U_e and the continuity of $t \mapsto U(t)$: the selection theorem of Michael [47] asserts that any lower semicontinuous multifunction from $[t_0, t_1]$ to \mathbb{R}^k with nonempty closed convex values has a continuous selection. Thus there actually exist pairs (u, u_e) in \mathcal{U} with u continuous rather than just \mathcal{L}^1 .

The mapping $(u, u_e) \mapsto x$ from $\mathcal{L}^1([t_0, t_1], \mathbb{R}^k) \times \mathbb{R}^{k_e}$ into $\mathcal{C}([t_0, t_1], \mathbb{R}^n)$ is affine and continuous, even compact:

$$(4.6) \quad x(t) = M(t) \left(B_e u_e + b_e + \int_{t_0}^t M(\tau)^{-1} [B(\tau)u(\tau) + b(\tau)] d\tau \right),$$

where $M(t)$ is the matrix with the property that $\xi(t) = M(t)x_0$ is the solution to $\dot{\xi}(t) = A(t)\xi(t)$, $\xi(t_0) = x_0$. The terms

$$\int_{t_0}^{t_1} [p(t) \cdot u(t) - c(t) \cdot x(t)] dt + [p_e \cdot u_e - c_e \cdot x(t_1)]$$

in $\mathcal{F}(u, u_e)$ therefore give a continuous, affine functional of (u, u_e) . The mapping that takes a pair (u, u_e) in $\mathcal{L}^1([t_0, t_1], \mathbb{R}^k) \times \mathbb{R}^{k_e}$ into the pair (s, s_e) in $\mathcal{L}^1([t_0, t_1], \mathbb{R}^\ell) \times \mathbb{R}^{\ell_e}$ given by

$$(4.7) \quad s(t) = q(t) - C(t)x(t) - D(t)u(t), \quad s_e = q_e - C_e x(t_1) - D_e u_e,$$

is affine and continuous too.

It remains only to show that the expressions

$$\begin{aligned} I_1(u, u_e) &= \int_{t_0}^{t_1} u(t) \cdot P(t)u(t) dt + u_e \cdot P_e u_e, \\ I_2(s, s_e) &= \int_{t_0}^{t_1} \rho_{V(t), Q(t)}(s(t)) dt + \rho_{V_e, Q_e}(s_e) \end{aligned}$$

give well defined, lower semicontinuous, convex functionals on $\mathcal{L}^1([t_0, t_1], \mathbf{R}^k) \times \mathbf{R}^{k_e}$ and $\mathcal{L}^1([t_0, t_1], \mathbf{R}^\ell) \times \mathbf{R}^{\ell_e}$ respectively, with values that are finite or ∞ . Certainly the continuity of $P(t)$ in t and the lower semicontinuity of $\rho_{V(t), Q(t)}(s)$ jointly in t and s (proved in Proposition 4.1) ensure that the integrands for I_1 and I_2 are measurable in t .

All the terms in the formula for I_1 are nonnegative and convex, because $P(t)$ and P_e are positive semidefinite. Therefore I_1 is a well defined convex functional with values in $[0, \infty]$. Its lower semicontinuity follows from Fatou's lemma, since every norm-convergent sequence in $\mathcal{L}^1([t_0, t_1], \mathbf{R}^k)$ has a subsequence that converges pointwise almost everywhere.

The argument for I_2 is the same, after a normalization. We showed at the outset of this proof that \mathcal{U} contains a pair (u, u_e) with u actually continuous. The same applies to \mathcal{V} . Taking (v, v_e) to be such a pair in \mathcal{V} and observing from the definition of the ρ functions that then

$$\rho_{V(t), Q(t)}(s) \geq s(t) \cdot v(t), \quad \rho_{V_e, Q_e}(s_e) \geq s_e \cdot v_e,$$

we can write

$$I_2(s, s_e) = I_3(s, s_e) + \int_{t_0}^{t_1} s(t) \cdot v(t) dt + s_e \cdot v_e,$$

where

$$I_3(s, s_e) = \int_{t_0}^{t_1} [\rho_{V(t), Q(t)}(s(t)) - s(t) \cdot v(t)] dt + [\rho_{V_e, Q_e}(s_e) - s_e \cdot v_e].$$

Thus I_2 differs by only a continuous linear functional from a functional I_3 whose terms are all convex and nonnegative. As with I_1 we can see that I_3 is well defined with values in $[0, \infty]$ and is convex and lower semicontinuous. Therefore I_2 has these required properties, except that its values will generally be in $(-\infty, \infty]$. \square

It is evident that in the minimization in (\mathcal{P}) we are really interested only in the controls $(u, u_e) \in \mathcal{U}$ yielding $\mathcal{F}(u, u_e) < \infty$. Such controls have to satisfy

$$(4.8) \quad q(t) - C(t)x(t) - D(t)u(t) \in L(t) \text{ a.e. and } q_e - C_e x(t_1) - D_e u_e \in L_e,$$

where $L(t)$ and L_e are the effective domains of $\rho_{V(t), Q(t)}$ and ρ_{V_e, Q_e} (cf. Proposition 2.3). Similarly, in the maximization in (\mathcal{Q}) we are really interested only in the controls (v, v_e) yielding $\mathcal{G}(v, v_e) > -\infty$, and these have to satisfy

$$(4.9) \quad B^*(t)y(t) + D^*(t)v(t) - p(t) \in K(t) \text{ a.e. and } B_e^* y(t_0) + D_e^* v_e - p_e \in K_e,$$

where $K(t)$ and K_e are the effective domains of $\rho_{U(t), P(t)}$ and ρ_{U_e, P_e} . These implicit constraints can be regarded as "linear," incidentally, since the sets $L(t)$, L_e , $K(t)$ and K_e are polyhedral convex cones (Propositions 2.3 and 2.4).

As stated in §1, our approach in this paper to such implicit constraints involving the states $x(t)$ and $y(t)$ is to skirt them when convenient by adopting alternative problem formulations where they have no force, specifically because $L(t)$ and L_e are all of \mathbf{R}^ℓ and \mathbf{R}^{ℓ_e} , or $K(t)$ and K_e are all of \mathbf{R}^k and \mathbf{R}^{k_e} . Accordingly the following type of assumption will sometimes be of importance to us.

We shall say that the *primal finiteness condition* is satisfied if the functions $\rho_{V(t), Q(t)}$ and ρ_{V_e, Q_e} are finite everywhere (i.e. $L(t) = \mathbf{R}^\ell$ and $L_e = \mathbf{R}^{\ell_e}$). Likewise, the *dual finiteness condition* is satisfied if the functions $\rho_{U(t), P(t)}$ and ρ_{U_e, P_e} are finite everywhere (i.e. $K(t) = \mathbf{R}^k$ and $K_e = \mathbf{R}^{k_e}$). Criteria for this are furnished by Proposition 2.4.

PROPOSITION 4.3. *If the primal finiteness condition is satisfied, then $\mathcal{F}(u, u_e)$ in (\mathcal{P}) is finite for all $(u, u_e) \in \mathcal{U}$ with $u \in \mathcal{L}^\infty$.*

Likewise, if the dual finiteness condition is satisfied, then $\mathcal{G}(v, v_e)$ in (\mathcal{Q}) is finite for all $(v, v_e) \in \mathcal{V}$ with $v \in \mathcal{L}^\infty$.

Proof. Under the primal finiteness condition the convex functions $\rho_{V(t), Q(t)}$ and ρ_{V_e, Q_e} are finite on \mathbb{R}^ℓ and \mathbb{R}^{ℓ_e} and therefore continuous on these spaces, inasmuch as a convex function on a finite-dimensional space is continuous on any open set where it is finite [36, §10]. Moreover $\rho_{V(t), Q(t)}(s)$ is continuous jointly in t and s by Proposition 4.1 and consequently is bounded above and below on $[t_0, t_1] \times S$ for any bounded subset $S \subset \mathbb{R}^\ell$. For the function $s(t)$ in (4.7), then, the expression $\rho_{V(t), Q(t)}(s(t))$ is \mathcal{L}^∞ in t when $u(t)$ is \mathcal{L}^∞ in t , as is the expression $u(t) \cdot P(t)u(t)$. All the integrals in the formula for $\mathcal{F}(u, u_e)$ are therefore finite when $u \in \mathcal{L}^\infty$. The argument for $\mathcal{G}(v, v_e)$ under the dual finiteness condition runs the same way. \square

The reader may wonder why we have formulated problems (\mathcal{P}) and (\mathcal{Q}) with control spaces involving \mathcal{L}^1 rather than \mathcal{L}^∞ . Matters would be simpler in some respects with \mathcal{L}^∞ , and for applications \mathcal{L}^∞ is apparently more natural. The work done in continuous-time programming uses \mathcal{L}^∞ too. Of course, our problems include the \mathcal{L}^∞ case by simple restriction. The real reason for taking \mathcal{L}^1 , however, is not extra generality but the need for allowing ample controls in order to close a possible duality gap between (\mathcal{P}) and (\mathcal{Q}) . The payoff will come in our result on strong duality, Theorem 6.3.

5. Special cases of the optimal control models. Our task now is to illuminate the scope of the problems (\mathcal{P}) and (\mathcal{Q}) introduced in §4. We explain how they cover the linear programming models (\mathcal{P}_1) and (\mathcal{Q}_1) in §1 and much more.

The treatment of endpoints $x(t_0)$ and $x(t_1)$ in (\mathcal{P}) and $y(t_0)$ and $y(t_1)$ in (\mathcal{Q}) departs from the traditional patterns in the literature on optimal control. We therefore begin by considering various important cases embedded in our formulation and the way they come to be dualized.

Example 5.1. (Problems with fixed endpoints.) How can one represent in terms of the endpoint provisions in the structure of (\mathcal{P}) a problem in which an integral

(5.1)

$$\int_{t_0}^{t_1} [p(t) \cdot u(t) + \frac{1}{2} u(t) \cdot P(t)u(t) - c(t) \cdot x(t) + \rho_{V(t), Q(t)}(q(t) - C(t)x(t) - D(t)u(t))] dt$$

is minimized over all pairs x, u , satisfying $u(t) \in U(t)$ a.e., $u \in \mathcal{L}^1$,

$$(5.2) \quad \dot{x}(t) = A(t)x(t) + B(t)u(t) + b(t) \quad \text{a.e.}, \quad x(t_0) = a_0, \quad x(t_1) = a_1,$$

where a_0 and a_1 are fixed points in \mathbb{R}^n ? The requirement $x(t_0) = a_0$ can be handled by setting $b_e = a_0$ and trivializing the u_e vector by taking \mathbb{R}^{k_e} to be *zero-dimensional* (so $U_e = \{0\}$, $B_e = 0$, $D_e = 0$, $p_e = 0$, $P_e = 0$). Only the term

$$(5.3) \quad \rho_{V_e, Q_e}(q_e - C_e x(t_1)) - c_e \cdot x(t_1)$$

remains then in the endpoint expression for (\mathcal{P}) . This can be made to represent the requirement $x(t_1) = a_1$ as follows. First choose $V_e = \mathbb{R}^n$ and $Q_e = 0$, so that

$$(5.4) \quad \rho_{V_e, Q_e}(s_e) = \sigma_{\mathbb{R}^n}(s_e) = \begin{cases} 0 & \text{if } s_e = 0, \\ \infty & \text{if } s_e \neq 0. \end{cases}$$

Then

$$(5.5) \quad \rho_{V_e, Q_e}(q_e - C_e x(t_1)) = \begin{cases} 0 & \text{if } C_e x(t_1) = q_e, \\ \infty & \text{if } C_e x(t_1) \neq q_e. \end{cases}$$

Now all one has to do is take $C_e = I$, $q_e = a_1$, $c_e = 0$.

Note that the dual problem (Q) in this case has as its endpoint term

$$(5.6) \quad q_e \cdot v_e - \frac{1}{2} v_e \cdot Q_e v_e - b_e \cdot y(t_0) - \rho_{V_e, P_e}(B_e^* y(t_0) + D_e^* v_e - p_e) = a_1 \cdot y(t_1) - a_0 \cdot y(t_0).$$

In (Q) , therefore, one maximizes

$$(5.7) \quad \int_{t_0}^{t_1} [q(t) \cdot v(t) - \frac{1}{2} v(t) \cdot Q(t) v(t) - b(t) \cdot y(t) - \rho_{U(t), P(t)}(B^*(t) y(t) + D^*(t) v(t) - p(t))] dt \\ + a_1 \cdot y(t_1) - a_0 \cdot y(t_0)$$

over all pairs y, v , such that $v(t) \in V(t)$ a.e., $v \in \mathcal{L}^1$, and

$$(5.8) \quad -\dot{y}(t) = A^*(t) y(t) + C^*(t) v(t) + c(t) \quad \text{a.e.}$$

(with no restriction on the endpoints $y(t_0)$ and $y(t_1)$).

Of course one can stop with (5.4), (5.5), and have in place of $x(t_1) = a_1$ the more general constraint $C_e x(t_1) = q_e$ for some matrix C_e and vector q_e . In (Q) this would correspond to replacing the term $a_1 \cdot y(t_1)$ in (5.7) by $q_e \cdot v_e$, where v_e is unrestricted but $y(t_1) = C_e^* v_e$ in (5.8) (if $c_e = 0$ still).

If we only want $x(t_0) = a_0$ in (5.2), so that $x(t_1)$ is a free endpoint in (P) , and correspondingly want to incorporate a term $-d_1 \cdot x(t_1)$ in the objective (5.1), we can represent this by trivializing the vector v_e too, i.e. by taking \mathbb{R}^{l_e} to be zero-dimensional (so that $V_e = \{0\}$, $C_e = 0$, $q_e = 0$, $Q_e = 0$), and setting $c_e = d_1$. Then the term (5.3) reduces to $-d_1 \cdot x(t_1)$. In the corresponding version of (Q) the term $a_1 \cdot y(t_1)$ drops from (5.7) but $y(t_1) = d_1$ is added to (5.8). Thus (Q) is a problem of the same type but with $y(t_1)$ fixed and $y(t_0)$ free.

Example 5.2. (General linear constraints on endpoints.) Instead of fixed endpoints let us consider a much more general case where the functional (5.1) is to be minimized over all pairs x, u , satisfying $u(t) \in U(t)$ a.e., $u \in \mathcal{L}^1$,

$$\dot{x}(t) = A(t)x(t) + B(t)u(t) + b(t) \quad \text{a.e.}$$

and a constraint system of the form

$$(5.9) \quad A_0 x(t_0) + A_1 x(t_1) \geq a$$

on the endpoints, with $a \in \mathbb{R}^d$. This can be placed in the form of (P) by choosing $B_e = I$ and $b_e = 0$ (so that $x(t_0) = u_e$ in (P)) and then setting $U_e = \mathbb{R}^n$, $D_e = A_0$, $C_e = A_1$, $q_e = a$, $V_e = \mathbb{R}_+^d$, $Q_e = 0$. Then

$$(5.10) \quad \rho_{V_e, Q_e}(q_e - C_e x(t_1) - D_e u_e) = \begin{cases} 0 & \text{if (5.9) holds,} \\ \infty & \text{otherwise.} \end{cases}$$

Taking $p_e = 0$, $P_e = 0$, $c_e = 0$, we get all the endpoint terms other than (5.9) to drop out, and (P) then represents the problem as specified.

The corresponding dual problem (Q) maximizes

$$\int_{t_0}^{t_1} [q(t) \cdot v(t) - \frac{1}{2} v(t) \cdot Q(t) v(t) - b(t) \cdot y(t) - \rho_{U(t), P(t)}(B^*(t) y(t) + D^*(t) v(t) - p(t))] dt + a \cdot v_e$$

over all y, v, v_e satisfying $v(t) \in V(t)$ a.e., $v \in \mathcal{L}^\infty$, $v_e \in \mathbb{R}_+^d$,

$$(5.11) \quad -\dot{y}(t) = A^*(t)y(t) + C^*(t)v(t) + c(t) \text{ a.e.}, \quad y(t_0) = -A_0^*v_e, \quad y(t_1) = A_1^*v_e.$$

Obviously the inequality in (5.9) can be converted to an equation by taking $V_e = \mathbb{R}^d$ instead of \mathbb{R}_+^d . For a particularly interesting case of this, let $A_0 = -I$, $A_1 = I$, $a = 0$. Then (5.9) reduces to the requirement that $x(t_0) = x(t_1)$, and the endpoint conditions in (5.11) reduce correspondingly to $y(t_0) = y(t_1)$ ("periodic" boundary conditions).

Example 5.3. (Basic intertemporal linear programming.) Problems (\mathcal{P}) and (\mathcal{Q}) turn into the basic linear programming models (\mathcal{P}_1) and (\mathcal{Q}_1) described in §1 when $P(t), P_e, Q(t)$ and Q_e are zero matrices and

$$(5.12) \quad U(t) = \mathbb{R}_+^k, \quad U_e = \mathbb{R}_+^{k_e}, \quad V(t) = \mathbb{R}_+^\ell, \quad V_e = \mathbb{R}_+^{\ell_e}$$

in the pattern of Example 3.1. By choosing products of orthants and subspaces in (5.12) instead of merely orthants, one obtains the versions of these problems having a mixture of equality and inequality constraints. Neither the primal nor the dual finiteness condition (as defined in the last section, before Proposition 4.3) is satisfied in any such formulation, however.

The endpoint conditions in Examples 5.1 and 5.2 all fit into the mold of this example, since only linear constraints are involved.

Example 5.4. (Bounded intertemporal linear programming.) With $P(t), P_e, Q(t)$ and Q_e still taken to be zero matrices as in the preceding example, replace (4.8) by a choice of vectorial intervals giving upper and lower bounds on the various control vectors:

$$(5.13)$$

$$U(t) = [\hat{u}^-(t), \hat{u}^+(t)], \quad U_e = [\hat{u}_e^-, \hat{u}_e^+], \quad V(t) = [\hat{v}^-(t), \hat{v}^+(t)], \quad V_e = [\hat{v}_e^-, \hat{v}_e^+].$$

The assumption of continuous dependence of $U(t)$ and $V(t)$ on t is satisfied if the vectors $\hat{u}^-(t)$, $\hat{u}^+(t)$, $\hat{v}^-(t)$ and $\hat{v}^+(t)$ depend continuously on t . In this case *the primal and dual finiteness conditions are both satisfied*. In the notation introduced in Example 3.4 the objective in (\mathcal{P}) is to minimize

$$\begin{aligned} \mathcal{F}(u, u_e) = & \int_{t_0}^{t_1} [p(t) \cdot u(t) - c(t) \cdot x(t)] dt + [p_e \cdot u_e - c_e \cdot x(t_1)] \\ & + \int_{t_0}^{t_1} \hat{v}^-(t) \cdot [q(t) - C(t)x(t) - D(t)u(t)]_- dt + \hat{v}_e^- \cdot [q_e - C_e x(t_1) - D_e u_e]_- \\ & + \int_{t_0}^{t_1} \hat{v}^+(t) \cdot [q(t) - C(t)x(t) - D(t)u(t)]_+ dt + \hat{v}_e^+ \cdot [q_e - C_e x(t_1) - D_e u_e]_+, \end{aligned}$$

while the objective in (\mathcal{Q}) is to maximize

$$\begin{aligned} \mathcal{G}(v, v_e) = & \int_{t_0}^{t_1} [q(t) \cdot v(t) - b(t) \cdot y(t)] dt + [q_e \cdot v_e - b_e \cdot y(t_0)] \\ & - \int_{t_0}^{t_1} \hat{u}^-(t) \cdot [B^*(t)y(t) + D^*(t)v(t) - p(t)]_- dt - \hat{u}_e^- \cdot [B_e^* y(t_0) + D_e^* v_e - p_e]_- \\ & - \int_{t_0}^{t_1} \hat{u}^+(t) \cdot [B^*(t)y(t) + D^*(t)v(t) - p(t)]_+ dt - \hat{u}_e^+ \cdot [B_e^* y(t_0) + D_e^* v_e - p_e]_+. \end{aligned}$$

For instance, by taking

$$(5.14) \quad V(t) = [-\lambda \mathbf{1}, \lambda \mathbf{1}], \quad V_e = [-\lambda_e \mathbf{1}, \lambda_e \mathbf{1}],$$

where $\mathbf{1}$ denotes a vector $(1, 1, \dots, 1)$ of appropriate dimension, we obtain in (\mathcal{P}) the objective

$$(5.15) \quad \begin{aligned} \mathcal{F}(u, u_e) = & \int_{t_0}^{t_1} [p(t) \cdot u(t) - c(t) \cdot x(t)] dt + [p_e \cdot u_e - c_e \cdot x(t_1)] \\ & + \lambda \int_{t_0}^{t_1} \|q(t) - C(t)x(t) - D(t)u(t)\|_1 dt + \lambda_e \|q_e - C_e x(t_1) - D_e u_e\|_1, \end{aligned}$$

where

$$(5.16) \quad \|s\|_1 = \|(s_1, \dots, s_\ell)\|_1 = |s_1| + \dots + |s_\ell|.$$

This corresponds to a mathematical model in which constraints of the form

$$(5.17) \quad C(t)x(t) + D(t)u(t) = q(t) \quad \text{a.e.}, \quad C_e x(t_1) + D_e u_e = q_e,$$

are to be enforced by linear penalties with parameter values $\lambda > 0$ and $\lambda_e > 0$ sufficiently high.

These ideas are useful in particular in penalty representations of endpoint constraints like the ones discussed in Examples 5.1 and 5.2. Thus a condition $x(t_1) = a_1$ can be modeled by a term $\lambda \|(x(t_1) - a_1)\|_1$ in the objective (the case of $C_e = I$, $D_e = 0$ and $q_e = a_1$ in (5.15) and (5.17)). A condition $x(t_0) = a_0$ corresponds of course to a trivial interval $U_e = [0, 0]$ and needs no penalty representation.

Example 5.5. (Intertemporal piecewise linear programming.) In the general case where $P(t) = 0$, $P_e = 0$, $Q(t) = 0$ and $Q_e = 0$, one minimizes in (\mathcal{P}) the objective

$$\begin{aligned} \mathcal{F}(u, u_e) = & \int_{t_0}^{t_1} [p(t) \cdot u(t) - c(t) \cdot x(t)] dt + [p_e \cdot u_e - c_e \cdot x(t_1)] \\ & + \int_{t_0}^{t_1} \sigma_{V(t)}(q(t) - C(t)x(t) - D(t)u(t)) dt + \sigma_{V_e}(q_e - C_e x(t_1) - D_e u_e) \end{aligned}$$

and one maximizes in (\mathcal{Q}) the objective

$$\begin{aligned} \mathcal{G}(v, v_e) = & \int_{t_0}^{t_1} [q(t) \cdot v(t) - b(t) \cdot y(t)] dt + [q_e \cdot v_e - b_e \cdot y(t_0)] \\ & - \int_{t_0}^{t_1} \sigma_{U(t)}(B^*(t)y(t) + D^*(t)v(t) - p(t)) dt - \sigma_{U_e}(B_e^* y(t_0) + D_e^* v_e - p_e), \end{aligned}$$

where the σ terms are support functions defined by (2.7) and are polyhedral convex (piecewise linear).

There are two different ways of using this general piecewise linear model, beyond those already covered in Examples 5.3 and 5.4, that deserve emphasis here. The first is in problems where the objective directly involves piecewise linear terms expressed as the pointwise maximum of finite collections of affine functions. This case corresponds to the patterns in Example 3.3 and need not be written out in detail. One has

$$V(t) = [\text{simplex in } \mathbf{R}^{\ell_1}] \times [\text{orthant or interval in } \mathbf{R}^{\ell_2}],$$

and similarly for V_e . Note that in taking in an interval for the second term in each product one has a case where $V(t)$ and V_e are both bounded, so *the primal finiteness condition is satisfied*.

The other way of using this model is less obvious but important in reaching formulations of intertemporal linear programming problems that satisfy the primal and dual finiteness conditions. As already noted in Example 5.3, those conditions are never fulfilled in the basic case of (\mathcal{P}_1) and (\mathcal{Q}_1) , but they can be brought to bear by passing to a bounded linear programming formulation as in Example 5.4. A more subtle approach is possible, however, in which only *some* of the constraints receive a linear penalty representation, namely those that definitely involve the state $x(t)$ (or $y(t)$). This might turn out to be a valuable consideration in the application of numerical methods for finding solutions.

For example, suppose we are dealing with a problem initially in the (\mathcal{P}_1) format but with constraints partitioned to clarify the involvement of $x(t)$:

minimize

$$\int_{t_0}^{t_1} [p(t) \cdot u(t) - c(t) \cdot x(t)] dt + [p_e \cdot u_e - c_e \cdot x(t_1)]$$

subject to

$$C_1(t)x(t) + D_1(t)u(t) \geq q_1(t),$$

$$D_2(t)u(t) \geq q_2(t), \quad u(t) \geq 0,$$

$$C_{e1}x(t_1) + D_{e1}u_e \geq q_{e1},$$

$$D_{e2}u_e \geq q_{e2}, \quad u_e \geq 0,$$

where $q_1(t) \in \mathbb{R}^{\ell_1}$, $q_2(t) \in \mathbb{R}^{\ell_2}$, $q_{e1} \in \mathbb{R}^{\ell_{e1}}$, $q_{e2} \in \mathbb{R}^{\ell_{e2}}$. The (\mathcal{P}_1) format corresponds to choosing

$$C(t) = \begin{bmatrix} C_1(t) \\ 0 \end{bmatrix}, \quad D(t) = \begin{bmatrix} D_1(t) \\ D_2(t) \end{bmatrix}, \quad C_e = \begin{bmatrix} C_{e1} \\ 0 \end{bmatrix}, \quad D_e = \begin{bmatrix} D_{e1} \\ D_{e2} \end{bmatrix},$$

$$U(t) = \mathbb{R}_+^k, \quad U_e = \mathbb{R}_+^{\ell_e}, \quad V(t) = \mathbb{R}_+^{\ell}, \quad V_e = \mathbb{R}_+^{\ell_e}$$

(with $\ell = \ell_1 + \ell_2$ and $\ell_e = \ell_{e1} + \ell_{e2}$). An alternative formulation, however, is to take

$$C(t) = C_1(t), \quad D(t) = D_1(t), \quad q(t) = q_1(t),$$

$$U(t) = \{u \geq 0 \mid D_2(t)u \geq q_2(t)\}, \quad V(t) = \mathbb{R}_+^{\ell_1},$$

$$C_e = C_{e1}, \quad D_e = D_{e1}, \quad q_e = q_{e1},$$

$$U_e = \{u_e \geq 0 \mid D_{e2}u_e \geq q_{e2}\}, \quad V_e = \mathbb{R}_+^{\ell_{e1}}.$$

If $U(t)$ and U_e happen to be bounded sets, we *have the dual boundedness condition satisfied* in this formulation even though it was not satisfied in the formulation as (\mathcal{P}_1) .

What effect does this alternative have on the nature of the dual problem? One maximizes the expression

$$\begin{aligned} & \int_{t_0}^{t_1} [q_1(t) \cdot v_1(t) - b(t) \cdot y(t)] dt + [q_{e1} \cdot v_{e1} - b_e \cdot y(t_0)] \\ & - \int_{t_0}^{t_1} \sigma_{U(t)}(B^*(t)y(t) + D_1^*(t)v_1(t) - p(t)) dt - \sigma_{U_e}(B_e^*y(t_0) + D_{e1}^*v_{e1} - p_e) \end{aligned}$$

subject to $v_1(t) \in \mathbb{R}_+^{k_1}$ and $v_{e1} \in \mathbb{R}_+^{\ell_{e1}}$. One has

$$\begin{aligned} -\sigma_{U(t)}(r) &:= -\sup\{r \cdot u \mid u \geq 0, D_2(t)u \geq q_2(t)\} \\ &= \inf\{-r \cdot u \mid u \geq 0, D_2(t)u \geq q_2(t)\} \\ &= \sup\{q_2(t) \cdot v_2 \mid v_2 \geq 0, D_2^*(t)v_2 \leq -r\} \end{aligned}$$

by finite-dimensional linear programming duality, so that

$$\begin{aligned} & -\sigma_{U(t)}(B^*(t)y(t) + D_1^*(t)v_1(t) - p(t)) \\ & = \sup\{q_2(t) \cdot v_2 \mid v_2 \geq 0, B^*(t)y(t_0) + D_1^*(t)v_1(t) + D_2^*(t)v_2 \leq p(t)\}. \end{aligned}$$

Similarly

$$\begin{aligned} & -\sigma_{U_e}(B_e^*y(t_0) + D_{e1}^*v_{e1} - p_e) \\ & = \sup\{q_{e2} \cdot v_{e2} \mid v_{e2} \geq 0, B_e^*y(t) + D_{e1}^*v_{e1} + D_{e2}^*v_{e2} \leq p_e\}. \end{aligned}$$

The dual problem for the alternative approach is therefore essentially the same as (\mathcal{Q}_1) , except that the v_2 and v_{e2} components in \mathbf{R}^{ℓ_2} and $\mathbf{R}^{\ell_{e2}}$ have been “maximized out.” These components can ultimately be recovered if necessary, but in the meantime we do not need to worry about them in connection with theorems about optimality conditions, existence and duality, in particular the \mathcal{L}^∞ requirement on $v(t)$ in (\mathcal{Q}) .

Of course, in order for this approach to work, we must also be able to verify the assumption of continuous dependence of $U(t)$ on t . When $U(t) = \{u \geq 0 \mid D_2(t)u \geq q_2(t)\}$, this is satisfied for instance if $D_2(t)$ and $q_2(t)$ do not actually depend on t , or if there is a *continuous* function u such that $u(t) \geq 0$ and $D_2(t)u(t) > q_2(t)$ (strict inequality in every component). For the latter and also more general cases involving a possible mixture of equality and inequality constraints, see Rockafellar [48, Cor. 3.3].

Similar ideas can be applied to a partitioning of the constraints of a problem (\mathcal{Q}_1) into those that affect $y(t)$ and those that do not. In (\mathcal{P}_1) this would correspond to dynamics $\dot{x} = Ax + Bu + b$, $x(t_0) = B_e u_e + b_e$, where $B = [B_1, 0]$ and $B_e = [B_{e1}, 0]$, i.e. not all components of u and u_e are directly active in the dynamics.

Example 5.6. (Linear-quadratic regulator problem and generalizations.) Consider now a classical type of problem having the form

$$\begin{aligned} & \text{minimize} \quad \frac{1}{2} \int_{t_0}^{t_1} [u(t) \cdot P(t)u(t) + (x(t) - \hat{x}(t)) \cdot R(t)(x(t) - \hat{x}(t))] dt \\ & \quad + \frac{1}{2} (x(t_1) - a_1) \cdot R_e(x(t_1) - a_1) \\ & \text{subject to} \quad u \in \mathcal{L}^\infty([t_0, t_1], \mathbf{R}^k), \\ & \quad \dot{x}(t) = A(t)x(t) + B(t)u(t) + b(t) \text{ a.e.}, \quad x(t_0) = a_0, \end{aligned} \tag{5.18}$$

where a_0 and a_1 are given points, \hat{x} is a given function (continuous), $P(t)$ is positive *definite*, and $R(t)$ and R_e are positive *semidefinite*. This can be formulated as a problem (\mathcal{P}) by introducing factorizations

$$R(t) = C^*(t)Q(t)^{-1}C(t) \quad \text{and} \quad R_e = C_e^*Q_e^{-1}C_e, \tag{5.19}$$

where $Q(t)$ and Q_e are positive *definite*. (If $R(t)$ and R_e themselves are positive definite, one can of course take $C(t) = I$, $Q(t) = R(t)^{-1}$, $C_e = I$, $Q_e = R_e^{-1}$, in (5.19).) Set

$$p(t) = 0, \quad c(t) = 0, \quad D(t) = 0, \quad q(t) = C(t)\hat{x}(t), \quad U(t) = \mathbf{R}^k, \quad V(t) = \mathbf{R}^\ell.$$

Then in the general format of (\mathcal{P}) the terms

$$p(t) \cdot u(t) \cdot P(t)u(t) - c(t) \cdot x(t) + \rho_{V(t), Q(t)}(q(t) - C(t)x(t) - D(t)u(t))$$

reduce to

$$\frac{1}{2}u(t) \cdot P(t)u(t) + \frac{1}{2}(x(t) - \hat{x}(t)) \cdot R(t)(x(t) - \hat{x}(t)).$$

For the endpoints, trivialize u_e by taking \mathbf{R}^{k_e} to be zero-dimensional (so $U_e = \{0\}$, $p_e = 0$, $P_e = 0$, $B_e = 0$, $D_e = 0$) and let $b_e = a_0$, $q_e = C_e a_1$, $c_e = 0$. The terms

$$p_e \cdot u_e + \frac{1}{2} u_e \cdot P_e u_e - c_e \cdot x(t_1) + \rho_{v_e, Q_e}(q_e - C_e x(t_1) - D_e u_e)$$

in (\mathcal{P}) then reduce to

$$\frac{1}{2}(x(t_1) - a_1) \cdot R_e(x(t_1) - a_1),$$

and we get the desired problem (5.18) as a special case of (\mathcal{P}) . The corresponding dual (\mathcal{Q}) has the form

$$\begin{aligned} (5.20) \quad & \text{minimize} \quad \int_{t_0}^{t_1} [\hat{x}(t) \cdot C^*(t)v(t) - b(t) \cdot y(t)]dt + [a_1 \cdot C_e^* v_e - a_0 \cdot y(t_0)] \\ & \quad - \frac{1}{2} \int_{t_0}^{t_1} [v(t) \cdot Q(t)v(t) + y(t) \cdot S(t)y(t)]dt - \frac{1}{2} v_e \cdot Q_e v_e \\ & \text{subject to} \quad v \in \mathcal{L}^\infty([t_0, t_1], \mathbf{R}^\ell), \quad v_e \in \mathbf{R}^{\ell_e}, \\ & \quad -\dot{y}(t) = A^*(t)y(t) + C^*(t)v(t) \text{ a.e.}, \quad y(t_1) = C_e^* v_e, \end{aligned}$$

where

$$(5.21) \quad S(t) = B(t)P(t)^{-1}B^*(t).$$

Note that in this example *the primal and dual boundedness conditions are both satisfied*.

Generalizations of the linear-quadratic regulator problem can be made in several directions without going beyond the format of our problem (\mathcal{P}) . For instance, instead of letting $u(t)$ be a free vector in \mathbf{R}^k one can insist on bounds $\hat{u}^-(t) \leq u(t) \leq \hat{u}^+(t)$. Dually one can introduce bounds

$$-\lambda \mathbf{1} \leq v(t) \leq \lambda \mathbf{1} \quad \text{and} \quad -\lambda_e \mathbf{1} \leq v_e \leq \lambda_e \mathbf{1}$$

for parameter values $\lambda > 0$, $\lambda_e > 0$. The effect of this on the formulation of the original problem (5.18) is to replace the purely quadratic penalty expressions by terms that are quadratic near the origin but eventually grow at a linear rate. Thus for example if

$$R(t) = \mu I \quad \text{and} \quad R_e = \mu_e I \quad \text{for} \quad \mu > 0, \mu_e > 0$$

(corresponding in (5.19) to $C(t) = I$, $Q(t) = \mu^{-1}I$, $C_e = I$, $Q_e = \mu_e^{-1}I$) one has terms

$$(\mu/2) \int_{t_0}^{t_1} |x(t) - \hat{x}(t)|^2 dt + (\mu_e/2) |x(t_1) - a_1|^2$$

in (5.18) that are replaced by

$$\sum_{i=1}^n \left[\int_{t_0}^{t_1} \psi(|x_i(t) - \hat{x}_i(t)|) dt + \psi_e(|x_i(t_1) - a_{1i}|) \right],$$

where $x_i(t)$ and a_{1i} are the i th components of $x(t)$ and a_1 and ψ is the growth function defined by

$$\psi(\tau) = \begin{cases} (\mu/2)\tau^2 & \text{when } 0 \leq \tau \leq \lambda/\mu, \\ \lambda(\tau - (\lambda/\mu)) + (\lambda^2/2\mu) & \text{when } \tau \geq \lambda/\mu, \end{cases}$$

and similarly ψ_e in terms of λ_e and μ_e .

Still other generalizations of the linear-quadratic regulator problem are covered by the patterns in the next example.

Example 5.7. (Bounded intertemporal quadratic programming.) This corresponds to the finite-dimensional bounded quadratic programming models in Example 3.5 in the same way that Example 5.4 corresponds to the finite-dimensional bounded linear programming models in Example 3.4. Due to all the notation involved, we shall not write these problems out in full. The point is, however, that these are formulations of considerable versatility which allow for quadratic terms without damaging the explicit, symmetric nature of the dualization.

Example 5.8. (Problems whose duals are essentially finite-dimensional.) Suppose in problem (\mathcal{P}) that $C(t) \equiv 0$. Then in (\mathcal{Q}) the trajectory y is uniquely determined from v_e alone. Although $v(t)$ still appears in the objective in (\mathcal{Q}) , it does so in a very simple way: the value chosen for $v(t)$ has no connection to past or future. At each time t one can just take $v(t)$ to maximize the expression

$$q(t) \cdot v(t) - \frac{1}{2}v(t) \cdot Q(t)v(t) - b(t) \cdot y(t) - \rho_{V(t), Q(t)}(B^*(t)y(t) + D^*(t)v(t) - p(t))$$

over $V(t)$, where $y(t)$ is already fixed. In this sense (\mathcal{Q}) is really a problem in v_e alone and is therefore finite-dimensional. (Of course $v(t)$ must ultimately be an \mathcal{L}^1 function of t .)

6. Saddle points and optimality. The duality between problems (\mathcal{P}) and (\mathcal{Q}) will be established by associating them with an infinite-dimensional saddle point problem. This will lead to the principal results of this paper, which concern the existence and optimality properties of solutions to (\mathcal{P}) and (\mathcal{Q}) .

The saddle point representation we aim at follows the general guidelines at the beginning of §2. We take the control spaces \mathcal{U} and \mathcal{V} already introduced in §4 (which are nonempty by Theorem 4.2) and define on $\mathcal{U} \times \mathcal{V}$ a certain functional J , namely

$$(6.1) \quad J(u, u_e; v, v_e) = \int_{t_0}^{t_1} J(t, u(t), v(t))dt + J_e(u_e, v_e) - [(u, u_e), (v, v_e)]$$

under the convention $\infty - \infty = \infty$ (see below), where

$$(6.2) \quad J(t, u, v) = p(t) \cdot u + q(t) \cdot v + \frac{1}{2}u \cdot P(t)u - \frac{1}{2}v \cdot Q(t)v - v \cdot D(t)u,$$

$$(6.3) \quad J_e(u_e, v_e) = p_e \cdot u_e + q_e \cdot v_e + \frac{1}{2}u_e \cdot P_e u_e - \frac{1}{2}v_e \cdot Q_e v_e - v_e \cdot D_e u_e,$$

and

$$(6.4) \quad \begin{aligned} [(u, u_e), (v, v_e)] &= \int_{t_0}^{t_1} y(t) \cdot [B(t)u(t) + b(t)]dt + y(t_0) \cdot [B_e u_e + b_e] \\ &= \int_{t_0}^{t_1} x(t) \cdot [C^*(t)v(t) + c(t)]dt + x(t_1) \cdot [C_e^* v_e + c_e]. \end{aligned}$$

The common value of the two expressions for $[(u, u_e), (v, v_e)]$ in (6.4) stems from the integration-by-parts formula

$$\int_{t_0}^{t_1} y(t) \cdot \dot{x}(t)dt + y(t_0) \cdot x(t_0) = - \int_{t_0}^{t_1} x(t) \cdot \dot{y}(t)dt + x(t_1) \cdot y(t_1).$$

The term $[(u, u_e), (v, v_e)]$, which is affine in (u, u_e) for fixed (v, v_e) and affine in (v, v_e) for fixed (u, u_e) , as well as continuous with respect to all arguments, embodies the fundamental connection between the control systems in (\mathcal{P}) and (\mathcal{Q}) .

The convention $\infty - \infty = \infty$ mentioned in the definition (6.1) of J refers to possible ambiguities in the value of the integral of $J(t, u(t), v(t))$. In general, since

$u(t)$ and $v(t)$ are only \mathcal{L}^1 in t , the integral of the term $u(t) \cdot P(t)u(t)$ might be ∞ , the integral of $v(t) \cdot Q(t)v(t)$ might be $-\infty$, and the integral of $v(t) \cdot D(t)u(t)$ might be either. We use $\infty - \infty = \infty$ to resolve any dilemmas in extended arithmetic that might arise. This amounts to taking the integral term in (6.1) to be ∞ if $J(t, u(t), v(t))$ is not majorized by any \mathcal{L}^1 function of t . Of course if $J(t, u(t), v(t)) \leq \alpha(t)$ for an \mathcal{L}^1 function α , then the integral has an unambiguous value which is finite or $-\infty$, whereas if $J(t, u(t), v(t)) \geq \beta(t)$ for an \mathcal{L}^1 function β , it is finite or ∞ . Actually there is no difficulty at all if $u \in \mathcal{L}^\infty$ or $v \in \mathcal{L}^\infty$: one has

$$(6.5) \quad J(u, u_e; v, v_e) < \infty \quad \text{when} \quad u \in \mathcal{L}^\infty,$$

$$(6.6) \quad J(u, u_e; v, v_e) > -\infty \quad \text{when} \quad v \in \mathcal{L}^\infty,$$

and therefore $J(u, u_e; v, v_e)$ finite when both $u \in \mathcal{L}^\infty$ and $v \in \mathcal{L}^\infty$.

Anyway, under the specified convention J is a well-defined functional on $\mathcal{U} \times \mathcal{V}$ which is quadratic convex in (u, u_e) and quadratic concave in (v, v_e) . The convention $\infty - \infty = -\infty$ could have been used instead and would have led to a functional \tilde{J} that would serve our purposes in equivalent fashion; we shall occasionally make use of \tilde{J} in our proofs. Obviously from (6.5) and (6.6), J and \tilde{J} agree whenever $u \in \mathcal{L}^\infty$ or $v \in \mathcal{L}^\infty$.

THEOREM 6.1. *Problems (\mathcal{P}) and (\mathcal{Q}) are the primal and dual optimization problems associated with the saddle point problem for J on $\mathcal{U} \times \mathcal{V}$. Thus the functional \mathcal{F} which in (\mathcal{P}) is minimized over \mathcal{U} is given by*

$$(6.7) \quad \mathcal{F}(u, u_e) = \sup_{(v, v_e) \in \mathcal{V}} J(u, u_e; v, v_e),$$

whereas the functional \mathcal{G} which in (\mathcal{Q}) is maximized over \mathcal{V} is given by

$$(6.8) \quad \mathcal{G}(v, v_e) = \inf_{(u, u_e) \in \mathcal{U}} J(u, u_e; v, v_e).$$

Proof. In establishing (6.7) we take the second of the expressions in (6.4) for the term $[(u, u_e), (v, v_e)]$ in the definition (6.1) of J , so that

$$(6.9) \quad \begin{aligned} J(u, u_e; v, v_e) = & \int_{t_0}^{t_1} [p(t) \cdot u(t) + \tfrac{1}{2}u(t) \cdot P(t)u(t) - c(t) \cdot x(t)]dt \\ & + \int_{t_0}^{t_1} (v(t) \cdot [q(t) - C(t)x(t) - D(t)u(t)] - \tfrac{1}{2}v(t) \cdot Q(t)v(t))dt \\ & + [p_e \cdot u_e + \tfrac{1}{2}u_e \cdot P_e u_e - c_e \cdot x(t_1)] \\ & + v_e \cdot [q_e - C_e x(t_1) - D_e u_e] - \tfrac{1}{2}v_e \cdot Q_e v_e. \end{aligned}$$

From the definition (2.5) of the functions $\rho_{V(t), Q(t)}$ and ρ_{V_e, Q_e} it is clear that

$$(6.10) \quad \mathcal{F}(u, u_e) \geq J(u, u_e; v, v_e) \quad \text{for all } (u, u_e) \in \mathcal{U}, (v, v_e) \in \mathcal{V},$$

and that the desired equation (6.7) can be verified by showing that the equation

$$(6.11) \quad \sup_{\substack{v \in \mathcal{L}^\infty \\ v(t) \in V(t)}} \int_{t_0}^{t_1} [v(t) \cdot s(t) - \tfrac{1}{2}v(t) \cdot Q(t)v(t)]dt = \int_{t_0}^{t_1} \rho_{V(t), Q(t)}(s(t))dt$$

holds for arbitrary $s \in \mathcal{L}^1$. This equation can be written as

$$(6.12) \quad \sup_{v \in \mathcal{L}^\infty} \int_{t_0}^{t_1} [v(t) \cdot s(t) - \varphi_t(v(t))]dt = \int_{t_0}^{t_1} \varphi_t^*(s(t))dt$$

for the convex function $\varphi_t(v) = j_{Q(t)}(v) + \delta_{V(t)}(v)$ utilized in the proofs of Proposition 2.3, 2.4, and 4.1. It holds by [49, Thm. 2] (or [50, Thm. 3C]) if $\varphi(t, v) = \varphi_t(v)$ is a so-called *normal integrand* and the left side of (6.12) is not $-\infty$. Actually $\varphi(t, v)$ is lower semicontinuous jointly in t and v , inasmuch as $Q(t)$ and $V(t)$ depend continuously on t , whereas normality merely requires $\varphi(t, v)$ to be lower semicontinuous in v for fixed t and measurable in (t, v) with respect to the σ -algebra in $[t_0, t_1] \times \mathbb{R}^\ell$ generated by the Lebesgue sets in $[t_0, t_1]$ and the Borel sets in \mathbb{R}^ℓ [50, Thm. 24]. Thus φ is normal. Furthermore the left side of (6.12), or equivalently of (6.11), cannot be $-\infty$, because the integral is finite when $v \in \mathcal{L}^\infty$, and we do know (from the proof of Theorem 4.2) that \mathcal{V} contains at least one pair (v, v_e) with v actually continuous.

Our argument has not only verified (6.7) but shown that the same would be true if J were replaced by the alternative functional \tilde{J} using $\infty - \infty = -\infty$ instead of $\infty - \infty = \infty$. Indeed, (6.10) still holds for \tilde{J} , since $J \geq \tilde{J}$. Everything else is unchanged, because we relied only on $v \in \mathcal{L}^\infty$, and for such v the values of J and \tilde{J} agree. This symmetry is all we need to conclude that (6.8) is valid too. \square

THEOREM 6.2 (Weak Duality). *For the optimal control problems (\mathcal{P}) and (\mathcal{Q}) it is always true that*

$$\inf(\mathcal{P}) \geq \sup(\mathcal{Q}).$$

Furthermore a pair $((\bar{u}, \bar{u}_e), (\bar{v}, \bar{v}_e))$ is a saddle point of J on $\mathcal{U} \times \mathcal{V}$ if and only if (\bar{u}, \bar{u}_e) solves (\mathcal{P}) , (\bar{v}, \bar{v}_e) solves (\mathcal{Q}) , and $\min(\mathcal{P}) = \max(\mathcal{Q})$ (finite).

Proof. This is just a repeat of the general facts in Proposition 2.1 for the specific case in Theorem 6.1. \square

A stronger result is obtained by appealing to the *finiteness conditions* for (\mathcal{P}) and (\mathcal{Q}) that were introduced at the end of §4. We wish to emphasize again, as in §1, that this is by no means the most general result on strong duality. Rather, it is presented as a relatively simple result which is easy to work with and already capable of covering many important cases, especially in view of the modeling possibilities explained in §5.

THEOREM 6.3 (Strong Duality). *If the primal finiteness condition is satisfied, then*

$$(6.13) \quad \inf(\mathcal{P}) = \max(\mathcal{Q}) < \infty,$$

and moreover the dual objective \mathcal{G} is weakly sup-compact relative to \mathcal{V} , i.e. all level sets of the form

$$(6.14) \quad \{(v, v_e) \in \mathcal{V} \mid \mathcal{G}(v, v_e) \geq \alpha\} \quad \text{for } \alpha \in \mathbb{R}$$

are weakly compact in $\mathcal{L}^1([t_0, t_1], \mathbb{R}^\ell) \times \mathbb{R}^{\ell_e}$.

Likewise, if the dual finiteness condition is satisfied, then

$$(6.15) \quad \min(\mathcal{P}) = \sup(\mathcal{Q}) > -\infty,$$

and moreover the primal objective (\mathcal{P}) is weakly inf-compact relative to \mathcal{U} , i.e. all level sets of the form

$$(6.16) \quad \{(u, u_e) \in \mathcal{U} \mid \mathcal{F}(u, u_e) \leq \alpha\} \quad \text{for } \alpha \in \mathbb{R}$$

are weakly compact in $\mathcal{L}^1([t_0, t_1], \mathbb{R}^k) \times \mathbb{R}^{k_e}$.

Thus if both finiteness conditions are satisfied, solutions exist to both (\mathcal{P}) and (\mathcal{Q}) , and

$$(6.17) \quad \min(\mathcal{P}) = \max(\mathcal{Q}) \quad (\text{finite}).$$

Proof. Our proof of the formulas (6.7) and (6.8) in Theorem 6.1 gave something slightly stronger that will now be of use: if we denote by \mathcal{U}^∞ and \mathcal{V}^∞ the subsets of \mathcal{U} and \mathcal{V} having $u \in \mathcal{L}^\infty$ and $v \in \mathcal{L}^\infty$, then

$$(6.18) \quad \mathcal{F}(u, u_e) = \sup_{(v, v_e) \in \mathcal{V}^\infty} J(u, u_e; v, v_e) \quad \text{for all } (u, u_e),$$

$$(6.19) \quad \mathcal{G}(v, v_e) = \inf_{(u, u_e) \in \mathcal{U}^\infty} J(u, u_e; v, v_e) \quad \text{for all } (v, v_e).$$

In order to obtain (6.13) it will be enough by this to demonstrate

$$(6.20) \quad \inf_{\mathcal{U}^\infty} \sup_{\mathcal{V}} J = \max_{\mathcal{V}} \inf_{\mathcal{U}^\infty} J,$$

since the inequalities

$$\inf_{\mathcal{U}^\infty} \sup_{\mathcal{V}} J \geq \inf_{\mathcal{U}} \sup_{\mathcal{V}} J \geq \sup_{\mathcal{V}} \inf_{\mathcal{U}} J$$

hold trivially. The one-sided minimax theorem of Moreau [51] will justify (6.20) provided we can show that under the primal finiteness condition $J(u, u_e; v, v_e)$ is weakly sup-compact in (v, v_e) relative to \mathcal{V} when $(u, u_e) \in \mathcal{U}^\infty$. The latter will also give us the claimed sup-compactness of \mathcal{G} via (6.19).

Fix $(u, u_e) \in \mathcal{U}^\infty$. Taking J as expressed in (6.9) and introducing $s(t)$ and s_e as in (4.5), we have

$$(6.21) \quad J(u, u_e; v, v_e) = \int_{t_0}^{t_1} [v(t) \cdot s(t) - \frac{1}{2} v(t) \cdot Q(t) v(t)] dt + [v_e \cdot s_e - \frac{1}{2} v_e \cdot Q_e v_e] + \text{const.}$$

for all $(v, v_e) \in \mathcal{V}$, where $s(t)$ is \mathcal{L}^∞ in t . The required sup-compactness property of J is the weak compactness of the level sets

$$\{(v, v_e) \in \mathcal{V} \mid J(u, u_e; v, v_e) \geq \alpha\} \quad \text{for } \alpha \in \mathbb{R}.$$

We recognize now that this is the same as the weak compactness of the level sets

$$(6.22) \quad \{(v, v_e) \in \mathcal{V} \mid \frac{1}{2} \int_{t_0}^{t_1} v(t) \cdot Q(t) v(t) dt + \frac{1}{2} v_e \cdot Q_e v_e - \langle (v, v_e), (s, s_e) \rangle \leq \beta\}$$

for $\beta \in \mathbb{R}$, where

$$(6.23) \quad \langle (v, v_e), (s, s_e) \rangle = \int_{t_0}^{t_1} v(t) \cdot s(t) dt + v_e \cdot s_e.$$

Once again the convex function

$$\varphi_t(v) = j_{Q(t)}(v) + \delta_{V(t)}(v) = \begin{cases} \frac{1}{2} v \cdot Q(t) v & \text{if } v \in V(t), \\ \infty & \text{if } v \notin V(t) \end{cases}$$

will be useful, together with

$$\varphi_e(v_e) = j_{Q_e}(v_e) + \delta_{V_e}(v_e) = \begin{cases} \frac{1}{2} v_e \cdot Q_e v_e & \text{if } v_e \in V_e, \\ \infty & \text{if } v_e \notin V_e. \end{cases}$$

The convex functional

$$I(v, v_e) = \int_{t_0}^{t_1} \varphi_t(v(t)) dt + \varphi_e(v_e)$$

is well defined on $\mathcal{L}^1([t_0, t_1], \mathbb{R}^\ell) \times \mathbb{R}^{\ell_e}$ with values in $[0, \infty)$, and in terms of it the set (6.22) can be written as

$$(6.24) \quad \{(v, v_e) \in \mathcal{L}^1([t_0, t_1], \mathbb{R}^\ell) \times \mathbb{R}^{\ell_e} \mid I(v, v_e) - \langle (v, v_e), (s, s_e) \rangle \leq \beta\}.$$

We shall be able to establish the weak compactness of this set for arbitrary $(s, s_e) \in \mathcal{L}^\infty([t_0, t_1], \mathbf{R}^\ell) \times \mathbf{R}^{\ell_e}$ and $\beta \in R$ by means of the theory of integral functional conjugate to each other [49], [50].

Let us think of the spaces $\mathcal{L}^1([t_0, t_1], \mathbf{R}^\ell) \times \mathbf{R}^{\ell_e}$ and $\mathcal{L}^\infty([t_0, t_1], \mathbf{R}^\ell) \times \mathbf{R}^{\ell_e}$ as dual to each other under the pairing (6.23). The pairing formula and the formula for I can actually be viewed as integrals over a measure space that is the union of $[t_0, t_1]$ and an atom $\{e\}$ of measure 1. In this sense I is an integral functional, pure and simple. The functional

$$I^*(s, s_e) = \int_{t_0}^{t_1} \varphi_t^*(s(t)) dt + \varphi_e^*(s_e)$$

on $\mathcal{L}^\infty([t_0, t_1], \mathbf{R}^\ell) \times \mathbf{R}^{\ell_e}$, where φ_t^* and φ_e^* are conjugate to φ_t and φ_e , is an integral functional too, and I and I^* are conjugate to each other by [49, Thm. 2] (or [50, Thm. 3C]) with respect to the pairing (6.23). Indeed $\varphi_t^* = \rho_{V(t), Q(t)}$ and $\varphi_e^* = \rho_{V_e, Q_e}$, so φ_t^* and φ_e^* are finite convex functions on \mathbf{R}^ℓ under the primal finiteness condition we are assuming. Furthermore $\varphi_t^*(s)$ is for each $s \in \mathbf{R}^\ell$ continuous in t by Proposition 4.1, hence integrable over $[t_0, t_1]$. These properties for I^* plug into the weak inf-compactness criterion of [49, p. 538] for integral functional on \mathcal{L}^1 -type spaces and prove the required weak compactness of all level sets of the form (6.24) for the conjugate functional $I = (I^*)^*$.

The proof of (6.15) and the weak compactness of the sets (6.16) follows now by symmetry. \square

COROLLARY 6.4. *Suppose the primal and dual finiteness conditions both hold. Then in order that (\bar{u}, \bar{u}_e) solve (\mathcal{P}) and (\bar{v}, \bar{v}_e) solve (\mathcal{Q}) , it is both necessary and sufficient that $(\bar{u}, \bar{u}_e), (\bar{v}, \bar{v}_e)$ be a saddle point of J on $\mathcal{U} \times \mathcal{V}$.*

Proof. According to Theorem 6.2 the saddle point condition is always sufficient, and if $\min(\mathcal{P}) = \max(\mathcal{Q})$ it is also necessary. Necessity therefore follows from the primal and dual finiteness conditions by the result just proved in Theorem 6.3. \square

The saddle point condition in Corollary 6.4 means that $(\bar{u}, \bar{u}_e) \in \mathcal{U}, (\bar{v}, \bar{v}_e) \in \mathcal{V}$, and

$$(6.25) \quad J(\bar{u}, \bar{u}_e; v, v_e) \leq J(\bar{u}, \bar{u}_e; \bar{v}, \bar{v}_e) \leq J(u, u_e; \bar{v}, \bar{v}_e)$$

for all $(u, u_e) \in \mathcal{U}$ and $(v, v_e) \in \mathcal{V}$. This “global” condition actually decomposes, as we show next, into an “instantaneous” saddle point condition at each time t and an “endpoint” saddle point condition.

THEOREM 6.5 (Minimaximum Principle). *For $((\bar{u}, \bar{u}_e), (\bar{v}, \bar{v}_e))$ to be a saddle point of J on $\mathcal{U} \times \mathcal{V}$, it is necessary and sufficient that the following conditions hold (in addition to $\bar{u}(t)$ and $\bar{v}(t)$ being \mathcal{L}^1 in t). For almost every $t \in [t_0, t_1]$*

$$(6.26) \quad (\bar{u}(t), \bar{v}(t)) \text{ is a saddlepoint relative to } U(t) \times V(t) \text{ for } J(t, u, v) - u \cdot B^*(t)\bar{y}(t) - v \cdot C(t)\bar{x}(t),$$

and also

$$(6.27) \quad (\bar{u}_e, \bar{v}_e) \text{ is a saddlepoint relative to } U_e \times V_e \text{ for } J_e(u_e, v_e) - u_e \cdot B_e^*\bar{y}(t_0) - v_e \cdot C_e\bar{x}(t_1),$$

where \bar{x} and \bar{y} are the primal and dual state functions corresponding to (\bar{u}, \bar{u}_e) and (\bar{v}, \bar{v}_e) .

Proof. The saddle point condition (6.25) for J on $\mathcal{U} \times \mathcal{V}$ is equivalent by Theorem 6.1 to the condition

$$(6.28) \quad \mathcal{F}(\bar{u}, \bar{u}_e) = J(\bar{u}, \bar{u}_e; \bar{v}, \bar{v}_e) = \mathcal{G}(\bar{v}, \bar{v}_e).$$

Let us write this as

$$(6.29) \quad \mathcal{F}(\bar{u}, \bar{u}_e) + \bar{\alpha} = J(\bar{u}, \bar{u}_e; \bar{v}, \bar{v}_e) + \bar{\alpha} = \mathcal{G}(\bar{v}, \bar{v}_e) + \bar{\alpha},$$

where

$$\bar{\alpha} = \int_{t_0}^{t_1} [c(t) \cdot \bar{x}(t) + b(t) \cdot \bar{y}(t)] dt + c_e \cdot \bar{x}(t_1) + b_e \cdot \bar{y}(t_0) - [(\bar{u}, \bar{u}_e), (\bar{v}, \bar{v}_e)].$$

The alternative expressions for $[(\bar{u}, \bar{u}_e), (\bar{v}, \bar{v}_e)]$ in (6.4) give

$$\begin{aligned} \bar{\alpha} &= \int_{t_0}^{t_1} [c(t) \cdot \bar{x}(t) - \bar{u}(t) \cdot B^*(t) \bar{y}(t)] dt + [c_e \cdot \bar{x}(t_1) - \bar{u}_e \cdot B_e^* \bar{y}(t_0)] \\ &= \int_{t_0}^{t_1} [b(t) \cdot \bar{y}(t) - \bar{v}(t) \cdot C(t) \bar{x}(t)] dt + [b_e \cdot \bar{y}(t_0) - \bar{v}_e \cdot C_e \bar{x}(t_1)]. \end{aligned}$$

Using these along with the formulas defining \mathcal{F}, \mathcal{G} , and J , we get expressions of the form

$$\begin{aligned} \mathcal{F}(\bar{u}, \bar{u}_e) &= \int_{t_0}^{t_1} \bar{f}_t(\bar{u}(t)) dt + \bar{f}_e(\bar{u}_e), \\ \mathcal{G}(\bar{v}, \bar{v}_e) + \bar{\alpha} &= \int_{t_0}^{t_1} \bar{g}_t(\bar{v}(t)) dt + \bar{g}_e(\bar{v}_e), \\ J(\bar{u}, \bar{u}_e; \bar{v}, \bar{v}_e) + \alpha &= \int_{t_0}^{t_1} \bar{J}_t(\bar{u}(t), \bar{v}(t)) dt + \bar{J}_e(u_e, v_e), \end{aligned}$$

where

$$(6.30) \quad \bar{f}_t(u) = [p(t) - B^*(t) \bar{y}(t)] \cdot u + \frac{1}{2} u \cdot P(t) u + \rho_{V(t), Q(t)}(q(t) - C(t) \bar{x}(t) - D(t) u),$$

$$(6.31) \quad \bar{f}_e(u_e) = [p_e - B_e^* \bar{y}(t_0)] \cdot u + \frac{1}{2} u_e \cdot P_e u_e + \rho_{V_e, Q_e}(q_e - C_e \bar{x}(t_1) - D_e u_e),$$

$$(6.32) \quad \bar{g}_t(v) = [q(t) - C(t) \bar{x}(t)] \cdot v - \frac{1}{2} v \cdot Q(t) v - \rho_{V(t), P(t)}(B^*(t) \bar{y}(t) + D^*(t) v - p(t)),$$

$$(6.33) \quad \bar{g}_e(v_e) = [q_e - C_e \bar{x}(t_1)] \cdot v_e - \frac{1}{2} v_e \cdot Q_e v_e - \rho_{V_e, P_e}(B_e^* \bar{y}(t_0) + D_e^* v_e - p_e),$$

$$(6.34) \quad \bar{J}_t(u, v) = J(t, u, v) - u \cdot B^*(t) \bar{y}(t) - v \cdot C(t) \bar{x}(t),$$

$$(6.35) \quad \bar{J}_e(u_e, v_e) = J_e(u_e, v_e) - u_e \cdot B_e^* \bar{y}(t_0) - v_e \cdot C_e \bar{x}(t_1).$$

The saddle point condition on $((\bar{u}, \bar{u}_e), (\bar{v}, \bar{v}_e))$, written as (6.27), is equivalent under this formulation to

$$\begin{aligned} (6.36) \quad \int_{t_0}^{t_1} \bar{f}_t(\bar{u}(t)) dt + \bar{f}_e(\bar{u}_e) &= \int_{t_0}^{t_1} \bar{J}_t(\bar{u}(t), \bar{v}(t)) dt + \bar{J}_e(\bar{u}_e, \bar{v}_e) \\ &= \int_{t_0}^{t_1} \bar{g}_t(\bar{v}(t)) dt + \bar{g}_e(\bar{v}_e). \end{aligned}$$

But

$$(6.37a) \quad \bar{f}_t(u) = \sup_{v \in V(t)} \bar{J}_t(u, v), \quad \bar{g}_t(v) = \inf_{u \in U(t)} \bar{J}_t(u, v),$$

$$(6.37b) \quad \bar{f}_e(u_e) = \inf_{v \in V_e} \bar{J}_e(u_e, v_e), \quad \bar{g}_e(v_e) = \sup_{u \in U_e} \bar{J}_e(u_e, v_e),$$

by the definition of the ρ terms in (6.30) – (6.33), so

$$\begin{aligned} f_t(u) &\geq J_t(u, v) \geq g_t(v) \quad \text{for all } u \in U(t), v \in V(t), \\ f_e(u_e) &\geq J_e(u_e, v_e) \geq g_e(v_e) \quad \text{for all } u_e \in U_e, v_e \in V_e. \end{aligned}$$

Since the left side of (6.36) cannot be $-\infty$, whereas the right side cannot be ∞ (from the corresponding facts about $\mathcal{F}(u, u_e)$ and $\mathcal{G}(v, v_e)$ in Theorem 6.1), condition (6.36) holds if and only if

$$\bar{f}_t(\bar{u}(t)) = \bar{J}_t(\bar{u}(t), \bar{v}(t)) = \bar{g}_t(\bar{v}(t)) \quad \text{a.e.,} \quad f_e(\bar{u}_e) = \bar{J}_e(\bar{u}_e, \bar{v}_e) = \bar{g}_e(\bar{v}_e).$$

In view of (6.37a) and (6.37b) these are precisely the “instantaneous” and “endpoint” saddle point conditions asserted in the theorem. \square

Theorem 6.5 has an interesting interpretation in the context of the finite-dimensional linear-quadratic programming problems in §2, as revealed by its proof. We shall formulate this as a corollary.

Corresponding to the trajectories \bar{x} and \bar{y} , consider the “instantaneous” primal and dual problems associated with the linear-quadratic form $\bar{J}_t(u, v)$ on $U(t) \times V(t)$, where \bar{J}_t is given by (6.34), namely:

$$\begin{aligned} (\mathcal{P}_t(\bar{x}, \bar{y})) \quad & \text{minimize } \bar{f}_t(u) \text{ over } u \in V(t) \quad \text{where } \bar{f}_t \text{ is given by (6.30),} \\ (\mathcal{Q}_t(\bar{x}, \bar{y})) \quad & \text{maximize } \bar{g}_t(v) \text{ over } v \in V(t) \quad \text{where } \bar{g}_t \text{ is given by (6.31).} \end{aligned}$$

Consider too the “endpoint” primal and dual problems associated with the linear-quadratic form $\bar{J}_e(u_e, v_e)$ on $U_e \times V_e$, where \bar{J}_e is given by (6.35), namely:

$$\begin{aligned} (\mathcal{P}_e(\bar{x}, \bar{y})) \quad & \text{minimize } \bar{f}_e(u_e) \text{ over } u_e \in U_e \quad \text{where } \bar{f}_e \text{ is given by (6.32),} \\ (\mathcal{Q}_e(\bar{x}, \bar{y})) \quad & \text{maximize } \bar{g}_e(v_e) \text{ over } v_e \in V_e \quad \text{where } \bar{g}_e \text{ is given by (6.33).} \end{aligned}$$

COROLLARY 6.6. *For $((\bar{u}, \bar{u}_e), (\bar{v}, \bar{v}_e))$ to be a saddle point of J on $\mathcal{U} \times \mathcal{V}$, it is necessary and sufficient that the following conditions hold (in addition to $u(t)$ and $v(t)$ being \mathcal{L}^1 in t). For almost every $t \in [t_0, t_1]$*

$$(6.38) \quad \begin{aligned} \bar{u}(t) &\text{ solves the instantaneous primal } (\mathcal{P}_t(\bar{x}, \bar{y})), \text{ and} \\ \bar{v}(t) &\text{ solves the instantaneous dual } (\mathcal{Q}_t(\bar{x}, \bar{y})), \end{aligned}$$

and furthermore

$$(6.39) \quad \begin{aligned} \bar{u}_e &\text{ solves the endpoint primal } (\mathcal{P}_e(\bar{x}, \bar{y})), \text{ and} \\ \bar{v}_e &\text{ solves the endpoint dual } (\mathcal{Q}_e(\bar{x}, \bar{y})). \end{aligned}$$

Proof. Because the instantaneous and endpoint problems fall in the category of finite-dimensional linear-quadratic programming, we can apply Theorem 2.2 to them and see that (6.38) entails

$$\min(\mathcal{P}_t(\bar{x}, \bar{y})) = \max(\mathcal{Q}_t(\bar{x}, \bar{y})),$$

and (6.39) entails

$$\min(\mathcal{P}_e(\bar{x}, \bar{y})) = \max(\mathcal{Q}_e(\bar{x}, \bar{y})).$$

It follows then from Proposition 2.1 that (6.38) is equivalent to (6.26), whereas (6.39) is equivalent to (6.27). \square

Our final result extends Theorem 2.5 to the infinite-dimensional case. It provides a basis for the idea that in intertemporal linear-quadratic programming as well as in finite-dimensional linear-quadratic programming, a given pair of problems (\mathcal{P}) and (\mathcal{Q}) can often be remodeled, at least for computational purposes, by a more tractable pair $(\hat{\mathcal{P}})$ and $(\hat{\mathcal{Q}})$ in the pattern of *bounded* linear or quadratic programming as in Examples 5.4 and 5.7.

THEOREM 6.7. Consider along with (\mathcal{P}) and (\mathcal{Q}) an auxiliary pair of problems $(\hat{\mathcal{P}})$ and $(\hat{\mathcal{Q}})$ under the same assumptions and defined by the same data, except with the control sets $U(t)$, U_e , $V(t)$, and V_e replaced by sets

$$(6.40) \quad \hat{U}(t) \subset U(t), \quad \hat{U}_e \subset U_e, \quad \hat{V}(t) \subset V(t), \quad \hat{V}_e \subset V_e.$$

Suppose $\min(\hat{\mathcal{P}}) = \max(\hat{\mathcal{Q}})$, as would be true in particular by Theorem 6.3 if the sets $\hat{U}(t)$, \hat{U}_e , $\hat{V}(t)$ and \hat{V}_e are all bounded.

(a) If (\bar{u}, \bar{u}_e) and (\bar{v}, \bar{v}_e) satisfy the instantaneous and endpoint conditions in Theorem 6.5 (or Corollary 6.6) and also are such that

$$\bar{u}(t) \in \hat{U}(t) \text{ a.e.}, \quad \bar{u}_e \in \hat{U}_e, \quad \bar{v}(t) \in \hat{V}(t) \text{ a.e.}, \quad \bar{v}_e \in \hat{V}_e,$$

then (\bar{u}, \bar{u}_e) solves not only (\mathcal{P}) but $(\hat{\mathcal{P}})$, and (\bar{v}, \bar{v}_e) solves not only (\mathcal{Q}) but $(\hat{\mathcal{Q}})$.

(b) If (\bar{u}, \bar{u}_e) solves $(\hat{\mathcal{P}})$ and (\bar{v}, \bar{v}_e) solves $(\hat{\mathcal{Q}})$, and if $U(t)$ and $V(t)$ coincide with $\hat{U}(t)$ and $\hat{V}(t)$ around $\bar{u}(t)$ and $\bar{v}(t)$ for almost every t , while \hat{U}_e and \hat{V}_e coincide with U_e and V_e around \bar{u}_e and \bar{v}_e , then actually (\bar{u}, \bar{u}_e) solves (\mathcal{P}) and (\bar{v}, \bar{v}_e) solves (\mathcal{Q}) .

(The terminology about "coinciding" is defined in the statement of Theorem 2.5.)

Proof. Under the assumptions in (a), (\bar{u}, \bar{u}_e) and (\bar{v}, \bar{v}_e) give a saddle point of J on $\mathcal{U} \times \mathcal{V}$ (by Theorem 6.5, or as the case may be, Corollary 6.6), and this saddle point happens to lie in $\hat{\mathcal{U}} \times \hat{\mathcal{V}}$ (where $\hat{\mathcal{U}}$ and $\hat{\mathcal{V}}$ are the control spaces corresponding to $(\hat{\mathcal{P}})$ and $(\hat{\mathcal{Q}})$). Then $((\bar{u}, \bar{u}_e), (\bar{v}, \bar{v}_e))$ is also a saddle point for J relative to $\hat{\mathcal{U}} \times \hat{\mathcal{V}}$. Theorem 6.2, applied to both pairs of problems, yields the conclusions.

Under the assumptions in (b) we know by Theorem 6.2, as applied to $(\hat{\mathcal{P}})$ and $(\hat{\mathcal{Q}})$, that $((\bar{u}, \bar{u}_e), (\bar{v}, \bar{v}_e))$ is a saddle point of J relative to $\hat{\mathcal{U}} \times \hat{\mathcal{V}}$. The instantaneous conditions and endpoint conditions in Theorem 6.5 must therefore be satisfied relative to $\hat{U}(t) \times \hat{V}(t)$ and $\hat{U}_e \times \hat{V}_e$. But by Theorem 2.5 and our hypothesis about the sets coinciding locally, the same conditions are then satisfied relative to $U(t) \times V(t)$ and $U_e \times V_e$. Theorem 6.5 tells us now that $((\bar{u}, \bar{u}_e), (\bar{v}, \bar{v}_e))$ is a saddle point also for J relative to $\mathcal{U} \times \mathcal{V}$. Then (\bar{u}, \bar{u}_e) and (\bar{v}, \bar{v}_e) are optimal for (\mathcal{P}) and (\mathcal{Q}) by Theorem 6.2. \square

To make the best use of Theorem 6.7 in the manner outlined at the end of §3 for the finite-dimensional case, it would be helpful to have criteria under which (\mathcal{P}) and (\mathcal{Q}) have solutions (\bar{u}, \bar{u}_e) and (\bar{v}, \bar{v}_e) with \bar{u} and \bar{v} actually in \mathcal{L}^∞ . Then, for example, Theorem 6.7 can be applied with the subsets (6.40) taken to be intervals adequately large. Such criteria can be developed, but we shall not address the issue here. Results of this nature for the cases covered by continuous-time programming may be gleaned from Grinold [5], [6] and Reiland [8], [15].

REFERENCES

- [1] R.E. BELLMAN, *Dynamic Programming*, Princeton Univ. Press, 1957.
- [2] W.F. TYNDALL, *A duality theorem for a class of continuous linear programming problems*, SIAM J. Appl. Math., 13 (1965), pp. 644–666.
- [3] —, *An extended duality theorem for continuous linear programming problems*, SIAM J. Appl. Math., 15 (1967), pp. 1294–1298.
- [4] N. LEVINSON, *A class of continuous linear programming problems*, J. Math. Anal. Appl., 16 (1966), pp. 73–83.
- [5] R. GRINOLD, *Symmetric duality for continuous linear programs*, SIAM J. Appl. Math., 18 (1970), pp. 84–97.
- [6] —, *Continuous programming part one: linear objectives*, J. Math. Anal. Appl., 28 (1969), pp. 32–51.

- [7] M. SCHECTER, *Duality in continuous linear programming*, J. Math. Anal. Appl., 37 (1972), pp. 130–141.
- [8] T.W. REILAND, *Optimality conditions and duality in continuous programming*, II: *The linear problem revisited*, J. Math. Anal. Appl., 77 (1980), pp. 329–343.
- [9] R. MEIDAN AND A.F. PEROLD, *Optimality conditions and strong duality in abstract and continuous-time linear programming*, J. Optim. Theory Appl., 40 (1983), pp. 61–76.
- [10] M.A. HANSON, *Duality for a class of infinite programming problems*, SIAM J. Appl. Math., 16 (1968), pp. 318–323.
- [11] M.A. HANSON AND B. MOND, *A class of continuous convex programming problems*, J. Math. Anal. Appl., 22 (1968), pp. 427–437.
- [12] R. GRINOLD, *Continuous programming part two: nonlinear objectives*, J. Math. Anal. Appl., 27 (1969), pp. 639–655.
- [13] W.H. FARR AND M.A. HANSON, *Continuous-time programming with nonlinear constraints*, J. Math. Anal. Appl., 45 (1974), pp. 96–115.
- [14] T.W. REILAND AND M.A. HANSON, *Generalized Kuhn–Tucker conditions and duality for continuous nonlinear programming problems*, J. Math. Anal. Appl., 74 (1980), pp. 578–598.
- [15] T.W. REILAND, *Optimality conditions and duality in continuous programming*, I: *Convex programs and a theorem of the alternative*, J. Math. Anal. Appl., 77 (1980), pp. 297–325.
- [16] R.T. ROCKAFELLAR AND R.J.-B. WETS, *A dual solution procedure for quadratic stochastic programs with simple recourse*, in Numerical Methods, V. Pereyra and A. Reinoza, eds., Lecture Notes in Math. 1005, Springer-Verlag, Berlin, New York, 1983, pp. 252–265.
- [17] —, *A Lagrangian finite generation technique for solving linear-quadratic problems in stochastic programming*, Math. Programming Stud., 28 (1986), pp. 63–93.
- [18] —, *Linear-quadratic programming problems with stochastic penalties: the finite generation algorithm*, in Numerical Techniques for Stochastic Optimization Problems, Y. Ermoliev and R.J.-B. Wets, eds., Springer-Verlag, Berlin, New York, 1986.
- [19] A.F. PEROLD, *Fundamentals of a continuous-time simplex method*, Technical Report SOL 78–26 (1978), Dept. of Operations Research, Stanford Univ., Stanford, CA.
- [20] —, *Extreme points and basic feasible solutions in continuous-time linear programming*, this Journal, 19 (1981), pp. 52–63.
- [21] K.M. ANSTREICHER, *Generation of feasible descent directions in continuous-time linear programming*, Technical Report SOL 83–18 (1983), Dept. of Operations Research, Stanford Univ., Stanford, CA.
- [22] F.H. CLARKE, *Optimization and Nonsmooth Analysis*, Wiley–Interscience, New York, 1983.
- [23] R.T. ROCKAFELLAR, *Conjugate convex functions in optimal control and the calculus of variations*, J. Math. Anal. Appl., 32 (1970), pp. 174–222.
- [24] —, *Generalized Hamiltonian equations for convex problems of Lagrange*, Pacific J. Math., 33 (1970), pp. 411–428.
- [25] —, *Existence and duality theorems for convex problems of Bolza*, Trans. Amer. Math. Soc., 159 (1971), pp. 1–40.
- [26] —, *State constraints in convex problems of Bolza*, SIAM J. Control, 10 (1972), pp. 691–715.
- [27] —, *Semigroups of convex bifunctions generated by Lagrange problems in the calculus of variations*, Math. Scand., 36 (1975), pp. 137–158.
- [28] —, *Dual problems of Lagrange for arcs of bounded variation*, in Calculus of Variations and Control Theory, D.L. Russell, ed., Academic Press, New York, 1976, pp. 155–192.
- [29] —, *Optimality conditions for convex control problems with nonnegative states and the possibility of jumps*, in Game Theory and Mathematical Economics, O. Moeschlin, ed., North-Holland, Amsterdam, New York, 1981, pp. 339–349.
- [30] —, *Duality in optimal control*, in Mathematical Control Theory, W.A. Coppel, ed., Lecture Notes in Math., Springer-Verlag, Berlin, New York, 680 (1978), pp. 219–257.
- [31] J.M. MURRAY, *Some existence and regularity results for dual linear control problems*, J. Math. Anal. Appl., 112 (1985), pp. 190–209.
- [32] R.T. ROCKAFELLAR, *Conjugate Duality and Optimization*, Regional Conference Series in Appl. Math., 61, Society for Industrial and Applied Mathematics, Philadelphia, 1974.
- [33] W.S. DORN, *Duality in quadratic programming*, Quart. Appl. Math., 18 (1960), pp. 155–162.
- [34] R.W. COTTLE, *Symmetric dual quadratic programs*, Quart. Appl. Math., 21 (1963), pp. 237–243.
- [35] M. FRANK AND P. WOLFE, *An algorithm for quadratic programming*, Naval Res. Logist. Quart., 3 (1956), pp. 95–110.
- [36] R.T. ROCKAFELLAR, *Convex Analysis*, Princeton Univ. Press, Princeton, NJ, 1970.

- [37] S.M. ROBINSON, *Some continuity properties of polyhedral multifunctions*, Math. Programming Stud., 14 (1981), pp. 206–214.
- [38] R. FLETCHER, *Penalty functions*, in Mathematical Programming: The State of the Art, A. Bachem et al., eds, Springer-Verlag, 1983, pp. 87–114.
- [39] R. FOURER, *A simplex algorithm for piecewise-linear programming*, I: *Derivation and proof*, Math. Programming, 33 (1985), pp. 204–233.
- [40] R.T. ROCKAFELLAR, *Network Flows and Monotropic Optimization*, Wiley-Interscience, New York, 1984.
- [41] A. KING, R.T. ROCKAFELLAR, L. SOMYODY AND R. J.-B. WETS, *Lake eutrophication management: the Lake Balaton project*, in Numerical Techniques for Stochastic Optimization Problems, Y. Ermoliev and R. J.-B. Wets, eds., Springer-Verlag, Berlin, New York, 1986.
- [42] G. SALINETTI AND R.J.-B. WETS, *On the convergence of sequences of convex sets in finite dimensions*, SIAM Rev., 21 (1979), pp. 18–33.
- [43] R.T. ROCKAFELLAR, *Integrals which are convex functionals*, II, Pacific J. Math., 39 (1971), pp. 439–469.
- [44] R.A. WIJSMAN, *Convergence of sequences of convex sets, cones and functions*, II, Trans. Amer. Math. Soc., 123 (1966), pp. 32–45.
- [45] R.J.-B. WETS, *Convergence of convex functions, variational inequalities and convex optimization problems*, in Variational Inequalities and Complementarity Problems, R.W. Cottle et al., eds., John Wiley, New York, 1980, pp. 375–403.
- [46] L. MCLINDEN AND R.C. BERGSTROM, *Preservation of convergence of convex sets and functions*, Trans. Amer. Math. Soc., 268 (1981), pp. 127–142.
- [47] E. MICHAEL, *Continuous selections*, I, Ann. of Math., 63 (1956), pp. 361–382.
- [48] R.T. ROCKAFELLAR, *Lipschitzian properties of multifunctions*, J. Nonlin. Anal. Th. Meth. Appl., 9 (1985), pp. 867–885.
- [49] R.T. ROCKAFELLAR, *Integrals which are convex functionals*, Pacific J. Math., 24 (1968), pp. 525–539.
- [50] R.T. ROCKAFELLAR, *Integral functionals, normal integrands and measurable selections*, in Nonlinear Operators and the Calculus of Variations, L. Waelbroeck, ed., Lecture Notes in Math., Springer-Verlag, Berlin, New York, 543, 1976, pp. 157–207.
- [51] J.-J. MOREAU, *Théorèmes ‘inf-sup,’* C.R. Acad. Sci. Paris, 258 (1964), pp. 2720–2722.

LINEAR CONTROL THEORY WITH AN H_∞ OPTIMALITY CRITERION*

BRUCE A. FRANCIS† AND JOHN C. DOYLE‡

Abstract. This expository paper sets out the principal results in H_∞ control theory in the context of continuous-time linear systems. The focus is on the mathematical theory rather than computational methods.

Key words. H_∞ control theory, linear systems

AMS(MOS) subject classification. 93

1. Introduction. The subject of this paper is a general regulator problem: a controller is to be designed to regulate the output of a plant subjected to exogenous inputs, such as disturbances, sensor noises and reference signals. A theory for the regulator problem begins by specifying a model of the plant (the model may be a set, to reflect uncertainty), a model of the exogenous inputs, the performance requirements of the controlled system and the allowable class of controllers. For example, two typical regulator theories are the algebraic approach of Pernebo [70] and the Wiener-Hopf approach of Youla, Jabr and Bongiorno [93]. In both these theories the plant is a known time-invariant finite-dimensional linear system and the controller is required to be of this type too. In the algebraic approach the exogenous signal is (after prefiltering) an unknown initial condition, or equivalently a signal of the form $\delta(t)x$, where x is an unknown vector, and the performance requirements are internal stability and asymptotic regulation. In the Wiener-Hopf approach the exogenous signal is (again, after prefiltering) standard white noise, and the performance requirements are internal stability and minimization of the mean-square value of some signal.

In a seminal paper [96], [97], Zames introduced a new theory for the regulator problem. To describe this theory we need a few preliminary mathematical concepts [25], [82]. The Hardy space H_∞ is the class of matrix-valued functions which are analytic and bounded in the open right half-plane, the H_∞ -norm of such a function, say $F(s)$, being defined as

$$\|F\|_\infty := \sup \sigma_{\max}[F(s)].$$

Here σ_{\max} denotes maximum singular value and the supremum is over all s in the open right half-plane, $\operatorname{Re} s > 0$. For such a function the boundary value

$$F(j\omega) := \lim_{\xi \downarrow 0} F(\xi + j\omega)$$

exists for almost all ω and the boundary function is of class L_∞ (Fatou's theorem). As a consequence of the maximum modulus principle the H_∞ -norm of $F(s)$ equals the L_∞ -norm of the boundary function, i.e.,

$$\|F\|_\infty = \operatorname{ess\,sup}_\omega \sigma_{\max}[F(j\omega)].$$

For example, suppose $F(s)$ is scalar-valued, analytic and bounded in $\operatorname{Re} s > 0$, and continuous on the imaginary axis. Then $\|F\|_\infty$ equals the distance in the complex plane from the origin to the farthest point on the Nyquist plot of F .

* Received by the editors February 2, 1985; accepted for publication (in revised form) April 3, 1986.

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

‡ Department of Electrical Engineering, California Institute of Technology, Pasadena, California 91125.

A fundamental fact is that the $L_2[0, \infty)$ -gain of a causal time-invariant linear system equals the H_∞ -norm of its transfer function. To state this more precisely introduce the space $L_2[0, \infty)$ of vector-valued square-integrable functions. The norm on $L_2[0, \infty)$ is

$$\|x\|_2 := \left[\int_0^\infty x(t)^* x(t) dt \right]^{1/2},$$

where $*$ denotes complex-conjugate transpose. The Laplace transform of $x(t)$ in $L_2[0, \infty)$, denoted with abuse of notation by $x(s)$, belongs to the Hardy space H_2 of functions analytic in $\operatorname{Re} s > 0$ and satisfying the condition

$$\sup_{\xi > 0} \int_{-\infty}^{\infty} x(\xi + j\omega)^* x(\xi + j\omega) d\omega < \infty.$$

Such functions also have boundary values almost everywhere and the H_2 -norm is

$$\|x\|_2 := \left[(2\pi)^{-1} \int_{-\infty}^{\infty} x(j\omega)^* x(j\omega) d\omega \right]^{1/2}.$$

The Laplace transform is a Hilbert space isomorphism from $L_2[0, \infty)$ onto H_2 (the Paley-Wiener theorem). Now consider a causal time-invariant linear system having a transfer matrix $F(s)$. Suppose $F \in H_\infty$. Then $Fx \in H_2$ whenever $x \in H_2$, and moreover,

$$(1) \quad \|F\|_\infty = \sup\{\|Fx\|_2 : x \in H_2, \|x\|_2 = 1\}.$$

The H_∞ -norm arises in the regulator problem primarily under two circumstances: when there are sets of exogenous signals and when there is plant uncertainty.

Consider first an example of a tracking problem in which a plant output is to track a reference signal. Suppose, for simplicity, that these two signals are scalar-valued, and let $F(s)$ denote the transfer function from the reference input to the tracking error (reference minus output). Assume the system is stable in the sense that $F \in H_\infty$. Control designs are often based on test inputs, sinusoids being the natural ones in the frequency domain. Suppose the reference signal is allowed to be any sinusoid of amplitude no greater than 1 and of frequency belonging to some interval Ω . An appropriate performance measure might then be

$$\operatorname{ess\,sup}_{\omega \in \Omega} |F(j\omega)|,$$

this equaling the maximum amplitude of the tracking error. Let $W(s)$ be an H_∞ -function such that

$$|W(j\omega)| = 1, \quad \omega \in \Omega, \quad |W(j\omega)| = \varepsilon, \quad \omega \notin \Omega.$$

For small ε the performance measure is approximated by the H_∞ -norm $\|WF\|_\infty$. (For a nontrivial function to be analytic in the right half-plane, its magnitude cannot be zero on a subset of the imaginary axis of positive measure (F. and M. Riesz' theorem); hence the necessity of introducing ε .)

The previous example shows how an H_∞ -norm performance measure can arise from consideration of a set of exogenous inputs, namely sinusoids. Another way to arrive at the same performance measure is with inputs belonging to $L_2[0, \infty)$. Continuing with scalar-valued signals, suppose the reference input is allowed to be any function in the class

$$\{x : x = Wv \text{ for some } v \in H_2, \|v\|_2 \leq 1\},$$

where $W, W^{-1} \in H_\infty$; that is, the reference signal class consists of all x in H_2 such that

$$(2) \quad (2\pi)^{-1} \int_{-\infty}^{\infty} |x(j\omega)|^2 |W(j\omega)|^{-2} d\omega \leq 1.$$

This inequality can be interpreted as a constraint on the weighted energy of x : the energy-density spectrum $|x(j\omega)|^2$ is weighted by the factor $|W(j\omega)|^{-2}$. For example, if $|W(j\omega)|$ were relatively large on a certain frequency band and relatively small off it, then (2) would generate a class of signals having their energy concentrated on that band. This could be useful in representing, for example, a class of narrowband signals whose spectra are confined to a common frequency band. If $F(s)$ again denotes the transfer function from x to the tracking error, then by virtue of (1), $\|WF\|_\infty$ equals the maximum H_2 -norm of the tracking error (i.e., the square root of its energy).

The problem of robust stabilization can also lead to an H_∞ criterion. In this introductory section we consider a simplified version of the problem; a fuller account is given in § 2. The block diagram in Fig. 1(a) shows a plant and a controller with transfer matrices $P(s) + \Delta P(s)$ and $K(s)$ respectively; P represents the nominal plant and ΔP an unknown perturbation, usually caused by unmodeled dynamics or parameter variations. Suppose, for simplicity, that P , ΔP , and K are rational, P and ΔP are strictly proper, K is proper, and P and ΔP are analytic in $\text{Re } s \geq 0$. Suppose also that the system is internally stable for $\Delta P = 0$. How large can ΔP be so that internal stability is maintained?

One method which is used to obtain a transfer function model of a physical system is a frequency response experiment. This yields gain and phase estimates at several frequencies, which in turn provide an upper bound for the norm of $\Delta P(j\omega)$ at several values of ω . Suppose r is a scalar-valued H_∞ -function such that

$$\sigma_{\max}[\Delta P(j\omega)] < |r(j\omega)| \quad \text{for all } \omega,$$

or equivalently

$$(3) \quad \|r^{-1}\Delta P\|_\infty < 1.$$

How large can r be so that internal stability is maintained?

Simple loop transformations lead from Fig. 1(a) to Fig. 1(b) to Fig. 1(c). Since the nominal feedback system is internally stable, $K(I - PK)^{-1} \in H_\infty$. The small gain theorem [78], [95] says that the system in Fig. 1(c) will be internally stable provided the loop gain is less than unity, i.e.,

$$(4) \quad \|\Delta PK(I - PK)^{-1}\|_\infty < 1.$$

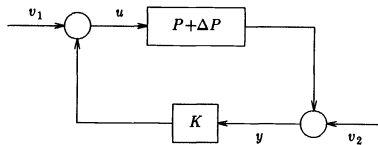


FIG. 1(a). Feedback system with perturbed plant.

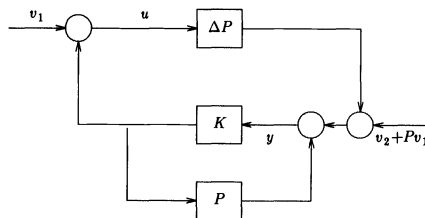


FIG. 1(b). Loop transformation.

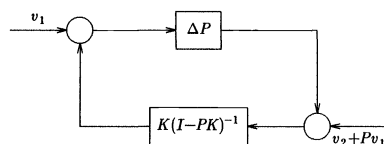


FIG. 1(c). Loop transformation.

In view of (3) a sufficient condition for (4) is

$$(5) \quad \|rK(I - PK)^{-1}\|_{\infty} \leq 1.$$

We conclude that an H_{∞} -norm bounded on a weighted closed-loop transfer matrix is sufficient for robust stability. (Condition (5) is actually necessary for internal stability for all perturbations satisfying (3) ([10], [19]).)

The problem treated in this paper concerns the system in Fig. 2. The signals w , u , z and y are vector-valued and denote, respectively, the exogenous signal (disturbances, sensor noises, reference inputs, etc.), the control signal, the signal to be regulated (tracking errors, plant outputs to be attenuated, weighted actuator outputs, etc.) and the measured signal. The transfer matrices G and K represent the plant and controller respectively. It is assumed that G is real-rational, proper and given; a real-rational proper K is sought to minimize the H_{∞} -norm of the transfer matrix from w to z under the constraint of internal stability.

For ease of reference let us call the problem just stated the *standard (H_{∞}) problem*. It must be emphasized that a controller is designed for a given nominal G ; uncertainty in G is not a consideration. (However, it may already be evident, and will be shown in § 2, that the robust stabilization problem can be recast as a standard problem.) There now exists a reasonably complete solution to the standard problem. The purpose of this paper is to set out the principal results in the context of continuous-time linear systems. The focus is on the mathematical theory rather than computational methods. For the latter the reader may consult [21], [50].

Inclusion of plant uncertainty into the H_{∞} problem increases its difficulty considerably. Let us suppose that uncertainty is introduced in the following general way: G can be any element in a family \mathbf{G} . We could then try to find a controller to minimize the maximum H_{∞} -norm of the transfer matrix from w to z , the maximum taken over all G in \mathbf{G} . For this problem Zames [97] has obtained qualitative results for a simple feedback configuration, showing how performance degrades as uncertainty increases, and Doyle [20] has introduced the concept of structured uncertainty, where the elements of \mathbf{G} have specified structures as well as norm constraints; the H_{∞} problem with structured uncertainty can be reduced to a family of standard problems, thus providing further motivation for the latter.

This introductory section concludes with a brief survey of the literature. The first papers on the subject of H_{∞} -norm optimization of systems are those of Helton [47],

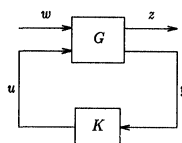


FIG. 2. The standard configuration.

Tannenbaum [83] and Zames [96]. The important papers of Sarason [79] and Adamjan, Arov and Krein [1] established connections between operator theory and complex function theory, in particular, H_∞-functions; Helton [47] showed that these two mathematical subjects have useful applications in electrical engineering, namely, in broadband matching. Tannenbaum [83], [84] used (Nevanlinna–Pick) interpolation theory to attack the problem of stabilizing a plant with an unknown gain. And Zames [96] formulated the problem of sensitivity reduction by feedback as an optimization problem with an operator norm, in particular, an H_∞-norm. The latter paper was amplified in the seminal paper [97].

The fact that the L₂-gain of a system equals the H_∞-norm of its transfer function (i.e. (1)) was used by Zames [94] in his pioneering work on nonlinear system theory. This fact is also central in L₂-stability theory, such as the circle criterion (see e.g. [17]).

Motivation for the H_∞ approach with regard to modeling the exogenous signals is discussed in [20], [21], [24], [90], [97], [98] and with regard to plant uncertainty in [21], [62], [73], [97]. Classical frequency-domain performance specifications also lead to an H_∞ criterion as shown in [48], [75]. The robust stabilization problem discussed above is treated in [22], [24], [44], [54], [58], [73], [84], [85], [88]–[90]. Various versions of the standard problem for finite-dimensional time-invariant systems are covered in [35], [41], [56], [62], [86], [90], [98] in the single-input/single-output case and in [8], [9], [12], [13], [20]–[24], [32]–[34], [36], [37], [48], [49], [60], [61], [71], [74], [77], [87], [90]–[92], [97], [99] in the multivariable case. The mathematical tools primarily used in these references are Nevanlinna–Pick interpolation theory [16], the operator theory of Sarason [79] and Adamjan, Arov and Krein [1], [2] and the geometric theory of Ball and Helton [4]. The standard problem is extended beyond the finite-dimensional time-invariant case in [14], [15], [26]–[31], [42], [54], [55], [57]. References [7], [23], [35], [38]–[40], [51], [66]–[69] present performance bounds for systems designed according to an H_∞ criterion. Algorithms for computing optimal controllers are contained in [21], [50], [76], [80]. Part of the computation in [21] involves solving a special model-reduction problem, for which state-space algorithms are presented in [5], [6], [43], [59], [81]. Finally, the H_∞ approach is compared with the Wiener–Hopf approach in [21], [45], [97], [98].

2. The standard problem. The standard problem pertains to Fig. 2. It is *assumed* that G is real-rational and proper (analytic at $s = \infty$). Partition it as

$$G = \begin{bmatrix} G_{11} & G_{12} \\ G_{21} & G_{22} \end{bmatrix},$$

so that the equations corresponding to Fig. 2 are

$$z = G_{11}w + G_{12}u, \quad y = G_{21}w + G_{22}u, \quad u = Ky.$$

Now eliminate u and y to get that the transfer matrix from w to z is a linear fractional transformation of K :

$$z = [G_{11} + G_{12}K(I - G_{22}K)^{-1}G_{21}]w.$$

It simplifies the theory to guarantee that the rational matrix $I - G_{22}K$ is invertible for every proper real-rational K . A simple sufficient condition for this is that G_{22} be strictly proper (equal to zero at $s = \infty$). Accordingly, this will be *assumed* hereafter.

To define what it means for K to stabilize G , introduce two fictitious inputs v_1 and v_2 as in Fig. 3. It is easy to show that the nine transfer matrices from the three

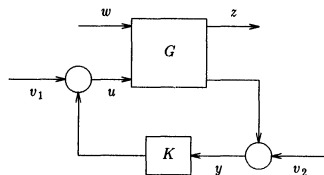


FIG. 3. System for definition of stability.

inputs w , v_1 , v_2 to the three signals z , u , y exist and are proper; if they belong to \mathbf{H}_∞ , then K stabilizes G . This is the usual notion of internal stability. An equivalent definition in terms of state-space models is as follows. Take minimal state-space realizations of G and K and in Fig. 2 set the input w to zero. Then K stabilizes G if and only if the state vectors of G and K tend to zero from every initial condition.

The *standard problem* is this: find a real-rational proper K to minimize the \mathbf{H}_∞ -norm of the transfer matrix from w to z under the constraint that K stabilize G .

Following are three examples of the standard problem.

2.1. A model-matching problem. In Fig. 4 the transfer matrix T_1 represents a “model” which is to be matched by the cascade T_2QT_3 of three transfer matrices T_2 , T_3 and Q . Here, T_i ($i = 1-3$) are given and the “controller” Q is to be designed. Let \mathbf{RH}_∞ denote the space of real-rational matrices in \mathbf{H}_∞ , that is, the space of real-rational proper matrices which are stable, i.e., analytic in $\text{Re } s \geq 0$. It is assumed that $T_i \in \mathbf{RH}_\infty$ ($i = 1-3$) and it is required that $Q \in \mathbf{RH}_\infty$. Thus the four blocks in Fig. 4 represent stable linear systems.

For our purposes the *model-matching criterion* is

$$\sup \{\|z\|_2 : w \in \mathbf{H}_2, \|w\|_2 \leq 1\} = \text{minimum.}$$

Thus the energy of the error z is to be minimized for the worst input w of unit energy. In view of (1) an equivalent criterion is

$$\|T_1 - T_2QT_3\|_\infty = \text{minimum.}$$

This model-matching problem can be cast as a standard problem by defining

$$G := \begin{bmatrix} T_1 & T_2 \\ T_3 & 0 \end{bmatrix}, \quad K := -Q,$$

so that Fig. 4 becomes equivalent to Fig. 2. The constraint that K stabilize G is then equivalent to the constraint that $Q \in \mathbf{RH}_\infty$.

This version of the model-matching problem is not very important per se; its significance in the context of this paper arises from the fact that the standard problem can be transformed to the model-matching problem (§ 3), which is simpler.

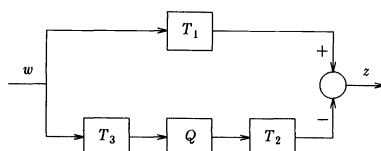


FIG. 4. Model-matching.

2.2. A tracking problem [90], [91]. Figure 5 shows a plant P whose output, v , is to track a reference signal r . The plant input, u , is generated by passing r and v through controllers C_1 and C_2 respectively. It is postulated that r is not a known fixed signal, but, as in the introduction, may be modeled as belonging to the class

$$\{r: r = Ww \text{ for some } w \in \mathbf{H}_2, \|w\|_2 \leq 1\}.$$

Here P and W are given and C_1 and C_2 are to be designed. These four transfer matrices are assumed to be real-rational and proper.

The tracking error signal is $r - v$. Let us take the cost function to be

$$(6) \quad (\|r - v\|_2^2 + \|\rho u\|_2^2)^{1/2},$$

where ρ is a positive scalar weighting factor. The reason for including ρu in (6) is to ensure the existence of an optimal proper controller; for $\rho = 0$ "optimal" controllers tend to be improper. Note that (6) equals the \mathbf{H}_2 -norm of

$$z := \begin{bmatrix} r - v \\ \rho u \end{bmatrix}.$$

Thus the *tracking criterion* is taken to be

$$\sup \{\|z\|_2: w \in \mathbf{H}_2, \|w\|_2 \leq 1\} = \text{minimum}.$$

The equivalent standard problem is obtained by defining

$$(7a) \quad y := \begin{bmatrix} r \\ v \end{bmatrix}, \quad K := \begin{bmatrix} C_1 & C_2 \end{bmatrix},$$

$$G := \begin{bmatrix} G_{11} & G_{12} \\ G_{21} & G_{22} \end{bmatrix},$$

$$(7b) \quad G_{11} := \begin{bmatrix} W \\ 0 \end{bmatrix}, \quad G_{12} := \begin{bmatrix} -P \\ \rho I \end{bmatrix},$$

$$(7c) \quad G_{21} := \begin{bmatrix} W \\ 0 \end{bmatrix}, \quad G_{22} := \begin{bmatrix} 0 \\ P \end{bmatrix}.$$

2.3. A robust stabilization problem [58], [73], [88]. This example has already been discussed in the Introduction. The system under consideration is shown in Fig. 1(a). Assume P is a strictly proper nominal plant and let r be a scalar-valued (radius) function in \mathbf{RH}_∞ . Now define a family \mathbf{P} of neighboring plants to be the set of all strictly proper real-rational matrices $P + \Delta P$ having the same number (in terms of McMillan degree) of poles in $\text{Re } s \geq 0$ as P has, where the perturbation ΔP satisfies the bound

$$\sigma_{\max}[\Delta P(j\omega)] < |r(j\omega)| \quad \text{for all } \omega.$$

For a real-rational proper K the *robust stability criterion* is that K stabilize all plants in \mathbf{P} . Stability means internal stability, that the four transfer matrices in Fig. 1(a) from v_1, v_2 to u, y all belong to \mathbf{RH}_∞ .

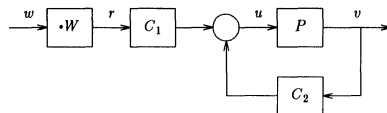


FIG. 5. Tracking.

We saw in the Introduction that robust stability is guaranteed by a small gain condition.

LEMMA 2.1 [10], [19]. *A real-rational proper K stabilizes all plants in \mathbf{P} if and only if K stabilizes the nominal plant P and*

$$\|rK(I - PK)^{-1}\|_{\infty} \leq 1.$$

We can convert to the set-up of the standard problem by defining G so that in Fig. 2 the transfer matrix from w to z equals $rK(I - PK)^{-1}$. This is accomplished by

$$G := \begin{bmatrix} 0 & rI \\ I & P \end{bmatrix}.$$

Then Lemma 2.1 implies that the following two conditions are equivalent: K achieves robust stability for the original system (Fig. 1(a)); in Fig. 2 K stabilizes G and puts the transfer matrix from w to z inside the closed unit ball of \mathbf{H}_{∞} .

There are several other examples of the standard problem and several other problems which are equivalent to the standard problem [21].

3. From the standard problem to a model-matching problem. In this section it is shown that the standard problem can be reduced to the model-matching problem of the first example. The procedure is to parametrize, via a parameter matrix Q in \mathbf{RH}_{∞} , all K 's which stabilize G . Then Fig. 2 can be transformed into Fig. 4, where the T_i 's depend only on G . The parametrization employed in this section is due to Youla et al. [93] as modified by Desoer et al. [18]. Previous work on stability theory for the system in Fig. 2 was carried out by Pernebo [70], Cheng and Pearson [11], and Antoulas [3]; Nett [65] treated a more general setup (K has an additional input and output).

When K stabilizes G can be characterized in terms of coprime factorizations over the ring \mathbf{RH}_{∞} of stable proper real-rational functions. Factor G and K as

$$G = NM^{-1} = \tilde{M}^{-1}\tilde{N}, \quad K = UV^{-1} = \tilde{V}^{-1}\tilde{U}.$$

The matrices N and M belong to \mathbf{RH}_{∞} and are *right-coprime*. This means that, if X is a square matrix in \mathbf{RH}_{∞} which is a right divisor of both N and M , i.e.,

$$N = YX, \quad M = ZX \quad \text{for some } Y, Z \text{ in } \mathbf{RH}_{\infty},$$

then X is a unit of \mathbf{RH}_{∞} , i.e. $X^{-1} \in \mathbf{RH}_{\infty}$. Such N and M constitute a *right-coprime factorization* (rcf) of G . Analogously, \tilde{N} and \tilde{M} are left-coprime and constitute a left-coprime factorization (lcf) of G . Similar remarks apply to the factorization of K . Such factorizations are known to exist [90].

PROPOSITION 3.1. *The following are equivalent statements about the proper transfer matrix K :*

(i) K stabilizes G ,

(ii) $\begin{bmatrix} M & \begin{bmatrix} 0 \\ I \end{bmatrix} U \\ [0 \ I] N & V \end{bmatrix}$ is a unit of \mathbf{RH}_{∞} ,

(iii) $\begin{bmatrix} \tilde{M} & \tilde{N} \begin{bmatrix} 0 \\ I \end{bmatrix} \\ \tilde{U} [0 \ I] & \tilde{V} \end{bmatrix}$ is a unit of \mathbf{RH}_{∞} .

The idea underlying the equivalence of (i) and (ii) is simply that the determinant of the matrix in (ii) is the least common denominator (in \mathbf{RH}_∞) of all the transfer functions from w, v_1, v_2 to z, u, y ; hence the determinant must be a unit for all these transfer functions to belong to \mathbf{RH}_∞ , and conversely.

The transfer matrix G is *stabilizable* if there exists a K which stabilizes it. Not every G is stabilizable; an obvious nonstabilizable G is $G_{12}=0, G_{21}=0, G_{22}=0, G_{11}$ unstable. Proposition 3.1 provides a test for stabilizability. For example, in terms of N and M, G is stabilizable if and only if suitable U and V exist satisfying condition (ii). Such a consideration readily leads to the following.

PROPOSITION 3.2. *The following are equivalent:*

- (i) G is stabilizable;
- (ii) $M, [0 \ I]N$ are right-coprime and $M, \begin{bmatrix} 0 \\ I \end{bmatrix}$ are left-coprime;
- (iii) $\tilde{M}, \tilde{N} \begin{bmatrix} 0 \\ I \end{bmatrix}$ are left-coprime and $M, [0 \ I]$ are right-coprime.

In terms of a state-space model G is stabilizable if and only if, roughly speaking, its unstable modes are controllable from u and observable from y (Fig. 2). For example, right-coprimeness of $M, [0 \ I]N$ can be interpreted as a frequency-domain stabilizability condition.

Hereafter, G will be *assumed* to be stabilizable. To recap, G is assumed so far to be real-rational, proper with G_{22} strictly proper, and stabilizable. Intuitively, stabilizability of G implies that G and G_{22} share the same unstable modes, so that to stabilize G it is enough to stabilize G_{22} . The controller K stabilizes G_{22} if, in Fig. 6, the four transfer matrices from v_1, v_2 to u, y are stable.

PROPOSITION 3.3. *K stabilizes G if and only if K stabilizes G_{22} .*

The next step is to parametrize all K 's stabilizing G_{22} . For this it is convenient to introduce a special rcf and lcf of G_{22} :

$$(8) \quad G_{22} = N_2 M_2^{-1} = \tilde{M}_2^{-1} \tilde{N}_2,$$

$$(9) \quad \begin{bmatrix} \tilde{X}_2 & -\tilde{Y}_2 \\ -\tilde{N}_2 & \tilde{M}_2 \end{bmatrix} \begin{bmatrix} M_2 & Y_2 \\ N_2 & X_2 \end{bmatrix} = I.$$

The eight matrices introduced in (8) and (9) all belong to \mathbf{RH}_∞ ; their existence is proved in § 4. Equation (9) is known as a generalized Bezout identity; its satisfaction guarantees that N_2, M_2 are right-coprime and \tilde{N}_2, \tilde{M}_2 are left-coprime. Equations (8) and (9) constitute a generalization of the usual polynomial matrix-fraction description [18], [53], [90].

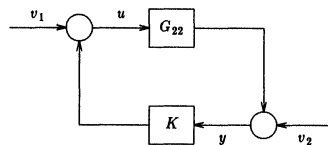


FIG. 6. Part of standard configuration.

THEOREM 3.1 [18], [90], [93]. *The following formulas parametrize all proper K 's which stabilize G_{22} :*

$$(10) \quad K = (Y_2 - M_2 Q)(X_2 - N_2 Q)^{-1}$$

$$(11) \quad = (\tilde{X}_2 - Q \tilde{N}_2)^{-1} (\tilde{Y}_2 - Q \tilde{M}_2), \quad Q \in \mathbf{RH}_\infty.$$

The right-hand sides of (10) and (11) constitute an rcf and lcf of K respectively; the inverses exist for every Q in \mathbf{RH}_∞ (because G_{22} is strictly proper). As Q varies over all stable proper matrices, the formulas generate all possible stabilizing K 's.

The final step is to determine the transfer matrix from w to z in Fig. 2 when K is given by formulas (10) and (11). Define

$$(12a) \quad T_1 := G_{11} + G_{12} Y_2 \tilde{M}_2 G_{21},$$

$$(12b) \quad T_2 := G_{12} M_2,$$

$$(12c) \quad T_3 := \tilde{M}_2 G_{21}.$$

It can be proved that $T_i \in \mathbf{RH}_\infty$ ($i = 1-3$).

THEOREM 3.2 [21]. *With K as in (10), (11), the transfer matrix from w to z equals $T_1 - T_2 Q T_3$.*

We conclude from this theorem that *the standard problem reduces to the model-matching problem* of finding matrices Q in \mathbf{RH}_∞ to minimize $\|T_1 - T_2 Q T_3\|_\infty$. A solution Q to the model-matching problem yields a solution K to the standard problem via formulas (10) and (11).

A special case is when G is itself stable. In (8) and (9) we may then take

$$N_2 = \tilde{N}_2 = G_{22},$$

$$M_2, \tilde{M}_2, X_2, \tilde{X}_2 \text{ all} = I,$$

$$Y_2 = 0, \quad \tilde{Y}_2 = 0,$$

in which case (10) and (11) become [97]

$$\begin{aligned} K &= -Q(I - G_{22}Q)^{-1} \\ &= -(I - QG_{22})^{-1}Q \end{aligned}$$

and (12) produces

$$T_1 = G_{11}, \quad T_2 = G_{12}, \quad T_3 = G_{21}.$$

4. State-space computations. The reduction in the previous section was developed using transfer matrix models. However, as a practical matter the computations are quite easily and reliably performed using state-space models. This section describes how to obtain state-space realizations of the matrices T_i ($i = 1-3$) starting from a state-space realization of G . The formulas are very simple and they provide a fundamental link (Theorem 4.1) between the stability result of Theorem 3.1 and observer-based stability theory. Some of the results of this section are contained in [64].

We begin with a minimal realization of $G(s)$,

$$G(s) = D + C(s - A)^{-1}B, \quad A, B, C, D \text{ real matrices.}$$

It is convenient to introduce a new data structure:

$$[A, B, C, D] := D + C(s - A)^{-1}B.$$

Since the input and output of G are partitioned as

$$\begin{bmatrix} w \\ u \end{bmatrix} \quad \text{and} \quad \begin{bmatrix} z \\ y \end{bmatrix},$$

the matrices B , C , and D have corresponding partitions,

$$B = [B_1 \quad B_2], \quad C = \begin{bmatrix} C_1 \\ C_2 \end{bmatrix}, \quad D = \begin{bmatrix} D_{11} & D_{12} \\ D_{21} & D_{22} \end{bmatrix}.$$

Thus

$$G_{ij}(s) = [A, B_j, C_i, D_{ij}], \quad i, j = 1, 2.$$

Note that $D_{22} = 0$ because G_{22} is strictly proper. It follows from the stabilizability of G that the pair (A, B_2) is stabilizable and the pair (C_2, A) is detectable.

The objective now is to specify state-space realizations of transfer matrices N_2 , M_2 , etc. belonging to \mathbf{RH}_∞ and satisfying (8) and (9). Denote the state, input and output vectors of G_{22} by x , u and y respectively, so that $y = G_{22}u$ and

$$(13a) \quad \dot{x} = Ax + B_2u,$$

$$(13b) \quad y = C_2x.$$

Next, choose a real matrix F so that $A_F := A + B_2F$ is stable (all eigenvalues in $\text{Re } s < 0$) and define the vector $v := u - Fx$. Then from (13) we get

$$\dot{x} = A_Fx + B_2v, \quad u = Fx + v, \quad y = C_2x.$$

The transfer matrix from v to u is

$$(14) \quad M_2(s) := [A_F, B_2, F, I]$$

and that from v to y is

$$(15) \quad N_2(s) := [A_F, B_2, C_2, 0].$$

Therefore $G_{22} = N_2M_2^{-1}$. Similarly, by choosing a real matrix H so that $A_H := A + HC_2$ is stable and defining

$$\tilde{M}_2(s) := [A_H, H, C_2, I], \quad \tilde{N}_2(s) := [A_H, B_2, C_2, 0],$$

we obtain $G_{22} = \tilde{M}_2^{-1}\tilde{N}_2$. Thus (8) is satisfied.

Now introduce an observer-based controller, denoted by $K_o(s)$, to stabilize G_{22} . The familiar equations for K_o are

$$\dot{\hat{x}} = A\hat{x} + B_2u + H(C_2\hat{x} - y), \quad u = F\hat{x},$$

or equivalently,

$$\dot{\hat{x}} = (A + B_2F + HC_2)\hat{x} - Hy, \quad u = F\hat{x}.$$

Thus

$$K_o(s) = [A + B_2F + HC_2, -H, F, 0].$$

Obtaining factorizations of K_o analogous to those just obtained for G_{22} , we arrive at the equations

$$(16) \quad K_o = Y_2X_2^{-1} = \tilde{X}_2^{-1}\tilde{Y}_2,$$

where

$$(17) \quad X_2(s) := [A_F, -H, C_2, I],$$

$$(18) \quad Y_2(s) := [A_F, -H, F, 0],$$

$$\tilde{X}_2(s) := [A_H, -B_2, F, I], \quad \tilde{Y}_2(s) := [A_H, -H, F, 0].$$

Routine algebra verifies (9).

Formula (16) provides just one controller which stabilizes G_{22} , whereas formulas (10) and (11) in Theorem 3.1, having the additional stable matrix Q , generate all stabilizing controllers. These two results can be used to show that every stabilization procedure amounts to adding stable dynamics to the plant and then using an observer-based controller to stabilize the result. The precise statement is as follows.

THEOREM 4.1 [21]. *Suppose K stabilizes*

$$G_{22}(s) = [A, B_2, C_2, 0].$$

Then G_{22} can be embedded in a system $[A_e, B_e, C_e, 0]$, where

$$A_e := \begin{bmatrix} A & 0 \\ 0 & A_a \end{bmatrix}, \quad B_e := \begin{bmatrix} B_2 \\ 0 \end{bmatrix}, \quad C_e := [C_2 \ 0]$$

and A_a is stable, such that K has the form

$$K(s) = [A_e + B_e F_e + H_e C_e, -H_e, F_e, 0],$$

where $A_e + B_e F_e$ and $A_e + H_e C_e$ are stable.

Now consider the transfer matrices T_i defined in (12). We have obtained realizations of all the matrices on the right-hand sides of (12). Algebraic manipulations lead to the realizations

$$\begin{aligned} (19a) \quad T_1(s) &= [A_1, \underline{B}_1, \underline{C}_1, D_{11}], \\ A_1 &= \begin{bmatrix} A_F & -B_2 F \\ 0 & A_H \end{bmatrix}, \quad \underline{B}_1 = \begin{bmatrix} B_1 \\ B_1 + H D_{21} \end{bmatrix}, \\ \underline{C}_1 &= [C_1 + D_{12} F \quad -D_{12} F], \\ (19b) \quad T_2(s) &= [A_F, B_2, C_1 + D_{12} F, D_{12}], \\ (19c) \quad T_3(s) &= [A_H, B_1 + H D_{21}, C_2, D_{21}]. \end{aligned}$$

5. Model-matching theory. This section treats the model-matching problem finding matrices Q in \mathbf{RH}_∞ to minimize the model-matching error $\|T_1 - T_2 Q T_3\|_\infty$ where $T_i \in \mathbf{RH}_\infty$ ($i = 1-3$).

5.1. Existence of a solution. Define the *infimal model-matching error*

$$(20) \quad \alpha := \inf \{ \|T_1 - T_2 Q T_3\|_\infty : Q \in \mathbf{RH}_\infty \}.$$

The natural question to answer first is when is this infimum achieved. The following provides a mild sufficient condition.

THEOREM 5.1 (e.g. [33]). *The infimum in (20) is achieved if the ranks of the two matrices $T_2(j\omega)$ and $T_3(j\omega)$ are constant for all $0 \leq \omega \leq \infty$.*

These rank conditions will be *assumed* to hold for the remainder of § 5. (In applications they do hold for well-defined problems.) In general there is a family of optimal Q 's, i.e., Q 's achieving the infimum. Moreover, the model-matching error cannot be reduced by a nonrational Q , i.e.,

$$\min \{ \|T_1 - T_2 Q T_3\|_\infty : Q \in \mathbf{RH}_\infty \} = \min \{ \|T_1 - T_2 Q T_3\|_\infty : Q \in \mathbf{H}_\infty \}.$$

The latter minimum equals the distance in \mathbf{H}_∞ from T_1 to the subspace

$$T_2 \mathbf{H}_\infty T_3 := \{ T_2 Q T_3 : Q \in \mathbf{H}_\infty \}.$$

Hence

$$(21) \quad \alpha = \text{dist}(T_1, T_2 \mathbf{H}_\infty T_3).$$

The rank conditions in Theorem 5.1 suffice to make $T_2\mathbf{H}_\infty T_3$ closed in the weak-star topology of \mathbf{H}_∞ , and this in turn guarantees the existence of a matrix in $T_2\mathbf{H}_\infty T_3$ closest to an arbitrary T_1 .

The model-matching problem is relatively easy when the T_i 's are scalar-valued. The main results can be summarized as follows. A function $f(s)$ in \mathbf{RH}_∞ is said to be an *inner function* if $f(s)f(-s) = 1$. The zeros of such a function all lie in $\text{Re } s > 0$; the number of its zeros will be called its *degree*.

PROPOSITION 5.1 [98] (Scalar-valued case). *The infimum in (20) is achieved if $T_2 T_3$ has no zeros on the extended imaginary axis. In this case the optimal Q is unique and is uniquely determined by the following property: $T_1 - T_2 Q T_3$ is a scalar multiple of an inner function of degree less than the number of zeros of $T_2 T_3$ in $\text{Re } s > 0$.*

5.2. A formula for the minimal model-matching error. This subsection shows that α equals the norm of a certain operator, a Hankel operator in a special case. First, it will be shown that α can be expressed in the form

$$\alpha = \text{dist} \left(R, \begin{bmatrix} I \\ 0 \end{bmatrix} \mathbf{H}_\infty \begin{bmatrix} I & 0 \end{bmatrix} \right)$$

where the matrix R belongs to \mathbf{RL}_∞ (the space of real-rational \mathbf{L}_∞ -matrices) and depends only on the T_i 's. For this maneuver we need the concepts of inner and outer matrices [82].

Introduce the notation $F^\sim(s) := F(-s)'$ for a matrix-valued function $F(s)$, where prime denotes transpose. A matrix F in \mathbf{RH}_∞ is an *inner matrix* if $F^\sim F = I$ and an *outer matrix* if it has full row rank in $\text{Re } s > 0$; it is termed *co-inner* or *co-outer* if its transpose is inner or outer, respectively.

LEMMA 5.1. *For each matrix F in \mathbf{RH}_∞ there exist inner, outer, co-inner and co-outer matrices F_i, F_o, F_{ci}, F_{co} respectively, such that*

$$F = F_i F_o = F_{co} F_{ci}.$$

If F has constant rank on the extended imaginary axis, then F_o has a right-inverse in \mathbf{RH}_∞ and F_{co} has a left-inverse in \mathbf{RH}_∞ .

Returning to the model-matching problem, introduce inner, outer, co-inner and co-outer matrices as follows:

$$T_2 = (T_2)_i (T_2)_o, \quad T_3 = (T_3)_{co} (T_3)_{ci}.$$

(In going from the standard problem to the model-matching problem, it is possible to arrange that T_2 is automatically inner and T_3 is automatically co-inner [21].) Lemma 5.1, together with the assumptions on T_2 and T_3 , implies that $(T_2)_o$ is right-invertible over \mathbf{RH}_∞ and $(T_3)_{co}$ is left-invertible over \mathbf{RH}_∞ . Thus the mapping

$$Q \rightarrow (T_2)_o Q (T_3)_{co}$$

on \mathbf{H}_∞ is surjective. Hence from (21) we get the expression

$$\alpha = \text{dist} (T_1, (T_2)_i \mathbf{H}_\infty (T_3)_{ci}).$$

To simplify notation define

$$U_i := (T_2)_i, \quad U_{ci} := (T_3)_{ci}$$

so that

$$(22) \quad \alpha = \text{dist} (T_1, U_i \mathbf{H}_\infty U_{ci}).$$

Being inner, U_i has the property that $U_i^\sim U_i = I$. This in turn implies that

$$(23) \quad E^\sim E = I,$$

where

$$E := \begin{bmatrix} U_i^\sim \\ I - U_i U_i^\sim \end{bmatrix}.$$

It is a consequence of (23) that the norm of an \mathbf{L}_∞ -matrix is unchanged if the matrix is pre-multiplied by E . Similarly, $L^\sim L = I$, where

$$L := \begin{bmatrix} U_{ci} \\ I - U_{ci}^\sim U_{ci} \end{bmatrix},$$

and post-multiplication by L^\sim preserves the \mathbf{L}_∞ -norm. Hence (22) yields

$$\alpha = \text{dist}(ET_1 L^\sim, EU_i \mathbf{H}_\infty U_{ci} L^\sim).$$

But

$$EU_i = \begin{bmatrix} I \\ 0 \end{bmatrix}, \quad U_{ci} L^\sim = [I \quad 0].$$

Defining

$$(24) \quad R := ET_1 L^\sim,$$

we conclude that

$$(25) \quad \alpha = \text{dist}\left(R, \begin{bmatrix} I \\ 0 \end{bmatrix} \mathbf{H}_\infty [I \quad 0]\right).$$

An interesting special case occurs when T_2 has full row rank and T_3 has full column rank over the field of rational functions. Then U_i and U_{ci} are both square, so that

$$U_i^\sim = U_i^{-1}, \quad U_{ci}^\sim = U_{ci}^{-1},$$

and R has the form

$$R = \begin{bmatrix} R_1 & 0 \\ 0 & 0 \end{bmatrix}$$

where $R_1 := U_i^{-1} T_1 U_{ci}^{-1}$. Then (25) simplifies to the expression

$$\alpha = \text{dist}(R_1, \mathbf{H}_\infty).$$

In this case α equals the distance in \mathbf{L}_∞ from R_1 to the nearest \mathbf{H}_∞ -matrix.

Let us concentrate on this simpler problem of finding the distance from an \mathbf{L}_∞ -matrix R to the nearest \mathbf{H}_∞ -matrix. (The subscript on R_1 has been temporarily dropped.) Let \mathbf{L}_2 denote the Hilbert space of vector-valued square-integrable functions on the imaginary axis. The Hardy space \mathbf{H}_2 is a closed subspace of \mathbf{L}_2 ; let \mathbf{H}_2^\perp denote its orthogonal complement. The \mathbf{L}_∞ -norm of R equals the norm of the corresponding Laurent operator on \mathbf{L}_2 (cf. (1)); denote this operator by M_R :

$$M_R f = Rf, \quad f \in \mathbf{L}_2.$$

For an H_∞ -matrix X the Laurent operator M_X leaves the subspace H_2 of L_2 invariant, i.e.,

$$(26) \quad \text{if } f \in H_2 \quad \text{then } M_X f = Xf \in H_2.$$

If $\Pi: L_2 \rightarrow H_2^\perp$ denotes the *orthogonal projection*, then (26) is equivalent to the condition that

$$\Pi M_X|_{H_2} = 0.$$

Thus we have that

$$(27) \quad \begin{aligned} \text{dist}(R, H_\infty) &= \min \{\|R - X\|_\infty: X \in H_\infty\} \\ &= \min \{\|M_R - M_X\|: X \in H_\infty\} \\ &\geq \min \{\|\Pi(M_R - M_X)|_{H_2}\|: X \in H_\infty\} \\ &= \|\Pi M_R|_{H_2}\|. \end{aligned}$$

In fact, equality holds in (27).

THEOREM 5.2. *The distance from a matrix R in L_∞ to the nearest matrix in H_∞ equals the norm of the operator $\Pi M_R|_{H_2}$ from H_2 to H_2^\perp .*

This result is generally known as Nehari's theorem [63] and the operator $\Pi M_R|_{H_2}$ is called the *Hankel operator with symbol R* [72]. As a concrete example, consider the scalar-valued function $R(s) = (s-1)^{-1}$. For g in H_2 we have

$$R(s)g(s) = (s-1)^{-1}g(1) + (s-1)^{-1}[g(s) - g(1)].$$

The first function on the right-hand side belongs to H_2^\perp and the second to H_2 . Hence the Hankel operator maps $g(s)$ in H_2 into $(s-1)^{-1}g(1)$ in H_2^\perp .

The Hankel operator has a time-domain version (e.g. [43]). Suppose $R(s)$ is analytic in a strip containing the imaginary axis; such would be the case if $R(s)$ were rational, for example. Taking the region of convergence to be this strip, let $r(t)$ denote the inverse bilateral Laplace transform of $R(s)$. The linear system with impulse response $r(t)$ is therefore $L_2(-\infty, \infty)$ -stable, but noncausal in general. The Hankel operator in the time-domain maps a function u in $L_2[0, \infty)$ into a function y in $L_2(-\infty, 0]$ according to the convolution equation

$$y(t) = \int_0^\infty r(t-\tau)u(\tau) d\tau, \quad t \leq 0.$$

Since the bilateral Laplace transformation is an isomorphism from $L_2[0, \infty)$ onto H_2 and from $L_2(-\infty, 0]$ onto H_2^\perp , the Hankel operators in the two domains have equal norms. A causal system leaves $L_2[0, \infty)$ invariant, so its Hankel operator equals zero. Interpreted in the time-domain, Theorem 5.2 states that the distance from the noncausal system with impulse response $r(t)$ to the nearest causal system equals the norm of the Hankel operator; in other words, the Hankel operator's norm is a measure of noncausality. Here the distance is the norm of the error system considered as a mapping on $L_2(-\infty, \infty)$.

It is a useful fact that the norm of a Hankel operator with a rational symbol can be computed by state-space methods. Let R be a matrix in \mathbf{RL}_∞ and let $C(s-A)^{-1}B$ be a minimal realization of its antistable part, i.e.,

$$(28) \quad R(s) = C(s-A)^{-1}B + (\text{a matrix in } \mathbf{RH}_\infty),$$

and the eigenvalues of A lie in $\operatorname{Re} s > 0$. Introduce the controllability and observability gramians

$$(29) \quad L_c := \int_{-\infty}^0 e^{At} B B' e^{A't} dt,$$

$$(30) \quad L_o := \int_{-\infty}^0 e^{A't} C' C e^{At} dt.$$

Thus L_c and L_o are the unique solutions of the Lyapunov equations

$$(31) \quad A L_c + L_c A' = B B',$$

$$(32) \quad A' L_o + L_o A = C' C.$$

It can be proved [81] that $L_c L_o$ has only real, nonnegative eigenvalues.

LEMMA 5.2 (e.g. [43]). *The Hankel operator $\Gamma := \Pi M_R|_{\mathbf{H}_2}$ with rational symbol R has finite rank. The operator $\Gamma^* \Gamma$ and the matrix $L_c L_o$ share the same nonzero eigenvalues. In particular, the norm of Γ equals the square-root of the largest eigenvalue of $L_c L_o$.*

Now let us return to the general case. From (25) α equals the distance in \mathbf{L}_∞ from R to the subspace

$$\begin{bmatrix} I \\ 0 \end{bmatrix} \mathbf{H}_\infty [I \ 0].$$

A matrix in this subspace has the form

$$(33) \quad \begin{bmatrix} I \\ 0 \end{bmatrix} X [I \ 0] = \begin{bmatrix} X & 0 \\ 0 & 0 \end{bmatrix} =: X$$

for some X in \mathbf{H}_∞ , and the corresponding Laurent operator is M_X . This operator acts on \mathbf{L}_2 -vectors partitioned conformably with the partitioning in (33):

$$M_X \begin{bmatrix} f \\ g \end{bmatrix} = X \begin{bmatrix} f \\ g \end{bmatrix} = \begin{bmatrix} Xf \\ 0 \end{bmatrix}.$$

Hence the domain of this operator is the product space (external direct sum) $\mathbf{L}_2 \times \mathbf{L}_2$. Moreover, since $X \in \mathbf{H}_\infty$, the subspace $\mathbf{H}_2 \times \mathbf{L}_2$ is invariant under this operator, or equivalently

$$\Pi M_X |_{(\mathbf{H}_2 \times \mathbf{L}_2)} = 0,$$

where Π denotes the orthogonal projection from $\mathbf{L}_2 \times \mathbf{L}_2$ onto $\mathbf{H}_2^\perp \times \mathbf{L}_2$. As in (27) we conclude that

$$\operatorname{dist} \left(R, \begin{bmatrix} I \\ 0 \end{bmatrix} \mathbf{H}_\infty [I \ 0] \right) \cong \|\Pi M_R |_{(\mathbf{H}_2 \times \mathbf{L}_2)}\|.$$

Again, equality holds.

To recap, let R be defined as in (24), let M_R denote the Laurent operator on $\mathbf{L}_2 \times \mathbf{L}_2$ of multiplication by R , and let Π denote the orthogonal projection from $\mathbf{L}_2 \times \mathbf{L}_2$ onto $\mathbf{H}_2^\perp \times \mathbf{L}_2$.

THEOREM 5.3 [29]. *The minimal model-matching error α equals the norm of the operator*

$$(34) \quad \Pi M_R |_{(\mathbf{H}_2 \times \mathbf{L}_2)}$$

from $\mathbf{H}_2 \times \mathbf{L}_2$ to $\mathbf{H}_2^\perp \times \mathbf{L}_2$.

Operator (34) is not a Hankel operator (by definition) and at present there is unfortunately no direct procedure for computing its norm. It is easy to get crude bounds for its norm: an upper bound is $\|R\|_\infty$ and a lower bound is the maximum of

$$\|[0 \ I]R\|_\infty, \quad \left\| R \begin{bmatrix} 0 \\ I \end{bmatrix} \right\|_\infty.$$

5.3. Nearly optimal solutions. The value of α cannot be computed directly, as we just noted, but it is possible to compute, by iteration, an upper bound which is as close to α as desired. To do this, let α_1 and α_2 be, respectively, any lower and upper bounds for α , for example, those given at the end of the previous subsection. A number γ satisfies the inequality $\alpha < \gamma$ if and only if (see (21))

$$(35) \quad \text{dist}(T_1, T_2 H_\infty T_3) < \gamma.$$

Thus testing if (35) holds for several values of γ in the interval $[\alpha_1, \alpha_2]$ serves to locate α for any desired accuracy. A bisection search could be used.

This subsection treats the problem of checking if (35) is true for a prespecified γ and, when it is true, of finding all Q 's in \mathbf{RH}_∞ which achieve the inequality

$$\|T_1 - T_2 Q T_3\|_\infty < \gamma;$$

when γ is a bit larger than α , such Q 's are nearly optimal. It will be shown that (35) is equivalent to the following three conditions:

$$\|Y\|_\infty < \gamma, \quad \|Z\|_\infty < 1, \quad \text{dist}(R, H_\infty) < 1.$$

Here¹ R , Y and Z are \mathbf{RL}_∞ -matrices which are computed from T_i ($i = 1-3$) and γ (Theorem 5.4 below). Notice that the distance from R to H_∞ can be readily computed (Theorem 5.2, Lemma 5.2).

We require two definitions. Let F be an \mathbf{RL}_∞ -matrix and let η be a positive number. The inequality

$$(36) \quad \eta > \|F\|_\infty$$

is equivalent to the condition

$$\eta^2 I - F(j\omega)^* F(j\omega) > 0 \quad \text{for all } 0 \leq \omega \leq \infty.$$

This latter condition implies that the matrix $\eta^2 I - F^* F$ has a spectral factorization:

$$\eta^2 I - F^* F = F_o^* F_o, \quad F_o, F_o^{-1} \in \mathbf{RH}_\infty.$$

Such a matrix F_o will be called a *spectral factor* of $\eta^2 I - F^* F$. (It is outer, hence the subscript o .) The same inequality, (36), implies that

$$\eta^2 I - F F^* = F_{co} F_{co}^*$$

for some $F_{co}, F_{co}^{-1} \in \mathbf{RH}_\infty$; F_{co} will be called a *co-spectral factor* of $\eta^2 I - F F^*$.

Let us consider, first, condition (35) under the simplifying assumption that $T_3 = I$. Write $T_2 = U_i U_o$ where U_i is inner and U_o is outer, and let $Q \in \mathbf{RH}_\infty$. Then, as in the previous subsection, the matrices $T_1 - T_2 Q$ and

$$\begin{bmatrix} U_i^* \\ I - U_i U_i^* \end{bmatrix} (T_1 - T_2 Q) = \begin{bmatrix} U_i^* T_1 - U_o Q \\ (I - U_i U_i^*) T_1 \end{bmatrix}$$

¹ The matrix R in this subsection is defined somewhat differently from that in the previous one.

have equal norms. Defining

$$Y := (I - U_i U_i^\sim) T_1,$$

we obtain that

$$\|T_1 - T_2 Q\|_\infty < \gamma$$

if and only if

$$(37) \quad \left\| \begin{bmatrix} U_i^\sim T_1 - U_o Q \\ Y \end{bmatrix} \right\|_\infty < \gamma.$$

It can be shown that (37) holds if and only if $\|Y\|_\infty < \gamma$ and

$$\|(U_i^\sim T_1 - U_o Q) Y_o^{-1}\|_\infty < 1,$$

where Y_o is a spectral factor of $\gamma^2 I - Y^\sim Y$. We conclude that

$$\text{dist}(T_1, T_2 \mathbf{H}_\infty) < \gamma$$

if and only if $\|Y\|_\infty < \gamma$ and

$$\text{dist}(U_i^\sim T_1 Y_o^{-1}, \mathbf{H}_\infty) < 1.$$

The general result is as follows.

THEOREM 5.4 [20], [21]. *Let $Q \in \mathbf{RH}_\infty$ and $\gamma > 0$. Then*

$$\|T_1 - T_2 Q T_3\|_\infty < \gamma$$

if and only if

$$(38) \quad \|Y\|_\infty < \gamma, \quad \|Z\|_\infty < 1, \quad \|R - X\|_\infty < 1,$$

where R, Y, Z are \mathbf{RL}_∞ -matrices and X is an \mathbf{RH}_∞ -matrix defined as follows:

$$(39) \quad T_2 = U_i U_o, \quad U_i \text{ inner}, \quad U_o \text{ outer},$$

$$(40) \quad Y := (I - U_i U_i^\sim) T_1,$$

$$Y_o = \text{spectral factor of } \gamma^2 I - Y^\sim Y,$$

$$(41) \quad T_3 Y_o^{-1} = V_{co} V_{ci}, \quad V_{co} \text{ co-outer}, \quad V_{ci} \text{ co-inner},$$

$$(42) \quad Z := U_i^\sim T_1 Y_o^{-1} (I - V_{ci}^\sim V_{ci}),$$

$$(43) \quad Z_{co} = \text{co-spectral factor of } I - Z Z^\sim,$$

$$(44) \quad R := Z_{co}^{-1} U_i^\sim T_1 Y_o^{-1} V_{ci}^\sim,$$

$$(45) \quad X := Z_{co}^{-1} U_o Q V_{co}.$$

Observe that $Z_{co}^{-1} U_o$ is right-invertible over \mathbf{RH}_∞ and V_{co} is left-invertible over \mathbf{RH}_∞ . Thus (45) provides a linear relation between Q 's satisfying (35) and X 's satisfying (38).

Let us recap. Suppose the objective is to compute an upper bound γ for α such that $\gamma - \alpha$ is less than a prespecified number, and then to determine a Q in \mathbf{RH}_∞ satisfying

$$\|T_1 - T_2 Q T_3\|_\infty < \gamma.$$

To accomplish this, first determine lower and upper bounds α_1 and α_2 for α . Then, select a trial value for γ in the interval $[\alpha_1, \alpha_2]$. Next, test to see if the following conditions hold:

$$\|Y\|_\infty < \gamma, \quad \|Z\|_\infty < 1, \quad \text{dist}(R, \mathbf{H}_\infty) < 1.$$

If so, reduce the value of γ ; if not, increase it. When a sufficiently accurate upper bound is obtained, find an X in \mathbf{RH}_∞ such that $\|R - X\|_\infty < 1$. Finally, solve (45) for a Q in \mathbf{RH}_∞ .

The part of the procedure which remains to be described is how to find such an X . This is the next topic.

5.4. Best approximation. Let R be an \mathbf{RL}_∞ -matrix. The problem of best approximation is that of finding one, some, or all matrices X in \mathbf{RH}_∞ such that

$$\|R - X\|_\infty = \text{dist}(R, \mathbf{RH}_\infty).$$

Such X 's will be termed *optimal*. This problem has an extensive theory, involving several different approaches. In this paper there is space to describe only two of them.

The first approach, due to Adamjan, Arov and Krein [1], is applicable only to the special case where R and X are scalar-valued; then the optimal X is unique and $R - X$ is a scalar times an inner function (cf. Prop. 5.1). As in Lemma 5.2 let Γ denote the Hankel operator with symbol R , and consider the self-adjoint operator

$$\Gamma\Gamma^*: \mathbf{H}_2^\perp \rightarrow \mathbf{H}_2^\perp.$$

Let λ^2 denote the maximum eigenvalue of $\Gamma\Gamma^*$, let f be a corresponding eigenvector, and define $g := \lambda^{-1}\Gamma^*f$. Observe that f and g satisfy the equations

$$\Gamma g = \lambda f, \quad \Gamma^* f = \lambda g;$$

such vectors form what is called a Schmidt pair for Γ .

THEOREM 5.5 [1]. *The optimal X equals $R - \lambda f/g$.*

Silverman and Bettayeb [81] employed this formula together with state-space realizations to get a simple way to compute the optimal X . As in (28)–(30) let $C(s - A)^{-1}B$ be a minimal realization of the antistable part of $R(s)$ and let L_c and L_o denote the controllability and observability gramians. By Lemma 5.2 λ^2 (defined above) equals the maximum eigenvalue of $L_c L_o$; let w be a corresponding eigenvector and define $v := \lambda^{-1}L_o w$.

COROLLARY 5.1 [81]. *The optimal X is given*

$$X(s) = R(s) - \lambda[A, w, C, 0]/[-A', v, B', 0].$$

The second approach, due to Ball and Helton [4], applies to the general matrix-valued problem. The theory of suboptimal X 's is simpler than the theory of optimal ones, so a characterization will be presented of all X 's in \mathbf{RH}_∞ such that

$$\|R - X\|_\infty \leq \beta,$$

where β can be any positive number greater than $\text{dist}(R, \mathbf{RH}_\infty)$. To simplify notation slightly, scale R and X by the factor β^{-1} ; now $\text{dist}(R, \mathbf{RH}_\infty) < 1$ and the problem is to find all X 's in \mathbf{RH}_∞ such that $\|R - X\|_\infty \leq 1$. Or, in terms of $S := R - X$, the problem is to find all S 's in \mathbf{RL}_∞ such that $R - S \in \mathbf{RH}_\infty$ and $\|S\|_\infty \leq 1$.

An outline of the Ball-Helton theory takes three steps. First, instead of looking at \mathbf{RL}_∞ -matrices, we look at graphs of operators. Consider the restriction to \mathbf{H}_2 of the Laurent operator induced by R :

$$M_R|_{\mathbf{H}_2}: \mathbf{H}_2 \rightarrow \mathbf{L}_2, \quad f \rightarrow Rf.$$

The graph of this operator, which we shall call the *graph* of R , is the set of ordered pairs (Rf, f) in $\mathbf{L}_2 \times \mathbf{H}_2$. Let us write these ordered pairs as 2-vectors $\begin{pmatrix} Rf \\ f \end{pmatrix}$, so that the graph has the representation

$$\mathbf{G}_R := \begin{bmatrix} R \\ I \end{bmatrix} \mathbf{H}_2.$$

If $R - S \in \mathbf{RH}_\infty$ and $\|S\|_\infty \leq 1$, what properties does the graph of S have? First, the condition $R - S \in \mathbf{RH}_\infty$ restricts the graph to being contained in a certain subspace of $\mathbf{L}_2 \times \mathbf{H}_2$, namely

$$\mathbf{W} := \begin{bmatrix} I & R \\ 0 & I \end{bmatrix} (\mathbf{H}_2 \times \mathbf{H}_2).$$

This is easily seen as follows:

$$\begin{aligned} \mathbf{G}_S &= \begin{bmatrix} S \\ I \end{bmatrix} \mathbf{H}_2 \\ &= \begin{bmatrix} I & R \\ 0 & I \end{bmatrix} \begin{bmatrix} S - R \\ I \end{bmatrix} \mathbf{H}_2 \\ &\subset \mathbf{W}. \end{aligned}$$

Second, the graph of S is *steep*, meaning that

$$(46) \quad \text{if } \begin{pmatrix} f \\ g \end{pmatrix} \in \mathbf{G}_S, \quad \text{then } \|f\|_2 \leq \|g\|_2.$$

(The slope of the line in the plane through the origin and the point $(\|f\|_2, \|g\|_2)$ is at least 45 degrees.) Fact (46) follows immediately from the condition $\|S\|_\infty \leq 1$. These two properties characterize the graph.

LEMMA 5.3 [4]. *Let S be an \mathbf{RL}_∞ -matrix (of the same dimensions as R). Then $\|S\|_\infty \leq 1$ and $R - S \in \mathbf{RH}_\infty$ if and only if the graph of S is steep and is contained in \mathbf{W} .*

The second step in the theory is to represent \mathbf{W} in a way which is useful for characterizing steep graphs. For this representation introduce the matrix

$$J := \begin{bmatrix} I & 0 \\ 0 & -I \end{bmatrix}.$$

(The dimension of the upper left unit matrix equals the number of rows of R ; the dimension of the lower right, the number of columns.) A square matrix L in \mathbf{RL}_∞ (of appropriate dimensions) is *J-unitary* if $L^* J L = J$.

LEMMA 5.4 [4]. *There exists a J-unitary matrix L such that $L(\mathbf{H}_2 \times \mathbf{H}_2) = \mathbf{W}$.*

Lemma 5.4 is a generalization of Beurling's theorem [46]. The useful fact about L is that, because it is *J-unitary*, it maps a steep subspace of $\mathbf{H}_2 \times \mathbf{H}_2$ into a steep subspace of \mathbf{W} (in fact, it is a one-to-one correspondence between such subspaces). Thus, if $Y \in \mathbf{RH}_\infty$ and $\|Y\|_\infty \leq 1$, then LG_Y is a steep subspace of \mathbf{W} .

The third step in the theory is to combine Lemmas 5.3 and 5.4 to obtain the main result, which is stated in terms of X .

THEOREM 5.6 [4]. *The set of all X 's in \mathbf{RH}_∞ such that $\|R - X\|_\infty \leq 1$ is parametrized by the formula*

$$\begin{aligned} X &= R - X_1 X_2^{-1}, & \begin{bmatrix} X_1 \\ X_2 \end{bmatrix} &= L \begin{bmatrix} Y \\ I \end{bmatrix}, \\ Y &\in \mathbf{RH}_\infty, & \|Y\|_\infty &\leq 1. \end{aligned}$$

Ball and Ran [5], [6] showed how to obtain a state-space realization of L . A summary of their procedure is as follows. As above, let $[A, B, C, 0]$ be a minimal realization of the antistable part of $R(s)$ and let L_c and L_o denote the controllability and observability gramians. (Recall that R has been scaled so that its distance to \mathbf{RH}_∞ is less than 1.) Then a realization of L is

$$L(s) = [A, B, \underline{C}, I],$$

where

$$\begin{aligned} \underline{A} &:= \begin{bmatrix} A & 0 \\ 0 & -A' \end{bmatrix}, \\ \underline{B} &:= \begin{bmatrix} -L_c & I \\ I & -L_o \end{bmatrix} \begin{bmatrix} I - L_o L_c & 0 \\ 0 & I - L_c L_o \end{bmatrix}^{-1} \begin{bmatrix} C' & 0 \\ 0 & B \end{bmatrix}, \\ \underline{C} &:= \begin{bmatrix} C & 0 \\ 0 & -B' \end{bmatrix}. \end{aligned}$$

Other approaches to the best approximation problem are those of Glover [43] (based on [2]) and Chang and Pearson [8] (based on [16]).

6. A numerical example. The purpose of this section is to elucidate the theory of §§ 3–5 by carrying out a numerical example of the tracking problem of § 2. With reference to Fig. 5 we take the unstable nonminimum phase plant

$$P(s) = \frac{s-1}{s(s-2)}.$$

The weighting factor ρ in (6) and the weighting filter W in Fig. 5 are as follows:

$$\rho = 1, \quad W(s) = \frac{s+1}{10s+1}.$$

The Bode magnitude plot of W is nearly 0 db up to the frequency .1, so this choice of W reflects a family of reference signals having their energy concentrated in the frequency band $[0, .1]$. With this choice of P , W and ρ , the transfer matrix G in Fig. 2 is determined from (7) to be

$$\begin{aligned} G(s) &= \begin{bmatrix} G_{11}(s) & G_{12}(s) \\ G_{21}(s) & G_{22}(s) \end{bmatrix} \\ &= \left[\begin{array}{c|c} \frac{s+1}{10s+1} & -\frac{s-1}{s(s-2)} \\ 0 & 1 \\ \hline \frac{s+1}{10s+1} & 0 \\ 0 & \frac{s-1}{s(s-2)} \end{array} \right]. \end{aligned}$$

The first step in computing a controller is to obtain the matrices T_i ($i = 1-3$) in the equivalent model-matching problem. We shall use the state-space method of § 4

(although this is not the easiest way for this simple example). We begin with a minimal realization of G :

$$G(s) = [A, B, C, D],$$

$$A = \begin{bmatrix} -1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 2 \end{bmatrix}, \quad B = [B_1 \ B_2] = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 1 \end{bmatrix},$$

$$C = \begin{bmatrix} C_1 \\ C_2 \end{bmatrix} = \begin{bmatrix} .09 & -.5 & -.5 \\ 0 & 0 & 0 \\ .09 & 0 & 0 \\ 0 & .5 & .5 \end{bmatrix}, \quad D = \begin{bmatrix} D_{11} & D_{12} \\ D_{21} & D_{22} \end{bmatrix} = \begin{bmatrix} .1 & 0 \\ 0 & 1 \\ .1 & 0 \\ 0 & 0 \end{bmatrix}.$$

Now F and H are chosen so that the matrices

$$A_F := A + B_2 F, \quad A_H := A + H C_2$$

are stable. The exact locations of the eigenvalues are not important for the purpose at hand; the choice

$$F = [0 \ .5 \ -4.5], \quad H = \begin{bmatrix} 0 & 0 \\ 0 & 1 \\ 0 & -9 \end{bmatrix}$$

yields

$$A_F = \begin{bmatrix} -1 & 0 & 0 \\ 0 & .5 & -4.5 \\ 0 & .5 & -2.5 \end{bmatrix}, \quad A_H = \begin{bmatrix} -1 & 0 & 0 \\ 0 & .5 & .5 \\ 0 & -4.5 & -2.5 \end{bmatrix},$$

which both have spectrum $\{-1, -1, -1\}$. Then (19) produces

$$T_1(s) = T_3(s) = \begin{bmatrix} \frac{s+1}{10s+1} \\ 0 \end{bmatrix}, \quad T_2(s) = \begin{bmatrix} -\frac{s-1}{(s+1)^2} \\ \frac{s(s-2)}{(s+1)^2} \end{bmatrix}.$$

Theorem 5.1 can now be used to show that there does indeed exist an optimal proper controller: $T_2(j\omega)$ and $T_3(j\omega)$ both have rank 1 for all $0 \leq \omega \leq \infty$.

The second step is to compute an upper bound γ for the infimal model-matching error α defined in (20). Recall from Theorem 5.4 that $\gamma > \alpha$ if and only if the following three conditions hold:

$$(47) \quad \|Y\|_\infty < \gamma, \quad \|Z\|_\infty < 1, \quad \text{dist}(R, H_\infty) < 1.$$

The matrices Y , Z and R are computed as in Theorem 5.4. A simple way to do the inner-outer factorization (39) of T_2 is first to get

$$\tilde{T}_2(s) T_2(s) = \frac{s^4 - 5s^2 + 1}{(-s+1)^2(s+1)^2}.$$

Solving $\tilde{T}_2 T_2 = U_o^\sim U_o$ for an outer function U_o gives

$$U_o(s) = \frac{s^2 + \sqrt{7}s + 1}{(s+1)^2}.$$

Then the inner factor U_i is

$$U_i = T_2 U_o^{-1}, \quad U_i(s) = \frac{1}{s^2 + \sqrt{7}s + 1} \begin{bmatrix} -s + 1 \\ s(s-2) \end{bmatrix}.$$

The matrix Y defined in (40) is determined next:

$$(48) \quad Y(s) = \frac{s+1}{(10s+1)(s^4-5s^2+1)} \begin{bmatrix} s^2(s^2-4) \\ -s(s+1)(s-2) \end{bmatrix}.$$

From (47) the value of γ must be at least $\|Y\|_\infty$. From (48) we get

$$Y^\sim(s) Y(s) = \frac{s^2(s^2-4)(s^2-1)}{(100s^2-1)(s^4-5s^2+1)}.$$

The spectral factor of $Y^\sim Y$ is

$$(49) \quad \frac{s(s+2)(s+1)}{(10s+1)(s^2+\sqrt{7}s+1)}.$$

Thus $\|Y\|_\infty$ equals the H_∞ -norm of the function (49), which can be read off its Bode magnitude plot. This yields $\|Y\|_\infty = .1683$. Thus γ must be at least .1683. It turns out that for $\gamma = .2$ the distance from R to H_∞ is greater than 1, violating (47). We shall show that (47) holds for $\gamma = .3$.

The spectral factor of $\gamma^2 - Y^\sim Y$ is

$$Y_o(s) = \frac{2.828s^3 + 7.615s^2 + 3.165s + .3}{(10s+1)(s^2+\sqrt{7}s+1)}.$$

Since $T_3 Y_o^{-1}$ is already co-outer, in (41) we can take $V_{co} = T_3 Y_o^{-1}$ and $V_{ci} = 1$. Then from (42) $Z = 0$, and from (43) $Z_{co} = 1$. Finally, from (44) we get

$$R(s) = \frac{(s+1)^2(s^2+\sqrt{7}s+1)}{(s^2-\sqrt{7}s+1)(2.828s^3 + 7.615s^2 + 3.165s + .3)}.$$

We shall compute the distance from R to H_∞ using Theorem 5.2 and Lemma 5.2. The antistable part of R is

$$\frac{.9267}{s-2.189} - \frac{.8217}{s-.4569},$$

which has the realization

$$A = \begin{bmatrix} 2.189 & 0 \\ 0 & .4569 \end{bmatrix}, \quad B = \begin{bmatrix} .9267 \\ -.8217 \end{bmatrix}, \quad C = [1 \quad 1].$$

The Lyapunov equations (31) and (32) are readily solved, giving

$$L_c = \begin{bmatrix} .1962 & -.2878 \\ -.2878 & .7389 \end{bmatrix}, \quad L_o = \begin{bmatrix} .2284 & .3780 \\ .3780 & 1.094 \end{bmatrix}.$$

The distance equals the square root of the largest eigenvalue of $L_c L_o$:

$$\text{dist}(R, H_\infty) = .7907.$$

Thus (47) is satisfied for $\gamma = .3$ and we conclude that $.2 < \alpha < .3$. We could at this point find a better estimate for α by reducing γ and checking (47) again, but we shall instead complete the computation with $\gamma = .3$.

The third step is to find the closest H_∞ -function X to R . Since these functions are scalar-valued, we can use Proposition 5.2. We already have that $\lambda^2 (= .7907^2)$ equals the maximum eigenvalue of $L_c L_o$; a corresponding eigenvector is

$$w = [1 \quad -2.862]'$$

Then $v := \lambda^{-1} L_o w$ equals

$$[-1.079 \quad -3.483]'$$

The formula in Proposition 5.2 yields

$$X(s) = \frac{(s^2 + \sqrt{7}s + 1)(.7170s^2 + 1.912s + .7628)}{(.3206s + 1)(2.828s^3 + 7.615s^2 + 3.165s + .3)}$$

The fourth step is to solve (45) for a Q in \mathbf{RH}_∞ . The matrix Q has dimensions 1×2 :

$$Q = [Q_1 \quad Q_2].$$

Equation (45) determines Q_1 uniquely:

$$Q_1(s) = \frac{(s+1)(.7170s^2 + 1.912s + .7628)}{(.3206s + 1)(s^2 + \sqrt{7}s + 1)},$$

whereas Q_2 is unconstrained, and hence may be taken to be zero.

Finally, the controller K is computed using Theorem 3.1. We can determine M_2 , N_2 , X_2 and Y_2 from (14), (15), (17), (18) and then get K from (10):

$$K = [C_1 \quad C_2],$$

$$C_1(s) = -\frac{(s+1)^3(.7170s^2 + 1.912s + .7628)}{(.3206s + 1)(s^2 + \sqrt{7}s + 1)(s^2 + 6s - 23)},$$

$$C_2(s) = -\frac{41s - 1}{s^2 + 6s - 23}.$$

(The reader will have noticed that C_1 is unstable, so the controller cannot be implemented as shown in Fig. 5. However, the theory guarantees that C_2 contains the unstable factor of C_1 . This common unstable factor would be moved past the summing junction into the loop.)

The properties of this design are illustrated in the Bode magnitude plots of Fig. 7. The transfer function, say H_1 , from reference r to tracking error $r - v$ has magnitude less than -10 db over the frequency band $[0, .1]$ of r (smaller tracking error could be obtained by reducing the weighting ρ on control energy), it peaks to about 4 db outside the operating band, and it rolls off to 0 db at high frequency, as it must for a proper controller. This sort of shape is characteristic of H_∞ designs. The transfer function, say H_2 , from r to u has a zero at $s = 0$ because P has a pole there. The actual quantity being minimized in this example is the H_∞ -norm of the transfer matrix

$$\begin{bmatrix} H_1 W \\ H_2 W \end{bmatrix}$$

from w to $(r - v, u)'$. This norm equals the supremum of

$$(50) \quad (|H_1(j\omega)|^2 + |H_2(j\omega)|^2)^{1/2} |W(j\omega)|$$

over all ω . For our design the function (50) is very nearly flat at -12 db (the supremum of (50) must be greater than -14 db corresponding to the bound $\alpha > .2$).

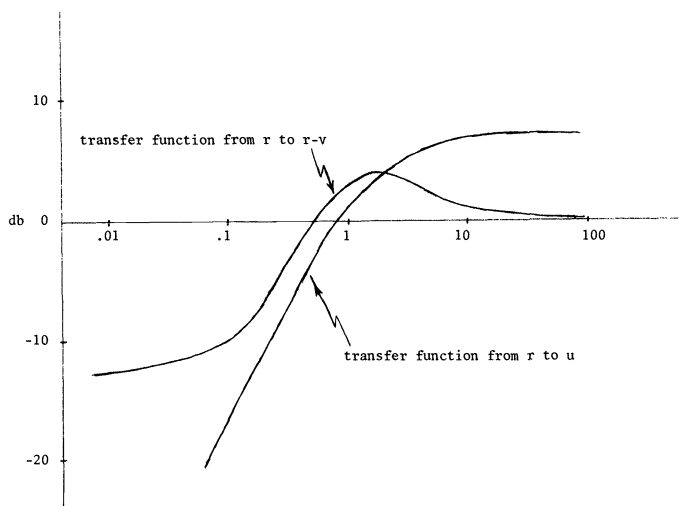


FIG. 7. Bode magnitude plots.

7. Achievable performance. For some simple examples of the standard problem it is possible to obtain useful bounds on achievable performance, sometimes even to characterize achievable performance exactly. Such results can shed light on what properties of a system affect its performance. This section presents three illustrative examples.

Figure 8 shows a feedback system with a disturbance signal w referred to the output of the plant P . As usual, P is strictly proper and K is proper. The transfer matrix from w to y is the *sensitivity matrix* $S := (I - PK)^{-1}$.

Suppose first that the spectrum of w is confined to a prespecified interval of frequencies $[-\omega_1, \omega_1]$, $\omega_1 > 0$. Then the problem of attenuating the effect of w on the output y of the plant is equivalent to the problem of making $\sigma_{\max}[S(j\omega)]$ uniformly small on the interval $[-\omega_1, \omega_1]$. Let χ denote the characteristic function of this interval, i.e.,

$$\chi(j\omega) = 1, \quad |\omega| \leq \omega_1, \quad \chi(j\omega) = 0, \quad |\omega| > \omega_1.$$

Then the maximum value of $\sigma_{\max}[S(j\omega)]$ over the interval $[-\omega_1, \omega_1]$ equals the L_∞ -norm $\|\chi S\|_\infty$. It may happen that as we try to make $\|\chi S\|_\infty$ smaller and smaller, the global bound $\|S\|_\infty$ becomes larger and larger. This is unpleasant because a large value of $\|S\|_\infty$ means the system has poor stability margin. (Think of the scalar-valued case: if $\|S\|_\infty$ is large, then the Nyquist plot of PK passes near the critical point.)

The first result says that if P is minimum phase, then $\|\chi S\|_\infty$ can be made as small as desired while $\|S\|_\infty$ is simultaneously maintained less than any bound δ . Of course δ must be greater than unity since $\|S\|_\infty \geq 1$ for every stabilizing K .

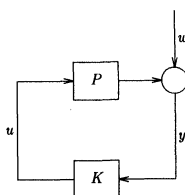


FIG. 8. Disturbance attenuation.

THEOREM 7.1 [67], [99]. *If P has a right-inverse which is analytic in $\operatorname{Re} s \geq 0$, then for every $\varepsilon > 0$ and $\delta > 1$ there exists a stabilizing K such that*

$$\|\chi S\|_\infty < \varepsilon, \quad \|S\|_\infty < \delta.$$

On the other hand, if P has a zero in the right half-plane, then $\|S\|_\infty$ must necessarily increase without limit if $\|\chi S\|_\infty$ tends to zero. This might be described as the “waterbed effect.”

THEOREM 7.2 [33], [35], [39]. *If at some point in $\operatorname{Re} s > 0$ the rank of P is less than the number of its rows, then there exists a positive real number a such that for every stabilizing K*

$$\|\chi S\|_\infty \|S\|_\infty^a > 1.$$

For the third result consider, with regard to Fig. 8 again, the problem of attenuating the effect of w (no longer restricted to be bandlimited) on the control signal u ; that is, the problem is to achieve feedback stability by a controller which limits as much as possible the control effort. The transfer matrix from w to u equals KS , so the objective is to minimize $\|KS\|_\infty$. The case where P is stable is trivial: an optimal K is $K = 0$. So we suppose P is not stable. For technical reasons it is assumed that P has no poles on the imaginary axis; thus P belongs to \mathbf{RL}_∞ but not \mathbf{RH}_∞ . Let Γ denote the Hankel operator with symbol P and let $\sigma_{\min}(\Gamma)$ denote the smallest (nonzero) singular value of Γ .

THEOREM 7.3 [44], [87]. *If P belongs to \mathbf{RL}_∞ but not \mathbf{RH}_∞ , then the minimum value of $\|KS\|_\infty$ over all stabilizing K 's equals the reciprocal of $\sigma_{\min}(\Gamma)$.*

8. Comparison with the Wiener–Hopf approach. In the model-matching problem posed in § 2 and solved in § 5, the criterion is to minimize the \mathbf{H}_∞ -norm of the error transfer matrix $T_1 - T_2 Q T_3$. It is evident from § 5 that, at least in the matrix case, optimal Q 's are not very easy to compute. By way of contrast the Wiener–Hopf approach to the model-matching problem is to minimize the \mathbf{H}_2 -norm of the error transfer matrix. (In Fig. 4, if w is standard white noise, then the root-mean-square value of z equals the \mathbf{H}_2 -norm of the transfer matrix from w to z .) It is relatively easy to compute optimal Q 's for the \mathbf{H}_2 -criterion. Therefore it is perhaps legitimate to ask if the computational effort required for the \mathbf{H}_∞ approach is worthwhile. How much better is the \mathbf{H}_∞ solution than the \mathbf{H}_2 solution?

To give one possible answer to this question we consider for simplicity the case where T_i ($i = 1-3$) are scalar-valued, in which case we may assume that $T_3 = 1$ by redefining T_2 . The function T_2 is assumed not to be zero anywhere on the extended imaginary axis (so that the following two optima exist). Let Q_2 denote the optimal solution for the \mathbf{H}_2 criterion

$$\|T_1 - T_2 Q\|_2 = \text{minimum},$$

and let Q_∞ denote the optimal solution for the \mathbf{H}_∞ criterion

$$\|T_1 - T_2 Q\|_\infty = \text{minimum}.$$

If we were to use the \mathbf{H}_2 solution, the supremal value of $\|z\|_2$ over all $\|w\|_2 \leq 1$ would equal $\|T_1 - T_2 Q_2\|_\infty$. Thus the above question can be rephrased as follows: How large can the ratio

$$(51) \quad \|T_1 - T_2 Q_2\|_\infty / \|T_1 - T_2 Q_\infty\|_\infty$$

be?

Let k denote the number of zeros of T_2 in the right half-plane.

PROPOSITION 8.1. *The supremum of the ratio (51) equals $2k$.*

Here the supremum is over all T_i 's in \mathbf{RH}_∞ such that T_2 has k zeros in $\operatorname{Re} s > 0$ (and no zeros on the extended imaginary axis). The idea of the proof of Proposition 8.1 is as follows. We may suppose without loss of generality that T_2 is an inner function: just absorb the outer factor into Q . Then for Q in \mathbf{H}_2 the \mathbf{H}_2 -norm of $T_1 - T_2 Q$ equals the \mathbf{L}_2 -norm of $T_1 T_2^{-1} - Q$. Hence Q_2 equals the projection of $T_1 T_2^{-1}$ onto \mathbf{H}_2^\perp . Denote this projection by T . Thus the \mathbf{H}_∞ -norm of $T_1 - T_2 Q_2$ equals the \mathbf{L}_∞ -norm of T . Similarly, the \mathbf{H}_∞ -norm of $T_1 - T_2 Q_\infty$ equals the norm of the Hankel operator with symbol T . Denote this operator by Γ_T . The function T has at most k poles in $\operatorname{Re} s > 0$. Glover proved ([43, Cor. 9.3]) that

$$\|T\|_\infty / \|\Gamma_T\| \leq 2k,$$

and Jonckheere et al. [52] showed that this bound could be approached as closely as desired by suitable choice of T .

We conclude that the improvement in the \mathbf{H}_∞ solution over the \mathbf{H}_2 solution, for the \mathbf{H}_∞ criterion, can be arbitrarily great.

9. Summary. The standard problem, which includes the robust stabilization problem, is to minimize the \mathbf{H}_∞ -norm of the closed-loop transfer function for a fixed known plant. Under parametrization of the controller, the standard problem reduces to one of model-matching: minimize the \mathbf{H}_∞ -norm of an affine function. The model-matching problem reduces in turn to a sequence of best approximation problems: approximate in \mathbf{L}_∞ -norm an unstable transfer function by a stable one.

10. Conclusion. The standard problem is well understood and software exists for its solution [12], [21]; the software uses standard routines (such as singular-value decompositions and solving Lyapunov equations). The \mathbf{H}_∞ problem with plant uncertainty, how to achieve frequency-domain performance specifications in the face of plant uncertainty, is a current area of research.

REFERENCES

- [1] V. M. ADAMJAN, D. Z. AROV AND M. G. KREIN, *Analytic properties of Schmidt pairs for a Hankel operator and the generalized Schur-Takagi problem*, Math. USSR-Sb., 15 (1971), pp. 31-73.
- [2] ———, *Infinite block Hankel matrices and related extension problems*, Amer. Math. Soc. Transl., 111 (1978), pp. 133-156.
- [3] A. C. ANTOUNLAS, *A new approach to synthesis problems in linear system theory*, IEEE Trans. Automat. Control, AC-30 (1985), pp. 465-473.
- [4] J. A. BALL AND J. W. HELTON, *A Beurling-Lax theorem for the Lie group $U(m, n)$ which contains most classical interpolation theory*, J. Operator Theory, 9 (1983), pp. 107-142.
- [5] J. A. BALL AND A. C. M. RAN, *Optimal Hankel norm model reductions and Wiener-Hopf factorizations I: the canonical case*, Tech. Report, Dept. of Mathematics, Virginia Polytechnic Inst. and State Univ., Blacksburg, VA, 1985.
- [6] ———, *Hankel norm approximation of a rational matrix function in terms of its realization*, Tech. Report, Dept. of Mathematics, Virginia Tech., 1985.
- [7] S. BOYD AND C. A. DESOER, *Subharmonic functions and performance bounds on linear time-invariant feedback systems*, Memo. No. UCB/ERL M84/51, Electronic Research Lab., Univ. of California, Berkeley, CA, 1984.
- [8] B. C. CHANG AND J. B. PEARSON, *Optimal disturbance reduction in linear multivariable systems*, IEEE Trans. Automat. Control, AC-29 (1984), pp. 880-887.
- [9] B. S. CHEN, *Controller synthesis of optimal sensitivity: multivariable case*, Proc. IEEE part D, 131 (1984), pp. 47-51.
- [10] M. J. CHEN AND C. A. DESOER, *Necessary and sufficient condition for robust stability of linear distributed feedback systems*, Internat. J. Control, 35 (1982), pp. 255-267.
- [11] L. CHENG AND J. B. PEARSON, *Frequency domain synthesis of multivariable linear regulators*, IEEE Trans. Automat. Control, AC-23 (1978), pp. 3-15.

- [12] C. C. CHU, *H_∞ optimization and robust multivariable control*, Ph.D. thesis, Dept. of Electrical Engineering, Univ. of Minnesota, Minneapolis, MN, 1985.
- [13] C. C. CHU AND J. C. DOYLE, *The general distance problem in H_∞ synthesis*, Proc. Conference on Decision and Control, 1985.
- [14] R. F. CURTAIN AND K. GLOVER, *Robust stabilization of infinite-dimensional systems by finite-dimensional controllers*, Systems Control Lett., to appear.
- [15] ———, *Robust stabilization of infinite-dimensional systems by finite-dimensional controllers: derivations and examples*, Proc. Symposium on Math. Theory of Networks and Systems, Stockholm, 1984.
- [16] P. DELSARTE, Y. GENIN AND Y. KAMP, *The Nevanlinna–Pick problem for matrix-valued functions*, SIAM J. Appl. Math., 36 (1979), pp. 47–61.
- [17] C. A. DESOER AND M. VIDYASAGAR, *Feedback systems: input–output properties*, Academic Press, New York, 1975.
- [18] C. A. DESOER, R. W. LIU, J. MURRAY AND R. SAEKS, *Feedback system design: the fractional representation approach*, IEEE Trans. Automat. Control, AC-25 (1980), pp. 399–412.
- [19] J. C. DOYLE AND G. STEIN, *Multivariable feedback design: concepts for a classical modern synthesis*, IEEE Trans. Automat. Control, AC-26 (1981), pp. 4–16.
- [20] J. C. DOYLE, *Synthesis of robust controllers and filters*, Proc. Conference on Decision and Control, 1983.
- [21] ———, *Lecture Notes in Advances in Multivariable Control*, Office of Naval Research/Honeywell Workshop, Minneapolis, MN, 1984.
- [22] ———, *Robust stability with structured perturbations*, Proc. Symposium on Math. Theory of Networks and Systems, Stockholm, 1985.
- [23] J. C. DOYLE AND C. C. CHU, *Matrix interpolation and H_∞ performance bounds*, Proc. American Control Conf., 1985.
- [24] J. C. DOYLE, *Structured uncertainty in control systems*, IFAC Workshop on Model Error Concepts and Compensation, Boston, MA, 1985.
- [25] P. L. DUREN, *Theory of H_p Spaces*, Academic Press, New York, 1970.
- [26] A. FEINTUCH, P. KHARGONEKAR AND A. TANNENBAUM, *On the sensitivity minimization problem for linear time-varying systems*, this Journal, to appear.
- [27] A. FEINTUCH AND B. A. FRANCIS, *Uniformly optimal control of linear time-varying systems*, Systems Control Lett., 5 (1984), pp. 67–71.
- [28] A. FEINTUCH AND A. TANNENBAUM, *Gain optimization for distributed systems*, Systems Control Lett., 6 (1986), pp. 295–302.
- [29] A. FEINTUCH AND B. A. FRANCIS, *Uniformly optimal control of linear systems*, Automatica, 21 (1986), pp. 563–574.
- [30] D. S. FLAMM AND S. K. MITTER, *Progress on H_∞ optimal sensitivity for delay systems*, Tech. Report LIDS-P-1513, Massachusetts Inst. of Technology, Cambridge, MA, 1985.
- [31] C. FOIAS, A. TANNENBAUM AND G. ZAMES, *Weighted sensitivity minimization for delay systems*, Tech. Report, Dept. of Electrical Engineering, McGill Univ., Montreal, 1985.
- [32] Y. K. FOO AND I. POSTLETHWAITE, *An H_∞-minimax approach to the design of robust control systems*, Systems Control Lett., 5 (1984), pp. 81–88.
- [33] B. A. FRANCIS, *Notes on H_∞-optimal linear feedback systems*, 1983, lectures given at Linköping Univ., Linköping, Sweden.
- [34] B. A. FRANCIS AND G. ZAMES, *Design of H_∞-optimal multivariable feedback systems*, Proc. Conference on Decision and Control, 1983.
- [35] ———, *On H_∞-optimal sensitivity theory for siso feedback systems*, IEEE Trans. Automat. Control, AC-29 (1984), pp. 9–16.
- [36] B. A. FRANCIS, J. W. HELTON AND G. ZAMES, *H_∞-optimal feedback controllers for linear multivariable systems*, IEEE Trans. Automat. Control, 29 (1984), pp. 888–900.
- [37] B. A. FRANCIS, *Optimal disturbance attenuation with control weighting*, in Lecture Notes in Control and Information Sci., Vol. 66, Springer-Verlag, Berlin, New York, 1985, Proc. 1984 Twente Workshop on Systems and Optimization.
- [38] J. FREUDENBERG, *Issues in frequency domain feedback control*, Ph.D. thesis, Dept. of Electrical Engineering, Univ. of Illinois, Urbana, IL, 1985.
- [39] J. FREUDENBERG AND D. LOOZE, *Right half-plane poles and zeros and design trade-offs in feedback systems*, IEEE Trans. Automat. Control, AC-30 (1985), pp. 555–565.
- [40] ———, *An analysis of H_∞-optimization design methods*, IEEE Trans. Automat. Control, to appear.
- [41] C. GANESH AND J. B. PEARSON, *Design of optimal control systems with stable feedback*, Tech. Report #8514, Dept. of Electrical Engineering, Rice Univ., Houston, TX, 1985.
- [42] T. GEORGIU AND P. KHARGONEKAR, *A constructive algorithm for sensitivity optimization of periodic systems*, this Journal, 25 (1987), pp. 334–340.

- [43] K. GLOVER, *All optimal Hankel-norm approximations of linear multivariable systems and their L_∞ -error bounds*, Internat. J. Control, 39 (1984), pp. 1115-1193.
- [44] ———, *Robust stabilization of linear multivariable systems: relations to approximation*, Internat. J. Control, 43 (1986), pp. 741-766.
- [45] M. J. GRIMBLE, *Optimal H_∞ robustness and the relationship to LQG design problems*, Tech. Report ICU/55, ICU, Univ. of Strathclyde, Glasgow, 1984.
- [46] P. R. HALMOS, *A Hilbert Space Problem Book*, Springer-Verlag, Berlin, New York, 1982.
- [47] J. W. HELTON, *Operator theory and broadband matching*, Proc. Allerton Conf., 1976.
- [48] ———, *Worst case analysis in the frequency-domain: an H_∞ approach to control*, IEEE Trans. Automat. Control, AC-30 (1985), pp. 1154-1170.
- [49] ———, *Lecture notes*, NSF-CBMS Conf. on Optimization in Operator Theory, Analytic Function Theory and Electrical Engineering, Lincoln, NB, 1985.
- [50] J. W. HELTON AND D. F. SCHWARTZ, *A primer on the H_∞ disk method in frequency-domain design: control*, Tech. Report, Dept. of Mathematics, Univ. of California, San Diego, CA, 1985.
- [51] E. JONCKHEERE AND M. VERMA, *A spectral characterization of H_∞ -optimal feedback performance: the multivariable case*, Tech. Report, Dept. of Electrical Engineering, Univ. of Southern California, Los Angeles, CA, 1986.
- [52] E. A. JONCKHEERE, M. G. SAFONOV AND L. M. SILVERMAN, *Topology induced by the Hankel norm in the space of transfer matrices*, Proc. Conference on Decision and Control, 1981.
- [53] T. KAILATH, *Linear Systems*, Prentice-Hall, Englewood Cliffs, NJ, 1980.
- [54] P. KHARGONEKAR AND K. POOLLA, *Robust stabilization of distributed systems*, Automatica, to appear.
- [55] P. KHARGONEKAR, K. POOLLA AND A. TANNENBAUM, *Robust control of linear time-invariant plants using periodic compensation*, IEEE Trans. Automat. Control, AC-30 (1985), pp. 1088-1098.
- [56] P. KHARGONEKAR AND A. TANNENBAUM, *Noneuclidean metrics and the robust stabilization of systems with parameter uncertainty*, IEEE Trans. Automat. Control, AC-30 (1985), pp. 1005-1013.
- [57] P. KHARGONEKAR AND K. POOLLA, *Uniformly optimal control of linear time-invariant plants: nonlinear time-varying controllers*, Systems Control Lett., 6 (1986), pp. 303-308.
- [58] H. KIMURA, *Robust stabilization for a class of transfer functions*, IEEE Trans. Automat. Control, AC-29 (1984), pp. 788-793.
- [59] S. Y. KUNG AND D. W. LIN, *Optimal Hankel-norm model reductions: multivariable systems*, IEEE Trans. Automat. Control, AC-26 (1981), pp. 832-852.
- [60] H. KWAKERNAAK, *Minimax frequency-domain optimization of multivariable linear feedback systems*, IFAC World Congress, Budapest, 1984.
- [61] ———, *A polynomial approach to minimax frequency domain optimization of multivariable systems*, Tech. Rept. 529, Dept. Appl. Math., Twente Univ. Tech., Twente, 1985.
- [62] ———, *Minimax frequency domain performance and robustness optimization of linear feedback systems*, IEEE Trans. Automat. Control, AC-30 (1985), pp. 994-1004.
- [63] Z. NEHARI, *On bounded bilinear forms*, Ann. of Math. (2), 65 (1957), pp. 153-162.
- [64] C. N. NETT, C. A. JACOBSON AND M. J. BALAS, *A connection between state-space and doubly coprime fractional representations*, IEEE Trans. Automat. Control, AC-29 (1984), pp. 831-832.
- [65] C. N. NETT, *Algebraic aspects of linear control system stability*, Proc. Conference on Decision and Control, 1985.
- [66] S. O'YOUNG, *Performance trade-offs in the design of multivariable controllers*, Ph.D. thesis, Dept. of Electrical Engineering, Univ. of Waterloo, Waterloo, Ontario, Canada, 1985.
- [67] S. O'YOUNG AND B. A. FRANCIS, *Sensitivity trade-offs for multivariable plants*, IEEE Trans. Automat. Control, AC-30 (1985), pp. 625-632.
- [68] ———, *Optimal performance and robust stabilization*, Automatica, 1986, to appear.
- [69] A. PASCOAL AND P. KHARGONEKAR, *Remarks on weighted sensitivity minimization*, Tech. Report, Dept. of Electrical Engineering, Univ. of Minnesota, Minneapolis, MN, 1986.
- [70] L. PERNEBO, *An algebraic theory for the design of controllers for linear multivariable systems, parts I & II*, IEEE Trans. Automat. Control, AC-26 (1981), pp. 171-194.
- [71] I. POSTLETHWAITE AND Y. K. FOO, *All solutions, all-pass form solutions, and the "best" solutions to an H_∞ optimization problem in robust control*, Symposium on Math. Theory of Networks and Systems, Stockholm, 1985.
- [72] S. C. POWER, *Hankel Operators on Hilbert Space*, Pitman, London, 1982.
- [73] M. SAFONOV AND M. ATHANS, *A multiloop generalization of the circle criterion for stability margin analysis*, IEEE Trans. Automat. Control, AC-26 (1981), pp. 415-422.
- [74] M. G. SAFONOV AND B. S. CHEN, *Multivariable stability margin optimization with decoupling and output regulation*, Proc. IEE, Part D, 129 (1982), pp. 276-282.

- [75] M. G. SAFONOV, *L_∞ -optimal sensitivity versus stability margin*, Proc. Conference on Decision and Control, 1983.
- [76] ———, *Optimal diagonal scaling for infinity norm optimization*, Proc. American Control Conference, Boston, MA, 1985.
- [77] M. G. SAFONOV AND M. S. VERMA, *L_∞ sensitivity optimization and Hankel approximation*, IEEE Trans. Automat. Control, AC-30 (1985), 279–280.
- [78] I. W. SANDBERG, *An observation concerning the application of the contraction mapping fixed-point theorem and a result concerning the norm-boundedness of solutions of nonlinear functional equations*, Bell System Tech. J., 44 (1965), pp. 1809–1812.
- [79] D. SARASON, *Generalized interpolation in H_∞* , Trans. Amer. Math. Soc., 127 (1967), pp. 179–203.
- [80] A. SIDERIS AND M. G. SAFONOV, *Design of linear control systems for robust stability and performance*, Proc. IFAC Workshop on Model Error Concepts and Compensation, Boston, MA, 1985.
- [81] L. SILVERMAN AND M. BETTAYEB, *Optimal approximation of linear systems*, Proc. Joint Automat. Control Conf., 1980.
- [82] B. SZ.-NAGY AND C. FOIAS, *Harmonic Analysis of Operators on Hilbert Space*, North-Holland, Amsterdam, New York, 1970.
- [83] A. TANNENBAUM, *On the blending problem and parameter uncertainty in control theory*, Tech. Report, Dept. of Mathematics, Weizmann Inst. of Science, Rehovot, Israel, 1977.
- [84] ———, *Feedback stabilization of linear dynamical plants with uncertainty in the gain factor*, Internat. J. Control, 32 (1980), pp. 1–16.
- [85] ———, *Modified Nevanlinna-Pick interpolation of linear plants with uncertainty in the gain factor*, Internat. J. Control, 36 (1982), pp. 331–336.
- [86] M. VERMA AND E. JONCKHEERE, *L_∞ -compensation with mixed sensitivity as a broadband matching problem*, Systems Control Lett., 4 (1984), pp. 125–130.
- [87] M. S. VERMA, *Synthesis of infinity-norm optimal linear feedback systems*, Ph.D. thesis, Dept. of Electrical Engineering, Univ. of Southern California, Los Angeles, CA, 1985.
- [88] M. VIDYASAGAR AND H. KIMURA, *Robust controllers for uncertain linear multivariable systems*, Automatica, to appear.
- [89] M. VIDYASAGAR, *The graph metric for unstable plants and robustness estimates for feedback stability*, IEEE Trans. Automat. Control, AC-29 (1984), pp. 403–418.
- [90] ———, *Control System Synthesis: A Factorization Approach*, MIT Press, Cambridge, MA, 1985.
- [91] ———, *Filtering and robust regulation using a two-parameter controller*, Proc. Symposium on Math. Theory of Networks and Systems, Stockholm, 1985.
- [92] Z. Z. WANG AND J. B. PEARSON, *Regulation and optimal error reduction in linear multivariable systems*, Proc. IFAC World Congress, Budapest, 1984.
- [93] D. C. YOULA, H. A. JABR AND J. J. BONGIORNO JR., *Modern Wiener-Hopf design of optimal controllers: part II*, IEEE Trans. Automat. Control, AC-21 (1976), pp. 319–338.
- [94] G. ZAMES, *Nonlinear operators for system analysis*, Tech. Report 370, Research Lab. of Electronics, Massachusetts Inst. of Technology, Cambridge, MA, 1960.
- [95] ———, *On the input-output stability of nonlinear time-varying feedback systems, parts I & II*, IEEE Trans. Automat. Control, AC-11 (1966), pp. 228–238, pp. 465–477.
- [96] ———, *Optimal sensitivity and feedback; weighted seminorms, approximate inverses, and plant invariant schemes*, Proc. Allerton Conf., 1979.
- [97] ———, *Feedback and optimal sensitivity: model reference transformations, multiplicative seminorms, and approximate inverses*, IEEE Trans. Automat. Control, AC-23 (1981), pp. 301–320.
- [98] G. ZAMES AND B. A. FRANCIS, *Feedback, minimax sensitivity, and optimal robustness*, IEEE Trans. Automat. Control, AC-28 (1983), pp. 585–601.
- [99] G. ZAMES AND D. BENSOUSSAN, *Multivariable feedback, sensitivity, and decentralized control*, IEEE Trans. Automat. Control, AC-28 (1983), pp. 1030–1035.

OPTIMAL ADAPTIVE CONTROL AND CONSISTENT PARAMETER ESTIMATES FOR ARMAX MODEL WITH QUADRATIC COST*

HAN-FU CHEN† AND LEI GUO†

Abstract. We consider the multidimensional ARMAX model

$$A(z)y_n = B(z)u_n + C(z)w_n$$

with loss function

$$J(u) = \overline{\lim}_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n (y_i^T Q_1 y_i + u_i^T Q_2 u_i)$$

where the coefficients in the matrix polynomials $A(z)$, $B(z)$ and $C(z)$ are unknown. Conditions used here are: 1) stability of $A(z)$ and full rank of A_p ; 2) strictly positive realness of $C(z) - \frac{1}{2}I$, and 3) controllability and observability of a matrix triple consisting of coefficients in $A(z)$, $B(z)$ and Q_1 . On the basis of the estimates given by the stochastic gradient algorithm for unknown parameters an adaptive control law is recursively defined. It is proved that the parameter estimates are strongly consistent and the quadratic loss function reaches its minimum. This paper also includes some general theorems on parameter estimation, on which the results about adaptive control are essentially based.

Key words. stochastic systems, ARMAX model, stochastic adaptive control, quadratic cost, parameter estimation

AMS(MOS) subject classification. 93C40

1. Introduction and statement of problem. In recent years there has been considerable research effort on the parameter estimation and adaptive control problem for linear stochastic systems (see e.g. Goodwin et al. (1984)). Ljung (1977), Solo (1979), Chen (1981), (1982) and Lai and Wei (1982) showed various conditions guaranteeing strong consistency of parameter estimates given by different algorithms for stochastic systems without monitoring, while Goodwin et al. (1981) and Sin and Goodwin (1982) gave adaptive control making the system global stable and the tracking error minimal, but the parameter estimates given there in general, as shown by Becker et al. (1985), are inconsistent. The first step towards getting both consistency of estimates and asymptotic minimality of tracking errors was made by Caines and Lafortune (1984), Chen (1984) and Chen and Caines (1985). In their results the parameter estimates are proved strongly consistent but the tracking error is no longer minimal because of the disturbance artificially introduced to the reference signal. Recently, Chen and Guo (1985a), (1985b) have given an adaptive control under which not only the parameter estimates are strongly consistent, but also the long run average of tracking error reaches its minimum.

For stochastic adaptive control when a general quadratic loss function is considered, Kumar (1983), Hijab (1983) and Caines and Chen (1985) are concerned with the case where the unknown parameters are valued in a finite set, Chen and Caines (1984) and Chen (1985) deal with systems for which the consistent parameter estimates are available, and Samson (1983) considers bounded disturbance case. Recently for systems in state space representation with state completely observed, Chen and Guo

* Received by the editors July 1, 1985; accepted for publication (in revised form) March 20, 1986. This work was supported by the Science Fund of the Chinese Academy of Sciences.

† Institute of Systems Science, Academia Sinica, Beijing, People's Republic of China.

(1986) have given the optimal stochastic LQ control based on the least squares estimates for unknown parameters which may take arbitrary values in the Euclidean spaces of compatible dimensions.

In this paper we consider the general stochastic MIMO system (ARMAX model):

$$(1) \quad A(z)y_n = B(z)u_n + C(z)w_n$$

with quadratic loss function

$$(2) \quad J(u) = \overline{\lim}_{n \rightarrow \infty} \frac{1}{n} \sum_{i=0}^{n-1} (y_i^T Q_1 y_i + u_i^T Q_2 u_i),$$

where $Q_1 \geq 0$, $Q_2 > 0$ and the matrix polynomials in shift-back operator z

$$(3) \quad A(z) = I + A_1 z + \cdots + A_p z^p, \quad p \geq 0,$$

$$(4) \quad B(z) = B_1 z + B_2 z^2 + \cdots + B_q z^q, \quad q \geq 1,$$

$$(5) \quad C(z) = I + C_1 z + \cdots + C_r z^r, \quad r \geq 0$$

are of known orders p , q and r , respectively, and with unknown parameter θ denoting

$$(6) \quad \theta^T = [-A_1 \cdots -A_p \ B_1 \cdots B_q \ C_1 \cdots C_r]$$

by definition. We emphasize that A_i , B_j , C_k ($i = 1 \cdots p$, $j = 1 \cdots q$, $k = 1 \cdots r$) may be any matrices of compatible dimensions.

Let dimensions for y_n , u_n and w_n be m , l and m , respectively, $y_i = 0$, $u_i = 0$, $w_i = 0$ for $i < 0$, and let $\{w_n\}$ be a martingale difference sequence with respect to a family $\{\mathcal{F}_n\}$ of increasing σ -algebras, i.e., w_n is \mathcal{F}_n -measurable and $E(w_n | \mathcal{F}_{n-1}) = 0$. In addition, we assume that

$$(7) \quad \lim_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n w_i w_i^T = Q > 0$$

and

$$(8) \quad \sup_n E[\|w_n\|^2 | \mathcal{F}_{n-1}] < \infty \quad \text{a.s.}$$

where and hereafter $\|X\|$ denotes the maximum singular value of X .

At any time n , by use of the past input-output data $\{u_i, y_j, 0 \leq i \leq n-1, 0 \leq j \leq n\}$ we want 1) to estimate the unknown parameter θ and 2) to define adaptive control u_n^a minimizing the loss function (2). In this paper, for the case where $A(z)$ is stable, we give a complete solution of this problem in the sense that the consistency of parameter estimates and minimality of the loss function are achieved simultaneously. Although the results are established for adaptive control based on parameter estimates given by the stochastic gradient algorithm, the same results also hold for the case where the extended least squares algorithm is applied.

In § 2 we describe the optimal control for system (1) and (2) with known parameters, and in § 3 we define the algorithm for both parameter estimation and adaptive control and formulate the main theorem of this paper. For its proof we start with some general theorems on strong consistency of parameter estimates for systems without monitoring (§ 4). Then in § 5 we prove that they can be applied to the adaptive control system defined in § 3, and show that the loss function is really minimized.

2. Optimal control for systems with known parameters. The adaptive control law given later on is inspired by the optimal control for system (1), (2) with known parameters. So we first rewrite (1) in the state space form

$$(9) \quad x_{k+1} = Ax_k + Bu_k + Cw_{k+1},$$

$$(10) \quad y_k = Hx_k, \quad x_0^T = [y_0^T 0 \cdots 0]$$

and give a solution of optimal control, where

$$(11) \quad A = \begin{bmatrix} -A_1 & I & 0 & \cdots & 0 \\ & 0 & \ddots & & \vdots \\ \vdots & \vdots & \ddots & \ddots & 0 \\ -A_s & 0 & \cdots & & I \end{bmatrix}, \quad B = \begin{bmatrix} B_1 \\ \vdots \\ B_s \end{bmatrix},$$

$$(12) \quad C^T = [I \ C_1^T \ \cdots \ C_{s-1}^T], \quad H = \underbrace{[I \ 0 \ \cdots \ 0]}_{ms} m$$

with $s = p \vee q \vee (r+1)$ and $A_i = 0, B_j = 0, C_k = 0$ for $i > p, j > q, k > r$.

We note at once that the nonzero eigenvalues of A coincide with the reciprocals of zeros of $\det A(z)$ (Chen (1985)).

All conditions used in this paper are listed here.

(a) A_p is of full rank ($A_0 = I$ by definition) and $A(z)$ is stable, i.e. all zeros of $\det A(z)$ lie outside the closed unit disk.

(b) $C(z) - \frac{1}{2}I$ is strictly positive real, i.e.

$$C(e^{i\varphi}) + C^T(e^{-i\varphi}) - I > 0 \quad \forall \varphi \in [0, 2\pi].$$

(c) (A, B, D) is controllable and observable, where D is any matrix such that $D^T D = H^T Q_1 H$.

We first explain these conditions.

(1) The full rank of A_p is used to ensure $\deg(\det A(z)) = mp$ for identifiability.

(2) For the uncorrelated noise case $r = 0, C(z) = I$, condition (b) is automatically satisfied.

(3) Condition (c) implies that $A(z)$ and $B(z)$ have no common left factor, i.e. there are matrix polynomials $M(z)$ and $N(z)$ such that

$$(13) \quad A(z)M(z) + B(z)N(z) = I;$$

this is a consequence of Theorem 6.2-6 of Kailath (1980, p. 366). Also, condition (c) implies either A_s or B_s is not zero, which implies $r+1 \leq \max(p, q)$. So under condition (c) $s = p \vee q$.

(4) If condition (c) is fulfilled (stability of $A(z)$ is not required here), then there is a unique positive definite matrix solution S in the class of nonnegative definite matrices for the Riccati algebraic equation

$$(14) \quad S = A^T S A - A^T S B (Q_2 + B^T S B)^{-1} B^T S A + H^T Q_1 H,$$

and the matrix $A + BL$ is stable with

$$(15) \quad L = -(Q_2 + B^T S B)^{-1} B^T S A$$

(see, e.g. Anderson and Moore (1971)).

(5) Instead of condition (c), which is rather restrictive, we can directly assume (14), (15) for which the weaker conditions are sufficient and assume that $A(z), B(z)$ and $C(z)$ have no common left factor which is a natural condition for identifiability of the system.

The following lemma is not concerned with adaptive control but it shows the minimal value of the loss function and hints the form of adaptive control.

Throughout the paper, the relationship between two random quantities may have an exceptional set with probability 0, but sometimes we omit to write "a.s."

LEMMA 1. *If conditions (a) and (c) hold, then*

$$(16) \quad J(u) = \text{tr } SCQC^\tau + \overline{\lim}_{n \rightarrow \infty} \frac{1}{n} \sum_{i=0}^{n-1} (u_i - Lx_i)^\tau (Q_2 + B^\tau SB)(u_i - Lx_i) \quad \text{a.s.}$$

whenever u_i is \mathcal{F}_i -measurable and $\{u_i\} \in U$ with

$$(17) \quad U = \left\{ u: \sum_{i=1}^n \|u_i\|^2 = O(n), \quad \|u_n\|^2 = o(n), \text{ as } n \rightarrow \infty \quad \text{a.s.} \right\}.$$

The proof is given in Appendix 1.

This lemma tells us that the optimal control is $u_n = Lx_n$ and that the lower bound to the loss is

$$\min_{u \in U} J(u) = \text{tr } SCQC^\tau.$$

We now give a multidimensional version of a result from Lai and Wei (1982) which is used in the proof of Lemma 1 and will be repeatedly used in the sequel.

LEMMA 2. *Let f_i be \mathcal{F}_i -measurable random vectors and let $\{w_i, \mathcal{F}_i\}$ be a martingale difference sequence satisfying (8). Then as $n \rightarrow \infty$*

$$\sum_{i=1}^n f_i w_{i+1}^\tau = O(s_n^{1/2} \log^{(1/2)+\eta}(s_n + e)) \quad \forall \eta > 0 \quad \text{with } s_n \triangleq \sum_{i=1}^n \|f_i\|^2.$$

The proof is given in Appendix 1.

3. Main theorem. For estimating the unknown parameter θ we use the stochastic gradient algorithm defined by

$$(18) \quad \theta_{n+1} = \theta_n + \frac{\varphi_n}{r_n} (y_{n+1}^\tau - \varphi_n^\tau \theta_n),$$

$$(19) \quad \varphi_n^\tau = [y_n^\tau, \dots, y_{n-p+1}^\tau, u_n^\tau, \dots, u_{n-q+1}^\tau, y_n^\tau - \varphi_{n-1}^\tau \theta_{n-1}, \dots, y_{n-r+1}^\tau - \varphi_{n-r}^\tau \theta_{n-r}],$$

$$(20) \quad r_n = 1 + \sum_{i=1}^n \|\varphi_i\|^2, \quad r_0 = 1.$$

Denote by A_{in}, B_{jn}, C_{kn} the estimates given by θ_n for A_i, B_j, C_k , respectively, $i = 1 \dots p, j = 1 \dots q, k = 1 \dots r$. The state x_n is estimated by the adaptive filter

$$(21) \quad \begin{aligned} \hat{x}_{n+1} &= \hat{A}_n \hat{x}_n + \hat{B}_n u_n + \hat{C}_n (y_{n+1} - H \hat{A}_n \hat{x}_n - H \hat{B}_n u_n), \\ \hat{x}_0 &= [y_0^\tau 0 \dots 0]^\tau \end{aligned}$$

where \hat{A}_n, \hat{B}_n and \hat{C}_n are defined by (11) and (12) with A_i, B_j, C_k replaced by their estimates A_{in}, B_{jn}, C_{kn} , respectively, $i = 1 \dots p, j = 1 \dots q, k = 1 \dots r$.

Set

$$(22) \quad L_n = -(\hat{B}_n^\tau S_n \hat{B}_n + Q_2)^{-1} \hat{B}_n^\tau S_n \hat{A}_n,$$

where S_n is recursively defined by

$$(23) \quad S_n = \hat{A}_n^\tau S_{n-1} \hat{A}_n - \hat{A}_n^\tau S_{n-1} \hat{B}_n (Q_2 + \hat{B}_n^\tau S_{n-1} \hat{B}_n)^{-1} \hat{B}_n^\tau S_{n-1} \hat{A}_n + H^\tau Q_1 H,$$

with an arbitrary initial value $S_0 \geq 0$.

It is natural to guess that $L_n \hat{x}_n$ is something we should take as adaptive control, but, in fact, it may lead to an inconsistent estimate for θ . To avoid this trouble we use the randomly varying truncation technique and the attenuating excitation technique similar to those used in Chen and Guo (1986).

Take an arbitrary l -dimensional i.i.d. sequence $\{\varepsilon_n\}$ independent of $\{w_n\}$ and with properties

$$(24) \quad E\varepsilon_1 = 0, \quad E\varepsilon_1 \varepsilon_1^\tau = I, \quad E\|\varepsilon_1\|^3 < \infty.$$

Without loss of generality we assume $\mathcal{F}_n = \sigma\{w_i, i \leq n, \varepsilon_j, j \leq n\}$.

Then the random sequence $\{v_n\}$ will serve as the source of attenuating excitation, where by definition

$$(25) \quad v_1 = 0, \quad v_n = \frac{\varepsilon_n}{\log^{\varepsilon/2} n} \quad \forall n \geq 2, \quad \varepsilon \in \left(0, \frac{1}{4s(m+2)}\right).$$

From Theorem 3, which is stated later on, we shall see that for strong consistency of parameter estimates besides conditions on system structure there is a growth rate requirement for system input when the attenuating excitation is applied to the control. But $L_n \hat{x}_n$ may not meet this requirement. This is the motivation to truncate the control at randomly varying bounds which we describe right now.

We partition the time axis by a sequence of stopping times

$$1 = \tau_1 < \sigma_1 < \tau_2 < \sigma_2 < \dots$$

at which the control is truncated in order to keep the required growth rate.

From the random time τ_k we define adaptive control u_n^a as $L_n \hat{x}_n$ excited by v_n as far as $n < \sigma_k$, where σ_k is the first time when the growth rate of $1/(j-1) \sum_{i=\tau_k}^{j-1} \|L_i \hat{x}_i\|^2$ is greater, roughly speaking, than $\log^\delta(j-1)$; and from the random time σ_k we define adaptive control as a pure disturbance v_n until $n < \tau_{k+1}$ where τ_{k+1} indicates the time when $(1/n) \sum_{j=1}^n \|\hat{x}_j\|^2$ is less than $\log^{\delta/2} n$ and when some other technical conditions are satisfied. To be precise, we define

$$(26) \quad \sigma_k = \sup \left\{ t > \tau_k : \sum_{i=\tau_k}^{j-1} \|L_i \hat{x}_i\|^2 \leq (j-1) \log^\delta(j-1) + \|L_{\tau_k} \hat{x}_{\tau_k}\|^2, \quad \forall j \in (\tau_k, t] \right\},$$

$$(27) \quad \tau_{k+1} = \inf \left\{ t > \sigma_k : \sum_{i=\tau_k}^{\sigma_k-1} \|L_i \hat{x}_i\|^2 \leq \frac{t \log^\delta t}{2^k}; \sum_{j=1}^t \|\hat{x}_j\|^2 \leq t \log^{\delta/2} t; \frac{\|L_t \hat{x}_t\|^2}{t \log^\delta t} \leq 1 \right\}$$

with any but fixed δ such that

$$(28) \quad \delta \in \left(0, \frac{\frac{1}{4} - (m+2)s\varepsilon}{2 + (m+1)s}\right).$$

Clearly, for any $\varepsilon \in (0, 1/(4s(m+2)))$ the interval for δ is not empty and the upper bound for δ is chosen to ensure an important inequality, which will be used later on:

$$(29) \quad \frac{1}{4} - 2\delta - \varepsilon - (mp + s)(\varepsilon + \delta) > 0.$$

On the right-hand side of the inequality in definition (26) the term $\|L_{\tau_k} \hat{x}_{\tau_k}\|^2$ is added to ensure the existence of σ_k , while in definition (27) the first and the last inequalities are rather technical and are used in the proof of Lemma 4 for considering case (3).

The adaptive control is defined by

$$(30) \quad u_n^a = L_n^0 \hat{x}_n + v_n$$

with

$$(31) \quad L_n^0 = \begin{cases} L_1 & \text{if } n \text{ belongs to some } [\tau_k, \sigma_k), \\ 0 & \text{if } n \text{ belongs to some } [\sigma_k, \tau_{k+1}). \end{cases}$$

We note at once that u_n^a can be recursively computed in real time and this makes the results developed here practically applicable. It is not difficult to see that u_n^a is indeed \mathcal{F}_n -measurable, and it will be shown in § 5 that $\{u_n^a\} \in U$ defined by (17).

We now formulate our main result.

THEOREM 1. *If conditions (a)–(c) are satisfied, then the adaptive control $u^a = \{u_n^a\}$ given by (30) is optimal in the following sense: that for system (1) with $\{u_n^a\}$ applied the parameter estimate θ_n given by (18) is strongly consistent and the loss function (2) attains its minimum, i.e.,*

$$\theta_n \xrightarrow[n \rightarrow \infty]{} \theta \quad \text{a.s.}$$

and

$$J(u^a) = \text{tr } SCQC^\tau \quad \text{a.s.}$$

The proof of Theorem 1 is given in § 5.

Obviously, the optimal adaptive control is not unique; it may differ first by a different choice of excitation source $\{v_n\}$, second by various estimation schemes applied to θ . For example, we can use the least squares algorithm. In this case, instead of (18)–(20) we take

$$(32) \quad \theta_{n+1} = \theta_n + a_n P_n \varphi_n (y_{n+1}^\tau - \varphi_n^\tau \theta_n),$$

$$(33) \quad P_{n+1} = P_n - a_n P_n \varphi_n \varphi_n^\tau P_n, \quad a_n = (1 + \varphi_n^\tau P_n \varphi_n)^{-1},$$

$$(34) \quad \varphi_n^\tau = [y_n^\tau \cdots y_{n-p+1}^\tau, u_n^\tau \cdots u_{n-q+1}^\tau, y_n^\tau - \varphi_{n-1}^\tau \theta_n, \cdots, y_{n-r+1}^\tau - \varphi_{n-r}^\tau \theta_{n-r+1}],$$

and we change $\log^{\varepsilon/2} n$ in (25) to $n^{\varepsilon/2}$, $\log^\delta(j-1)$ in (26) to $(j-1)^\delta$ and finally $\log^\delta t$ and $\log^{\delta/2} t$ in (27) to t^δ and $t^{\delta/2}$, respectively, then Theorem 1 can be modified to the following.

THEOREM 1'. *Assume that conditions (a) and (c) are satisfied and $C^{-1}(z) - \frac{1}{2}I$ is strictly positive real. If the parameter estimates are given by (32)–(34) and in the definition of adaptive control (25)–(31) $\log i$ is replaced by i for all i , then*

$$\theta_n \xrightarrow[n \rightarrow \infty]{} \theta \quad \text{and} \quad J(u^a) = \text{tr } SCQC^\tau \quad \text{a.s.}$$

The proof of this theorem can be carried out along the lines of that of Theorem 1. In the sequel by θ_n we always mean the estimate given by (18)–(20).

4. Consistency theorems. In this section we give some theorems on the strong consistency of parameter estimates.

In the sequel we always denote, respectively, by $\lambda_{\max}(X)$ and $\lambda_{\min}(X)$ the maximum and the minimum eigenvalues of a matrix X . We first give a result on matrix production; it plays a crucial role in the proof of Theorem 2.

LEMMA 3. *Let $\{f_i\}$ be a sequence of deterministic vectors of dimension d and let $F(n+1, i)$ be recursively defined by,*

$$(35) \quad F(n+1, i) = \left(I - \frac{f_n f_n^\tau}{r_n^f} \right) F(n, i), \quad F(i, i) = I,$$

$$(36) \quad r_n^f = 1 + \sum_{i=1}^n \|f_i\|^2, \quad r_0^f = 1.$$

If $r_n^f \xrightarrow{m \rightarrow \infty} \infty$ and for some $a \in [0, \frac{1}{4}]$ there are constants N_0 and M such that for all $n \geq N_0$

$$r_{n+1}^f / r_n^f \leq M(\log r_n^f)^a,$$

and

$$\frac{\lambda_{\max}\left(\sum_{i=1}^n f_i f_i^\tau + \frac{1}{d} I\right)}{\lambda_{\min}\left(\sum_{i=1}^n f_i f_i^\tau + \frac{1}{d} I\right)} \leq M(\log r_n^f)^{(1/4)-a}$$

then

$$F(n, 0) \xrightarrow{n \rightarrow \infty} 0.$$

The proof of Lemma 3 is given in Appendix 1.

Set

$$(37) \quad r_n^0 = 1 + \sum_{i=1}^n \|\varphi_i^0\|^2, \quad r_0^0 = 1,$$

$$(38) \quad \varphi_n^{0\tau} = [y_n^\tau, \dots, y_{n-p+1}^\tau, u_n^\tau, \dots, u_{n-q+1}^\tau, w_n^\tau, \dots, w_{n-r+1}^\tau],$$

which is obtained from φ_n with $y_i^\tau - \varphi_{i-1}^\tau \theta_{i-1}$ replaced by w_i^τ , $i = n \cdots n-r+1$.

THEOREM 2. If condition (b) holds, $r_n^0 \rightarrow \infty$ and if there are $a \in [0, \frac{1}{4}]$, N_0 and M possibly depending upon ω such that for any $n \geq N_0 - 1$

$$(39) \quad r_{n+1}^0 / r_n^0 \leq M(\log r_n^0)^a \quad \text{a.s.},$$

$$(40) \quad \frac{\lambda_{\max}\left(\sum_{i=1}^n \varphi_i^0 \varphi_i^{0\tau} + \frac{1}{d} I\right)}{\lambda_{\min}\left(\sum_{i=1}^n \varphi_i^0 \varphi_i^{0\tau} + \frac{1}{d} I\right)} \leq M(\log r_n^0)^{1/4-a} \quad \text{a.s.},$$

with $d = mp + lq + mr$, then

$$\theta_n \xrightarrow{n \rightarrow \infty} \theta \quad \text{a.s.}$$

The theorem holds true if in its conditions φ_i^0 and r_i^0 are replaced by φ_i and r_i respectively.

Proof. We rewrite $F(n, i)$ defined in Lemma 3 to $\Phi(n, i)$ and $\Phi^0(n, i)$ if f_i is replaced by φ_i and φ_i^0 respectively. We know that $\Phi(n, 0) \rightarrow 0$ is equivalent to $\Phi^0(n, 0) \rightarrow 0$ if condition (b) holds (Chen and Guo (1985a), (1985b)). Then by Lemma 3 under the conditions of the theorem we have $\Phi^0(n, 0) \rightarrow 0$; hence $\theta_n \rightarrow \theta$ as shown in Chen and Guo (1985a), (1985b), (1987).

For consistency of parameter estimates we now give a theorem that translates conditions on φ_n and φ_n^0 to conditions on u_n alone. This is a basic step for proving our main result and is interesting by itself.

THEOREM 3. Suppose that for system (1) $A(z)$, $B(z)$ and $C(z)$ have no common left factor and conditions (a) and (b) are satisfied and that

$$(41) \quad u_n = u_n^s + v_n$$

and

$$(42) \quad \frac{1}{n} \sum_{i=1}^n \|u_i^s\|^2 = O(\log^\delta n)$$

for some δ satisfying (28), where v_n is given by (25) and u_n^s is any \mathcal{F}'_{n-1} -measurable random vector with \mathcal{F}'_{n-1} being σ -algebra generated by $\{w_i, i \leq n, v_j, j \leq n-1\}$, $\forall n \geq 1$. Then θ_n is strongly consistent:

$$\theta_n \xrightarrow[n \rightarrow \infty]{} \theta \quad a.s.$$

The proof is given in Appendix 2.

5. Proof of the main theorem. The proof of Theorem 1 is separated into several lemmas.

LEMMA 4. Under conditions of Theorem 1 the estimate θ_n is strongly consistent:

$$\theta_n \xrightarrow[n \rightarrow \infty]{} \theta \quad a.s.$$

and

$$L_n \xrightarrow[n \rightarrow \infty]{} L \quad a.s.,$$

where L and L_n are defined by (15) and (22) respectively.

Proof. We first prove consistency of θ_n .

(1) If $\tau_k < \infty$, $\sigma_k = \infty$ for some k , then $L_i^0 = L_i$ for $i \geq \tau_k$ and by definition (26) for σ_k we have

$$\frac{1}{n} \sum_{i=1}^n \|L_i \hat{x}_i\|^2 = O(\log^\delta n).$$

Then by (30) and (31) we see that Theorem 3 can be applied, since $L_i^0 \hat{x}_i$ is obviously \mathcal{F}'_{i-1} -measurable. Hence $\theta_n \xrightarrow[n \rightarrow \infty]{} \theta$ a.s.

(2) If $\sigma_k < \infty$, $\tau_{k+1} = \infty$ for some k , then by (30) and (31) $u_n^a = v_n$ for $n \geq \sigma_k$, and again Theorem 3 leads to the conclusion of the lemma.

(3) If $\sigma_k < \infty$, $\tau_k < \infty$, for all k , then by (26), (27) and (31) we have for all $k \geq 1$

$$\begin{aligned} & \sup_{\tau_k \leq n < \tau_{k+1}} \frac{1}{n \log^\delta n} \sum_{i=1}^n \|L_i^0 \hat{x}_i\|^2 \\ &= \sup_{\tau_k \leq n \leq \sigma_{k-1}} \frac{1}{n \log^\delta n} \sum_{i=\tau_1}^n \|L_i^0 \hat{x}_i\|^2 \\ &= \sup_{\tau_k \leq n \leq \sigma_{k-1}} \frac{1}{n \log^\delta n} \left[\left(\sum_{i=\tau_1}^{\sigma_1-1} + \sum_{i=\tau_2}^{\sigma_2-1} + \cdots + \sum_{i=\tau_{k-1}}^{\sigma_{k-1}-1} + \sum_{i=\tau_k}^n \right) \|L_i^0 \hat{x}_i\|^2 \right] \\ &\leq \frac{1}{\tau_2 \log^\delta \tau_2} \sum_{i=\tau_1}^{\sigma_1-1} \|L_i \hat{x}_i\|^2 + \cdots + \frac{1}{\tau_k \log^\delta \tau_k} \sum_{i=\tau_{k-1}}^{\sigma_{k-1}-1} \|L_i \hat{x}_i\|^2 \\ &\quad + \sup_{\tau_k \leq n \leq \sigma_{k-1}} \frac{1}{n \log^\delta n} \sum_{i=\tau_k}^n \|L_i \hat{x}_i\|^2 \\ &\leq \sum_{i=1}^{k-1} \frac{1}{2^i} + \sup_{\tau_k \leq n \leq \sigma_{k-1}} \frac{1}{n \log^\delta n} (n \log^\delta n + \|L_{\tau_k} \hat{x}_{\tau_k}\|^2) \leq 3 \quad \forall k \geq 1. \end{aligned}$$

Hence in this case Theorem 3 can also be applied. Thus we have established the strong consistency of θ_n . The second assertion follows from Lemma 5.

In the proof of Lemmas 5, 6 and 7 we need the following fact; If matrices Ω_n converge to a stable matrix, then there are constants $0 < \mu < 1$ and c_2 such that (Chen (1985, p. 191))

$$(43) \quad \|\Omega_k \Omega_{k-1} \cdots \Omega_{i+1}\| \leq c_2 \mu^{k-i} \quad \forall k > i, \quad \forall i \geq 0.$$

LEMMA 5. If $\theta_n \xrightarrow{n \rightarrow \infty} \theta$ and condition (c) holds, then S_n defined by (23) tends to the solution S of (14) as $n \rightarrow \infty$.

The proof is given in Appendix 1.

We now write x_n given by (9) and \hat{x}_n given by (21) in the vector component forms

$$(44) \quad x_n = [x_n^{1\tau}, \cdots, x_n^{s\tau}]^\tau, \quad \hat{x}_n = [\hat{x}_n^{1\tau}, \cdots, \hat{x}_n^{s\tau}]^\tau,$$

where x_n^i and \hat{x}_n^i are m -dimensional, $i = 1 \cdots s$.

Set

$$(45) \quad z_n = [x_n^{2\tau} \cdots x_n^{s\tau}]^\tau, \quad \hat{z}_n = [\hat{x}_n^{2\tau} \cdots \hat{x}_n^{s\tau}]^\tau.$$

From (21) we have

$$(46) \quad \begin{aligned} \hat{x}_{n+1}^1 &= A_{1n} \hat{x}_n^1 + \hat{x}_n^2 + B_{1n} u_n + (H A x_n + H B u_n + w_{n+1} - H \hat{A}_n \hat{x}_n - H \hat{B}_n u_n) \\ &= A_1 x_n^1 + x_n^2 + B_1 u_n + w_{n+1} = x_{n+1}^1. \end{aligned}$$

Then

$$(47) \quad \begin{aligned} &\hat{C}_n (y_{n+1} - H \hat{A}_n \hat{x}_n - H \hat{B}_n u_n) \\ &= \hat{C}_n H A (x_n - \hat{x}_n) + \hat{C}_n H (A - \hat{A}_n) \hat{x}_n + \hat{C}_n H (B - \hat{B}_n) u_n + \hat{C}_n w_{n+1} \\ &= \hat{C}_n^0 (z_n - \hat{z}_n) + \hat{C}_n H (A - \hat{A}_n) \hat{x}_n + \hat{C}_n H (B - \hat{B}_n) u_n + \hat{C}_n w_{n+1}, \end{aligned}$$

with

$$\hat{C}_n^0 = [\underbrace{\hat{C}_n, 0}_{(s-1)m}] sm.$$

Consequently, by taking $u_n = u_n^a$ we can write (21) in the following form:

$$(48) \quad \begin{aligned} \hat{x}_{n+1} &= (\hat{A}_n + \hat{B}_n L_n^0) \hat{x}_n + \hat{B}_n v_n + \hat{C}_n^0 (z_n - \hat{z}_n) + \hat{C}_n H (A - \hat{A}_n) \hat{x}_n \\ &\quad + \hat{C}_n H (B - \hat{B}_n) L_n^0 \hat{x}_n + \hat{C}_n H (B - \hat{B}_n) v_n + \hat{C}_n w_{n+1} \\ &= [\hat{A}_n + \hat{B}_n L_n^0 + \hat{C}_n H (A - \hat{A}_n) + \hat{C}_n H (B - \hat{B}_n) L_n^0] \hat{x}_n \\ &\quad + \hat{C}_n^0 (z_n - \hat{z}_n) + [\hat{B}_n + \hat{C}_n H (B - \hat{B}_n)] v_n + \hat{C}_n w_{n+1}. \end{aligned}$$

From (9), (21) and (47) we obtain

$$\begin{aligned} x_{n+1} - \hat{x}_{n+1} &= A (x_n - \hat{x}_n) + (A - \hat{A}_n) \hat{x}_n + (B - \hat{B}_n) u_n^a + (C - \hat{C}_n) w_{n+1} \\ &\quad - \hat{C}_n^0 (z_n - \hat{z}_n) - \hat{C}_n H (A - \hat{A}_n) \hat{x}_n - \hat{C}_n H (B - \hat{B}_n) u_n^a, \end{aligned}$$

and from here and (46)

$$(49) \quad \begin{aligned} z_{n+1} - \hat{z}_{n+1} &= G_n (z_n - \hat{z}_n) + (A' - \hat{A}'_n) \hat{x}_n + (B' - \hat{B}'_n) u_n^a + (C' - \hat{C}'_n) w_{n+1} \\ &\quad - \hat{C}'_n H (A - \hat{A}_n) \hat{x}_n - C'_n H (B - \hat{B}_n) u_n^a \\ &= G_n (z_n - \hat{z}_n) + [A' - \hat{A}'_n + (B' - \hat{B}'_n) L_n^0 - \hat{C}'_n H (A - \hat{A}_n) \\ &\quad - \hat{C}'_n H (B - \hat{B}_n) L_n^0] \hat{x}_n \\ &\quad + [B' - \hat{B}'_n - \hat{C}'_n H (B - \hat{B}_n)] v_n + (C' - \hat{C}'_n) w_{n+1}, \end{aligned}$$

where

$$G_n = \begin{bmatrix} -\hat{C}'_n & I \\ 0 & 0 \end{bmatrix} \begin{matrix} (s-2)m \\ m \end{matrix}$$

and X' denotes the matrix obtained from X by deleting its first m rows, for example, $B' = [B'_2 \cdots B'_s]^T$.

Finally, (48) and (49) give us a useful representation:

$$(50) \quad \begin{pmatrix} \hat{x}_{n+1} \\ z_{n+1} - \hat{z}_{n+1} \end{pmatrix} = \Phi_n \begin{pmatrix} \hat{x}_n \\ z_n - \hat{z}_n \end{pmatrix} + \begin{pmatrix} \hat{B}_n + \hat{C}_n H(B - \hat{B}_n) \\ B' - \hat{B}'_n - \hat{C}'_n H(B - \hat{B}_n) \end{pmatrix} v_n + \begin{pmatrix} \hat{C}_n \\ C - \hat{C}_n \end{pmatrix} w_{n+1},$$

where

$$(51) \quad \Phi_n = \begin{pmatrix} \hat{A}_n + \hat{B}_n L_n^0 + \hat{C}_n H(A - \hat{A}_n) + \hat{C}_n H(B - \hat{B}_n) L_n^0 & \hat{C}_n^0 \\ A' - \hat{A}'_n + (B' - \hat{B}'_n) L_n^0 - \hat{C}'_n H(A - \hat{A}_n) - \hat{C}'_n H(B - \hat{B}_n) L_n^0 & G_n \end{pmatrix}.$$

LEMMA 6. *If conditions of Theorem 1 hold then there is a k such that*

$$\tau_k < \infty, \quad \sigma_k = \infty.$$

Proof. Since $1 = \tau_1 < \sigma_1 < \tau_2 < \sigma_2 < \cdots$, we only need to prove the impossibility of the following two cases:

- (1) $\sigma_k < \infty, \tau_{k+1} = \infty$ for some k ;
- (2) $\tau_k < \infty, \sigma_k < \infty$ for all k .

By (50) we have for $n \geq \sigma_k$

$$(52) \quad \begin{pmatrix} \hat{x}_{n+1} \\ z_{n+1} - \hat{z}_{n+1} \end{pmatrix} = \prod_{j=\sigma_k}^n \Phi_j \begin{pmatrix} \hat{x}_{\sigma_k} \\ z_{\sigma_k} - \hat{z}_{\sigma_k} \end{pmatrix} + \sum_{i=\sigma_k}^n \prod_{j=i+1}^n \Phi_j \left\{ \begin{pmatrix} \hat{B}_i + \hat{C}_i H(B - \hat{B}_i) \\ B' - \hat{B}'_i - \hat{C}'_i H(B - \hat{B}_i) \end{pmatrix} v_i + \begin{pmatrix} \hat{C}_i \\ C - \hat{C}_i \end{pmatrix} w_{i+1} \right\}$$

where by definition

$$\prod_{j=i+1}^n \Phi_j = \begin{cases} \Phi_n \cdots \Phi_{i+1} & \text{for } n > i, \\ I & \text{for } n = i. \end{cases}$$

In case (1) $L_n^0 = 0$ for $n \geq \sigma_k$ by (32); then by Lemma 4 we have

$$(53) \quad \Phi_n \xrightarrow{n \rightarrow \infty} \begin{pmatrix} A & C^0 \\ 0 & G \end{pmatrix} \triangleq \Phi \quad \text{with } C^0 = [C, 0] sm, \quad G = \begin{pmatrix} -C' & I \\ 0 & 0 \end{pmatrix}.$$

Notice that $A(z)$ is stable by condition (a), and $C(z)$ is also stable since $C(z)$ is strictly positive real by condition (b), so Φ is a stable matrix.

From (43), (52) and Lemma 4 we obtain for all $n \geq \sigma_k$

$$\begin{aligned} & \frac{1}{n} \sum_{i=\sigma_k}^n (\|\hat{x}_{i+1}\|^2 + \|z_{i+1} - \hat{z}_{i+1}\|^2) \\ &= O\left(\frac{1}{n} \sum_{i=\sigma_k}^n \mu^{i-\sigma_k}\right) + O\left(\frac{1}{n} \sum_{i=\sigma_k}^n \sum_{j=\sigma_k}^i \mu^{i-j} (\|w_{j+1}\|^2 + \|v_j\|^2)\right) = O(1). \end{aligned}$$

Then

$$\sum_{k=1}^n \|\hat{x}_k\|^2 = O(n) \quad \text{as } n \rightarrow \infty.$$

This means that τ_{k+1} must be finite by its definition (27) since $L_n \xrightarrow{n \rightarrow \infty} L$ by Lemma 4. Therefore case (1) cannot occur.

Now assume that $\tau_k < \infty$, $\sigma_k < \infty$ for all k . By definition τ_k is a sequence of monotonically increasing integers; then $\tau_k \xrightarrow{k \rightarrow \infty} \infty$.

By (31) $L_n^0 = L_n$ for $n \in [\tau_k, \sigma_k)$, and then by (51) and Lemma 4 we have

$$(54) \quad \Phi_n \xrightarrow[n \in [\tau_k, \sigma_k), k \rightarrow \infty]{} \begin{bmatrix} A + BL & C^0 \\ 0 & G \end{bmatrix},$$

where C^0 and G are defined in (53).

Since $A + BL$ is stable then for $n \in (\tau_k, \sigma_k - 1]$ by (43), (54) and Lemma 4 it immediately follows from (50) that

$$(55) \quad \sum_{i=\tau_k}^n \|\hat{x}_{i+1}\|^2 \leq c_3(\|\hat{x}_{\tau_k}\|^2 + \|z_{\tau_k} - \hat{z}_{\tau_k}\|^2) + c_4\sigma_k,$$

where here and hereafter c_i , $i = 3, 4, \dots$, denote constants free of k .

Similarly, from (49) we know that

$$(56) \quad \begin{aligned} \|z_{\tau_k} - \hat{z}_{\tau_k}\|^2 &= O(1) + O\left(\sum_{i=0}^{\tau_k} \|\hat{x}_i\|^2\right) + \left(\sum_{i=0}^{\tau_k} (\|w_{i+1}\|^2 + \|v_i\|^2)\right) \\ &\leq c_5\tau_k + c_6\tau_k \log^{\delta/2} \tau_k, \end{aligned}$$

where for the last inequality (27) is invoked.

Putting (56) into (55) and noticing the boundedness of L_i , we conclude that for sufficiently large k

$$\sum_{i=\tau_k}^{\sigma_k} \|L_i \hat{x}_i\|^2 \leq c_7\tau_k \log^{\delta/2} \tau_k + c_8\sigma_k \leq c_9\sigma_k \log^{\delta/2} \sigma_k < \sigma_k \log^{\delta} \sigma_k + \|L_{\tau_k} \hat{x}_{\tau_k}\|^2.$$

On the other hand, by definition (26) we have the converse inequality

$$\sum_{i=\tau_k}^{\sigma_k} \|L_i \hat{x}_i\|^2 > \sigma_k \log^{\delta} \sigma_k + \|L_{\tau_k} \hat{x}_{\tau_k}\|^2$$

since $\sigma_k < \infty$.

The obtained contradiction shows that case (2) cannot take place as well.

To finish the proof of Theorem 1 it remains to show that the loss function reaches its minimum when u_n^a given by (30) is applied. It is done in the next lemma.

LEMMA 7. *If conditions of Theorem 1 hold, then $\{u_n^a\} \in U$ defined by (17) and*

$$J(u_n^a) = \text{tr } SCQC^T.$$

Proof. By Lemma 6 and (31) there exists some k_0 such that

$$L_n^0 = L_n \quad \forall n \geq \tau_{k_0}.$$

By Lemma 4 we know that $\{\Phi_n\}$ converges to the matrix stated at the right-hand side of (54). Then by (43) from (50) it is easy to see that

$$(57) \quad \|\hat{x}_{k+1}\|^2 + \|z_{k+1} - \hat{z}_{k+1}\|^2 = O(1) + O\left(\sum_{i=1}^k \mu^{k-i} (\|w_{i+1}\|^2 + \|v_i\|^2)\right),$$

and

$$(58) \quad \begin{aligned} & \frac{1}{n} \sum_{k=1}^n (\|\hat{x}_{k+1}\|^2 + \|z_{k+1} - \hat{z}_{k+1}\|^2) \\ &= O\left(\frac{1}{n} \sum_{k=1}^n \mu^k\right) + O\left(\frac{1}{n} \sum_{k=1}^n \sum_{i=1}^k \mu^{k-i} (\|w_{i+1}\|^2 + \|v_i\|^2)\right) = O(1). \end{aligned}$$

Then by (A2), (25), (57) and (58) it follows that

$$(59) \quad \|\hat{x}_n\|^2 = o(n) \quad \text{and} \quad \sum_{k=1}^n \|\hat{x}_k\|^2 = O(n) \quad \text{a.s.}$$

Hence $\sum_{i=1}^n \|u_i^a\|^2 = O(n)$, $\|u_n^a\|^2 = o(n)$, and thus $\{u^a\} \in U$.

Using (59) and the consistency of θ_n and noticing that G_n in (49) converges to a stable matrix, then from (49) we are easily convinced of

$$(60) \quad \frac{1}{n} \sum_{k=1}^n \|z_{k+1} - \hat{z}_{k+1}\|^2 = o(1) \quad \text{a.s.,}$$

which together with (46) yields

$$(61) \quad \frac{1}{n} \sum_{i=1}^n \|x_i - \hat{x}_i\|^2 \xrightarrow{n \rightarrow \infty} 0.$$

From (59) and (61) we see

$$(62) \quad \frac{1}{n} \sum_{i=1}^n \|x_i\|^2 = O(1).$$

Finally, putting u_n^a into (16) and using (25), (61), (62) and the fact $L_n^0 \xrightarrow{n \rightarrow \infty} L$ we conclude that

$$\begin{aligned} J(u^a) &= \text{tr } SCQC^\tau + \overline{\lim}_{n \rightarrow \infty} \frac{1}{n} \sum_{i=0}^{n-1} [(L_i^0 - L)x_i + L_i^0(\hat{x}_i - x_i) + v_i]^\tau (Q_2 + B^\tau SB) \\ &\quad \cdot [(L_i^0 - L)x_i + L_i^0(\hat{x}_i - x_i) + v_i] \\ &= \text{tr } SCQC^\tau. \end{aligned}$$

This completes the proof for Lemma 7 as well as for Theorem 1.

6. Conclusion remark. 1) In Chen and Guo (1987) the authors have given the optimal stochastic control minimizing the tracking error and leading to consistency of estimates given by the stochastic gradient algorithm. It is natural to ask: Is it possible to give a unified adaptive control applicable to both problems of tracking and quadratic cost. This requires further consideration.

2) The stability condition on $A(z)$ is rather restrictive. It is desirable to weaken it.

Appendix 1.

Proof of Lemma 1. By a standard treatment (see e.g. Chen (1985)) from (9), (10), (14) and (15) we have

$$(A.1) \quad \begin{aligned} \sum_{i=0}^{n-1} (y_i^\tau Q_1 y_i + u_i^\tau Q_2 u_i) &= x_0^\tau S x_0 - x_n^\tau S x_n + \sum_{i=0}^{n-1} w_{i+1}^\tau C^\tau S C w_{i+1} \\ &\quad + 2 \sum_{i=0}^{n-1} (A x_i + B u_i)^\tau S C w_{i+1} \\ &\quad + \sum_{i=0}^{n-1} (u_i - L x_i)^\tau (Q_2 + B^\tau S B) (u_i - L x_i). \end{aligned}$$

From (7) it is clear that

$$(A.2) \quad \frac{\|w_n\|^2}{n} = \text{tr} \frac{w_n w_n^\tau}{n} = \text{tr} \left(\frac{\sum_{i=1}^n w_i w_i^\tau}{n} - \frac{\sum_{i=1}^{n-1} w_i w_i^\tau}{n} \right) \xrightarrow{n \rightarrow \infty} 0.$$

By stability of A there are constants c_0 and $\rho \in (0, 1)$ such that

$$(A.3) \quad \|A^k\| \leq c_0 \rho^k \quad \forall k \geq 0.$$

Then by (9) and (A.3) it follows that

$$\|x_n\|^2 \leq 3c_0^2 \rho^{2n} \|x_0\|^2 + 3c_0^2 \frac{\|B\|^2 + \|C\|^2}{1 - \rho} \sum_{j=0}^{n-1} \rho^{n-j} (\|u_j\|^2 + \|w_j\|^2).$$

Therefore by (7), (17) and (A.2) from here it is concluded that

$$(A.4) \quad \frac{\|x_n\|^2}{n} = o(1), \quad \sum_{i=1}^n \|x_i\|^2 = O(n) \quad \text{a.s.}$$

From (17), (A.1) and (A.4) the conclusion of the lemma will follow immediately if we can show that

$$\sum_{i=0}^{n-1} (Ax_i + Bu_i)^\tau SCw_{i+1} = O \left(\left[\sum_{i=1}^{n-1} (\|x_i\|^2 + \|u_i\|^2) \right]^{1/2} \log^{1/2+\eta} \left(\sum_{i=1}^{n-1} (\|x_i\|^2 + \|u_i\|^2) + e \right) \right) \\ \forall \eta > 0.$$

But this is a direct consequence of Lemma 2.

Proof of Lemma 2. By the martingale convergence theorem (Chow (1965)) $\sum_{i=1}^n f_i w_{i+1}^\tau$ is convergent on the set $V = \{\omega: s_\infty < \infty\}$; hence Lemma 2 obviously holds on V .

Further, for $\omega \in V^c$, without loss of generality we assume $f_1 \neq 0$; then we have

$$\begin{aligned} \sum_{i=2}^{\infty} E \left[\left\| \frac{f_i w_{i+1}^\tau}{s_i^{1/2} \log^{1/2+\eta} (s_i + e)} \right\|^2 \middle| \mathcal{F}_i \right] &\leq \sigma^2 \sum_{i=2}^{\infty} \frac{\|f_i\|^2}{s_i \log^{1+2\eta} (s_i + e)} \\ &\leq \sigma^2 \sum_{i=2}^{\infty} \left(\int_{s_{i-1}}^{s_i} dx \right) / s_i \log^{1+2\eta} (s_i + e) \\ &\leq \sigma^2 \sum_{i=2}^{\infty} \left(\int_{s_{i-1}}^{s_i} \frac{dx}{x \log^{1+2\eta} (x + e)} \right) \\ &= \sigma^2 \int_{s_1}^{\infty} \frac{dx}{x \log^{1+2\eta} (x + e)} \\ &< \infty, \end{aligned}$$

where $\sigma^2 = \sup_n E[\|w_{n+1}\|^2 | \mathcal{F}_n]$ by definition. Again by the martingale convergence theorem we see that

$$\sum_{i=2}^{\infty} f_i w_{i+1}^\tau / s_i^{1/2} \log^{1/2+\eta} (s_i + e)$$

is convergent on V^c . Then the Kronecker Lemma guarantees validity of Lemma 2 on V^c .

Proof of Lemma 3. Set

$$(A.5) \quad m(t) = \max [n: t_n \leq t],$$

$$(A.6) \quad t_n \triangleq \sum_{i=N_0}^{n-1} \frac{\|f_i\|^2}{r_i^f (\log r_{i-1}^f)^{1/4}}.$$

We note that $m(t)$ is nothing but the inverse function of t_n , which is defined such that it diverges in an appropriate rate.

It is easy to see that

$$\begin{aligned} t_n &\geq \frac{1}{M} \sum_{i=N_0}^{n-1} \frac{\|f_i\|^2}{r_{i-1}^f (\log r_{i-1}^f)^{1/4+a}} \geq \frac{1}{M} \sum_{i=N_0}^{n-1} \int_{r_{i-1}^f}^{r_i^f} \frac{dt}{t (\log t)^{1/4+a}} \\ &= \frac{4}{(3-4a)M} (\log^{3/4-a} r_{n-1}^f - \log^{3/4-a} r_{N_0-1}^f), \end{aligned}$$

which via (A.5) implies $t_n \rightarrow \infty$, $m(t) < \infty$ for all t , and

$$(A.7) \quad \log r_{m(N+k\alpha)-1}^f \leq \left[\frac{(3-4a)M}{4} (N+k\alpha) + \log^{3/4-a} r_{N_0-1}^f \right]^{4/(3-4a)} \quad \forall N \geq 1.$$

For sufficiently large N_0 we have

$$(A.8) \quad \log r_i^f \leq \log r_{i-1}^f + \log M + a \log \log r_{i-1}^f \leq 2 \log r_{i-1}^f \quad \forall i \geq N_0.$$

then

$$t_n \leq 2 \sum_{i=N_0}^{n-1} \frac{\|f_i\|^2}{r_i^f (\log r_i^f)^{1/4}} \leq \frac{8}{3} (\log^{3/4} r_{n-1}^f - \log^{3/4} r_{N_0-1}^f)$$

and hence

$$t \leq t_{m(t)+1} \leq \frac{8}{3} (\log^{3/4} r_{m(t)}^f - \log^{3/4} r_{N_0-1}^f)$$

or

$$(A.9) \quad \log^a r_{m(N+(k-1)\alpha)}^f \geq \left[\frac{3}{8} (N+(k-1)\alpha) + \log^{3/4} r_{N_0-1}^f \right]^{4a/3}.$$

Since $m(t) < \infty$ for all t , there exists N such that $m(N) \geq N_0$ and

$$(A.10) \quad \frac{(\log r_i^f)^{1/4-a}}{r_i^f} \leq \frac{1}{2M} \quad \forall i \geq m(N).$$

For any $k \geq 1$ by summation by parts and using (A.9) we obtain

$$\begin{aligned} \sum_{i=m(N+(k-1)\alpha)}^{m(N+k\alpha)-1} \frac{f_i f_i^T}{r_i^f} &\geq \sum_{i=m(N+(k-1)\alpha)}^{m(N+k\alpha)} \frac{1}{r_i^f} \left(\sum_{j=1}^i f_j f_j^T - \sum_{j=1}^{i-1} f_j f_j^T \right) - I \\ &\geq \sum_{i=m(N+(k-1)\alpha)+1}^{m(N+k\alpha)} \sum_{j=1}^{i-1} f_j f_j^T \frac{\|f_i\|^2}{r_{i-1}^f r_i^f} - 2I \\ &\geq \sum_{i=m(N+(k-1)\alpha)+1}^{m(N+k\alpha)} \left[\frac{\lambda_{\max}(\sum_{j=1}^{i-1} f_j f_j^T + (1/d)I)}{M(\log r_{i-1}^f)^{1/4-a}} - \frac{1}{d} \right] \frac{\|f_i\|^2}{r_i^f r_{i-1}^f} I - 2I \\ &\geq \frac{1}{d} (\log r_{m(N+(k-1)\alpha)}^f)^a \cdot \sum_{i=m(N+(k-1)\alpha)+1}^{m(N+k\alpha)} \\ &\quad \cdot \left(\frac{1}{M} - \frac{(\log r_{i-1}^f)^{1/4-a}}{r_{i-1}^f} \right) \frac{\|f_i\|^2}{r_i^f (\log r_{i-1}^f)^{1/4}} I - 2I \\ &\geq \frac{1}{2Md} \log^a r_{m(N+(k-1)\alpha)}^f (t_{m(N+k\alpha)+1} - t_{m(N+(k-1)\alpha)+1}) I - 2I \\ &\geq \left[\frac{\alpha-1}{2Md} \log^a r_{m(N+(k-1)\alpha)}^f - 2 \right] I \\ &\geq \left[\frac{\alpha-1}{2Md} \left(\frac{3}{8} (N-\alpha) + \frac{3\alpha}{8} k + \log^{3/4} r_{N-1}^f \right)^{4a/3} - 2 \right] I. \end{aligned}$$

We take N, α large enough so that $N > \alpha$ and

$$b = \frac{\alpha - 1}{2Md} \left(\frac{3\alpha}{8} \right)^{4a/3} - 2 > 0;$$

then

$$(A.11) \quad \sum_{i=m(N+(k-1)\alpha)}^{m(N+k\alpha)-1} \frac{f_i f_i^\tau}{r_i^f} \geq b(k^{4a/3})I \quad \forall k \geq 1.$$

Let ρ_k be the maximum eigenvalue of the matrix

$$F^\tau(m(N+k\alpha), m(N+(k-1)\alpha))F(m(N+k\alpha), m(N+(k-1)\alpha))$$

and let $x_{m(N+(k-1)\alpha)}$ be the corresponding normalized eigenvector. For $i \in [m(N+(k-1)\alpha), m(N+k\alpha)-1]$ recursively define x_i

$$(A.12) \quad x_{i+1} = \left(I - \frac{f_i f_i^\tau}{r_i^f} \right) x_i.$$

Then we have

$$(A.13) \quad \begin{aligned} x_{m(N+k\alpha)}^\tau x_{m(N+k\alpha)} &= x_{m(N+(k-1)\alpha)}^\tau F^\tau(m(N+k\alpha), m(N+(k-1)\alpha)) \\ &\quad \cdot F(m(N+k\alpha), m(N+(k-1)\alpha)) x_{m(N+(k-1)\alpha)} \\ &= x_{m(N+(k-1)\alpha)}^\tau \rho_k x_{m(N+(k-1)\alpha)} \end{aligned}$$

and

$$(A.14) \quad x_{i+1}^\tau x_{i+1} \leq x_i^\tau x_i - x_i^\tau \frac{f_i f_i^\tau}{r_i^f} x_i.$$

Summing up both sides of (A.14) we obtain that

$$(A.15) \quad \sum_{i=m(N+(k-1)\alpha)}^{m(N+k\alpha)-1} \frac{\|f_i^\tau x_i\|^2}{r_i^f} \leq \|x_{m(N+(k-1)\alpha}\|^2 - \|x_{m(N+k\alpha)}\|^2 = 1 - \rho_k.$$

For $i \in [m(N+(k-1)\alpha), m(N+k\alpha)-1]$ from (A.12) by Schwarz inequality and (A.6), (A.15) we see that

$$(A.16) \quad \begin{aligned} \|x_i - x_{m(N+(k-1)\alpha}\| &= \left\| \sum_{j=m(N+(k-1)\alpha)}^{i-1} \frac{f_j f_j^\tau}{r_j^f} x_j \right\| \\ &\leq \{\log r_{m(N+k\alpha)-1}^f\}^{1/8} \sum_{j=m(N+(k-1)\alpha)}^{m(N+k\alpha)-1} \frac{\|f_j\|}{(r_j^f)^{1/2} (\log r_{j-1}^f)^{1/8}} \cdot \frac{\|f_j^\tau x_j\|}{(r_j^f)^{1/2}} \\ &\leq \{\log r_{m(N+k\alpha)-1}^f\}^{1/8} \sqrt{1+\alpha} \cdot \sqrt{1-\rho_k}. \end{aligned}$$

Finally, by (A.7), (A.11), (A.15) and (A.16) we conclude that

$$\begin{aligned} bk^{4a/3} &\leq x_{m(N+(k-1)\alpha)}^\tau \sum_{i=m(N+(k-1)\alpha)}^{m(N+k\alpha)-1} \frac{f_i f_i^\tau}{r_i^f} (x_{m(N+(k-1)\alpha)} - x_i + x_i) \\ &\leq (\log r_{m(N+k\alpha)-1}^f)^{1/4} \sum_{i=m(N+(k-1)\alpha)}^{m(N+k\alpha)-1} \frac{\|f_i\|^2}{r_i^f (\log r_{i-1}^f)^{1/4}} \|x_{m(N+(k-1)\alpha)} - x_i\| \\ &\quad + \{\log r_{m(N+k\alpha)-1}^f\}^{1/8} \sum_{i=m(N+(k-1)\alpha)}^{m(N+k\alpha)-1} \frac{\|f_i\|}{(r_i^f)^{1/2} (\log r_{i-1}^f)^{1/8}} \cdot \frac{\|f_i^\tau x_i\|}{(r_i^f)^{1/2}} \\ &\leq \{(\log r_{m(N+k\alpha)-1}^f)^{3/8} (\alpha+1)^{3/2} + (\log r_{m(N+k\alpha)-1}^f)^{1/8} (\alpha+1)^{1/2}\} \sqrt{1-\rho_k} \end{aligned}$$

$$\leq \left\{ (\alpha+1)^{3/2} + (\alpha+1)^{1/2} \left[\frac{(3-4a)M}{4} (N+k\alpha) + \log^{3/4-a} r_{N_0-1}^f \right]^{-1/(3-4a)} \right\} \\ \times \left\{ \frac{(3-4a)M}{4} (N+k\alpha) + \log^{3/4-a} r_{N_0-1}^f \right\}^{3/2(3-4a)} \cdot \sqrt{1-\rho_k}.$$

It is clear that there is a constant $c_1 > 0$ such that

$$bk^{4a/3} \leq c_1 k^{3/2(3-4a)} (1-\rho_k)^{1/2} \quad \forall k \geq 1,$$

or

$$\rho_k \leq 1 - \frac{b^2}{c_1^2} \cdot \frac{1}{k^{3/(3-4a)-(8a/3)}}.$$

Then

$$\|F(m(N+k\alpha), 0)\| \leq \prod_{i=1}^k \|F(m(N+i\alpha), m(N+(i-1)\alpha))\| \cdot \|F(m(N), 0)\| \\ \leq \prod_{i=1}^k \sqrt{\rho_i} \xrightarrow[k \rightarrow \infty]{} 0$$

since

$$\frac{5}{6} \leq \frac{3}{3-4a} - \frac{8}{3}a \leq 1 \quad \text{for } a \in [0, \frac{1}{4}].$$

Notice that $\|F(n, 0)\|$ is nonincreasing; then the lemma follows immediately.

Proof of Lemma 5. For simplicity we denote by $P(A, B, S)$ the right-hand side of (14). By Theorem 14.3 of Lipster and Shirayev (1978) equation (14) can be solved recursively

$$(A.17) \quad \Gamma_{n+1} = P(A, B, \Gamma_n)$$

and $\Gamma_n \rightarrow S$ for any $\Gamma_0 \geq 0$. Γ_n with initial value $\Gamma_0 = 0$ is denoted by Γ_n^0 . In this theorem it is proved that for any vector x of compatible dimension

$$(A.18) \quad x^T \Gamma_n^0 x \leq x^T \Gamma_n x \leq x^T S x + \bar{x}_n^T (\Gamma_0 - S) \bar{x}_n,$$

or equivalently,

$$x^T (\Gamma_n^0 - S) x \leq x^T (\Gamma_n - S) x \leq \bar{x}_n^T (\Gamma_0 - S) \bar{x}_n,$$

where $\bar{x}_n \xrightarrow[n \rightarrow \infty]{} 0$ and $\Gamma_n^0 \rightarrow S$ and both \bar{x}_n and Γ_n^0 are independent of Γ_0 . Hence from (A.18) we see that the convergence $\Gamma_n \rightarrow S$ is uniform in Γ_0 for $\|\Gamma_0\| \leq c$ with c being any fixed constant.

From (23) we know that

$$S_n \leq \hat{A}_n^T S_{n-1} \hat{A}_n + H^T Q_1 H \quad \forall n \geq 1.$$

Then, taking into account (43) we have the boundedness of S_n :

$$(A.19) \quad \|S_n\| \leq \|\hat{A}_n^T S_{n-1} \hat{A}_n + H^T Q_1 H\| \leq \dots \\ \leq \left\| \sum_{i=2}^n (\hat{A}_i \hat{A}_{i+1} \cdots \hat{A}_n)^T H^T Q_1 H (\hat{A}_i \hat{A}_{i+1} \cdots \hat{A}_n) \right. \\ \left. + (\hat{A}_1 \cdots \hat{A}_n)^T S_0 (\hat{A}_1 \cdots \hat{A}_n) + H^T Q_1 H \right\| \\ \leq \|H^T Q_1 H\| + c_2^2 (\|H^T Q_1 H\| + \|S_0\|) \frac{1}{1-\mu} \triangleq c \quad \forall n \geq 1.$$

By strong consistency of θ_n and by boundedness of S_n it is easy to see

$$P(A, B, S_n) - P(\hat{A}_{n+1}, \hat{B}_{n+1}, S_n) \xrightarrow{n \rightarrow \infty} 0.$$

Hence for any $\varepsilon > 0$ we can find $N > 0$ such that

$$(A.20) \quad \|\Delta S_{n+k}\| \leq \varepsilon \quad \forall k \geq 0, \quad \forall n \geq N,$$

where

$$(A.21) \quad \Delta S_{n+k} = S_{n+k} - P(A, B, S_{n+k-1}).$$

For simplicity we set

$$P_1(\Gamma) \triangleq P(A, B, \Gamma), \quad P_n(\Gamma) \triangleq P_1(P_{n-1}(\Gamma)).$$

It is easy to show that there is a constant ζ such that

$$(A.22) \quad P_1(\Gamma + \Delta\Gamma) = P_1(\Gamma) + \overline{\Delta\Gamma}, \quad \|\overline{\Delta\Gamma}\| \leq \zeta\varepsilon$$

for matrices $\Gamma \geq 0$ with $\|\Gamma\| \leq c$ and $\Delta\Gamma$ with $\|\Delta\Gamma\| \leq \varepsilon$.

We now by induction prove that for any $n \geq N$ and $k \geq 1$

$$(A.23) \quad S_{n+k} = P_k(S_n) + Z_{nk}(\varepsilon) \quad \text{with} \quad \|Z_{nk}(\varepsilon)\| \leq c_k\varepsilon,$$

where c_k is a real number independent of n .

By (A.20), (A.21), we see that (A.23) is true for $k = 1$. Now assume (A.23) holds for k . By boundedness of $\|S_n\| \leq c$ for all n , the same argument as that used in (A.19) leads to the conclusion that $P_k(S_n)$ is uniformly bounded in $n \geq 0$ and $k \geq 1$. Then by (A.21)–(A.23) it follows that

$$\begin{aligned} S_{n+k+1} &= P_1(S_{n+k}) + \Delta S_{n+k+1} = P_1(P_k(S_n) + Z_{nk}(\varepsilon)) + \Delta S_{n+k+1} \\ &= P_{k+1}(S_n) + \overline{Z_{nk}}(\varepsilon) + \Delta S_{n+k+1} = P_{k+1}(S_n) + Z_{n,k+1}(\varepsilon), \end{aligned}$$

where, obviously, $\|Z_{n,k+1}(\varepsilon)\| \leq c_{k+1} \cdot \varepsilon$ with $c_{k+1} = \zeta c_k + 1$. Hence (A.23) holds for $k + 1$.

In the present notation

$$\Gamma_n = P_n(\Gamma_0)$$

where Γ_n is defined by (A.17). By the uniform convergence of Γ_n for any $\delta > 0$ we can take k_0 large enough such that

$$(A.24) \quad \|P_{k_0}(\Gamma_0) - S\| \leq \delta \quad \forall \Gamma_0: \|\Gamma_0\| \leq c.$$

For $\varepsilon \triangleq \delta / c_{k_0}$ take N such that

$$(A.25) \quad \|\Delta S_{n+k_0}\| \leq \varepsilon \quad \forall n \geq N.$$

Then from (A.23) we have

$$S_{n+k_0} = P_{k_0}(S_n) + Z_{nk_0}(\varepsilon), \quad \|Z_{nk_0}(\varepsilon)\| \leq c_{k_0}\varepsilon = \delta$$

and by (A.24) for all $n \geq N$

$$\|S_{n+k_0} - S\| \leq \|P_{k_0}(S_n) - S\| + \|Z_{nk_0}(\varepsilon)\| \leq 2\delta,$$

which yields the conclusion of the lemma.

Appendix 2.

Proof of Theorem 3. First we note that $\{v_n, \mathcal{F}_n\}$ is a martingale difference sequence. Then by Lemma 2 we have

$$(A.26) \quad \sum_{i=1}^n u_i^s v_i^t = O\left(\left(\sum_{i=1}^n \|u_i^s\|^2\right)^{1/2} \log^{1/2+\eta}\left(\sum_{i=1}^n \|u_i^s\|^2 + e\right)\right).$$

Further, by (24), (25) we know that for $\gamma \in (\frac{2}{3}, 1)$

$$\sum_{i=1}^{\infty} E \left[\left| \frac{\|v_i v_i^t - I / \log^{\epsilon} i\|^{3/2}}{i^{3\gamma/2}} \right| \middle| \mathcal{F}_{i-1}' \right] < \infty.$$

Hence $\sum_{i=1}^{\infty} [v_i v_i^t - (1/\log^{\epsilon} i)I] / i^{\gamma}$ is convergent by the martingale convergence theorem. Then from the Kronecker lemma it follows that

$$(A.27) \quad \lim_{n \rightarrow \infty} \frac{1}{n^{\gamma}} \sum_{i=1}^n \left(v_i v_i^t - \frac{1}{\log^{\epsilon} i} I \right) = 0 \quad \forall \gamma \in (\frac{2}{3}, 1).$$

It is clear that

$$\int_2^{n+1} \frac{dx}{\log^{\epsilon} x} \leq \sum_{i=2}^n \frac{1}{\log^{\epsilon} i} \leq \frac{1}{\log^{\epsilon} 2} + \int_2^n \frac{dx}{\log^{\epsilon} x}$$

and the l'Hôpital rule shows

$$(A.28) \quad \frac{\log^{\epsilon} n}{n} \sum_{i=2}^n \frac{1}{\log^{\epsilon} i} \xrightarrow{n \rightarrow \infty} 1;$$

hence by (A.27)

$$(A.29) \quad \frac{\log^{\epsilon} n}{n} \sum_{i=1}^n v_i v_i^t \xrightarrow{n \rightarrow \infty} I \quad \text{a.s.}$$

From (42), (A.26) and (A.29) we see

$$(A.30) \quad \frac{1}{n} \sum_{i=1}^n \|u_i\|^2 = O(\log^{\delta} n).$$

Then by condition (a)

$$(A.31) \quad \frac{1}{n} \sum_{i=1}^n \|y_i\|^2 = O(\log^{\delta} n);$$

hence

$$(A.32) \quad r_n^0 = O(n \log^{\delta} n),$$

which means

$$(A.33) \quad \lambda_{\max} \left(\sum_{i=1}^n \varphi_i^0 \varphi_i^{0\tau} + \frac{1}{d} I \right) = O(n \log^{\delta} n).$$

Again by (41), (42), (A.26), (A.29), and noting that $q \geq 1$, we have for all sufficiently large n

$$(A.34) \quad r_n^0 \geq \sum_{i=1}^n \|u_i\|^2 \geq \frac{1}{2} \sum_{i=1}^n \|v_i\|^2 \geq \frac{l}{4} \frac{n}{\log^{\epsilon} n}.$$

Then $r_n^0 \xrightarrow{n \rightarrow \infty} \infty$ a.s. and

$$\frac{r_{n+1}^0}{r_n^0} = O\left(\frac{(n+1) \log^{\delta} (n+1)}{n / \log^{\epsilon} n}\right) = O(\log^{\delta+\epsilon} n) = O((\log r_n^0)^{\delta+\epsilon}).$$

Comparing with conditions in Theorem 2 we find that $a = \delta + \varepsilon$ and by (A.33) and (A.34) for (40) to hold we only need to verify

$$(A.35) \quad \lim_{n \rightarrow \infty} \frac{(\log n)^{1/4-2\delta-\varepsilon}}{n} \lambda_{\min} \left(\sum_{i=1}^n \varphi_i^0 \varphi_i^{0\tau} + \frac{1}{d} I \right) \neq 0.$$

By condition (a) it is easy to see that

$$\begin{aligned} y_{n-i} &= A^{-1}(z)B(z)z^i u_n + A^{-1}(z)C(z)z^i w_n \\ &= z^i A^{-1}(z)[B(z), C(z)] \cdot \begin{bmatrix} u_n \\ w_n \end{bmatrix}. \end{aligned}$$

Then φ_n^0 can be written as

$$(A.36) \quad \varphi_n^0 = \begin{bmatrix} F_{n1}(z) \\ F_{n2}(z) \\ F_{n3}(z) \end{bmatrix} \cdot \begin{bmatrix} u_n \\ w_n \end{bmatrix},$$

where by definition

$$\begin{aligned} F_{n1}(z) &= \begin{bmatrix} A^{-1}(z)[B(z), C(z)] \\ zA^{-1}(z)[B(z), C(z)] \\ \vdots \\ z^{p-1}A^{-1}(z)[B(z), C(z)] \end{bmatrix}, & F_{n2}(z) &= \begin{bmatrix} [I_l, 0] \\ z[I_l, 0] \\ \vdots \\ z^{q-1}[I_l, 0] \end{bmatrix}, \\ F_{n3}(z) &= \begin{bmatrix} [0, I_m] \\ z[0, I_m] \\ \vdots \\ z^{r-1}[0, I_m] \end{bmatrix}, \end{aligned}$$

where I_x denotes the identity matrix of dimension x .

Set

$$(A.37) \quad \psi_n = [\det A(z)] \varphi_n^0$$

and notice that A_p is of full rank, then $\deg A(z) = p$, $\deg [\det A(z)] = mp$, and $\deg [\text{Adj } A(z)] = mp - p$, since $A(z)[\text{Adj } A(z)] = [\det A(z)] \cdot I$.

Let

$$\det A(z) = a_0 + a_1 z + \cdots + a_{mp} z^{mp}.$$

Since $\varphi_i^0 = 0$ for $i < 0$ we have

$$\begin{aligned} \lambda_{\min} \left(\sum_{i=1}^n \psi_i \psi_i^\tau \right) &= \inf_{\|x\|=1} \sum_{i=1}^n (x^\tau \psi_i)^2 \\ &= \inf_{\|x\|=1} \sum_{i=1}^n \left(\sum_{j=0}^{mp} a_j x^\tau \varphi_{i-j}^0 \right)^2 \\ &\leq \sum_{j=0}^{mp} a_j^2 \inf_{\|x\|=1} \sum_{i=1}^n \sum_{j=0}^{mp} (x^\tau \varphi_{i-j}^0)^2 \\ &\leq (mp+1) \sum_{j=0}^{mp} a_j^2 \inf_{\|x\|=1} \sum_{i=1}^n (x^\tau \varphi_i^0)^2 \\ &= (mp+1) \sum_{j=0}^{mp} a_j^2 \lambda_{\min} \left(\sum_{i=1}^n \varphi_i^0 \varphi_i^{0\tau} \right). \end{aligned}$$

Hence for (A.35) it is sufficient to prove

$$(A.38) \quad \lim_{n \rightarrow \infty} \frac{(\log n)^\lambda}{n} \lambda_{\min} \left(\sum_{i=1}^n \psi_i \psi_i^\tau \right) \neq 0 \quad \text{a.s.},$$

where for simplicity we set $\lambda = \frac{1}{4} - 2\delta - \varepsilon$.

Let D be the set on which (A.38) is not satisfied. Suppose that $P(D) > 0$. Then for any $\omega \in D$ there exist vectors

$$\eta_{n_k} = (\alpha_{n_k}^{0\tau} \cdots \alpha_{n_k}^{(p-1)\tau} \beta_{n_k}^{0\tau} \cdots \beta_{n_k}^{(q-1)\tau} \gamma_{n_k}^{0\tau} \cdots \gamma_{n_k}^{(r-1)\tau})^\tau \in \mathbb{R}^d,$$

where $\|\eta_{n_k}\| = 1$ such that

$$(A.39) \quad \frac{(\log n_k)^\lambda}{n_k} \sum_{i=1}^{n_k} (\eta_{n_k}^\tau \psi_i)^2 \xrightarrow[k \rightarrow \infty]{} 0.$$

Set

$$(A.40) \quad H_{n_k}(z) \triangleq \sum_{i=0}^{p-1} \alpha_{n_k}^{i\tau} z^i (\text{Adj } A(z)) [B(z), C(z)] + \sum_{i=0}^{q-1} \beta_{n_k}^{i\tau} z^i [\det A(z) I_l, 0]$$

$$+ \sum_{i=0}^{r-1} \gamma_{n_k}^{i\tau} z^i [0, \det A(z) I_m]$$

$$(A.41) \quad \triangleq \sum_{i=0}^t [h_{n_k}^{i\tau}, g_{n_k}^{i\tau}] z^i,$$

where $t = mp + s - 1$, and $h_{n_k}^i$ and $g_{n_k}^i$ are l - and m -dimensional vectors, respectively.

Since $\|\alpha_{n_k}^i\| \leq 1$, $\|\beta_{n_k}^i\| \leq 1$, $\|\gamma_{n_k}^i\| \leq 1$, for any $k \geq 1$, $i = 0 \cdots p-1$, $j = 0 \cdots q-1$, and $\nu = 0 \cdots r-1$, there exists a constant $c_1 > 0$ independent of k and i such that

$$(A.42) \quad \|h_{n_k}^i\| \leq c_1, \quad \|g_{n_k}^i\| \leq c_1 \quad \forall k \geq 1, \quad i = 0, \dots, t.$$

By (A.36), (A.37) and (A.41) we can rewrite (A.39) as

$$(A.43) \quad \frac{(\log n_k)^\lambda}{n_k} \sum_{i=1}^{n_k} (h_{n_k}^{0\tau} u_i + \cdots + h_{n_k}^{t\tau} u_{i-t} + g_{n_k}^{0\tau} w_i + \cdots + g_{n_k}^{t\tau} w_{i-t})^2 \xrightarrow[k \rightarrow \infty]{} 0,$$

or equivalently,

$$(A.44) \quad \frac{(\log n_k)^\lambda}{n_k} \left\{ \sum_{i=1}^{n_k} [(h_{n_k}^{0\tau} v_i)^2 + (h_{n_k}^{0\tau} u_i^s + h_{n_k}^{1\tau} u_{i-1} + \cdots + h_{n_k}^{t\tau} u_{i-t} + g_{n_k}^{0\tau} w_i + \cdots + g_{n_k}^{t\tau} w_{i-t})^2] \right. \\ \left. + 2h_{n_k}^{0\tau} \left(\sum_{i=1}^{n_k} u_i^s v_i^\tau \right) h_{n_k}^0 + 2 \sum_{j=1}^{n_k} h_{n_k}^{j\tau} \left(\sum_{i=1}^{n_k} u_{i-j} v_i^\tau \right) h_{n_k}^0 \right. \\ \left. + 2 \sum_{j=0}^t g_{n_k}^{j\tau} \left(\sum_{i=1}^{n_k} w_{i-j} v_i^\tau \right) h_{n_k}^0 \right\} \xrightarrow[k \rightarrow \infty]{} 0.$$

We now show that (A.44) implies

$$(A.45) \quad \|h_{n_k}^i\| \xrightarrow[k \rightarrow \infty]{} 0, \quad \|g_{n_k}^i\| \xrightarrow[k \rightarrow \infty]{} 0 \quad \forall i: 0 \leq i \leq t.$$

Applying Lemma 2 to $\sum_{i=1}^{n_k} w_{i-j} v_i^\tau$ and noticing (7), (A.42), we find that

$$\lim_{n \rightarrow \infty} \frac{\log^\lambda n_k}{n_k} \sum_{j=0}^t g_{n_k}^{j\tau} \left(\sum_{i=1}^{n_k} w_{i-j} v_i^\tau \right) h_{n_k}^0 \leq (1+t) c_1^2 \lim_{k \rightarrow \infty} \frac{\log^\lambda n_k}{n_k} O(n_k^{1/2} \log^{1/2+\eta}(n_k + e)) = 0$$

for any $\omega \in D$ with a possible exception set of probability zero. In the following discussion such a possible exception is always assumed. We note that no measurability of $h_{n_k}^i$ and $g_{n_k}^i$ is required.

Similarly, by applying Lemma 2 to $\sum_{i=1}^{n_k} u_i^s v_i^\tau$ and $\sum_{i=1}^{n_k} u_{i-j} v_i^\tau$ ($j \geq 1$) and by use of (41), (42) and (A.42) we conclude that for $\omega \in D$

$$\lim_{k \rightarrow \infty} \frac{\log^\lambda n_k}{n_k} \left[h_{n_k}^{0\tau} \left(\sum_{i=1}^{n_k} u_i^s v_i^\tau \right) h_{n_k}^0 + \sum_{j=1}^t h_{n_k}^{j\tau} \left(\sum_{i=1}^{n_k} u_{i-j} v_i^\tau \right) h_{n_k}^0 \right] = 0.$$

Hence from (A.44) we have

$$(A.46) \quad \frac{(\log n_k)^\lambda}{n_k} \sum_{i=1}^{n_k} (h_{n_k}^{0\tau} u_i^s + h_{n_k}^{1\tau} u_{i-1} + \cdots + h_{n_k}^{t\tau} u_{i-t} + g_{n_k}^{0\tau} w_i + \cdots + g_{n_k}^{t\tau} w_{i-t})^2 \xrightarrow[k \rightarrow \infty]{} 0,$$

and

$$(A.47) \quad \frac{(\log n_k)^\lambda}{n_k} \sum_{i=1}^{n_k} (h_{n_k}^{0\tau} v_i)^2 \xrightarrow[k \rightarrow \infty]{} 0 \quad \text{for } \omega \in D.$$

By (A.29) and (A.47) it is clear that

$$(A.48) \quad \|h_{n_k}^0\|^2 = o((\log n_k)^{-\lambda+\varepsilon}), \quad \omega \in D;$$

hence by (42)

$$\frac{(\log n_k)^{\lambda-(\varepsilon+\delta)}}{n_k} \sum_{i=1}^{n_k} (h_{n_k}^{0\tau} u_i^s)^2 = o(1), \quad \omega \in D.$$

Then from here and (A.46) we have for $\omega \in D$

$$(A.49) \quad \frac{(\log n_k)^{\lambda-(\varepsilon+\delta)}}{n_k} \sum_{i=1}^{n_k} (h_{n_k}^{1\tau} u_{i-1} + \cdots + h_{n_k}^{t\tau} u_{i-t} + g_{n_k}^{0\tau} w_i + \cdots + g_{n_k}^{t\tau} w_{i-t})^2 \xrightarrow[k \rightarrow \infty]{} 0.$$

Comparing (A.49) with (A.43) we see that in (A.49) we have deleted u_i by changing the order of $\log n_k$ from λ to $\lambda - (\varepsilon + \delta)$.

Generally, using the same treatment as described above we conclude that

$$(A.50) \quad \|h_{n_k}^i\|^2 = o((\log n_k)^{-\lambda+i(\varepsilon+\delta)+\varepsilon}), \quad 0 \leq i \leq t, \quad \omega \in D$$

and

$$(A.51) \quad \frac{(\log n_k)^{\lambda-(t+1)(\varepsilon+\delta)}}{n_k} \sum_{i=1}^{n_k} (g_{n_k}^{0\tau} w_i + \cdots + g_{n_k}^{t\tau} w_{i-t})^2 \xrightarrow[k \rightarrow \infty]{} 0, \quad \omega \in D.$$

The same argument applied to (A.51) by using (7) and (A.42) leads to

$$(A.52) \quad \|g_{n_k}^i\|^2 = o((\log n_k)^{-\lambda+(t+1)(\varepsilon+\delta)}) \quad \forall i: 0 \leq i \leq t.$$

Since $t = mp + s - 1$ and $\lambda = \frac{1}{4} - 2\delta - \varepsilon$, then by (29), (A.50) and (A.52) imply (A.45); hence we have

$$(A.53) \quad H_{n_k}(z) \xrightarrow[k \rightarrow \infty]{} 0, \quad \omega \in D.$$

Let $\{\eta_{m_k}\}$ be a convergent subsequence of $\{\eta_{n_k}\}$: $\eta_{m_k} \xrightarrow[k \rightarrow \infty]{} \eta$ with

$$(A.54) \quad \|\eta\| = 1, \quad \omega \in D, \\ \eta = (\alpha^{0\tau} \cdots \alpha^{(p-1)\tau}, \beta^{0\tau} \cdots \beta^{(q-1)\tau}, \gamma^{0\tau} \cdots \gamma^{(\gamma-1)\tau}).$$

Then by (A.40) and (A.53) we have

$$(A.55) \quad \sum_{i=0}^{p-1} \alpha^{ir} z^i (\text{Adj } A(z)) [B(z), C(z)] \\ = - \sum_{i=0}^{q-1} \beta^{ir} z^i [\det A(z) I_l, 0] - \sum_{i=0}^{r-1} \gamma^{ir} z^i [0, \det A(z) I_m].$$

Since $A(z)$, $B(z)$ and $C(z)$ have no common left factor, there are matrix polynomials $M(z)$, $N(z)$ and $L(z)$ such that

$$A(z)M(z) + B(z)N(z) + C(z)L(z) = I.$$

Then by (A.55) we see

$$(A.56) \quad \sum_{i=0}^{p-1} \alpha^{ir} z^i \text{Adj } A(z) \\ = \sum_{i=0}^{p-1} \alpha^{ir} z^i \text{Adj } A(z) \left(A(z)M(z) + [B(z), C(z)] \begin{bmatrix} N(z) \\ L(z) \end{bmatrix} \right) \\ = \det A(z) \left[\sum_{i=0}^{p-1} \alpha^{ir} z^i M(z) - \sum_{i=0}^{q-1} \beta^{ir} z^i N(z) - \sum_{i=0}^{r-1} \gamma^{ir} z^i L(z) \right], \quad \omega \in D.$$

But

$$\deg \left(\sum_{i=0}^{p-1} \alpha^{ir} z^i \text{Adj } A(z) \right) \leq p-1 + \deg (\text{Adj } A(z)) \\ = p-1 + mp - p < mp = \deg (\det A(z)),$$

so (A.56) implies

$$\sum_{i=0}^{p-1} \alpha^{ir} z^i \text{Adj } A(z) = 0, \quad \omega \in D.$$

Hence $\alpha^i = 0$, $i = 0, \dots, p-1$, and by (A.55) $\beta^i = 0$, $i = 0 \cdots q-1$, and $\gamma^j = 0$, $j = 1 \cdots r-1$ for $\omega \in D$. This conclusion contradicts with $\|\eta\| = 1$; therefore, $P(D) = 0$ and (A.38) is verified.

REFERENCES

- B. D. O. ANDERSON AND J. B. MOORE (1971), *Linear Optimal Control*, Prentice-Hall, Englewood Cliffs, NJ.
- A. BECKER, P. R. KUMAR AND C. Z. WEI (1985), *Adaptive control with the stochastic approximation algorithm, Geometry and convergence*, IEEE Trans. Automat. Control, AC-30, pp. 330-338.
- P. E. CAINES AND S. LAFORTUNE (1984), *Adaptive control with recursive identification for stochastic linear systems*, IEEE Trans. Automat. Control, AC-29, pp. 312-321.
- P. E. CAINES AND H. F. CHEN (1985), *Optimal adaptive LQG control for systems with finite state process parameter*, IEEE Trans. Automat. Control, AC-30, pp. 185-189.
- H. F. CHEN (1981), *Quasi-least squares identification and its strong consistency*, Internat. J. Control, 34, pp. 921-936.
- (1982), *Strong consistency and convergence rate of least squares identification*, Sci. Sinica. Ser. A, 25, pp. 771-784.
- (1984), *Recursive system identification and adaptive control by use of the modified least squares algorithm*, this Journal, 22, pp. 759-776.
- (1985), *Recursive Estimation and Control for Stochastic Systems*, John Wiley, New York.
- H. F. CHEN AND P. E. CAINES (1984), *Adaptive linear quadratic control for stochastic discrete-time systems*, Preprints of the 9th World Congress of the International Federation of Automatic Control, Vol. 12, Budapest, pp. 150-154.

- H. F. CHEN AND P. E. CAINES (1985), *Strong consistency of the stochastic gradient algorithm of adaptive control*, IEEE Trans. Automat. Control, AC-30, pp. 189-192.
- H. F. CHEN AND L. GUO (1985a), *Adaptive control with recursive identification for stochastic linear systems*, in Advances in Control and Dynamic Systems, C. T. Leondes, ed., Vol. 24, Academic Press, New York.
- (1985b), *Strong consistency of parameter estimates for discrete-time stochastic systems*, J. Systems Sci. Math. Sci., 5, pp. 81-93.
- (1986), *Optimal stochastic adaptive control with quadratic index*, Internat. J. Control, 43, pp. 869-881.
- (1987), *Asymptotically optimal adaptive control with consistent parameter estimates*, this Journal, 25, pp. 558-575.
- Y. S. CHOW (1965), *Local convergence of martingales and the law of large numbers*, Ann. Math. Stat., 36, pp. 552-558.
- G. C. GOODWIN, P. J. RAMADGE AND P. E. CAINES (1981), *Discrete-time stochastic adaptive control*, this Journal, 19, pp. 829-853.
- G. C. GOODWIN, D. J. HILL AND M. PALANISWAMI (1984), *A perspective on convergence of adaptive control algorithm*, Automatica J. IFAC, 20, pp. 519-531.
- O. B. HIJAB (1983), *The adaptive LQG problem, Part 1*, IEEE Trans. Automat. Control, AC-28, pp. 171-178.
- T. KAILATH (1980), *Linear Systems*, Prentice-Hall, Englewood Cliffs, NJ.
- P. R. KUMAR (1983), *Optimal adaptive control of linear-quadratic-Gaussian systems*, this Journal, 21, pp. 163-178.
- T. L. LAI AND C. Z. WEI (1982), *Least squares estimation in stochastic regression models with application to identification and control of dynamic systems*, Ann. Statist., 10, pp. 154-166.
- R. S. LIPSTER AND A. N. SHIRYAYEV (1978), *Statistics of Random Processes*, Springer, New York.
- L. LJUNG (1977), *Analysis of recursive stochastic algorithm*, IEEE Trans. Automat. Control, AC-22, pp. 551-575.
- C. SAMSON (1983), *Stability analysis of adaptively controlled systems subject to bounded disturbances*, Automatica—J IFAC, 19, pp. 81-86.
- K. S. SIN AND G. C. GOODWIN (1982), *Stochastic adaptive control using a modified least squares algorithm*, Automatica—J. IFAC, 18, pp. 315-321.
- V. SOLO (1979), *The convergence of AML*, IEEE Trans. Automat. Control, AC-24, pp. 958-962.

THE STRUCTURE OF TIME-OPTIMAL TRAJECTORIES FOR SINGLE-INPUT SYSTEMS IN THE PLANE: THE GENERAL REAL ANALYTIC CASE*

H. J. SUSSMANN†

Abstract. For arbitrary single-input real analytic systems in the plane, in which the control enters linearly, we prove that the time-optimal trajectories are finite concatenations of “bang-bang” and singular arcs, with local bounds on the number of switchings. No “nondegeneracy” assumption is made other than real-analyticity. The analysis proceeds by applying our previous results on nondegenerate C^∞ systems to study the behavior of the trajectories on a neighborhood of every point, except for the points in a discrete set of “branch points.” The branch points are then handled by a combination of control-theoretic arguments and the use of subanalytic set theory.

Key words. time-optimal control, two-dimensional systems, regular synthesis

AMS(MOS) subject classifications. 93C10, 93B15, 93B20

1. Introduction. This paper continues the analysis, begun in [A], of the time-optimal trajectories for systems

$$(1.1) \quad \dot{x} = F(x) + uG(x), \quad |u| \leq 1,$$

where x belongs to an open subset of the plane, and F and G are C^∞ vector fields. In [A] we proved theorems about the structure of the trajectories in regions where the system is sufficiently nonsingular. Here we will limit ourselves to real-analytic systems, but we will make no extra assumptions. It then turns out, as shown in § 2, that the analysis of [A] applies in a neighborhood of every point x of the state space, except for (a) certain “branch points,” that constitute a discrete set, and (b) systems which are so degenerate that, locally, *every* trajectory is time-optimal. The main part of the paper is § 3, where we analyse the structure of the trajectories in a neighborhood of a branch point. In § 4 we dispose of the other exceptional case mentioned above, and the analysis is complete. The main conclusion of our analysis is then stated in § 5.

We will be using all the definitions and notations introduced in [A].

Our work makes heavy use of the theory of subanalytic sets. It is a remarkable fact that this theory is not only used in the way mentioned earlier (i.e., to go from the local bounds on the number of switchings to the existence of regular synthesis) but also in the trajectory analysis itself. We provide an Appendix which gives all the definitions and results about analytic, semianalytic, and subanalytic sets which are used in the paper.

Our work partially overlaps with Baytman’s book [Ba]. Baytman proved existence of a regular synthesis for a class of smooth systems, without using the theory of subanalytic sets. However, he has to make various nondegeneracy assumptions which, in particular, fail to be satisfied for arbitrary analytic systems.

2. The analytic case. In [A] we studied the time-optimal trajectories in the neighborhood of near-ordinary points p , for an arbitrary smooth system Σ . We now consider the case when Σ is *analytic*, i.e., we assume that

$$(2.1) \quad F \text{ and } G \text{ are real analytic vector fields on } M.$$

* Received by the editors February 2, 1982; accepted for publication (in revised form) March 6, 1986. This work was partially supported by National Science Foundation grant no. DMS83-01678-01.

† Mathematics Department, Rutgers University, New Brunswick, New Jersey 08903.

Moreover, we will assume that

(2.II) M is connected.

A system Σ for which (2.I) and (2.II) hold will be called a *connected analytic system* (C.A.S.). If Σ is a C.A.S., then the functions $\Delta_A: M \rightarrow \mathbb{R}$ and $\Delta_B: M \rightarrow \mathbb{R}$ are real analytic and, therefore, each of these functions either vanishes everywhere on M or its set of zeros has an empty interior in M .

We let $L(F, G)$ denote the Lie algebra of vector fields on M which is generated by F and G , and we let $L_0(F, G)$ denote the ideal of $L(F, G)$ generated by G . Then $L_0(F, G)$ is spanned by $G, [F, G], [F, [F, G]], [G, [F, G]],$ etc. If \mathcal{S} is any set of vector fields on M , and $p \in M$, we write

$$(2.1) \quad \mathcal{S}(p) = \{V(p): V \in \mathcal{S}\}.$$

We say that Σ has *the accessibility property* (AP) at a point $p \in M$ if

$$(2.2) \quad \dim L(F, G)(p) = 2$$

and we say that Σ has *the strong accessibility property* (SAP) at p if

$$(2.3) \quad \dim L_0(F, G)(p) = 2.$$

It is well known (cf. [SuJ]) that Σ has the AP at p if and only if the set of points that can be reached from p by Σ -trajectories has a nonempty interior. Also, Σ has the SAP from p if and only if there is a $t > 0$ such that $A_t^\Sigma(p)$ has a nonempty interior, where $A_t^\Sigma(p)$ is the set of points that can be reached in time t by Σ -trajectories from p . If Σ has the SAP from p , then $A_t^\Sigma(p)$ has a nonempty interior for every $t > 0$.

We say that Σ has *the accessibility property* (AP) if Σ has the AP at p for all $p \in M$, and that Σ has *the dense accessibility property* (DAP) if it has the AP from p for all p in some dense subset of M . The definitions of the *strong accessibility property* (SAP), and of the *dense strong accessibility property* (DSAP) are similar.

LEMMA 2.1. *Let Σ be a C.A.S. on M . Then the following conditions are equivalent:*

- (a) Σ has the DAP,
- (b) Σ has the AP at some $p \in M$,
- (c) Δ_A does not vanish identically on M .

Proof. If $G \equiv 0$ on M , then it is clear that (a), (b) and (c) are all false, and so the equivalence holds. Therefore we may assume that G is not $\equiv 0$ on M . So there is an open dense subset M' of M such that $G(p) \neq 0$ for $p \in M'$ (because M is connected, and G is analytic). If $\Delta_A \equiv 0$, then there is a function $\alpha_F: M' \rightarrow \mathbb{R}$ such that $F = \alpha_F G$ on M' . Then it is easy to see, that, if L' is the set of all $E \in L(F, G)$ such that $E \upharpoonright M'$ is equal to $G \upharpoonright M'$ times some scalar function $\alpha_E: M' \rightarrow \mathbb{R}$, then L' is a Lie algebra. Since $F \in L'$ and $G \in L'$, we see that $L' = L(F, G)$. So, if $E \in L(F, G)$, then $E = \alpha_E G$ on M' , for some $\alpha_E: M' \rightarrow \mathbb{R}$. Therefore, if $p \in M'$, the space $L(F, G)(p)$ is the linear span of $G(p)$. Hence $L(F, G)(p)$ is one-dimensional, and so the AP fails to hold at p . If $q \in M$ and if the AP holds at q , then there are E_1, E_2 in $L(F, G)$ such that $E_1(q)$ and $E_2(q)$ are linearly independent. But then $E_1(q')$ and $E_2(q')$ must be independent for all q' in some neighborhood of q . Since M' is dense, there must be a $q' \in M'$ such that $E_1(q')$ and $E_2(q')$ are independent, and therefore $\dim L(F, G)(q') = 2$. But this contradicts the fact that the AP fails at all $q' \in M'$. So the AP fails at all $q \in M$. Therefore the negation of (c) implies the negation of (b), and so (b) \Rightarrow (c). That (a) \Rightarrow (b) is obvious. Finally, if (c) holds, then there is a dense $M' \subseteq M$ such that $F(p)$ and $G(p)$ are independent for every $p \in M'$. But then the AP holds at every $p \in M'$, and so (a) holds. So (c) \Rightarrow (a). \square

LEMMA 2.2. *Let Σ be a C.A.S. on M . Then the following are equivalent:*

- (a) Σ has the DSAP,
- (b) Σ has the SAP at some $p \in M$,
- (c) Δ_B does not vanish identically on M .

Proof. The proof that (c) \Rightarrow (a) is exactly like that of the analogous statement for Lemma 2.1. The implication (a) \Rightarrow (b) is also trivial. The proof that (b) \Rightarrow (c) is almost identical to the corresponding proof for Lemma 2.1. We may assume that $G \neq 0$ on a dense open $M' \subseteq M$. If $\Delta_B \equiv 0$, then $H = \alpha_H G$ for some $\alpha_H: M' \rightarrow \mathbb{R}$. We define L' exactly as in the proof of Lemma 2.1. Then it is easy to see that, if $E \in L'$, then $[F, E]$ and $[G, E]$ are in L' . So L' is an ideal in $L(F, G)$. Since $G \in L'$, we conclude that $L_0(F, G) \subseteq L'$, and so $L_0(F, G)(q)$ is one-dimensional for all $q \in M'$. Therefore the SAP fails at all $q \in M'$. Exactly as in the proof of Lemma 2.1, this implies that SAP fails at every $q \in M$. So (b) \Rightarrow (c). \square

We now consider a C.A.S. Σ that has the DSAP, and hence the DAP as well. Then the functions Δ_A and Δ_B both fail to vanish identically. We let $\text{NO}(\Sigma)$ denote the set of nonordinary points of Σ . Then $\text{NO}(\Sigma)$ is the set of zeros of the real analytic function $\Delta_A \Delta_B$, and so $\text{NO}(\Sigma)$ is an *analytic subset* of M . Moreover, since $\Delta_A \Delta_B$ does not vanish identically on M , and M is connected, the set $\text{NO}(\Sigma)$ is of dimension ≤ 1 . It follows from the general structure theory of analytic sets that $\text{NO}(\Sigma)$ is a locally finite union of arcs and points. More precisely, there is a locally finite partition \mathcal{P} of M into connected analytic embedded submanifolds of M that are semianalytic sets, such that $\text{NO}(\Sigma)$ is a union of members of \mathcal{P} , and that, if $S \in \mathcal{P}$, then $\text{Clos}(S) - S$ is also a union of members of \mathcal{P} , whose dimension is smaller than that of S . If $S \in \mathcal{P}$, and $S \subseteq \text{NO}(\Sigma)$, then S is either a point or an analytic arc. In the latter case, the function $\Delta_A \upharpoonright S$ either vanishes identically, or has a set of zeros S_1 which is discrete relative to S . Since S_1 is semianalytic, it follows that S_1 is locally finite (i.e., discrete) in M (so that S_1 has no accumulation points in $M - S$ either). Let $S'_1 = S_1$ if $\Delta_A \upharpoonright S$ does not vanish identically, and let $S'_1 = \emptyset$ if $\Delta_A \equiv 0$ on S . Define S_2, S'_2 similarly, using the function Δ_B instead of Δ_A . Also, define $S_3, S'_3, S_4, S'_4, S_5, S'_5$ using the vector-valued functions X, Y and G , respectively. Define S_6 to be the set of points $p \in S$ such that $X(p)$ is tangent to S . If X is not everywhere tangent to S , then S_6 is a discrete subset of S , and it is easy to show that S_6 is semianalytic in M , and so S_6 is discrete in M . (Indeed, suppose first that the functions $X^k(\Delta_A \Delta_B)$ vanish identically on S for all integers $k > 0$. Then it follows easily that $(\Delta_A \Delta_B)(\Phi_t^X(p)) = 0$ for all $p \in S$ and all $t \in \mathbb{R}$ for which $\Phi_t^X(p)$ is defined. If $p \in S$, then we claim that $\text{NO}(\Sigma) \cap U = S \cap U$ for some neighborhood U of p . To see this, suppose there were no such U . Then there would be a sequence $\{q_n\}$ such that $q_n \in \text{NO}(\Sigma)$, $q_n \notin S$, and $q_n \rightarrow p$ as $n \rightarrow \infty$. Since \mathcal{P} is locally finite, we may assume that all the q_n are in some $\tilde{S} \in \mathcal{P}$, which is necessarily one-dimensional and disjoint from S . But then $p \in \text{Clos}(\tilde{S}) - \tilde{S}$, and so $\{p\} \in \mathcal{P}$, contradicting the fact that $p \in S \in \mathcal{P}$, $\dim S = 1$. So U exists, and therefore there is an $\varepsilon > 0$ such that, if $|t| < \varepsilon$ and $\Phi_t^X(p) \in \text{NO}(\Sigma)$, it follows that $\Phi_t^X(p) \in S$. Since $(\Delta_A \Delta_B)(\Phi_t^X(p)) = 0$ for all t , it follows that $\Phi_t^X(p) \in \text{NO}(\Sigma)$ for all t , and then that $\Phi_t^X(p) \in S$ for small t . So $X(p)$ is tangent to S . But this is true for all p , which is a contradiction, since we are assuming that $S_6 \neq S$. So there is a smallest $k > 0$ with the property that $X^k(\Delta_A \Delta_B) \upharpoonright S$ is not identically zero. If we let $\eta = X^{k-1}(\Delta_A \Delta_B)$, then η is an analytic function on M , and $\eta \upharpoonright S \neq 0$. Let Q be the set of zeros of $X\eta$. Then Q is an analytic subset of M . So $Q \cap S$ is semianalytic in M . Since $Q \cap S$ is discrete in M , because $X\eta$ does not vanish identically on S , we conclude that $Q \cap S$ is discrete in M . If $p \in S_6$, then $(X\eta)(p) = 0$, because $\eta \upharpoonright S \neq 0$, and $X(p)$ is tangent to S . So $p \in Q \cap S$. Therefore $S_6 \subseteq Q \cap S$, and so S_6 is discrete in M .) We now define $S'_6 = S_6$ if $S_6 \neq S$, $S'_6 = \emptyset$ if

$S_6 = S$. We define S_7 and S'_7 similarly, using Y instead of X . If X is not everywhere tangent to S , then there is a $k \geq 0$ such that $X^k \Delta_B$ does not vanish identically on S , but that $X^i \Delta_B = 0$ on S for $0 \leq i < k$. (Indeed, if $X^k \Delta_B = 0$ for all k , then $X^k (\Delta_A \Delta_B) = 0$ for all k , and so X would be everywhere tangent to S , as shown above during the discussion of S_6 .) We let S'_8 be the set of zeros of $X^k \Delta_B$ on S , so that S'_8 is semianalytic in M . If X is everywhere tangent to S , let $S'_8 = \emptyset$. Define S'_9 similarly, using Y instead of X . Finally, let

$$(2.4) \quad S' = \bigcup_{i=1}^9 S'_i.$$

Then S' is discrete in M , for each $S \in \mathcal{P}$ such that $\dim S = 1$, $S \subseteq \text{NO}(\Sigma)$. Form a new partition \mathcal{P}' by letting \mathcal{P}' consist of: (a) all the connected components of $M - \text{NO}(\Sigma)$, (b) all the connected components of $S - S'$, for all the $S \in \mathcal{P}$ such that $S \subseteq \text{NO}(\Sigma)$, $\dim S = 1$, and (c) all the sets $\{p\}$, where $p \in S'$ for some $S \in \mathcal{P}$, $S \subseteq \text{NO}(\Sigma)$, $\dim S = 1$, or $\{p\} \in \mathcal{P}$, $p \in \text{NO}(\Sigma)$. Then \mathcal{P}' satisfies:

- (2.II.i) \mathcal{P}' is a locally finite partition of M , whose elements are embedded connected analytic submanifolds of M , which are also semianalytic subsets of M ;
- (2.II.ii) $\text{NO}(\Sigma)$ is the union of all the members of \mathcal{P}' which have dimension zero or one;
- (2.II.iii) If $P \in \mathcal{P}'$, then $(\text{Clos } P) - P$ is a union of members of \mathcal{P}' , of dimension strictly smaller than $\dim P$;
- (2.II.iv) Every one-dimensional $P \in \mathcal{P}'$ is a regular INOA, which is nondegenerate if it is of the antiturnpike type.

It follows from the preceding considerations that, with the exception of a discrete set, all the nonordinary points of Σ belong to regular INOA's which, if they are of the antiturnpike type, actually are nondegenerate. Precisely, let us say that a point $p \in \text{NO}(\Sigma)$ is *good* if $p \in S$ for some S such that: (a) S is a regular INOA, and (b) if S is of the antiturnpike type, then S is nondegenerate.

Let us say that p is *bad* if it is not good. Let us write $\text{NO}_g(\Sigma)$, $\text{NO}_b(\Sigma)$ for the sets of good and bad points $p \in \text{NO}(\Sigma)$, respectively. Then

$$(2.5) \quad \text{NO}(\Sigma) = \text{NO}_g(\Sigma) \cup \text{NO}_b(\Sigma)$$

and

$$(2.6) \quad \text{NO}_g(\Sigma) \cap \text{NO}_b(\Sigma) = \emptyset.$$

Moreover, the preceding remarks prove the following.

COROLLARY 2.3. *If Σ is a C.A.S. that has the DSAP, then the set $\text{NO}_b(\Sigma)$ consists of isolated points.*

In the terminology of § 6 of [A], $\text{NO}_b(\Sigma)$ is the complement of the set of near-ordinary points. Using Corollary 6.5 of [A], we conclude the following.

COROLLARY 2.4. *Let $p \in M$ be such that $p \notin \text{NO}_b(\Sigma)$, and suppose that Σ is a C.A.S. that has the DSAP. Then p has a neighborhood U such that*

$$\text{Opt}^1(\Sigma|U) \subseteq [\text{Traj}(X \vee Y \vee Z)]^5.$$

In order to complete the description of the local behavior of time-optimal trajectories, at least for systems which have the DSAP, we must study what happens in the neighborhood of each point in $\text{NO}_b(\Sigma)$. This will be done in § 3.

3. Bad nonordinary points. We now study the time-optimal trajectories in the neighborhood of a point $p \in \text{NO}_b(\Sigma)$, for a system Σ which is analytic and has the DSAP. As will be shown later (cf. Appendix A2), the case when $X(p) = Y(p) = 0$ (i.e., when $F(p) = G(p) = 0$) is fundamentally different. In this section we shall exclude this possibility, and we shall assume that $X(p) \neq 0$ or $Y(p) \neq 0$.

Precisely, here is a list of our assumptions for this section:

- (3.I.i) At least one of the vectors $X(p)$, $Y(p)$ is nonzero,
- (3.I.ii) X and Y are real analytic in a neighborhood of p ,
- (3.I.iii) The function $\Delta_A \Delta_B$ does not vanish identically in any neighborhood of p .

We shall distinguish three cases, namely

- (3.II.a) $X(p) = Y(p)$,
- (3.II.b) $X(p) \neq 0$, and $Y(p)$ is not of the form $rX(p)$ with $r \geq 1$, and
- (3.II.c) $Y(p) = rX(p)$, $r > 1$ or $X(p) = 0$.

If (3.II.c) holds, then we may interchange X and Y , and the resulting system satisfies (3.II.b). So, without loss of generality, we may assume that either (3.II.a) or (3.II.b) holds.

Our discussion of the two cases (3.II.a), (3.II.b) will be parallel, but there will be some significant differences. We begin by choosing an analytic coordinate chart $(U_0, (\xi, \eta))$, centred at p , such that U_0 is mapped by (ξ, η) onto a square $\text{Sq}(\varepsilon_0)$, that $\text{Clos}(U_0) \subseteq M$, that X and Y are analytic on U_0 , and that, relative to this chart, X has components $(1, 0)$. (This is possible because, if either (3.II.a) or (3.II.b) hold, then $X(p) \neq 0$.) If $X(p)$ and $Y(p)$ are linearly independent, we impose the additional requirement that $Y(p)$ have components $(0, 1)$.

Let α, β be the components of Y . If (3.II.a) holds, then $\alpha(p) = 1$, and so we may assume, by making ε_0 smaller, that $\alpha > 0$ throughout U_0 . If (3.II.b) holds, then either $X(p)$ and $Y(p)$ are linearly independent, in which case $\alpha(p) = 0$, or they are dependent, in which case $Y(p) = rX(p)$ for some constant $r < 1$. (Note that r may be negative or zero.) In that case, $\alpha(p) < 1$. In either case, we may assume, by making ε_0 smaller, that $\alpha < 1$ throughout U_0 .

We may also assume that $(\Delta_A \Delta_B)(p) = 0$. (Otherwise, p would be an ordinary point, in which case we already know that $\text{Traj}(X \vee Y \vee Z)$ is boundedly sufficient on some neighborhood of p .)

We now identify U_0 with $\text{Sq}(\varepsilon_0)$, and we list the assumptions made so far:

- (3.III.i) $U_0 = \text{Sq}(\varepsilon_0)$, $p = (0, 0)$, and $\text{Clos } U_0 \subseteq M$;
- (3.III.ii) $X = \partial_x$;
- (3.III.iii) $Y = \alpha \partial_x + \beta \partial_y$, where α and β are analytic functions on U_0 ;
- (3.III.iv) either $\alpha(0, 0) = 1$ and $\beta(0, 0) = 0$, in which case $\alpha(x, y) > 0$ for all $(x, y) \in U_0$, or $\alpha(x, y) < 1$ for all $(x, y) \in U_0$. The former situation occurs if (3.II.a) holds, and the latter if (3.II.b) holds;
- (3.III.v) $(\Delta_A \Delta_B)(p) = 0$, but $\Delta_A \Delta_B$ does not vanish identically on U_0 .

Our goal is to find an ε such that $0 < \varepsilon < \varepsilon_0$, and that $\text{Traj}(X \vee Y \vee Z)$ is boundedly sufficient on $\text{Sq}(\varepsilon)$. This will require that we construct an appropriate CASA stratification \mathcal{S} of $\text{Sq}(\varepsilon_0)$, and that we then restrict \mathcal{S} to $\text{Sq}(\varepsilon)$ in an appropriate sense.

We let $\text{Strat}_p(\varepsilon)$ denote the set of all CASA stratifications \mathcal{S} of $\text{Sq}(\varepsilon)$ such that $\{p\} \in \mathcal{S}$. We are interested in the subset $\text{Strat}_p^*(\varepsilon)$ that consists of those $\mathcal{S} \in \text{Strat}_p(\varepsilon)$

that are compatible with the vector fields X , Y , and ∂_y . Also, we are interested in the collection $\mathcal{A}(\varepsilon)$ of all subsets S of $\text{Sq}(\varepsilon)$ that satisfy

- (3.IV.i) S is a one-dimensional CASA subset of $\text{Sq}(\varepsilon)$,
- (3.IV.ii) S is a set of the form $\{(\psi(y), y) : a < y < b\}$ where $-\varepsilon \leq a < b \leq \varepsilon$, and $\psi :]a, b[\rightarrow]-\varepsilon, \varepsilon[$ is a real analytic function which is either constant or strictly monotonic.

Also, we let $\mathcal{A}'(\varepsilon)$ denote the set of all horizontal segments $\{(x, y) : a < x < b, y = y_0\}$, such that $-\varepsilon \leq a < b \leq \varepsilon$, $-\varepsilon < y_0 < \varepsilon$.

If $\mathcal{S} \in \text{Strat}_p(\varepsilon)$, and $i = 0, 1, 2$, we let \mathcal{S}_i denote the set of i -dimensional strata of \mathcal{S} other than $\{p\}$.

LEMMA 3.1. *If $\mathcal{S} \in \text{Strat}_p^*(\varepsilon)$, then*

$$(3.1) \quad \mathcal{S}_1 \subseteq \mathcal{A}(\varepsilon) \cup \mathcal{A}'(\varepsilon).$$

Proof. Let $S \in \mathcal{S}_1$. Since \mathcal{S}_1 is compatible with X , the vector field X is either everywhere tangent to S , or nowhere tangent to S . In the former case, $S \in \mathcal{A}'(\varepsilon)$. In the latter case, the function $(x, y) \rightarrow y$ has a nonzero directional derivative in the direction tangential to S , at every point of S . So S can be parametrized by the y -coordinate. Therefore S is of the form $\{(\psi(y), y) : a < y < b\}$, where $\psi :]a, b[\rightarrow \mathbb{R}$ is real analytic. Since $S \subseteq \text{Sq}(\varepsilon)$, it is clear that $-\varepsilon \leq a < b \leq \varepsilon$, and that ψ takes values in $]-\varepsilon, \varepsilon[$. Since \mathcal{S} is compatible with ∂_y , the derivative $d\psi/dy$ either vanishes for all y or for no y . In the former case, ψ is constant. In the latter case, it is strictly monotonic. \square

If $S \in \mathcal{A}(\varepsilon)$, and if $0 < \delta < \varepsilon$, we let $S^\delta = S \cap \text{Sq}(\delta)$. If S is the graph of $\psi :]a, b[\rightarrow]-\varepsilon, \varepsilon[$ as in (3.IV.ii), then S^δ is the graph of $\psi^\delta :]a^\delta, b^\delta[\rightarrow]-\delta, \delta[$, where $]a^\delta, b^\delta[$ is the intersection of the intervals $]-\delta, \delta[$ and $\{y : a < y < b, -\delta < \psi(y) < \delta\}$, and $\psi^\delta = \psi|_{]a^\delta, b^\delta[}$. So we have the following.

LEMMA 3.2. *If $S \in \mathcal{A}(\varepsilon)$, and $0 < \delta < \varepsilon$, then $S^\delta \in \mathcal{A}(\delta)$. \square*

Now suppose $\mathcal{S} \in \text{Strat}_p^*(\varepsilon_0)$. We say that ε is *good* for \mathcal{S} if

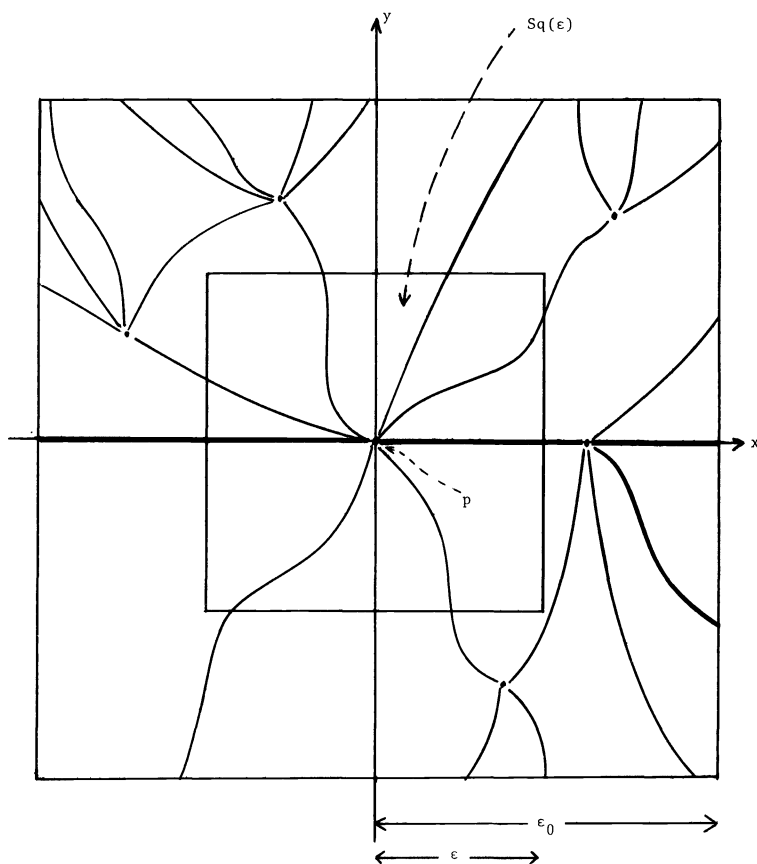
- (3.V.i) $0 < \varepsilon < \varepsilon_0$,
- (3.V.ii) the only zero-dimensional stratum of \mathcal{S} that is contained in $\text{Sq}(\varepsilon)$ is $\{p\}$,
- (3.V.iii) if $S \in \mathcal{S}_1$, and $S \cap \text{Sq}(\varepsilon) \neq \emptyset$, then $p \in \text{Clos } S$.

We then have (cf. Fig. 1) the following.

LEMMA 3.3. *For every $\mathcal{S} \in \text{Strat}_p^*(\varepsilon_0)$ there exists a good ε .*

Proof. Pick ε_1 such that $0 < \varepsilon_1 < \varepsilon_0$. Since $\text{Sq}(\varepsilon_1)$ is relatively compact in $\text{Sq}(\varepsilon_0)$, only a finite number of strata of \mathcal{S} meet $\text{Sq}(\varepsilon_1)$. Let P be the union of all the zero-dimensional strata $\{q\}$ of \mathcal{S} such that $q \neq p$, $q \in \text{Sq}(\varepsilon_0)$, and of the closures of all the one-dimensional strata $S \in \mathcal{S}$ such that $p \notin \text{Clos } S$. Then P is closed, and $p \notin P$. If $0 < \varepsilon < \varepsilon_1$, and $P \cap \text{Sq}(\varepsilon) = \emptyset$, then ε is good. \square

If $\mathcal{S} \in \text{Strat}_p^*(\varepsilon_0)$, and ε is good for \mathcal{S} , we let $\mathcal{S}_1(\varepsilon)$ denote the set of all sets S^ε , as S ranges over all the one-dimensional strata of \mathcal{S} such that $p \in \text{Clos } S$. If $T \in \mathcal{S}_1(\varepsilon)$, then T is of the form $\{(\psi(y), y) : a < y < b\}$ for some strictly monotonic or constant analytic ψ , or T is a horizontal segment (because, if $T = S^\varepsilon$, $S \in \mathcal{S}_1$, then $S \in \mathcal{A}(\varepsilon_0) \cup \mathcal{A}'(\varepsilon_0)$ by Lemma 3.1; if $S \in \mathcal{A}(\varepsilon_0)$ then $T \in \mathcal{A}(\varepsilon)$, and if $S \in \mathcal{A}'(\varepsilon_0)$ then $T \in \mathcal{A}'(\varepsilon)$). In the former case, the fact that $p \in \text{Clos } T$ but $p \notin T$ implies that either $a = 0$ or $b = 0$. (If $a < 0 < b$, $\psi(0) \neq 0$, then $p \notin \text{Clos } T$; if $a < 0 < b$, $\psi(0) = 0$, then $p \in T$, and if either $0 < a < b$ or $a < b < 0$, then $p \notin \text{Clos } T$.) If $a = 0$, then $\psi(y)$ must converge to zero as $y \rightarrow 0+$. If we let $L = \lim_{y \rightarrow b-} \psi(y)$, then the point (b, L) belongs to the closure of T

FIG. 1. A stratification in $\text{Strat}_p^*(\epsilon_0)$ and a good ϵ .

(in $\text{Sq}(\epsilon_0)$). If $T = S^\epsilon$, $S \in \mathcal{S}_1$, we either have $(b, L) \in S$ or $(b, L) \in (\text{Clos } S) - S$. In the former case, (b, L) cannot be in $\text{Sq}(\epsilon)$ (otherwise (b, L) would be in T). In the latter case, $\{(b, L)\}$ is a zero-dimensional stratum of \mathcal{S} , and $(b, L) \neq (0, 0)$, so that, again, $(b, L) \notin \text{Sq}(\epsilon)$. Therefore b must equal ϵ , or $|L|$ must be ϵ . In any case, we see that T disconnects $\text{Sq}(\epsilon)^+$, where

$$(3.2a) \quad \text{Sq}(\epsilon)^+ = \{(x, y): |x| < \epsilon, 0 < y < \epsilon\}.$$

If $b = 0$, a similar reasoning shows that $\psi(b-) = 0$, and that either $\psi(a+) = \pm\epsilon$, or $a = -\epsilon$. Also, if we let

$$(3.2b) \quad \text{Sq}(\epsilon)^- = \{(x, y): |x| < \epsilon, -\epsilon < y < 0\},$$

we see that T disconnects $\text{Sq}(\epsilon)^-$.

Finally, there is the possibility that T is a segment $\{(x, y): a < x < b, y = y_0\}$. Since $p \in \text{Clos } T$, we necessarily have $y_0 = 0$. Also, a or b must equal 0 and, if $a = 0$, b must equal ϵ whereas, if $b = 0$, then $a = -\epsilon$. So $T = \Gamma_\epsilon^+$ or $T = \Gamma_\epsilon^-$, where

$$(3.3a) \quad \Gamma_\epsilon^+ = \{(x, y): 0 < x < \epsilon, y = 0\},$$

$$(3.3b) \quad \Gamma_\epsilon^- = \{(x, y): -\epsilon < x < 0, y = 0\}.$$

Let us define $\mathcal{A}_p^+(\epsilon)$ to be the set of all CASA subsets of $\text{Sq}(\epsilon_0)$ that are of the form $\{(\psi(y), y): 0 < y < b\}$, where $0 < b \leq \epsilon$, and where ψ is an analytic function on

$]0, b[$, which is either constant or strictly monotonic, and satisfies $\psi(0+) = 0$, and either $\psi(b-) = \pm \varepsilon$ or $b = \varepsilon$. We define $\mathcal{A}_p^-(\varepsilon)$ similarly, except that, instead of an interval $]0, b[$, the domain of ψ is an interval of the form $]a, 0[$, and the behavior of ψ at the endpoints is given by $\psi(0-) = 0$, and $\psi(a+) = \pm \varepsilon$ or $a = -\varepsilon$. Then we can summarize our preceding remarks in the following.

LEMMA 3.4. *Let $\mathcal{S} \in \text{Strat}_p^*(\varepsilon_0)$, and let ε be good for \mathcal{S} . Then every $T \in \mathcal{S}_1(\varepsilon)$ is either one of the segments Γ_ε^\pm , or $T \in \mathcal{A}_p^+(\varepsilon)$, or $T \in \mathcal{A}_p^-(\varepsilon)$. If $T \in \mathcal{A}_p^+(\varepsilon)$, then T disconnects $\text{Sq}(\varepsilon)^+$ and, similarly, if $T \in \mathcal{A}_p^-(\varepsilon)$, then T disconnects $\text{Sq}(\varepsilon)^-$. \square*

Suppose that $T = S^\varepsilon$ for $S \in \mathcal{S}_1$, $\mathcal{S} \in \text{Strat}_p^*(\varepsilon_0)$, that ε is good for \mathcal{S} , and that $T \neq \Gamma_\varepsilon^+$, $T \neq \Gamma_\varepsilon^-$. Then $T \in \mathcal{A}_p^+(\varepsilon) \cup \mathcal{A}_p^-(\varepsilon)$, and T splits the half-square $\text{Sq}(\varepsilon)^+$ (or $\text{Sq}(\varepsilon)^-$) into two connected components, T_L and T_R , where T_R is the component into which X points, and T_L is the other one. We refer to T_R and T_L as the “right” and “left” sides of T , but a word of caution is needed: a vector v , at a point $q \in T$, may very well “point right” (i.e., have a positive x -component) and point to the “left” of T (i.e., to T_L). (This will happen if v has a larger slope than T at q .) Since \mathcal{S} is compatible with Y , the vector $Y(q)$ points to T_R for all q , or to T_L for all q , or is tangent to T for all q . In particular, we have the following.

LEMMA 3.5. *Let $\mathcal{S} \in \text{Strat}_p^*(\varepsilon_0)$, let ε be good for \mathcal{S} , and let $T = S^\varepsilon$ for some $S \in \mathcal{S}_1$. Suppose that $T \neq \Gamma_\varepsilon^+$ and $T \neq \Gamma_\varepsilon^-$. Let $U^T = \text{Sq}(\varepsilon)^+$ if $T \in \mathcal{A}_p^+(\varepsilon)$, $U^T = \text{Sq}(\varepsilon)^-$ if $T \in \mathcal{A}_p^-(\varepsilon)$. Then either (a) $Y(q)$ points towards T_L for all $q \in T$, or (b) T is a barrier in U^T . \square*

We now construct a particular stratification $\mathcal{S} \in \text{Strat}_p^*(\varepsilon_0)$. We let $E(\varepsilon)$ denote, for $0 < \varepsilon \leq \varepsilon_0$, the set of zeros of $\Delta_A \Delta_B$ in $\text{Sq}(\varepsilon)$ (i.e., $E(\varepsilon) = \text{Sq}(\varepsilon) \cap \text{NO}(\Sigma)$). Then $E(\varepsilon_0)$ is an analytic subset of $\text{Sq}(\varepsilon_0)$. Moreover, for $0 < \varepsilon \leq \varepsilon_0$:

$$E(\varepsilon) = E_1(\varepsilon) \cup E_2(\varepsilon) \cup E_3(\varepsilon),$$

where the (possibly overlapping) sets $E_i(\varepsilon)$ are defined by

$$(3.4a) \quad E_1(\varepsilon) = \{(x, y) \in \text{Sq}(\varepsilon) : \Delta_A(x, y) = 0\},$$

$$(3.4b) \quad E_2(\varepsilon) = \{(x, y) \in \text{Sq}(\varepsilon) : \Delta_B(x, y) = 0\},$$

$$(3.4c) \quad E_3(\varepsilon) = \{(x, y) \in \text{Sq}(\varepsilon) : G(x, y) = 0\}.$$

(Notice that, since

$$(3.5) \quad G = \frac{1}{2}((\alpha - 1)\partial_x + \beta\partial_y),$$

the set $E_3(\varepsilon_0)$ is necessarily empty if (3.II.b) holds.)

We let $E = E(\varepsilon_0)$, $E_i = E_i(\varepsilon_0)$. The sets E_i are analytic and have dimension zero or one. We pick a CASA stratification $\hat{\mathcal{S}}$ of $\text{Sq}(\varepsilon_0)$ which is compatible with the sets E_i , $i = 1, 2, 3$, with $\{p\}$, with $\text{Sq}(\varepsilon_1)$ (where ε_1 is some number such that $0 < \varepsilon_1 < \varepsilon_0$) and with the vector fields X , Y , and ∂_y . We then define \mathcal{S} by modifying $\hat{\mathcal{S}}$ as follows.

For each $S \in \hat{\mathcal{S}}_1$ which meets $\text{Sq}(\varepsilon_1)$, let $k^X(S)$ (resp. $k^Y(S)$) be the smallest nonnegative integer k such that $X^k \Delta_B$ (resp. $Y^k \Delta_B$) is not identically zero on S . (If no such k exists let $k^X(S)$ (resp. $k^Y(S)$) be $+\infty$.) Then let S_X (resp. S_Y) be the intersection of S with the zero set of $X^{k^X}(S) \Delta_B$ (resp. $Y^{k^Y}(S) \Delta_B$). If $k^X(S)$ (resp. $k^Y(S)$) is infinite, let S_X (resp. S_Y) be empty. Then S_X and S_Y are subanalytic subsets of $\text{Sq}(\varepsilon_0)$, and they are clearly discrete. If $S \in \hat{\mathcal{S}}_1$, and S meets $\text{Sq}(\varepsilon_1)$, the set $\check{S} = (S_X \cup S_Y) \cap \text{Sq}(\varepsilon_1)$ is therefore finite. Form a new collection S^* of sets whose elements are the connected components of $S - \check{S}$, and the sets $\{q\}$, $q \in \check{S}$. Then S^* is a partition of S . Form \mathcal{S} by replacing each $S \in \hat{\mathcal{S}}_1$ which meets $\text{Sq}(\varepsilon_1)$ by the collection

of all members of S^* . The stratification \mathcal{S} obtained in this fashion has the following properties:

- (3.VI.i) \mathcal{S} is a CASA stratification of $\text{Sq}(\varepsilon_0)$, compatible with the sets E_i , with $\{p\}$, and with the vector fields X , Y , and ∂_y .
- (3.VI.ii) If $S \in \mathcal{S}_1$, $p \in \text{Clos } S$, and if V is either X or Y , then either V is everywhere tangent to S or, if it is not, and if k is the first nonnegative integer j for which $V^j \Delta_B$ is not $\equiv 0$ on S , then $V^k \Delta_B$ never vanishes on S .

A stratification \mathcal{S} with properties (3.VI.i, ii) above will be called an *appropriate stratification* of $\text{Sq}(\varepsilon_0)$. The following trivial observation will be useful later.

LEMMA 3.6. *If \mathcal{S} is an appropriate stratification of $\text{Sq}(\varepsilon_0)$, and if \mathcal{S}' is a CASA stratification of $\text{Sq}(\varepsilon_0)$ which is a refinement of \mathcal{S} , then \mathcal{S}' is also appropriate.* \square

The remarks preceding the definition of an appropriate stratification prove that such stratifications exist. From now on, we let \mathcal{S} be a fixed appropriate stratification of $\text{Sq}(\varepsilon_0)$.

We now return to our search for an $\varepsilon > 0$ such that $\text{Traj}(X \vee Y \vee Z)^\infty$ is boundedly sufficient for $\text{Sq}(\varepsilon)$. We begin by proving some lemmas that reduce the problem to that of finding bounds for the number of switchings for some special families of trajectories in the half-squares $\text{Sq}(\varepsilon)^\pm$.

LEMMA 3.7. *Let ε be good for \mathcal{S} . Suppose that $\nu > 0$ is such that*

$$(3.6) \quad \text{Opt}^1(\Sigma \upharpoonright \text{Sq}(\varepsilon)^+) \cup \text{Opt}^1(\Sigma \upharpoonright \text{Sq}(\varepsilon)^-) \subseteq \text{Traj}(X \vee Y \vee Z)^\nu.$$

Then

$$(3.7) \quad \text{Opt}^1(\Sigma \upharpoonright (\text{Sq}(\varepsilon) - \{p\})) \subseteq \text{Traj}(X \vee Y \vee Z)^{3\nu+2}.$$

Proof. We first prove that, if $Q = \Gamma_\varepsilon^+$ or $Q = \Gamma_\varepsilon^-$, then Y is either everywhere tangent to Q or nowhere tangent to Q . It is clear that Y is tangent to Q at a point $q \in Q$ if and only if $\beta(q) = 0$, i.e. iff $\Delta_A(q) = 0$ (since $\Delta_A = \beta$). If Y is tangent to Q at some point $q \in Q$, then $\Delta_A(q) = 0$ and so $q \in E_1$. So $q \in S$ for an $S \in \mathcal{S}$ such that $S \subseteq E_1$. Therefore $\dim S = 0$ or 1 . Since ε is good, $\dim S \neq 0$. So $S \in \mathcal{S}_1$, and $q \in S^\varepsilon$. Since $q \in \Gamma_\varepsilon^- \cup \Gamma_\varepsilon^+$, the set S^ε cannot belong to $\mathcal{A}_p^+(\varepsilon) \cup \mathcal{A}_p^-(\varepsilon)$. So Lemma 3.4 implies that $S^\varepsilon = \Gamma_\varepsilon^-$ or $S^\varepsilon = \Gamma_\varepsilon^+$. Therefore $S^\varepsilon = Q$. We conclude that $\Delta_A \equiv 0$ on Q , i.e., that Y is everywhere tangent to Q .

Since X is everywhere tangent to Γ_ε^- and to Γ_ε^+ , we conclude that Γ_ε^- is a barrier in $\text{Sq}(\varepsilon) - (\{p\} \cup \Gamma_\varepsilon^+)$, and that Γ_ε^+ is a barrier in $\text{Sq}(\varepsilon) - (\{p\} \cup \Gamma_\varepsilon^-)$.

So, by Lemma 4.1 of [A], if γ is a trajectory in $\text{Sq}(\varepsilon) - (p)$, then γ cannot leave one of the segments Γ_ε^- , Γ_ε^+ and return to it unless it crosses the other one first.

Now suppose that (3.II.a) holds. Then $\alpha > 0$ throughout U_0 , and therefore every trajectory of Σ in U_0 goes from left to right. If a trajectory $\gamma \in \text{Traj}(\Sigma \upharpoonright (\text{Sq}(\varepsilon) - \{p\}))$ starts, say, at a point in $\text{Sq}(\varepsilon)^-$, then γ either stays in $\text{Sq}(\varepsilon)^-$ forever, in which case it is in $\text{Traj}(X \vee Y \vee Z)^\nu$, or it crosses into $\text{Sq}(\varepsilon)^+$ through Γ_ε^- or Γ_ε^+ (possibly after staying in Γ_ε^- or Γ_ε^+ for some time interval). If γ crosses through Γ_ε^+ , then it cannot cross again into $\text{Sq}(\varepsilon)^-$ unless it goes through Γ_ε^- . But this is impossible because γ goes from left to right. If γ crosses through Γ_ε^- , then it may cross again into $\text{Sq}(\varepsilon)^-$ through Γ_ε^+ , but once this crossing has taken place then no more crossings are possible. Therefore $\gamma \in \text{Traj}(X \vee Y \vee Z)^{3\nu+2}$. So, if (3.II.a) holds, we have proved that

$$\text{Opt}^1(\Sigma \upharpoonright (\text{Sq}(\varepsilon) - \{p\})) \subseteq \text{Traj}(X \vee Y \vee Z)^{3\nu+2}.$$

The case when (3.II.b) holds is slightly more difficult. The main observation for this case is that no trajectory $\gamma \in \text{Traj}(\Sigma \setminus (\text{Sq}(\varepsilon) - \{p\}))$ that goes from a $q_- \in \Gamma_\varepsilon^-$ to a $q_+ \in \Gamma_\varepsilon^+$ can belong to $\text{Opt}^1(\Sigma \setminus \text{Sq}(\varepsilon) - \{p\})$.

To prove this, we first show that $T(\gamma) > T(\gamma')$, where γ' is the X -trajectory from q_- to q_+ . Let $q_- = (x_-, 0)$, $q_+ = (x_+, 0)$ (so that $x_- < 0 < x_+$). Let $\gamma: [a, b] \rightarrow \text{Sq}(\varepsilon) - \{p\}$ have the form $t \rightarrow (x(t), y(t))$, and let γ correspond to the control $u(\cdot): [a, b] \rightarrow [-1, 1]$. Then

$$(3.8) \quad \dot{x}(t) = \frac{1}{2}[(1 - u(t)) + \alpha(x(t), y(t))(1 + u(t))]$$

for almost all t . Since $\alpha < 1$ throughout $\text{Sq}(\varepsilon)$, (3.8) gives

$$(3.9) \quad \dot{x}(t) \leq 1 \quad \text{for almost all } t.$$

Therefore

$$(3.10) \quad x_+ - x_- = \int_a^b \dot{x}(t) dt \leq b - a,$$

i.e., $T(\gamma') \leq T(\gamma)$. Moreover, equality is only possible if $\dot{x}(t) = 1$ for almost all t , i.e., if $u(t) = -1$ for almost all t . But in this case γ would be an X -trajectory, and then γ would go through p , contradicting the fact that γ is a trajectory in $\text{Sq}(\varepsilon) - \{p\}$. So $T(\gamma') < T(\gamma)$.

The inequality $T(\gamma') < T(\gamma)$ does not yet prove that γ is not in $\text{Opt}^1(\Sigma \setminus (\text{Sq}(\varepsilon) - \{p\}))$, because $\gamma' \notin \text{Traj}(\Sigma \setminus (\text{Sq}(\varepsilon) - \{p\}))$ (since γ' goes through p). However, it is clear that Y cannot be everywhere tangent to $\Gamma_\varepsilon^- \cup \Gamma_\varepsilon^+ \cup \{p\}$, and that Y cannot always point to the same side of this segment either. (Otherwise $\Gamma_\varepsilon^- \cup \Gamma_\varepsilon^+ \cup \{p\}$ would be a barrier in $\text{Sq}(\varepsilon)$, contradicting the fact that γ leaves it and later returns.) So Y points to one side of $\Gamma_\varepsilon^- \cup \Gamma_\varepsilon^+ \cup \{p\}$ for $q \in \Gamma_\varepsilon^-$, and to the other side for $q \in \Gamma_\varepsilon^+$. If $\rho > 0$ is very small, then we can form a trajectory γ'_ρ from q_- to q_+ by following first a Y -trajectory until we get to the line $y = +\rho$ or $y = -\rho$ (depending on whether $Y(q_-)$ points up or down), then an X -trajectory until we get to a point in the Y -trajectory through q_+ , and finally this Y -trajectory to q_+ . The curves γ'_ρ are in $\text{Traj}(\Sigma \setminus (\text{Sq}(\varepsilon) - \{p\}))$, and they converge to γ' as $\rho \rightarrow 0$. So, for sufficiently small ρ , $T(\gamma'_\rho) < T(\gamma)$, proving that $\gamma \notin \text{Opt}^1(\Sigma \setminus (\text{Sq}(\varepsilon) - \{p\}))$.

Now let $\gamma \in \text{Opt}^1(\Sigma \setminus (\text{Sq}(\varepsilon) - \{p\}))$. Then, once γ crosses Γ_ε^- , it cannot cross Γ_ε^- again unless it crosses Γ_ε^+ first, and it cannot cross Γ_ε^+ because then it would not be optimal. So the worst that can happen is that γ starts in $\text{Sq}(\varepsilon)^-$, or in $\text{Sq}(\varepsilon)^+$, then crosses Γ_ε^+ (possibly after staying in Γ_ε^+ for some time), then crosses Γ_ε^- (again, after staying there for some time) back to the region where it started, and then it stays in this region forever. So, again, γ must belong to $\text{Traj}(X \vee Y \vee Z)^{3\nu+2}$. \square

LEMMA 3.8. *If ε is good for \mathcal{S} , then there is an integer $\nu > 0$ such that every $\gamma \in \text{Opt}^1(\Sigma \setminus \text{Sq}(\varepsilon)^-) \cap \text{Traj}(Y \vee Z)^\infty$ or $\text{Opt}^1(\Sigma \setminus \text{Sq}(\varepsilon)^+) \cap \text{Traj}(Y \vee Z)^\infty$ is actually in $\text{Traj}(Y \vee Z)^\nu$.*

Proof. If $\gamma \in \text{Opt}^1(\Sigma \setminus \text{Sq}(\varepsilon)^+)$, and γ is a trajectory in $\text{Traj}(Y \vee Z)^\infty$, then γ is a concatenation of finitely many pieces which either are Y -trajectories or are entirely contained in some S^ε that is a turnpike, for some $S \in \mathcal{S}_1$. Then $S^\varepsilon \in \mathcal{A}_p^+(\varepsilon)$. For any given S^ε which is a turnpike, the complement of S^ε in $\text{Sq}(\varepsilon)^+$ has two connected components, and Y points into one of them. Therefore γ can never return to S^ε after it has left it. So γ cannot have more than μ Z -pieces, where μ is the number of sets S^ε , $S \in \mathcal{S}_1$, that are turnpikes. So $\gamma \in \text{Traj}(Y \vee Z)^\nu$, where $\nu = 2\mu + 1$.

A similar reasoning shows that, if $\gamma \in \text{Traj}(Y \vee Z)^\infty$ is an optimal trajectory in $\text{Sq}(\varepsilon)^-$, then $\gamma \in \text{Traj}(Y \vee Z)^\nu$. \square

A similar proof shows the following.

LEMMA 3.9. *If ε is good for \mathcal{S} , then there is an integer $\nu > 0$ such that*

$$[\text{Opt}^1(\Sigma \upharpoonright \text{Sq}(\varepsilon)^-) \cup \text{Opt}^1(\Sigma \upharpoonright \text{Sq}(\varepsilon)^+)] \cap \text{Traj}(X \vee Z)^\infty \subseteq \text{Traj}(X \vee Z)^\nu. \quad \square$$

If $\gamma \in \text{Traj}(X \vee Y \vee Z)^\infty$, let $\hat{I}_0(\gamma)$ denote the collection of all intervals J such that

- (3.VII.i) $J \subseteq \text{Dom}(\gamma)$;
- (3.VII.ii) J contains some interval J' , with nonempty interior, such that $\gamma \upharpoonright J' \in \text{Traj}(X)$;
- (3.VII.iii) $\gamma \upharpoonright J \in \text{Traj}(X \vee Z)^\infty$.

We let $\hat{I}(\gamma)$ denote the set of all maximal elements of $\hat{I}_0(\gamma)$, and we let $I(\gamma)$ denote the set of those $J \in \hat{I}(\gamma)$ that do not contain any of the endpoints of $\text{Dom}(\gamma)$. Then, if $\text{Dom}(\gamma) = [a, b]$, an interval J belongs to $I(\gamma)$ if and only if

- (3.VIII.i) $J = [t_0, t_1]$ for some t_0, t_1 such that $a < t_0 < t_1 < b$;
- (3.VIII.ii) $\gamma \upharpoonright J \in \text{Traj}(X \vee Z)^\infty$;
- (3.VIII.iii) there exist t'_0, t'_1 such that $t_0 \leq t'_0 < t'_1 \leq t_1$ and $\gamma \upharpoonright [t'_0, t'_1] \in \text{Traj}(X)$;
- (3.VIII.iv) there exists $\delta > 0$ such that $a \leq t_0 - \delta$, $b \geq t_1 + \delta$, and that $\gamma \upharpoonright [t_0 - \delta, t_0]$ and $\gamma \upharpoonright [t_1, t_1 + \delta]$ are Y -trajectories.

For each $\gamma \in \text{Traj}(X \vee Y \vee Z)^\infty$, let $m(\gamma)$ denote the number of intervals $J \in I(\gamma)$. For $0 < \varepsilon \leq \varepsilon_0$, let

$$(3.11a) \quad m_\varepsilon^+ = \sup \{m(\gamma) : \gamma \in \text{Opt}^1(\Sigma \upharpoonright \text{Sq}(\varepsilon)^+) \cap \text{Traj}(X \vee Y \vee Z)^\infty\};$$

$$(3.11b) \quad m_\varepsilon^- = \sup \{m(\gamma) : \gamma \in \text{Opt}^1(\Sigma \upharpoonright \text{Sq}(\varepsilon)^-) \cap \text{Traj}(X \vee Y \vee Z)^\infty\}.$$

The main fact to be proved is the following.

LEMMA 3.10. *If ε is sufficiently small, then m_ε^+ and m_ε^- are finite.*

Before we prove Lemma 3.10, let us show that this lemma solves our problem, i.e. let us prove the following.

LEMMA 3.11. *Suppose that ε is good for \mathcal{S} , and that m_ε^+ and m_ε^- are finite. Then $\text{Traj}(X \vee Y \vee Z)$ is boundedly sufficient for $\text{Sq}(\varepsilon)$.*

Proof. If $\gamma \in \text{Opt}^1(\Sigma \upharpoonright \text{Sq}(\varepsilon))$, then γ cannot have a loop. So γ goes through p at most once. Therefore γ is the concatenation of at most two trajectories, each of which is entirely contained in $\text{Sq}(\varepsilon) - \{p\}$ (except possibly for one of its endpoints). So our conclusion will follow if we prove that $\text{Traj}(X \vee Y \vee Z)$ is boundedly sufficient for $\text{Sq}(\varepsilon) - \{p\}$.

In view of Lemma 3.7, it is sufficient to prove that $\text{Traj}(X \vee Y \vee Z)$ is boundedly sufficient for $\text{Sq}(\varepsilon)^+$ and for $\text{Sq}(\varepsilon)^-$.

If ε is good for \mathcal{S} , then Lemmas 3.8 and 3.9 tell us that there is an integer ν such that, whenever γ is an optimal trajectory in $\text{Sq}(\varepsilon)^+$, or in $\text{Sq}(\varepsilon)^-$, such that $\gamma \in \text{Traj}(X \vee Z)^\infty$ or that $\gamma \in \text{Traj}(Y \vee Z)^\infty$, then $\gamma \in \text{Traj}(X \vee Y \vee Z)^\nu$. If γ is an arbitrary optimal trajectory in $\text{Sq}(\varepsilon)^+$ or in $\text{Sq}(\varepsilon)^-$, then γ must be in $\text{Traj}(X \vee Y \vee Z)^\infty$, because every point of $\text{Sq}(\varepsilon)^+ \cup \text{Sq}(\varepsilon)^-$ is either an ordinary point or a member of a regular, nondegenerate INOA. So γ is a concatenation of at most $m(\gamma) + 2$ pieces that are in $\text{Traj}(X \vee Z)^\infty$, and of at most $m(\gamma) + 1$ pieces that belong to $\text{Traj}(Y \vee Z)^\infty$. Each of these pieces belongs to $\text{Traj}(X \vee Y \vee Z)^\nu$, and so

$$\gamma \in \text{Traj}(X \vee Y \vee Z)^{\mu(\gamma)},$$

where $\mu(\gamma) = (2m(\gamma) + 3)\nu$.

If the numbers m_ε^+ , m_ε^- are finite, then $\mu < \infty$, where

$$\mu = \sup \{\mu(\gamma) : \gamma \in \text{Opt}^1(\Sigma \upharpoonright \text{Sq}(\varepsilon)^+) \cup \text{Opt}^1(\Sigma \upharpoonright \text{Sq}(\varepsilon)^-)\}.$$

Also, $\gamma \in \text{Traj}(X \vee Y \vee Z)^\mu$. Therefore $\text{Traj}(X \vee Y \vee Z)$ is boundedly sufficient for $\text{Sq}(\varepsilon)^+$ and for $\text{Sq}(\varepsilon)^-$. As explained before, our desired conclusion follows. \square

We now must prove Lemma 3.10. We let \mathcal{H} denote the set of all trajectories $\gamma \in \text{Traj}(X \vee Y \vee Z)^\infty$ such that, for some $a, b, t_0, t_1, t'_0, t'_1$, the following hold:

- (3.IX.i) $\text{Dom}(\gamma) = [a, b]$;
- (3.IX.ii) $a < t_0 \leq t'_0 < t'_1 \leq t_1 < b$;
- (3.IX.iii) $\gamma|_{[a, t_0]}$ and $\gamma|_{[t_1, b]}$ are Y -trajectories;
- (3.IX.iv) $\gamma|_{[t_0, t_1]} \in \text{Traj}(X \vee Z)$;
- (3.IX.v) $\gamma|_{[t'_0, t'_1]} \in \text{Traj}(X)$.

Lemma 3.10 will follow from the following.

LEMMA 3.12. *If ε is sufficiently small, then there exist finite sets $\mathcal{B}^+, \mathcal{B}^-$, such that*

- (3.X.a) *If $B \in \mathcal{B}^+$ then B is a barrier in $\text{Sq}(\varepsilon)^+$, and if $B \in \mathcal{B}^-$ then B is a barrier in \mathcal{B}^- ;*
- (3.X.b) *If $B_1 \neq B_2$ and both B_1 and B_2 are in $\mathcal{B}^+ \cup \mathcal{B}^-$, then $B_1 \cap B_2 = \emptyset$;*
- (3.X.c) *If $q \in B \in \mathcal{B}^+ \cup \mathcal{B}^-$, then $X(q)$ is not tangent to B at q , and*
- (3.X.d) *If $\gamma \in \mathcal{H}$, and γ is an optimal trajectory in $\text{Sq}(\varepsilon)^+$ or in $\text{Sq}(\varepsilon)^-$, then γ goes through some point of some $B \in \mathcal{B}^+ \cup \mathcal{B}^-$.*

Before we prove Lemma 3.12, let us show that this lemma implies Lemma 3.10. Let γ be an arbitrary optimal trajectory in $\text{Sq}(\varepsilon)^+$, or in $\text{Sq}(\varepsilon)^-$, such that $\gamma \in \text{Traj}(X \vee Y \vee Z)^\infty$. Let $J_1(\gamma), \dots, J_m(\gamma)(\gamma)$ be the members of $I(\gamma)$, ordered from left to right. Each $J_i(\gamma)$ is immediately preceded by a maximal interval $J_i(\gamma)_-$ such that $\gamma|_{J_i(\gamma)_-}$ is a Y -trajectory, and it is followed by another maximal interval $J_i(\gamma)_+$ such that $\gamma|_{J_i(\gamma)_+}$ is a Y -trajectory.

We let $J_i^*(\gamma) = J_i(\gamma)_- \cup J_i(\gamma) \cup J_i(\gamma)_+$. Then we define $\gamma_i = \gamma|_{J_i^*(\gamma)}$. The trajectories γ_i are in \mathcal{H} , and they are optimal. So each γ_i meets some $B \in \mathcal{B}^+ \cup \mathcal{B}^-$.

For each $B \in \mathcal{B}^+ \cup \mathcal{B}^-$, let $N(B)$ denote the set of indices $i \in \{1, 2, \dots, m(\gamma)\}$ such that γ_i meets B . We claim that $N(B)$ consists of at most two indices. Indeed, suppose there are three different indices i, j, k that belong to $N(B)$. Order them so that $i < j < k$. Then there are times t_i, t_k such that $t_i \in J_i^*(\gamma), t_k \in J_k^*(\gamma)$ and that $\gamma(t_i) \in B, \gamma(t_k) \in B$. Since γ is entirely contained in $\text{Sq}(\varepsilon)^+$ (or in $\text{Sq}(\varepsilon)^-$) and B is a barrier in $\text{Sq}(\varepsilon)^+$ (or in $\text{Sq}(\varepsilon)^-$), Lemma 4.1 of [A] implies that $\gamma|_{[t_i, t_k]}$ is entirely contained in B . On the other hand,

$$J_j(\gamma) \subseteq [t_i, t_k],$$

and therefore $\gamma|_{J_j(\gamma)}$ is entirely contained in B . However, the fact that $J_j(\gamma) \in I(\gamma)$ implies that $J_j(\gamma)$ contains some interval J' such that $\gamma|_{J'} \in \text{Traj}(X)$. But $\gamma|_{J'}$ must be contained in B . Since X is nowhere tangent to B , we have reached a contradiction.

So $N(B)$ has at most two elements for each $B \in \mathcal{B}^+ \cup \mathcal{B}^-$. Since every $i \in \{1, 2, \dots, m(\gamma)\}$ is in $N(B)$ for some B , it follows that $m(\gamma) \leq 2\nu$, where ν is the cardinality of $\mathcal{B}^+ \cup \mathcal{B}^-$. Therefore m_ε^+ and m_ε^- are finite. This concludes the proof that Lemma 3.10 follows from Lemma 3.12.

We now proceed to the rather long proof of Lemma 3.12. The main step of the proof is the construction of a stratification \mathcal{T} with some special properties. In order to construct \mathcal{T} , we must first study some properties of conjugate points. If q_1, q_2 are two points of $\text{Sq}(\varepsilon_0)$, we write $q_1 \sim q_2$ if q_1 and q_2 lie on the same X -trajectory and are conjugate along it. In coordinates, if $q_i = (x_i, y_i)$, the relation $q_1 \sim q_2$ holds if and only if

$$(3.12a) \quad y_1 = y_2$$

and

$$(3.12b) \quad \sigma(x_1, y_1; x_2, y_2) = 0,$$

where $\sigma(x_1, y_1; x_2, y_2)$ is equal to four times the determinant of the vectors $G(q_1)$, $G(q_2)$, i.e.

$$(3.13) \quad \sigma(x_1, y_1; x_2, y_2) = [\alpha(x_1, y_1) - 1]\beta(x_2, y_2) - [\alpha(x_2, y_2) - 1]\beta(x_1, y_1).$$

We define, for $0 < \varepsilon \leq \varepsilon_0$

$$(3.14) \quad Q(\varepsilon) = \{(q_1, q_2) \in \text{Sq}(\varepsilon) \times \text{Sq}(\varepsilon) : q_1 \sim q_2\}.$$

Then $Q(\varepsilon_0)$ is an analytic subset of $\text{Sq}(\varepsilon_0) \times \text{Sq}(\varepsilon_0)$, because it is the set of solutions of the pair of equations (3.12a), (3.12b). If $0 < \varepsilon < \varepsilon_0$, then $Q(\varepsilon)$ is a semianalytic, relatively compact subset of $\text{Sq}(\varepsilon_0) \times \text{Sq}(\varepsilon_0)$.

We now pick an ε_1 such that $0 < \varepsilon_1 < \varepsilon_0$ and that ε_1 is good for \mathcal{S} . From now on, and until the end of this section, ε_1 will be kept fixed. We consider the set \mathcal{P} of those strata $S \in \mathcal{S}$ such that: (a) S is one-dimensional, (b) S meets $\text{Sq}(\varepsilon_1)$ and (c) G does not vanish identically on S . Then \mathcal{P} is finite. For each $S \in \mathcal{P}$, we let S^\sim denote the set of all $q_2 \in \text{Sq}(\varepsilon_1)$ such that $q_2 \sim q_1$ for some $q_1 \in S \cap \text{Sq}(\varepsilon_1)$. If

$$\pi: \text{Sq}(\varepsilon_0) \times \text{Sq}(\varepsilon_0) \rightarrow \text{Sq}(\varepsilon_0)$$

denotes the projection $(q_1, q_2) \rightarrow q_1$, then

$$(3.15) \quad S^\sim = \pi(Q(\varepsilon_1) \cap (\text{Sq}(\varepsilon_1) \times S)).$$

So S^\sim is the image, under the analytic map π , of the relatively compact semianalytic subset $Q(\varepsilon_1) \cap (\text{Sq}(\varepsilon_1) \times S)$ of $\text{Sq}(\varepsilon_0) \times \text{Sq}(\varepsilon_0)$. Therefore S^\sim is a subanalytic subset of $\text{Sq}(\varepsilon_0)$, and $S^\sim \subseteq \text{Sq}(\varepsilon_1)$, so that S^\sim is relatively compact in $\text{Sq}(\varepsilon_0)$. (Actually, S^\sim is semianalytic, because every subanalytic subset of \mathbb{R}^2 is semianalytic; but we will not need to use this fact.)

It will be important later to know that, if $S \in \mathcal{P}$, then S^\sim cannot be two-dimensional. In order to prove this, we need the results of a calculation that will be useful for other reasons as well.

We have

$$(3.16) \quad [X, Y] = (\partial_x \alpha) \partial_x + (\partial_x \beta) \partial_y.$$

On the other hand,

$$(3.17) \quad (\partial_{x_2} \sigma)(x_1, y_1; x_2, y_2) = [\alpha(x_1, y_1) - 1](\partial_x \beta)(x_2, y_2) - (\partial_x \alpha)(x_2, y_2)\beta(x_1, y_1),$$

i.e.

$$(3.18) \quad (\partial_{x_2} \sigma)(x_1, y_1; x_2, y_2) = 2 \det(G(x_1, y_1), [X, Y](x_2, y_2)).$$

Now suppose S^\sim were two-dimensional for some $S \in \mathcal{P}$. Pick a q_2 in the interior of S^\sim , and let $W \subseteq \text{Sq}(\varepsilon_1)$ be open, and such that $q_2 \in W$, $W \subseteq S^\sim$. For each $q'_2 \in W$, there is a $q'_1 \in S \cap \text{Sq}(\varepsilon_1)$ such that $q'_1 \sim q'_2$. In particular, q'_1 must have the same y -coordinate as q'_2 , and so $S \cap \text{Sq}(\varepsilon_1)$ must contain points not on the x -axis. In view of Lemma 3.4, together with the fact that ε_1 is good for \mathcal{S} , the set $S \cap \text{Sq}(\varepsilon_1)$ is entirely contained in $\text{Sq}(\varepsilon_1)^+$ or in $\text{Sq}(\varepsilon_1)^-$. Pick a $q_1 \in S$ for which $q_2 \sim q_1$ and $q_1 \in \text{Sq}(\varepsilon_1)$. Let L denote the open segment $\{(x, y_1) : |x| < \varepsilon_1\}$, where we let $q_i = (x_i, y_i)$, $i = 1, 2$ (so that, in particular, $y_1 = y_2$). Since W is open, $W \cap L$ contains a nonempty open segment L' . Each $q' \in L'$ must satisfy $q' \sim q$ for some $q \in S \cap \text{Sq}(\varepsilon_1)$. In principle, the point q might depend on q' , but it is easy to see that it does not. Indeed, since $S \cap \text{Sq}(\varepsilon_1)$ is

in $\mathcal{A}_p^+(\varepsilon_1) \cup \mathcal{A}_p^-(\varepsilon_1)$, $S \cap L$ consists of at most one point, and so $S \cap L = \{q_1\}$. So $q' \sim q$ for all $q' \in L'$. Therefore the function

$$(3.19) \quad x \rightarrow \sigma(x_1, y_1; x, y_1)$$

vanishes for all x in some neighborhood of x_2 . Because the function (3.19) is analytic, we can conclude that

$$(3.20) \quad \sigma(x_1, y_1; x, y_1) = 0 \quad \text{for } |x| < \varepsilon_1.$$

So

$$(3.21) \quad (\partial_x \sigma)(x_1, y_1; x, y_1) = 0 \quad \text{for } |x| < \varepsilon_1,$$

and then, using (3.18), we get

$$(3.22) \quad \det(G(x_1, y_1), [X, Y](x, y_1)) = 0 \quad \text{for } |x| < \varepsilon_1.$$

Now recall that $S \in \mathcal{P}$, and therefore G does not vanish identically on S . Since $S \in \mathcal{S}$, which is compatible with E_3 , we see that G vanishes nowhere on S . So $G(x_1, y_1) \neq 0$. If $|x| < \varepsilon_1$, (3.20) shows that $G(x_1, y_1)$ and $G(x, y_1)$ are linearly dependent, and so $G(x, y_1) = \rho G(x_1, y_1)$ for some number ρ . Therefore

$$\begin{aligned} \Delta_B(x, y_1) &= \frac{1}{2} \det(G(x, y_1), [X, Y](x, y_1)) \\ &= \frac{\rho}{2} \det(G(x_1, y_1), [X, Y](x, y_1)) \\ &= 0. \end{aligned}$$

So Δ_B vanishes identically on L , and $L \subseteq E_2$. Since ε_1 is good for \mathcal{S} , and $y_1 \neq 0$, only a finite number of one-dimensional strata of \mathcal{S} meet L , and no zero-dimensional stratum of L meets L . So some point of L must belong to a two-dimensional stratum of \mathcal{S} . Since $L \subseteq E_2$ and \mathcal{S} is compatible with E_2 , it follows that E_2 must contain an open set, and so Δ_B vanishes identically, contradicting (3.III.v).

The contradiction arose from assuming that S^\sim was two-dimensional. So, we have proved the following.

LEMMA 3.13. *If $S \in \mathcal{P}$, then $\dim S^\sim \leq 1$. \square*

We now let $Q_0(\varepsilon)$ denote, for $0 < \varepsilon \leq \varepsilon_0$, the set of all pairs $(q_1, q_2) \in Q(\varepsilon)$ such that $q_1 \neq q_2$, that both q_1 and q_2 are ordinary points, and that $f(q_1)$ and $f(q_2)$ have opposite signs. Then $Q_0(\varepsilon)$ is the subset of $\text{Sq}(\varepsilon_0) \times \text{Sq}(\varepsilon_0)$ defined by the conditions $(q_1, q_2) \in Q(\varepsilon)$ and $(\Delta_A \Delta_B)(q_1) \cdot (\Delta_A \Delta_B)(q_2) < 0$. So $Q_0(\varepsilon)$ is semianalytic in $\text{Sq}(\varepsilon_0) \times \text{Sq}(\varepsilon_0)$. Moreover, $Q_0(\varepsilon)$ is relatively compact if $\varepsilon < \varepsilon_0$.

LEMMA 3.14. *For $0 < \varepsilon \leq \varepsilon_0$, the set $Q_0(\varepsilon)$ is an embedded two-dimensional submanifold of $\text{Sq}(\varepsilon_0) \times \text{Sq}(\varepsilon_0)$, and the projection $\pi: Q_0(\varepsilon) \rightarrow \text{Sq}(\varepsilon_0)$ is nonsingular at each point $(q_1, q_2) \in Q_0(\varepsilon)$.*

Proof. Pick $(\bar{q}_1, \bar{q}_2) \in Q_0(\varepsilon)$, and let $\bar{q}_i = (\bar{x}_i, \bar{y}_i)$, for $i = 1, 2$. Then $\bar{y}_1 = \bar{y}_2$, and $\sigma(\bar{x}_1, \bar{y}_1; \bar{x}_2, \bar{y}_2) = 0$. On the other hand, $\Delta_B(\bar{q}_2) \neq 0$, and so

$$(3.23) \quad \det(G(\bar{x}_2, \bar{y}_2), [X, Y](\bar{x}_2, \bar{y}_2)) \neq 0.$$

Since $\bar{q}_1 \sim \bar{q}_2$, the vectors $G(\bar{x}_1, \bar{y}_1)$ and $G(\bar{x}_2, \bar{y}_2)$ are linearly dependent. Since \bar{q}_1 and \bar{q}_2 are ordinary points, both $G(\bar{x}_1, \bar{y}_1)$ and $G(\bar{x}_2, \bar{y}_2)$ are nonzero. So (3.23) implies

$$(3.24) \quad \det(G(\bar{x}_1, \bar{y}_1), [X, Y](\bar{x}_2, \bar{y}_2)) \neq 0.$$

In view of (3.18), we get

$$(3.25) \quad \frac{\partial \sigma}{\partial x_2}(\bar{x}_1, \bar{y}_1; \bar{x}_2, \bar{y}_2) \neq 0.$$

Pick $\delta > 0$ such that, whenever $q_i = (x_i, y_i)$, $i = 1, 2$, and $|x_i - \bar{x}_i| < \delta$, $|y_i - \bar{y}_i| < \delta$, for $i = 1, 2$, it follows that $(\Delta_A \Delta_B)(q_1) \cdot (\Delta_A \Delta_B)(q_2) < 0$, and that $q_1 \in \text{Sq}(\varepsilon)$, $q_2 \in \text{Sq}(\varepsilon)$.

In view of (3.25), and of the fact that $\bar{y}_1 = \bar{y}_2$, we can apply the Implicit Function Theorem to find δ_1, δ_2 such that $0 < \delta_1 < \delta$, $0 < \delta_2 < \delta$, and a real analytic function k , defined for $|x_1 - \bar{x}_1| < \delta_1$, $|y_1 - \bar{y}_1| < \delta_1$, with the property that $|k(x_1, y_1) - \bar{x}_2| < \delta_2$ for $(x_1 - \bar{x}_1, y_1 - \bar{y}_1) \in \text{Sq}(\delta_1)$, and that, if $(x_1 - \bar{x}_1, y_1 - \bar{y}_1) \in \text{Sq}(\delta_1)$ and $|x - \bar{x}_2| < \delta_2$, then $\sigma(x_1, y_1; x, y_1) = 0$ if and only if $x = k(x_1, y_1)$.

Now let W denote the set

$$(3.26) \quad \{(x_1, y_1; x_2, y_2) : |x_1 - \bar{x}_1| < \delta_1, |y_1 - \bar{y}_1| < \delta_1, |x_2 - \bar{x}_2| < \delta_2, |y_2 - \bar{y}_2| < \delta_2\}.$$

Then $W \cap Q_0(\varepsilon)$ is the set of those $(x_1, y_1; x_2, y_2) \in W$ that satisfy $y_2 = y_1$, $x_2 = k(x_1, y_1)$. So $W \cap Q_0(\varepsilon)$ is an embedded submanifold of W . Since every $(\bar{q}_1, \bar{q}_2) \in Q_0(\varepsilon)$ has such a neighborhood W , it follows that $Q_0(\varepsilon)$ is an embedded manifold. Moreover, the map $(x_1, y_1) \rightarrow (x_1, y_1; k(x_1, y_1), y_1)$ is an inverse of $\pi|_{(Q_0(\varepsilon) \cap W)}$. So $\pi|_{Q_0(\varepsilon)}$ has rank two at every point of $Q_0(\varepsilon)$. \square

Suppose that V is a smooth vector field on $\text{Sq}(\varepsilon_0)$. Since $\pi : Q_0(\varepsilon_0) \rightarrow \pi(Q_0(\varepsilon_0))$ is a local diffeomorphism, there exists a unique vector field V^* on $Q_0(\varepsilon_0)$ such that

$$(3.27) \quad \pi_*(V^*(q_1, q_2)) = V(q_1)$$

for all $(q_1, q_2) \in Q_0(\varepsilon_0)$. (Here π_* is the differential of π .)

If $t \rightarrow \zeta(t)$ denotes the integral curve of V^* such that $\zeta(0) = (q_1, q_2)$, then $\zeta(t) = (\zeta_1(t), \zeta_2(t))$, where $\zeta_1(t) = \pi(\zeta(t))$, and therefore $\zeta_1(t) = \Phi_t^V(q_1)$. So $\zeta_1(0) = V(q_1)$. This shows that V^* is of the form

$$(3.28) \quad V^*(q_1, q_2) = (V(q_1), V^{\sim}(q_1, q_2))$$

for some vector-valued $V^{\sim} : Q_0(\varepsilon_0) \rightarrow \mathbb{R}^2$.

We now compute V^{\sim} . Pick $(\bar{q}_1, \bar{q}_2) \in Q_0(\varepsilon_0)$, and then pick $\delta, \delta_1, \delta_2, k(\cdot, \cdot)$ as in the proof of Lemma 3.14. Let K denote the map $(x_1, y_1) \rightarrow (k(x_1, y_1), y_1)$. Then it is clear that

$$(3.29) \quad \Phi_t^{V^*}(\bar{q}_1, \bar{q}_2) = (\Phi_t^V(\bar{q}_1), K(\Phi_t^V(\bar{q}_1)))$$

and therefore

$$(3.30) \quad V^{\sim}(\bar{q}_1, \bar{q}_2) = K_*(\bar{q}_1) V(\bar{q}_1),$$

where $K_*(\bar{q}_1)$ is the Jacobian matrix of K at \bar{q}_1 . An easy computation gives

$$(3.31) \quad K_*(\bar{q}_1) = \begin{bmatrix} (\partial_{x_1} k)(\bar{x}_1, \bar{y}_1) & \partial_{y_1} k(\bar{x}_1, \bar{y}_1) \\ 0 & 1 \end{bmatrix}.$$

On the other hand, the partial derivatives of k can be computed by implicit differentiation of

$$(3.32) \quad \sigma(x_1, y_1; k(x_1, y_1), y_1) = 0.$$

The result is

$$(3.33) \quad \partial_{x_1} k(x_1, y_1) = -\frac{\partial_{x_1} \sigma}{\partial_{x_2} \sigma}(x_1, y_1; k(x_1, y_1), y_1),$$

$$(3.34) \quad \partial_{y_1} k(x_1, y_1) = -\frac{\partial_{y_1} \sigma + \partial_{y_2} \sigma}{\partial_{x_2} \sigma}(x_1, y_1; k(x_1, y_1), y_1).$$

So

$$(3.35) \quad K_*(\bar{q}_1) = \frac{1}{\partial_{x_2}\sigma} \begin{bmatrix} -\partial_{x_1}\sigma & -(\partial_{y_1}\sigma + \partial_{y_2}\sigma) \\ 0 & \partial_{x_2}\sigma \end{bmatrix} (\bar{q}_1, \bar{q}_2).$$

In particular, if V is given by

$$(3.36) \quad V = v_1\partial_x + v_2\partial_y$$

we find that

$$(3.37) \quad V^\sim(q_1, q_2) = v_1^\sim(q_1, q_2)\partial_x + v_2^\sim(q_1, q_2)\partial_y,$$

where

$$(3.38) \quad v_1^\sim(q_1, q_2) = -\left(\frac{\partial_{x_1}\sigma}{\partial_{x_2}\sigma}\right)(q_1, q_2)v_1(q_1) - \left(\frac{\partial_{y_1}\sigma + \partial_{y_2}\sigma}{\partial_{x_2}\sigma}\right)(q_1, q_2)v_2(q_1)$$

and

$$(3.39) \quad v_2^\sim(q_1, q_2) = v_2(q_1).$$

Of particular interest to us are the vector fields X^* and Y^* . We have

$$(3.40) \quad X^\sim(q_1, q_2) = -\left(\frac{\partial_{x_1}\sigma}{\partial_{x_2}\sigma}\right)(q_1, q_2)\partial_x$$

and

$$(3.41) \quad Y^\sim(q_1, q_2) = \alpha^\sim(q_1, q_2)\partial_x + \beta^\sim(q_1, q_2)\partial_y$$

where

$$(3.42) \quad \alpha^\sim(q_1, q_2) = -\left(\frac{\partial_{x_1}\sigma}{\partial_{x_2}\sigma}\right)(q_1, q_2)\alpha(q_1) - \left(\frac{\partial_{y_1}\sigma + \partial_{y_2}\sigma}{\partial_{x_2}\sigma}\right)(q_1, q_2)\beta(q_2)$$

and

$$(3.43) \quad \beta^\sim(q_1, q_2) = \beta(q_1).$$

It will be important to know the direction of $X^\sim(q_1, q_2)$. The following lemma tells us all we need to know.

LEMMA 3.15. *Suppose that $(q_1, q_2) \in Q_0(\varepsilon_0)$ and that $\Delta_A(q_1)\Delta_A(q_2) > 0$. Then the vector $X^\sim(q_1, q_2)$ points to the left, i.e.,*

$$(3.44) \quad -\frac{\partial_{x_1}\sigma}{\partial_{x_2}\sigma}(q_1, q_2) < 0.$$

Proof. Suppose that $(q_1, q_2) \in Q_0(\varepsilon_0)$, and that $\Delta_A(q_1)\Delta_A(q_2) > 0$. Then $G(q_1) \neq 0$, $G(q_2) \neq 0$, and $G(q_1)$ is linearly independent from $G(q_2)$, so that

$$G(q_1) = \rho G(q_2)$$

for some nonzero number ρ . Formula (3.18) then implies that

$$(\partial_{x_2}\sigma)(q_1, q_2) = 2\rho \det(G(q_2), [X, Y](q_2)),$$

i.e., that

$$(\partial_{x_2}\sigma)(q_1, q_2) = 4\rho\Delta_B(q_2).$$

A similar calculation yields

$$\begin{aligned}(\partial_{x_1}\sigma)(q_1, q_2) &= -2 \det(G(q_2), [X, Y](q_1)) \\ &= -4\rho^{-1}\Delta_B(q_1).\end{aligned}$$

Therefore

$$-\frac{\partial_{x_1}\sigma}{\partial_{x_2}\sigma}(q_1, q_2) = \rho^{-2} \frac{\Delta_B(q_1)}{\Delta_B(q_2)}.$$

Since $\Delta_A(q_1)\Delta_B(q_1)$ and $\Delta_A(q_2)\Delta_B(q_2)$ have opposite signs (because $(q_1, q_2) \in Q_0(\varepsilon_0)$), but we are assuming that $\Delta_A(q_1)$ and $\Delta_A(q_2)$ have the same sign, it follows that $\Delta_B(q_1)$ and $\Delta_B(q_2)$ have opposite signs. So (3.44) follows. \square

We now single out, for $0 < \varepsilon \leq \varepsilon_0$, subsets $\hat{Q}_1(\varepsilon)$, $Q_1(\varepsilon)$ of $Q_0(\varepsilon)$ as follows. We let $(q_1, q_2) \in \hat{Q}_1(\varepsilon)$ if and only if (a) $(q_1, q_2) \in Q_0(\varepsilon_0)$ and (b) the vectors $Y(q_2)$ and $Y^\sim(q_1, q_2)$ are linearly dependent. We let $(q_1, q_2) \in Q_1(\varepsilon)$ iff (a) $(q_1, q_2) \in \hat{Q}_1(\varepsilon)$ but (b) there does not exist a $\delta > 0$ such that $\Phi_t^{X^*}(q_1, q_2) \in \hat{Q}_1(\varepsilon)$ for all $t \in]-\delta, \delta[$.

LEMMA 3.16. (a) $\hat{Q}_1(\varepsilon_0)$ is a semianalytic subset of $\text{Sq}(\varepsilon_0) \times \text{Sq}(\varepsilon_0)$, (b) $Q_1(\varepsilon)$ is a subanalytic subset of $\text{Sq}(\varepsilon_0) \times \text{Sq}(\varepsilon_0)$, of dimension not greater than one, if $0 < \varepsilon < \varepsilon_0$.

Proof. If $(q_1, q_2) \in \text{Sq}(\varepsilon_0) \times \text{Sq}(\varepsilon_0)$, then $(q_1, q_2) \in \hat{Q}_1(\varepsilon_0)$ if and only if $(q_1, q_2) \in Q_0(\varepsilon_0)$ and $\tau(q_1, q_2) = 0$, where τ is the determinant of $Y^\sim(q_1, q_2)$ and $Y(q_2)$. Formulas (3.41)–(3.43) imply that τ is a quotient τ_1/τ_2 , where τ_1 and τ_2 are analytic functions on $\text{Sq}(\varepsilon_0) \times \text{Sq}(\varepsilon_0)$, and τ_2 never vanishes on $Q_0(\varepsilon_0)$. Therefore $(q_1, q_2) \in \hat{Q}_1(\varepsilon_0)$ if and only if $(q_1, q_2) \in Q_0(\varepsilon_0)$ and $\tau_1(q_1, q_2) = 0$. So $\hat{Q}_1(\varepsilon_0)$ is the intersection of $Q_0(\varepsilon_0)$ with an analytic set, and is therefore semianalytic.

Now let $0 < \varepsilon < \varepsilon_0$. Define a new vector field X^* by

$$(3.45) \quad X^*(q_1, q_2) = \partial_{x_2}\sigma(q_1, q_2)X^*(q_1, q_2).$$

Then X^* is a well defined analytic vector field on the whole cube $\text{Sq}(\varepsilon_0) \times \text{Sq}(\varepsilon_0) \subseteq \mathbb{R}^4$, which is tangent to the submanifold $Q_0(\varepsilon_0)$ and which, restricted to this submanifold, is equal to a nonzero scalar multiple of X^* . If $(q_1, q_2) \in Q_0(\varepsilon_0)$, then the integral curve of X^* through (q_1, q_2) coincides with that of X^* , after a suitable reparametrization. Therefore the definition of $Q_1(\varepsilon)$ can be rephrased as follows:

- (3.XI) (q_1, q_2) belongs to $Q_1(\varepsilon)$ iff
 (3.XI.a) $(q_1, q_2) \in \hat{Q}_1(\varepsilon)$, and
 (3.XI.b) For every $\delta > 0$ there exists a t such that $|t| < \delta$ and that $\Phi_t^{X^*}(q_1, q_2) \notin \hat{Q}_1(\varepsilon)$.

Since X^* is an analytic vector field on $\text{Sq}(\varepsilon_0) \times \text{Sq}(\varepsilon_0)$, and $0 < \varepsilon < \varepsilon_0$, we can pick $\bar{\varepsilon}$, $\bar{\delta}$ such that $0 < \bar{\delta}$, that $\varepsilon < \bar{\varepsilon} < \varepsilon_0$, and that $\Phi_t^{X^*}(q_1, q_2)$ is defined and belongs to $\text{Sq}(\varepsilon_0) \times \text{Sq}(\varepsilon_0)$ for all $(t, q_1, q_2) \in W$, where

$$W =]-\bar{\delta}, \bar{\delta}[\times \text{Sq}(\bar{\varepsilon}) \times \text{Sq}(\bar{\varepsilon}).$$

Now let $\tilde{W} \subseteq \mathbb{R}^9$ be the open set $W \times \text{Sq}(\varepsilon_0) \times \text{Sq}(\varepsilon_0)$, and let $\tilde{\Phi} \subseteq \tilde{W}$ be the graph of the restriction to W of the flow of X , that is

$$(3.46) \quad \tilde{\Phi} = \{(t, q_1, q_2, q'_1, q'_2) \in \tilde{W} : (q'_1, q'_2) = \Phi_t^{X^*}(q_1, q_2)\}.$$

Then $\tilde{\Phi}$ is the set of zeros in \tilde{W} of four analytic real-valued functions defined on all of \tilde{W} , so that $\tilde{\Phi}$ is an analytic subset of \tilde{W} . Let $\tilde{\Phi}_1$ denote the set of those $(t, q_1, q_2, q'_1, q'_2) \in \tilde{\Phi}$ such that (q'_1, q'_2) does not belong to $\hat{Q}_1(\varepsilon)$. Then $\tilde{\Phi}_1$ is the

intersection of $\tilde{\Phi}$ with $\psi^{-1}([\text{Sq}(\varepsilon_0) \times \text{Sq}(\varepsilon_0)] - \hat{Q}_1(\varepsilon))$, where ψ is the projection $(t, q_1, q_2, q'_1, q'_2) \rightarrow (q'_1, q'_2)$. So $\tilde{\Phi}_1$ is a semianalytic subset of \tilde{W} . Now let $\tilde{W} \subseteq \mathbb{R}^{10}$ be the set $] -\bar{\delta}, \bar{\delta}[\times \tilde{W}$, and consider the subset $\tilde{\Phi}_1$ of those $(\delta, t, q_1, q_2, q'_1, q'_2)$ that satisfy $|t| < \delta$, $(t, q_1, q_2, q'_1, q'_2) \in \tilde{\Phi}_1$. Let $0 < \delta^* < \bar{\delta}$, and let $\tilde{\Phi}_1^*$ be the set of those $(\delta, t, q_1, q_2, q'_1, q'_2) \in \tilde{\Phi}_1$ such that $\delta < \delta^*$, $(q_1, q_2) \in \text{Sq}(\varepsilon) \times \text{Sq}(\varepsilon)$, and $(q'_1, q'_2) \in \text{Sq}(\varepsilon) \times \text{Sq}(\varepsilon)$. Then $\tilde{\Phi}_1$ is semianalytic in \tilde{W} , and $\tilde{\Phi}_1^*$ is semianalytic and relatively compact in \tilde{W} . Let $\nu: \tilde{W} \rightarrow]0, \bar{\delta}^*[\times \text{Sq}(\varepsilon) \times \text{Sq}(\varepsilon)$ be the map $(\delta, t, q_1, q_2, q'_1, q'_2) \rightarrow (\delta, q_1, q_2)$. Then (δ, q_1, q_2) is in $\nu(\tilde{\Phi}_1^*)$ if and only if

- (3.XII.a) $0 < \delta < \delta^*$;
- (3.XII.b) $(q_1, q_2) \in \text{Sq}(\varepsilon) \times \text{Sq}(\varepsilon)$;
- (3.XII.c) There is a t such that $|t| < \delta$ and that $\Phi_t^{X^*}(q_1, q_2)$ does not belong to $\hat{Q}_1(\varepsilon)$, but belongs to $\text{Sq}(\varepsilon) \times \text{Sq}(\varepsilon)$.

Therefore, a point

$$(\delta, q_1, q_2) \in]0, \delta^*[\times \text{Sq}(\varepsilon) \times \text{Sq}(\varepsilon)$$

will fail to belong to $\nu(\tilde{\Phi}_1^*)$ if and only if, for every t such that $|t| < \delta$, the point $\Phi_t^{X^*}(q_1, q_2)$ either belongs to $\hat{Q}_1(\varepsilon)$ or fails to belong to $\text{Sq}(\varepsilon) \times \text{Sq}(\varepsilon)$.

Let D denote the set of those $(\delta, q_1, q_2) \in]0, \delta^*[\times \text{Sq}(\varepsilon) \times \text{Sq}(\varepsilon)$ such that $(\delta, q_1, q_2) \notin \nu(\tilde{\Phi}_1^*)$, but $(q_1, q_2) \in Q_0(\varepsilon)$. Then D is relatively compact in W . Moreover, D is the intersection of the semianalytic subset $\hat{\psi}^{-1}(Q_0(\varepsilon))$ of W with the set $W - \nu(\tilde{\Phi}_1^*)$. (Here $\hat{\psi}$ is the projection $(\delta, q_1, q_2) \rightarrow (q_1, q_2)$.) Since $\tilde{\Phi}_1^*$ is semianalytic and relatively compact in \tilde{W} , the set $\nu(\tilde{\Phi}_1^*)$ is subanalytic in W . So D is a subanalytic subset of W , and it is clearly relatively compact in W . Moreover, D is the set of those (δ, q_1, q_2) such that $0 < \delta < \delta^*$, $(q_1, q_2) \in Q_0(\varepsilon)$, and $\Phi_t^{X^*}(q_1, q_2) \in \hat{Q}_1(\varepsilon)$ or $\Phi_t^{X^*}(q_1, q_2) \notin \text{Sq}(\varepsilon) \times \text{Sq}(\varepsilon)$ for $|t| < \delta$. Therefore $\hat{\psi}(D)$ is the set of those $(q_1, q_2) \in Q_0(\varepsilon)$ such that $\Phi_t^{X^*}(q_1, q_2) \in \hat{Q}_1(\varepsilon)$ for all sufficiently small t . So

$$Q_1(\varepsilon) = \hat{Q}_1(\varepsilon) - \hat{\psi}(D).$$

Since $\hat{\psi}(D)$ is subanalytic, it follows that $Q_1(\varepsilon)$ is subanalytic.

Finally, to prove that $\dim Q_1(\varepsilon) \leq 1$, observe that, if $\dim Q_1(\varepsilon)$ were equal to two, this would imply that $Q_1(\varepsilon)$ contains a subset S which is open in $Q_0(\varepsilon)$. If $(q_1, q_2) \in S$, then $\Phi_t^{X^*}(q_1, q_2) \in Q_1(\varepsilon)$ for small t , and so $\Phi_t^{X^*}(q_1, q_2) \in \hat{Q}_1(\varepsilon)$ for small t . But then $(q_1, q_2) \notin Q_1(\varepsilon)$, and we have reached a contradiction. \square

We now let $\pi_2: \text{Sq}(\varepsilon_0) \times \text{Sq}(\varepsilon_0) \rightarrow \text{Sq}(\varepsilon_0)$ denote the projection $(q_1, q_2) \rightarrow q_2$. The set $\pi_2(Q_1(\varepsilon_1))$ is therefore a subanalytic subset of $\text{Sq}(\varepsilon_0)$.

We choose a pair \mathcal{V}, \mathcal{T} of CASA stratifications of $\text{Sq}(\varepsilon_0) \times \text{Sq}(\varepsilon_0)$ and $\text{Sq}(\varepsilon_0)$, respectively, such that

- (3.XIII.i) \mathcal{V} is compatible with $\text{Sq}(\varepsilon_1) \times \text{Sq}(\varepsilon_1)$, with $Q_0(\varepsilon_0)$, with $Q_1(\varepsilon_1)$, and with $\{(p, p)\}$;
- (3.XIII.ii) \mathcal{T} is compatible with $\pi_2(Q_1(\varepsilon_1))$, with all the members of \mathcal{S} , and with all the sets S^- , for $S \in \mathcal{P}$;
- (3.XIII.iii) $(\mathcal{V}, \mathcal{T})$ is compatible with π over $\text{Sq}(\varepsilon_1) \times \text{Sq}(\varepsilon_1)$.

Since \mathcal{T} is a refinement of \mathcal{S} , it follows from Lemma 3.6 that \mathcal{T} is appropriate.

We now take $\varepsilon > 0$ so small that $0 < \varepsilon < \varepsilon_1$ and that ε is good for \mathcal{T} . We define \mathcal{B}_0 to be the set of all $T^\varepsilon = T \cap \text{Sq}(\varepsilon)$, as T ranges over all one-dimensional strata of \mathcal{T} that meet $\text{Sq}(\varepsilon)$ (i.e., equivalently, over all one-dimensional $T \in \mathcal{T}$ such that

$p \in \text{Clos } T$). By Lemma 3.4, every element of \mathcal{B}_0 is in $\mathcal{A}_p^+(\varepsilon)$, or in $\mathcal{A}_p^-(\varepsilon)$, or is one of the segments Γ_ε^+ , Γ_ε^- . We define

$$(3.47) \quad \mathcal{B}_0^+ = \mathcal{B}_0 \cap \mathcal{A}_p^+(\varepsilon),$$

$$(3.48) \quad \mathcal{B}_0^- = \mathcal{B}_0 \cap \mathcal{A}_p^-(\varepsilon).$$

We let \mathcal{B}_1^+ be the set obtained by adding to \mathcal{B}_0^+ the segment $\{0\} \times]0, \varepsilon[$. Similarly, we let

$$\mathcal{B}_1^- = \mathcal{B}_0^- \cup \{\{0\} \times]-\varepsilon, 0[\}.$$

Finally, we let \mathcal{B}^+ denote the set of all $B \in \mathcal{B}_1^+$ that are barriers in $\text{Sq}(\varepsilon)^+$, and we define \mathcal{B}^- similarly. It is clear that \mathcal{B}^+ and \mathcal{B}^- are finite sets, and that the members of \mathcal{B}^+ (resp. \mathcal{B}^-) are barriers in $\text{Sq}(\varepsilon)^+$ (resp. $\text{Sq}(\varepsilon)^-$). Moreover, the members of $\mathcal{B}^+ \cup \mathcal{B}^-$ are clearly pairwise disjoint. (Indeed, if B_1, B_2 are in $\mathcal{B}^+ \cup \mathcal{B}^-$, then $B_1 \cap B_2 = \emptyset$ if one of the B_i is in \mathcal{B}^+ and the other one in \mathcal{B}^- . If, say, both B_i are in \mathcal{B}^+ , then it is clear that $B_1 \cap B_2 = \emptyset$ if both B_i are in \mathcal{B}_0^+ . If, say, $B_1 \in \mathcal{B}_0^+$ and $B_2 = \{0\} \times]0, \varepsilon[$, then B_1 is the graph $\{(\psi(y), y)\}$ of a function ψ which is defined on some interval $]0, \delta[$, and $\psi(0+) = 0$. Moreover, ψ is either constant (in which case $\psi \equiv 0$) or strictly monotonic. In the latter case, ψ never takes the value zero, so that $B_1 \cap B_2 = \emptyset$. In the former case we necessarily have $\delta = \varepsilon$ and $B_1 = B_2$.) Finally, it is clear that, if $B \in \mathcal{B}^+ \cup \mathcal{B}^-$, then X is never tangent to B , because B is a graph $\{(\psi(y), y)\}$.

So \mathcal{B}^+ and \mathcal{B}^- satisfy all the conditions (3.X) of Lemma 3.12, except, possibly, for (3.X.d) (which is, naturally, the most important one). The rest of this section is devoted to the proof that \mathcal{B}^+ and \mathcal{B}^- satisfy (3.X.d) as well.

We consider a trajectory γ which belongs to \mathcal{K} , and which is time-optimal and entirely contained in $\text{Sq}(\varepsilon)^+$. We let $a, t_0, t'_0, t'_1, t_1, b$ be such that (3.XI.i, \dots, v) hold. It is clear that we can choose $[t'_0, t'_1]$ to be a maximal subinterval J of $[t_0, t_1]$ such that $\gamma|_J$ is an X -trajectory. Also, if there are several such maximal intervals, we can choose $[t'_0, t'_1]$ to be the one farthest to the left. With this choice, we may, and will, assume that

(3.XIV.a) Either $t'_0 = t_0$, or $\gamma|_{[t_0, t'_0]}$ is a Z -trajectory;

(3.XIV.b) Either $t'_1 = t_1$, or there is a $\delta > 0$ such that $\gamma|_{[t'_1, t'_1 + \delta]}$ is a Z -trajectory.

Finally, let us write

$$q_0 = \gamma(t_0), \quad q_1 = \gamma(t_1), \quad q'_0 = \gamma(t'_0), \quad q'_1 = \gamma(t'_1).$$

Our goal is to prove that γ necessarily meets one of the members of \mathcal{B}^+ . We will do it by considering several different possible cases.

CASE 1. $t_0 = t'_0, t'_1 = t_1$, and both q_0 and q_1 are ordinary points.

In this case, γ is a strict $Y * X * Y$ -trajectory. Moreover, the points q_0 and q_1 are conjugate along $\gamma|_{[t_0, t_1]}$, because γ is optimal. Since both q_0 and q_1 are ordinary, the quantities $\Delta_A(q_0), \Delta_A(q_1), \Delta_B(q_0), \Delta_B(q_1)$ are all nonzero. In particular (since $\Delta_A = \frac{1}{2}\beta$) we see that $\beta(q_0)$ and $\beta(q_1)$ are both nonzero.

We distinguish two subcases:

SUBCASE 1a. $\beta(q_0)\beta(q_1) < 0$,

SUBCASE 1b. $\beta(q_0)\beta(q_1) > 0$.

Subcase 1a can only occur if (3.II.a) holds. Indeed, if (3.II.b) holds, then $\alpha < 1$ throughout U_0 , and therefore the equality

$$(3.49) \quad (\alpha(q_0) - 1)\beta(q_1) - (\alpha(q_1) - 1)\beta(q_0) = 0$$

implies that $\beta(q_0)\beta(q_1) > 0$.

So, if we are in Subcase 1a, it follows that $\alpha > 0$ throughout U_0 . Since $\beta(\gamma(t_0))$ and $\beta(\gamma(t_1))$ have opposite signs, there is a point $t_2 \in [t_0, t_1]$ such that $\beta(\gamma(t_2)) = 0$. Let $q_2 = \gamma(t_2)$. Then $\Delta_A(q_2) = 0$, so that $q_2 \in E_1(\varepsilon)$. Therefore $q_2 \in S^\varepsilon$ for some stratum $S \in \mathcal{T}$ such that $S \subseteq E_1$. Since ε is good for \mathcal{T} , it follows that S^ε is in $\mathcal{A}_p^+(\varepsilon)$. (Clearly, S^ε cannot lie in $\mathcal{A}_p^-(\varepsilon) \cup \{\Gamma_\varepsilon^+\} \cup \{\Gamma_\varepsilon^-\}$, because $q_2 \in S \cap \text{Sq}(\varepsilon)^+$.) To show that S^ε is a barrier, it suffices to show that both X and Y point to the same side of S^ε . But this is trivial since, along S^ε , X has components $(1, 0)$, whereas Y has components $(\alpha, 0)$, and $\alpha > 0$. So S is a barrier. Since $S \in \mathcal{T}$, we conclude that $S^\varepsilon \in \mathcal{B}^+$.

We now turn to Subcase 1b. The facts that q_0 and q_1 are ordinary points, and that optimal switchings occur at q_0, q_1 , in opposite senses (i.e. from $Y-$ to $X-$ at q_0 , and from $X-$ to $Y-$ at q_1) imply that $f(q_0)$ and $f(q_1)$ have opposite signs, so that $(q_0, q_1) \in Q_0(\varepsilon)$. Moreover, $\Delta_A(q_0)$ and $\Delta_A(q_1)$ have the same sign, and so Lemma 3.15 applies, and we can conclude that the vector $X^\sim(q_0, q_1)$ points to the left.

Now let $\zeta(\cdot)$ be the maximal integral curve of X^* such that $\zeta(0) = (q_0, q_1)$. (We emphasize that X^* is, by definition, a vector field on $Q_0(\varepsilon_0)$, so that, if $\lim_{s \rightarrow \bar{s}-} \zeta(s) \notin Q_0(\varepsilon_0)$, then $\zeta(\bar{s})$ is not defined.) The curve $\zeta(\cdot)$ is defined on an interval $[0, s_{\max}[$, where s_{\max} is finite. (To see that $s_{\max} < \infty$, notice that, if

$$(3.50) \quad \zeta(s) = (\zeta_0(s), \zeta_1(s)),$$

then $\zeta_0(\cdot)$ is an X -trajectory, so that the point $\zeta_0(s)$ moves to the right as s increases. On the other hand, as long as $\zeta(s)$ stays in $Q_0(\varepsilon_0)$, the number $\Delta_A(\zeta_0(s))\Delta_A(\zeta_1(s))$ must remain strictly positive, because it is positive for $s = 0$, and it can never vanish, since $\zeta_0(s)$ and $\zeta_1(s)$ are ordinary points. So $X^\sim(\zeta_0(s), \zeta_1(s))$ points left. Therefore the point $\zeta_1(s)$ moves to the left. Since $\zeta_0(s)$ moves right with velocity 1, the points $\zeta_0(s)$ and $\zeta_1(s)$ would have to coincide for some finite s , if $s_{\max} = \infty$. But, if $\zeta_0(s) = \zeta_1(s)$, then $\zeta(s)$ would not be in $Q_0(\varepsilon_0)$, which is a contradiction. So $s_{\max} < \infty$.)

The preceding remarks also show that $\zeta(s_{\max}-) = (\hat{q}_0, \hat{q}_1)$ is well defined, that $s_{\max} < t_1 - t_0$, and that $\hat{q}_0 = \gamma(\hat{t}_0)$, $\hat{q}_1 = \gamma(\hat{t}_1)$ for some \hat{t}_0, \hat{t}_1 such that $t_0 < \hat{t}_0 \leq \hat{t}_1 < t_1$ (and that, moreover, $\hat{t}_0 = t_0 + s_{\max}$). It is clear that $\zeta_0(s) \sim \zeta_1(s)$ for all $s \in [0, s_{\max}[$, and so $\hat{q}_0 \sim \hat{q}_1$. On the other hand, the point (\hat{q}_0, \hat{q}_1) is not in $Q_0(\varepsilon_0)$. Therefore one of the equalities $\Delta_A(\hat{q}_0) = 0$, $\Delta_A(\hat{q}_1) = 0$, $\Delta_B(\hat{q}_0) = 0$, $\Delta_B(\hat{q}_1) = 0$ must hold. (Otherwise, if $(\Delta_A\Delta_B)(\hat{q}_0)$ and $(\Delta_A\Delta_B)(\hat{q}_1)$ were $\neq 0$, it would follow from the inequality $(\Delta_A\Delta_B)(\zeta_0(s)) \cdot (\Delta_A\Delta_B)(\zeta_1(s)) < 0$ that $(\Delta_A\Delta_B)(\hat{q}_0)$ and $(\Delta_A\Delta_B)(\hat{q}_1)$ have opposite signs, so that $(\hat{q}_0, \hat{q}_1) \in Q_0(\varepsilon_0)$.)

We now observe that

$$(3.XV) \quad \text{The vector } Y^\sim(q_0, q_1) \text{ is not a linear combination of } X(q_1) \text{ and } Y(q_1) \text{ with strictly positive coefficients.}$$

Indeed, suppose $Y^\sim(q_0, q_1)$ were a combination $\mu X(q_1) + \nu Y(q_1)$, with $\mu > 0$, $\nu > 0$. Let $\lambda(\cdot): [t_0 - \delta, t_0] \rightarrow Q_0(\varepsilon_0)$ be an integral curve of Y^* such that $\lambda(t_0) = (q_0, q_1)$. If $\lambda(t) = (\lambda_1(t), \lambda_2(t))$, then $\lambda_1(\cdot)$ is a Y -trajectory, and $\lambda_2(\cdot)$ satisfies

$$(3.51) \quad \dot{\lambda}_2(t) = Y^\sim(\lambda_1(t), \lambda_2(t)).$$

Therefore λ_2 is, after a suitable reparametrization, a trajectory of Σ . Notice that both $Y(\lambda_1(t))$ and $Y^\sim(\lambda_1(t), \lambda_2(t))$ have the same y component. Therefore, either both $\lambda_1(\cdot)$ and $\lambda_2(\cdot)$ go up, or both go down.

By choosing δ smaller, if necessary, we may assume that $\lambda_2(t)$ lies to the right of $\lambda_1(t)$ for all $t \in [t_0 - \delta, t_0]$.

Let λ_2^* be the result of reparametrizing λ_2 so that it becomes a trajectory of Σ . Let γ_2 be the concatenation of the piece of X -trajectory going from $\gamma(t_0 - \delta)$ to

$\lambda_2(t_0 - \delta)$, and of λ_2^* . Let $\gamma_1 = \gamma \upharpoonright [t_0 - \delta, t_1]$. Then both γ_1 and γ_2 go from the same initial point to the same terminal point. It is clear that γ_1 and γ_2 satisfy the hypotheses of Lemma 3.6 of [A], so that $T(\gamma_1) = T(\gamma_2)$. Since γ is time-optimal, it follows that γ_1 is time-optimal, and so γ_2 is time-optimal. Therefore λ_2^* is time-optimal. On the other hand, λ_2^* consists entirely of ordinary points. Since λ_2^* is not bang-bang, it cannot be time-optimal. This contradiction proves that (3.XV) holds.

Let us subdivide Subcase 1b into three sub-subcases.

SUB-SUBCASE 1b.i. $\zeta(s) \in Q_1(\varepsilon)$ for some $s \in [0, s_{\max}]$;

SUB-SUBCASE 1b.ii. $\zeta(s) \notin Q_1(\varepsilon)$ for all $s \in [0, s_{\max}]$;

SUB-SUBCASE 1b.iii. $\zeta(s) \in \bar{Q}_1(\varepsilon) - Q_1(\varepsilon)$ for some $s \in [0, s_{\max}]$.

We consider Sub-subcase 1b.i first.

It is clear from the definition of $Q_1(\varepsilon)$ that the set of those s for which $\zeta(s) \in Q_1(\varepsilon)$ is discrete. Let \bar{s} be the smallest element of this set. Let $\bar{q} = (\bar{q}_0, \bar{q}_1) = \zeta(\bar{s})$. Then $\bar{q} \in Q_1(\varepsilon)$, and therefore $\bar{q} \in Q_1(\varepsilon_1)$, so that \bar{q} belongs to a stratum V of \mathcal{V} . Moreover, V is entirely contained in $Q_1(\varepsilon_1)$, so that $\dim V \leq 1$. Since $(\mathcal{V}, \mathcal{T})$ is compatible with π over $\text{Sq}(\varepsilon_1) \times \text{Sq}(\varepsilon_1)$, it follows that $\pi(V) \in \mathcal{T}$. On the other hand, $\bar{q}_1 \in \pi(V) \cap \text{Sq}(\varepsilon)$, so that $\pi(V)$ meets $\text{Sq}(\varepsilon)$, and therefore $\pi(V)$ is one-dimensional, and $\pi(V) \cap \text{Sq}(\varepsilon) \in \mathcal{A}_p^+(\varepsilon)$ (because ε is good for \mathcal{T}).

Since \mathcal{T} is compatible with the subanalytic set $\pi_2(Q_1(\varepsilon_1))$, the point \bar{q}_1 must belong to a stratum W of \mathcal{T} , such that $W \subseteq \pi_2(Q_1(\varepsilon_1))$. Since W meets $\text{Sq}(\varepsilon)$, and ε is good for \mathcal{T} , we see that W is one-dimensional, and that $W \in \mathcal{A}_p^+(\varepsilon)$. The point \bar{q}_1 cannot belong to the closure of any set $T \cap \text{Sq}(\varepsilon)$, for $T \in \mathcal{T}_1$. So there is a neighborhood U of \bar{q}_1 such that $U \cap \pi_2(Q_1(\varepsilon_1)) = U \cap W$.

Now pick an open arc V_0 of V , such that $\bar{q} \in V_0$ and that $\pi_2(V_0) \subseteq U$. The map $\pi \upharpoonright V_0$ is a diffeomorphism from V_0 onto $\pi(V_0)$. Since $\pi(V_0) \subseteq \pi(V)$, the set $\pi(V_0)$ is of the form $\{(\psi(y), y) : w < y < w'\}$, where $w < \bar{y} < w'$ and $\psi :]w, w'[\rightarrow \mathbb{R}$ is analytic. (Here we let \bar{y} be the y -coordinate of \bar{q}_0 and of \bar{q}_1 .) Moreover, we can use the y -coordinate, for $y \in]w, w'[$, to parametrize V_0 . Then V_0 is the set of all points $(\psi(y), y, \psi'(y), \psi''(y))$, where ψ' , ψ'' are analytic functions on $]w, w'[$. Since $(\psi'(y), \psi''(y)) \sim (\psi(y), y)$ (because $V_0 \subseteq Q_1(\varepsilon_1) \subseteq Q_0(\varepsilon_0)$), we necessarily have $\psi''(y) = y$. On the other hand, the point $(\psi'(y), y)$ belongs to $\pi_2(V_0) \cap U$, and therefore to $\pi_2(Q_1(\varepsilon_1)) \cap U$. So $(\psi'(y), y) \in W$. Therefore, if $W = \{(\psi^*(y), y) : w^* < y < w^{*'}\}$, we see that $\psi' = \psi \upharpoonright]w, w'[$.

We now prove that one of the sets $\pi(V) \cap \text{Sq}(\varepsilon)$, $W \cap \text{Sq}(\varepsilon)$ is necessarily a barrier in $\text{Sq}(\varepsilon)^+$. It is clear that X points to the right of both sets. Since \mathcal{T} is compatible with Y , if $Y(\bar{q}_0)$ points to the right of $\pi(V) \cap \text{Sq}(\varepsilon)$, or is tangent to it, then the same will be true for $Y(q)$, for all $q \in \pi(V) \cap \text{Sq}(\varepsilon)$, proving that $\pi(V) \cap \text{Sq}(\varepsilon)$ is a barrier. A similar conclusion works for W . So, all we need is to prove that $Y(\bar{q}_0)$ and $Y(\bar{q}_1)$ cannot both point to the "left" side of $\pi(V) \cap \text{Sq}(\varepsilon)$, $W \cap \text{Sq}(\varepsilon)$, respectively. Suppose they do. The numbers $\beta(\bar{q}_0)$, $\beta(\bar{q}_1)$ both have the same sign. (Recall that we are in Subcase 1b, so that $\beta(q_0)\beta(q_1) > 0$. Moreover, as long as $s \in [0, s_{\max}]$, $\zeta_0(s)$ and $\zeta_1(s)$ are ordinary points, and so $\beta(\zeta_0(s))$ and $\beta(\zeta_1(s))$ cannot change sign.) So $Y(\bar{q}_0)$ and $Y(\bar{q}_1)$ both point up, or both point down. In the former case, let v be the tangent vector v_1 at \bar{q} to the curve $y \mapsto (\psi(y), y, \psi'(y), y)$. If both $Y(\bar{q}_0)$ and $Y(\bar{q}_1)$ point down, let $v = -v_1$. In either case, v is tangent to V , and therefore $\pi_*(v)$, $\pi_2^*(v)$ are tangent to $\pi(V) \cap \text{Sq}(\varepsilon)$, $W \cap \text{Sq}(\varepsilon)$, respectively. Moreover, we have $\pi_*(v) = \theta_1 Y(\bar{q}_0) + \theta_2 X(\bar{q}_0)$, and $\pi_2^*(v) = \theta_3 Y(\bar{q}_1) + \theta_4 X(\bar{q}_1)$, where the coefficients θ_i are strictly positive. On the other hand, using the fact that π is a local diffeomorphism on $Q_0(\varepsilon_0)$, we get $v = \theta_1 Y^*(\bar{q}) + \theta_2 X^*(\bar{q})$ and, projecting via π_2 , we get $\pi_2^*(v) = \theta_1 Y^{\sim}(\bar{q}) + \theta_2 X^{\sim}(\bar{q})$. Because $\bar{q} \in Q_1(\varepsilon_1)$, the vectors $Y(\bar{q}_1)$ and $Y^{\sim}(\bar{q})$ are linearly dependent. Since both

have the same y -component, we see that $Y^{\sim}(\bar{q}) = \theta_5 Y(\bar{q}_1)$ for some $\theta_5 > 0$. Also, since $\bar{q} \in Q_0(\varepsilon_0)$ and $\Delta_A(\bar{q}_0)\Delta_A(\bar{q}_1) > 0$, we see from Lemma 3.15 that $X^{\sim}(\bar{q}) = -\theta_6 X(\bar{q}_1)$, where $\theta_6 > 0$. So

$$(3.52) \quad \pi_{2*}(v) = \theta_1 \theta_5 Y(\bar{q}_1) - \theta_2 \theta_6 X(\bar{q}_1).$$

If we equate coefficients for the two expressions of $\pi_{2*}(v)$ as a linear combination of $Y(\bar{q}_1)$ and $X(\bar{q}_1)$, we get $\theta_3 = -\theta_2 \theta_6$, which contradicts the fact that all the θ_i are strictly positive.

This contradiction shows that one of the sets $\pi(V) \cap \text{Sq}(\varepsilon)$, $W \cap \text{Sq}(\varepsilon)$ is a barrier in $\text{Sq}(\varepsilon)^+$. Since both $\pi(V)$ and W are in \mathcal{T} , we see that $\pi(V) \cap \text{Sq}(\varepsilon) \in \mathcal{B}^+$ or that $W \cap \text{Sq}(\varepsilon) \in \mathcal{B}^+$. On the other hand, it is clear that all the points $\zeta_0(s)$, $\zeta_1(s)$, $0 \leq s < s_{\max}$, lie on γ . So γ meets a member of \mathcal{B}^+ . This concludes the discussion of Sub-subcase 1b.i.

We now consider Sub-subcases 1b.ii and 1b.iii. (cf. Fig. 2). We may have $\hat{q}_0 = \hat{q}_1$ or $\hat{q}_0 \neq \hat{q}_1$. If $\hat{q}_0 \neq \hat{q}_1$, then, as observed earlier, one of the equalities $\Delta_A(\hat{q}_0) = 0$, $\Delta_A(\hat{q}_1) = 0$, $\Delta_B(\hat{q}_0) = 0$, $\Delta_B(\hat{q}_1) = 0$ must hold, so that at least one of the \hat{q}_i is in E . If, say, $\hat{q}_0 \in E$, then $\hat{q}_0 \in S$ for an $S \in \mathcal{S}_1$, so that $\hat{q}_0 \in S$ for an $S \in \mathcal{P}$. Therefore $\hat{q}_1 \in S^{\sim}$, and so the stratum $T \in \mathcal{T}$ to which \hat{q}_1 belongs is one-dimensional. If $\hat{q}_0 = \hat{q}_1$ then, again, \hat{q}_0 cannot be an ordinary point, and so \hat{q}_0 belongs to an $S \in \mathcal{S}_1$. In either case, we see that the points \hat{q}_0 , \hat{q}_1 belong to one-dimensional strata S_0 , S_1 of \mathcal{T} . Moreover, these strata have the property that, whenever r_0 , r_1 are points in S_0 , S_1 which lie on the same horizontal line, then $r_0 \sim r_1$. (Reason: suppose, e.g., that $S_0 \subseteq S \in \mathcal{S}_1$. Since $\hat{q}_1 \in S^{\sim} \cap S_1$, it follows that $S_1 \subseteq S^{\sim}$. Therefore $r_1 \sim r'_0$ for some $r'_0 \in S$ which, of necessity, must lie on the same horizontal line as r_1 and r_0 . Since S intersects each horizontal line at most once, it follows that $r'_0 = r_0$.)

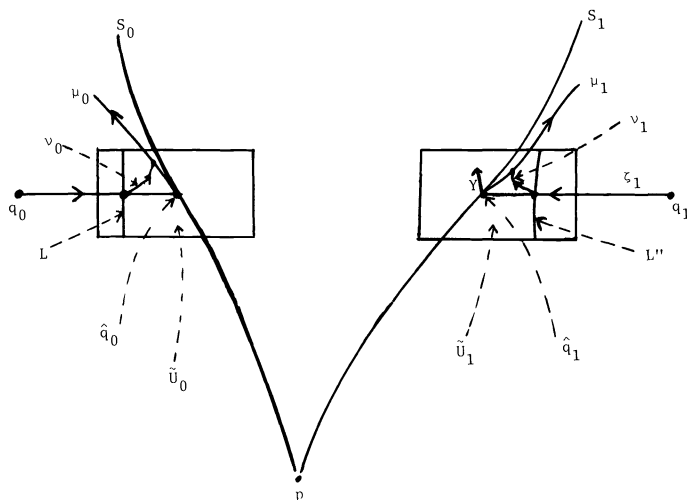


FIG. 2. (The situation depicted here is proved to be impossible, because $Y(\hat{q}_1)$ has to point to the right of μ_1 .)

We now prove that at least one of the sets $S_i \cap \text{Sq}(\varepsilon)$ is a barrier in $\text{Sq}(\varepsilon)^+$. As before, it is sufficient to prove that it is impossible for $Y(\hat{q}_i)$ to point to the “left” of S_i for $i = 0, 1$.

Suppose $Y(\hat{q}_0)$, $Y(\hat{q}_1)$ point to the “left” of S_0 , S_1 . Let $\hat{q}_i = (\hat{x}_i, \hat{y}_i)$ (so that $\hat{y}_0 = \hat{y}_1 = \hat{y}$). Pick rectangular neighborhoods \tilde{U}_i ($i = 0, 1$) of \hat{q}_i , of the form

$$(3.53) \quad \tilde{U}_i =]\hat{x}_i - \delta, \hat{x}_i + \delta[\times]\hat{y} - \delta', \hat{y} + \delta'[$$

such that

$$(3.54) \quad \tilde{U}_i \cap S_i = \{(\psi_i(y), y) : \hat{y} - \delta' < y < \hat{y} + \delta\},$$

where the ψ_i are analytic functions on $] \hat{y} - \delta, \hat{y} + \delta[$, and that \tilde{U}_i intersects no other one-dimensional stratum \mathcal{T} . Then each set $\tilde{U}_i - S_i$ is partitioned into two disjoint connected open sets $(\tilde{U}_i)_L, (\tilde{U}_i)_R$ (the “left” and right sides of S_i in U_i) defined by

$$(3.55a) \quad (x, y) \in (\tilde{U}_i)_L \Leftrightarrow (x, y) \in \tilde{U}_i \quad \text{and} \quad x < \psi_i(y),$$

$$(3.55b) \quad (x, y) \in (\tilde{U}_i)_R \Leftrightarrow (x, y) \in \tilde{U}_i \quad \text{and} \quad x > \psi_i(y).$$

The vectors $G(\hat{q}_0), G(\hat{q}_1)$ are both nonzero (otherwise we would have $Y(\hat{q}_i) = X(\hat{q}_i)$ for one of the i 's, and so $Y(\hat{q}_i)$ would point to the “right” of S_i). Since they are linearly dependent (because $\hat{q}_0 \sim \hat{q}_1$), it follows that either $\beta(\hat{q}_0) \neq 0$ and $\beta(\hat{q}_1) \neq 0$, or $\beta(\hat{q}_0) = \beta(\hat{q}_1) = 0$, in which case

$$\alpha(\hat{q}_0) - 1 \neq 0 \quad \text{and} \quad \alpha(\hat{q}_1) - 1 \neq 0.$$

By making δ and δ' smaller, if needed, we may assume that either

$$(3.XVI.1) \quad \beta \text{ never vanishes on } \tilde{U}_0 \cup \tilde{U}_1, \text{ or}$$

$$(3.XVI.2) \quad \alpha - 1 \text{ never vanishes on } \tilde{U}_0 \cup \tilde{U}_1.$$

Notice that, if $\hat{q}_0 = \hat{q}_1$, then $S_0 = S_1$ and $\tilde{U}_0 = \tilde{U}_1$. If $\hat{q}_0 \neq \hat{q}_1$, we shall assume that δ, δ' are so small that $\tilde{U}_0 \cap \tilde{U}_1 = \emptyset$.

Define $\tau: \tilde{U}_0 \cup \tilde{U}_1 \rightarrow \mathbb{R}$ by

$$(3.56a) \quad \tau = \frac{1 - \alpha}{\beta} \quad \text{if (3.XVI.1) holds,}$$

$$(3.56b) \quad \tau = \frac{\beta}{\alpha - 1} \quad \text{if (3.XVI.2) holds.}$$

(If both (3.XVI.1) and (3.XVI.2) hold, we may define τ either way.) In either case, τ is a well-defined analytic function on $\tilde{U}_0 \cup \tilde{U}_1$, and two points $(x_0, y) \in \tilde{U}_0, (x_1, y) \in \tilde{U}_1$ satisfy $(x_0, y) \sim (x_1, y)$ if and only if

$$\tau(x_0, y) = \tau(x_1, y).$$

In particular, this implies that

$$(3.57) \quad \tau(\psi_0(y), y) = \tau(\psi_1(y), y) \quad \text{for } |y - \hat{y}| < \delta'.$$

The x -derivative of τ is given by

$$(3.58) \quad \partial_x \tau = \frac{(\alpha - 1)\beta_x - \alpha_x \beta}{D},$$

where the denominator D equals β^2 or $(\alpha - 1)^2$, according to whether (3.XVI.1) or (3.XVI.2) holds. In any case, we see that $\partial_x \tau$ equals Δ_B times a strictly positive number, so that

$$(3.59) \quad (\partial_x \tau) \Delta_B > 0 \quad \text{wherever } \Delta_B \neq 0.$$

In particular, Δ_B never vanishes on $(\tilde{U}_0)_L$ or on $(\tilde{U}_1)_R$, and therefore

$$(3.60) \quad (\partial_x \tau) \Delta_B > 0 \quad \text{on } (\tilde{U}_0)_L \cup (\tilde{U}_1)_R.$$

On the other hand, the point $\zeta_0(s)$ is in $(\tilde{U}_0)_L$ for s near s_{\max} , while $\zeta_1(s) \in (\tilde{U}_1)_R$. Since $\Delta_A(\zeta_0(s))$ and $\Delta_A(\zeta_1(s))$ have the same sign for all $s \in [0, s_{\max}]$, and $(\Delta_A \Delta_B)(\zeta_0(s)) \cdot (\Delta_A \Delta_B)(\zeta_1(s)) < 0$ (because $\zeta(s) \in Q_0(\varepsilon_0)$), it follows that $\Delta_B(\zeta_0(s)) \cdot \Delta_B(\zeta_1(s)) < 0$ and so Δ_B has opposite signs on $(\tilde{U}_0)_L$ and on $(\tilde{U}_1)_R$. Therefore $\partial_x \tau$ has opposite signs on $(\tilde{U}_0)_L$ and on $(\tilde{U}_1)_R$.

Let $\tilde{\delta}$ be so small that $\zeta_i(s) \in \tilde{U}_i$ for $i = 0, 1$, $s > s_{\max} - \tilde{\delta}$. Let L be an open vertical segment, of height $2\delta^* \leq 2\delta'$, whose center is the point $\zeta_0(\bar{s})$, where \bar{s} is a number such that $s_{\max} - \tilde{\delta} < \bar{s} < s_{\max}$. Make δ^* smaller, if needed, and assume that $L \subseteq (\tilde{U}_0)_L$. Since $\zeta(\bar{s}) \in Q_0(\varepsilon_0)$, and π is a local diffeomorphism on $Q_0(\varepsilon_0)$, we may assume, by making δ^* smaller if necessary, that $L = \pi(L')$, where L' is a smooth arc contained in $Q_0(\varepsilon_0)$, such that $\zeta(\bar{s}) \in L'$. Let $L'' = \pi_2(L')$. Then

$$(3.61) \quad L'' = \{(\theta(y), y) : \hat{y} - \delta^* < y < \hat{y} + \delta^*\},$$

where $\theta :]\hat{y} - \delta^*, \hat{y} + \delta^*] \rightarrow \mathbb{R}$ is analytic. Now replace δ' by δ^* (i.e. shrink both \tilde{U}_i in the vertical direction, if needed). Then L is a vertical segment which goes from the lower to the upper edge of \tilde{U}_0 , and is entirely contained in $(\tilde{U}_0)_L$. Also, L'' is an arc which goes from the lower to the upper edge of \tilde{U}_1 . Moreover, L'' is never horizontal, because it is of the form $\{(\theta(y), y) : \hat{y} - \delta' < y < \hat{y} + \delta'\}$. Let $\theta_1 = \theta$, and let θ_0 be the function such that $L = \{(\theta_0(y), y) : |y - \hat{y}| < \delta'\}$ (so that θ_0 is actually a constant). Let τ' denote τ , if $\partial_x \tau < 0$ on $(\tilde{U}_0)_L$, and let τ' be $-\tau$ if $\partial_x \tau > 0$ on $(\tilde{U}_0)_L$. Then τ' is strictly decreasing as a function of x on $(\tilde{U}_0)_L$, and it is therefore strictly increasing as a function of x on $(\tilde{U}_1)_R$. Therefore $\tau'(\theta_0(y), y) > \tau'(\psi_0(y), y)$ for $|y - \hat{y}| < \delta'$. Since, by construction, $(\theta_0(y), y) \sim (\theta_1(y), y)$, we see that $\tau'(\theta_1(y), y) > \tau'(\psi_1(y), y)$. This implies, in particular, that L'' never meets $S_1 \cap \tilde{U}_1$. Since $\zeta_1(\bar{s}) \in L'' \cap (\tilde{U}_1)_R$, we see that $L'' \subseteq (\tilde{U}_1)_R$. If r is any point (x, y) such that $|y - \hat{y}| < \delta'$, $\theta_0(y) \leq x \leq \psi_0(y)$, it follows that there exists a unique point $K(r) = (k(x, y), y)$ such that $K(r) \sim r$ and that $\psi_1(y) \leq k(x, y) \leq \theta_1(y)$. If, in addition, $x < \psi_0(y)$, then it follows that $k(x, y) > \psi_1(y)$. In this case, both r and $K(r)$ are ordinary points. Moreover, we already know that Δ_A has the same sign on $(\tilde{U}_0)_L$ as on $(\tilde{U}_1)_R$, whereas Δ_B has opposite signs. So $(\Delta_A \Delta_B)(r)$ and $(\Delta_A \Delta_B)(K(r))$ have opposite signs, and therefore $(r, K(r)) \in Q_0(\varepsilon_0)$. Since $\pi(r, K(r)) = r$, and K is clearly continuous, the fact that π is a local diffeomorphism on $Q_0(\varepsilon_0)$ implies that K is analytic on $\{(x, y) : \theta_0(y) \leq x < \psi_0(y), |y - \hat{y}| < \delta'\}$. Now let $\mu_0(\cdot)$ denote the integral curve of Y which goes through \hat{q}_0 at time 0. Let t_{\max} be the largest t such that $\mu_0([0, t])$ is entirely contained in $\tilde{U}_0 \cap \{(x, y) : x > \theta_0(y)\}$. Then $\mu_0(t_{\max})$ is well defined, and belongs to L , or to the boundary of \tilde{U}_0 . Since we are assuming that Y points to the left of S_0 , the curve $\mu_0 \upharpoonright]0, t_{\max}[$ is entirely contained in $(\tilde{U}_0)_L$.

Now let $\mu_1(t) = K(\mu_0(t))$. We see that $\mu_1(0) = \hat{q}_1$, that $\mu_1 \upharpoonright]0, t_{\max}[$ is entirely contained in $(\tilde{U}_1)_R$, and that $(\mu_0(t), \mu_1(t)) \in Q_0(\varepsilon_0)$ for $0 < t < t_{\max}$. Let $\mu(t) = (\mu_0(t), \mu_1(t))$. Then $\mu \upharpoonright]0, t_{\max}[$ is a curve in $Q_0(\varepsilon_0)$, and $\pi \circ \mu = \mu_0$, which is a Y -trajectory. So $\mu \upharpoonright]0, t_{\max}[$ is a Y -trajectory, and therefore

$$(3.62) \quad \dot{\mu}_1(t) = Y^-(\mu_0(t), \mu_1(t)) \quad \text{for } 0 < t < t_{\max}.$$

As observed before, the vector $Y^-(q_0, q_1)$ is not a linear combination of $Y(q_1)$ and $X(q_1)$ with strictly positive coefficients. We claim that the same conclusion then follows for $Y^-(\mu_0(t), \mu_1(t))$, for $t \in]0, t_{\max}[$. To see this, let us consider, for a particular t , the curve ν_0 obtained by following ζ_0 from q_0 to $\zeta_0(\bar{s})$, and then some continuous curve $\tilde{\nu}$ in $(\tilde{U}_0)_L \cap \{(x, y) : x \geq \theta_0(y)\}$ up to $\mu_0(t)$. Then there is a continuous curve ν_1 (obtained by following ζ_1 from q_1 to $\zeta_1(\bar{s})$, and then $K \circ \tilde{\nu}$ to $\mu_1(t)$) such that, if

$\nu(s) = (\nu_0(s), \nu_1(s))$, then ν is a curve in $Q_0(\varepsilon_0)$, such that $\pi \circ \nu = \nu_0$. For each $s \in \text{Dom}(\nu)$, we can express $Y^-(\nu(s))$ in a unique way as a combination

$$(3.63) \quad \tilde{Y}(\nu(s)) = \theta_1(s)Y(\nu_1(s)) + \theta_2(s)X(\nu_1(s)).$$

The coefficients θ_1, θ_2 are continuous functions of s . We know that, initially, θ_1 and θ_2 are not both positive. We will use this to show that

$$(3.XVII) \quad \text{For every } s \in \text{Dom}(\nu) \text{ either } \theta_1(s) \leq 0 \text{ or } \theta_2(s) \leq 0.$$

In view of (3.XV), the desired conclusion holds when s is the left endpoint of $\text{Dom}(\nu)$.

We also know that $Y^-(r_0, r_1)$ has the same y -component as $Y(r_0)$ for all $(r_0, r_1) \in Q_0(\varepsilon_0)$. Since $\nu_0(s)$ and $\nu_1(s)$ are ordinary points for all $s \in \text{Dom}(\nu)$ (because ν is a curve in $Q_0(\varepsilon_0)$), the function $s \rightarrow \beta(\nu_0(s))\beta(\nu_1(s))$ cannot vanish for $s \in \text{Dom}(\nu)$. Since $\beta(q_0)\beta(q_1) > 0$ (because we are in Subcase 1b), we conclude that

$$(3.64) \quad \beta(\nu_0(s))\beta(\nu_1(s)) > 0 \quad \text{for all } s \in \text{Dom}(\nu).$$

So $Y^-(\nu_0(s), \nu_1(s))$ and $Y(\nu_1(s))$ have y -components of the same sign, for each $s \in \text{Dom}(\nu)$. In particular, this implies that

$$(3.XVIII) \quad \theta_1(s) > 0 \text{ whenever } \theta_2(s) = 0.$$

We now prove (3.XVII). We treat Sub-subcases 1b.ii and 1b.iii separately. Let us begin with Sub-subcase 1b.iii. If s^* is such that $s^* \in [0, s_{\max}[$, and $\zeta(s^*) \in \hat{Q}_1(\varepsilon) - Q_1(\varepsilon)$, then the definition of $Q_1(\varepsilon)$ implies that $\zeta(s) \in \hat{Q}_1(\varepsilon)$ for s near s^* and so, by analyticity, that $Y(\zeta(s))$ and $Y^-(\zeta_0(s), \zeta_1(s))$ are linearly dependent for all $s \in [0, s_{\max}[$. So, either

$$(3.XIX.a) \quad \nu(s) \in \hat{Q}_1(\varepsilon) \quad \text{for all } s \in \text{Dom}(\nu), \text{ or}$$

$$(3.XIX.b) \quad \nu(s) \notin \hat{Q}_1(\varepsilon) \quad \text{for some } s \text{ such that } \nu_0(s) \in (\tilde{U}_0)_L.$$

If (3.XIX.a) holds, then $\theta_2(s) = 0$ for all $s \in \text{Dom}(\nu)$, and so (3.XVII) holds. If (3.XIX.b) holds, then $\nu(s^*)$ must belong to the frontier of $\hat{Q}_1(\varepsilon_1)$ for some s such that $\nu_0(s^*) \in (\tilde{U}_0)_L$ (because $\zeta(\bar{s}) \in \hat{Q}_1(\varepsilon_1)$). So $\nu(s^*) \in V^*$, where V^* is a stratum of \mathcal{V} such that $\dim V^* \leq 1$. But then $\nu_0(s^*) \in \pi(V^*)$, which is a stratum of \mathcal{T} such that $\dim \pi(V^*) \leq 1$. But this contradicts the fact that $(U_0)_L$ does not meet any low-dimensional strata of \mathcal{T} . This contradiction proves that (3.XVII) holds, if we are in Sub-subcase 1b.iii.

We now prove (3.XVII) in Sub-subcase 1b.ii. If (3.XVII) were not true, there would have to be an $s^* \in \text{Dom}(\nu)$ such that either $\theta_1(s^*) = 0$ or $\theta_2(s^*) = 0$. The possibility that $\theta_1(s^*) = 0$ is excluded, because the y -component of $Y^-(\nu(x))$ is nonzero for all s . So $\theta_2(s^*) = 0$, and $\nu(s^*) \in \hat{Q}_1(\varepsilon_1)$. Since we are in Sub-subcase 1b.ii, so that $\zeta(s) \notin \hat{Q}_1(\varepsilon_1)$ for $s \in [0, s_{\max}[$, we necessarily have $\nu(s^*) \in (\tilde{U}_0)_L \times (\tilde{U}_1)_R$. Let V^* be the stratum of \mathcal{V} to which $\nu(s^*)$ belongs. Then $\nu_0(s^*) \in \pi(V^*)$, which is a stratum of \mathcal{T} . If $\dim V^* \leq 1$, then $\dim \pi(V^*) \leq 1$, and so $\nu_0(s^*)$ is in a stratum of \mathcal{T} of dimension < 2 , which contradicts the fact that $\nu_0(s^*) \in (\tilde{U}_0)_L$, and that $(\tilde{U}_0)_L$ does not meet any low-dimensional strata of \mathcal{T} . So $\dim V^* = 2$. Since \mathcal{V} is compatible with $\hat{Q}_1(\varepsilon_1)$ and with $Q_1(\varepsilon_1)$, and $\dim Q_1(\varepsilon_1) \leq 1$, we necessarily have $V^* \subseteq \hat{Q}_1(\varepsilon_1) - Q_1(\varepsilon_1)$. Since $\pi: Q_0(\varepsilon_0) \rightarrow \text{Sq}(\varepsilon_0)$ is a local diffeomorphism, and $\hat{Q}_1(\varepsilon_1) \subseteq Q_0(\varepsilon_0)$, we see that $\pi(V^*)$ is open in $\text{Sq}(\varepsilon_0)$. But then $\nu(s^*)$ has a neighborhood W with the property that $(r, K(r)) \in \hat{Q}_1(\varepsilon_1)$ for all $r \in W$. Therefore the vectors $Y(K(r))$ and $Y^-(r, K(r))$ are linearly dependent for all $r \in W$. By analyticity, it follows that $Y(K(r))$ and $Y^-(r, K(r))$ are dependent for all $r \in (U_0)_L$, i.e., that $(r, K(r)) \in \hat{Q}_1(\varepsilon_1)$ for all $r \in (U_0)_L$. But then, in particular, $\zeta(\bar{s}) \in \hat{Q}_1(\varepsilon_1)$, which contradicts the hypothesis that we are in Sub-subcase 1b.ii.

So (3.XVII) holds in Sub-subcase 1b.ii as well.

Having proved (3.XVII) we can conclude, in particular, that $\theta_1(s) \leq 0$ or $\theta_2(s) \leq 0$ if s is the right endpoint of $\text{Dom}(\nu)$. So, for each $t \in \text{Dom}(\mu)$

(3.XX) The vector $Y^-(\mu(t))$ is not a linear combination of $X(\mu_1(t))$ and $Y(\mu_1(t))$ with strictly positive coefficients.

Since the y -components of $Y(\mu_1(t))$ and $Y^-(\mu(t))$ have the same sign, (3.XX) implies that:

(3.XXI) $Y(\mu_1(t))$ is a linear combination of $X(\mu_1(t))$ and $\tilde{Y}(\mu(t))$ with nonnegative coefficients.

From (3.XXI) we conclude that, at each point $\mu_1(t)$, both $Y(\mu_1(t))$ and $X(\mu_1(t))$ point to the same side of μ_1 . Also, we know that β is either >0 throughout $(\tilde{U}_0)_L \cup (\tilde{U}_1)_R$, or that it is <0 throughout $(\tilde{U}_0)_L \cup (\tilde{U}_1)_R$. Consider the former case. (The other one is similar.) Then $Y^-(\mu_0(t), \mu_1(t))$ has a positive y -component for $t \in]0, t_{\max}[$. So the curve μ_1 starts off at $\hat{q}_1 \in S_1$, and then goes to the right and up, into $(\tilde{U}_1)_R$. On the other hand, the integral curve of Y through \hat{q}_1 goes up but enters $(\tilde{U}_1)_L$. So, for \hat{q}'_1 near \hat{q}_1 and to the right of it, the integral curve of Y through \hat{q}'_1 also enters $(\tilde{U}_1)_L$ after a while. Therefore this curve has to cross μ_1 from "right" to "left". At the crossing, Y and X must point to opposite sides of μ_1 , which is a contradiction. This contradiction completes the proof that either S_0^ε or S_1^ε is a barrier in $\text{Sq}(\varepsilon)^+$. This completes the analysis of Sub-subcase 1b.ii, which was the only missing sub-subcase of Subcase 1b, which was the only missing subcase of Case 1. So our analysis of Case 1 is complete.

We now proceed to Case 2.

CASE 2. $t_0 = t'_0$ and $t'_1 = t_1$, but at least one of q_0, q_1 is not an ordinary point.

We consider first the easiest subcase.

SUBCASE 2a. $\Delta_A(q_0)\Delta_A(q_1) = 0$.

If $\Delta_A(q_0) = 0$, then $q_0 \in S$ for some stratum $S \in \mathcal{S}_1$ such that Δ_A vanishes on S . We claim that $S \cap \text{Sq}(\varepsilon)$ is a barrier in $\text{Sq}(\varepsilon)^+$. To prove this, it suffices to exclude the possibility that $Y(q_0)$ points to the left of S . Since $Y(q_0)$ and $X(q_0)$ are linearly dependent, this possibility can only occur if $Y(q_0) = -\rho X(q_0)$ for some $\rho > 0$. On the other hand, it is clear that the X - and Y -curves through q_0 have a finite-order tangency at q_0 . So we can apply Lemma 5.1 of [A], and conclude that γ is not optimal. This contradiction shows that S^ε is indeed a barrier in $\text{Sq}(\varepsilon)^+$.

If $\Delta_A(q_0) \neq 0$, but $\Delta_A(q_1) = 0$, a similar reasoning shows that $q_1 \in S$ for some $S \in \mathcal{S}_1$, and that S^ε is a barrier in $\text{Sq}(\varepsilon)^+$. This concludes the analysis of Subcase 2a.

We now analyze the following.

SUBCASE 2b. $\Delta_A(q_0) \neq 0 \neq \Delta_A(q_1)$.

In this subcase, at least one of the numbers $\Delta_B(q_0), \Delta_B(q_1)$ must vanish. Therefore one of q_0, q_1 belongs to an $S \in \mathcal{S}_1$. This S is necessarily in \mathcal{P} , and so the other q_i is in S^- , and therefore in a one-dimensional stratum of \mathcal{T} . In either case, the points q_0, q_1 belong to one-dimensional strata S_0, S_1 of \mathcal{T} , with the property that, whenever $r_0 \in S_0, r_1 \in S_1$, and r_0 lies in the same horizontal line as r_1 , then $r_0 \sim r_1$. We distinguish two sub-subcases.

SUB-SUBCASE 2b.i. $\beta(q_0)\beta(q_1) < 0$,

SUB-SUBCASE 2b.ii. $\beta(q_0)\beta(q_1) > 0$.

Sub-subcase 2b.i is handled exactly like Sub-subcase 1b.i. Since $q_0 \sim q_1$, this sub-subcase can only occur if (3.II.a) holds, in which case $\alpha > 0$ throughout $\text{Sq}(\varepsilon_0)$. On the other hand, there must be a $t \in [t_0, t_1]$ such that $\beta(\gamma(t)) = 0$. Therefore $\gamma(t) \in E_1$, and so $\gamma(t) \in S$, where $S \in \mathcal{S}_1$ and $\Delta_A \equiv 0$ on S . Since $\alpha > 0$ on $\text{Sq}(\varepsilon_0)$, we see that $Y(q)$ is a positive multiple of $X(q)$ for each $q \in S$. Therefore S^ε is a barrier in $\text{Sq}(\varepsilon)$.

We now turn to Sub-subcase 2b.ii. In this sub-subcase, Δ_A never vanishes on $S_0 \cup S_1$, and so β is either positive throughout $S_0 \cup S_1$, or negative throughout $S_0 \cup S_1$. We consider the case when $\beta > 0$ on $S_0 \cup S_1$. (The other case is similar.)

Suppose that both S_0 and S_1 fail to be barriers. Then Y points to the “left” of both S_0 and S_1 . Therefore both S_0 and S_1 , if suitably reparametrized, are trajectories of Σ . Now pick a point \bar{p} in γ , of the form $\gamma(t)$ for some $t < t_0$, but close to t_0 . Let p_0, p_1 be the points where the horizontal line through \bar{p} meets S_0, S_1 . Then we can apply Lemma 3.6 of [A] and conclude that

(3.65)

$$T(p_0, p_1 q_1) = T(p_0 m_0 q_0 q_1)$$

(where, for any points r_1, r_2, \dots, r_m , we use $T(r_1 r_2 \dots r_m)$ to denote the time along the trajectory from r_1 to r_m which goes through $r_2 \dots r_{m-1}$ as shown in Fig. 3). On the other hand, we claim that

(3.66)

$$T(p_0 m_0 q_0) < T(p_0 \bar{p} q_0).$$

To see this, notice first that f never vanishes in the open region \mathcal{R} bounded by the arcs $p_0 \bar{p}$, $\bar{p} m_2$, $m_2 q_1$ and $q_0 p_0$. Since γ is time-optimal, $T(\bar{p} q_0 m_2) \leq T(\bar{p} m_1 m_2)$, and therefore Lemma 3.11 of [A] excludes the possibility that $f > 0$ on \mathcal{R} . Then $f < 0$ on \mathcal{R} , and therefore (3.66) holds.

If we add $T(q_0 q_1)$ to both sides, we find that

(3.67)

$$T(p_0 m_0 q_0 q_1) < T(p_0 \bar{p} q_0 q_1).$$

In view of (3.65), we have

(3.68)

$$T(p_0 \bar{p} p_1 q_1) < T(p_0 \bar{p} q_0 q_1).$$

Subtracting $T(p_0 \bar{p})$ from both sides, we get

(3.69)

$$T(\bar{p} p_1 q_1) < T(\bar{p} q_0 q_1).$$

Therefore the arc $\bar{p} q_0 q_1$ is not optimal. Since this arc is a piece of γ , we conclude that γ is not optimal. This contradiction establishes that S_0^e or S_1^e is a barrier. This completes the analysis of Sub-subcase 2b.ii, which concludes the analysis of Case 2.

We now consider Case 3.

CASE 3. $t_0 = t'_0$ but $t'_1 < t_1$.

In this case, the switching at q'_1 is necessarily from an X - to a Z -trajectory. Therefore $q'_1 \in E_2$, so that q'_1 belongs to a stratum $S \in \mathcal{S}_1$ such that $\Delta_B \equiv 0$ on S . (Actually,

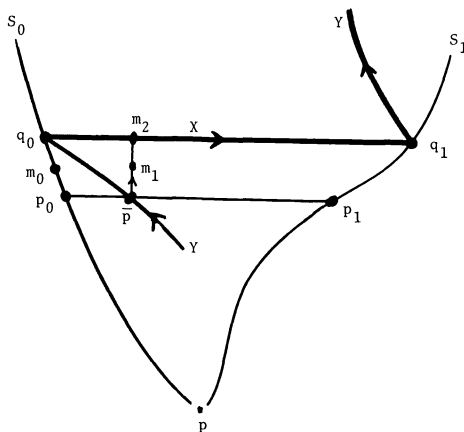


FIG. 3

S must be a turnpike.) So $q_0 \in S^\sim$, and therefore both q_0 and q_1 belong to one-dimensional strata S_0, S_1 of \mathcal{T} . If $\Delta_A(q_0) = 0$, then we prove that S_0^ε must be a barrier in $Sq(\varepsilon)^+$, using exactly the same reasoning as in Subcase 2a. The possibility that $\Delta_A(q'_1) = 0$ is excluded because S_1 is a turnpike.

Now suppose that $\Delta_A(q_0)$ and $\Delta_A(q'_1)$ are nonzero. Then we distinguish the subcases when $\Delta_A(q_0)\Delta_A(q'_1) < 0$, and when $\Delta_A(q_0)\Delta_A(q'_1) > 0$. The first case is handled exactly like Sub-subcase 2b.i, by showing that some point of γ , lying between q_0 and q'_1 , must belong to a one-dimensional $S \in \mathcal{T}$ such that S^ε is a barrier. Finally, if $\Delta_A(q_0)\Delta_A(q'_1) > 0$, then we proceed exactly as in Sub-subcase 2b.ii. We assume that $\beta > 0$ throughout $S_0 \cup S_1$ (the other case being similar). Then we are exactly in the same situation as in Fig. 3, except that now q_1 is replaced by q'_1 and that, after q'_1 , γ continues along S_1 , rather than along a Y -trajectory. Since the reasoning of Sub-subcase 2b.ii did not in any way depend on the nature of γ after it went through q'_1 , this reasoning applies here as well, and we reach again the conclusion that γ is not optimal. This concludes the analysis of Case 3.

Next, we consider Case 4.

CASE 4. $t_0 < t'_0$ but $t'_1 = t_1$.

This case is identical to Case 3. Both q'_0 and q_1 must belong to one-dimensional strata S_0, S_1 of \mathcal{T} . Since S_0 is a turnpike, we necessarily have $\Delta_A(q'_0) = 0$. If $\Delta_A(q_1)$ vanishes, we see that S_1^ε must be a barrier in $Sq(\varepsilon)^+$. If both $\Delta_A(q'_0)$ and $\Delta_A(q_1)$ are nonzero, but they have opposite signs, then there has to be an $S \in \mathcal{T}_1$ which crosses γ between q'_0 and q_1 , and is a barrier in $Sq(\varepsilon)^+$. Finally, if $\Delta_A(q'_0)\Delta_A(q_1) > 0$, then we may assume that $\beta > 0$ on $S_0 \cup S_1$. (The other case is similar.) Suppose that neither S_0^ε nor S_1^ε are barriers. Then we are in the situation shown in Fig. 4. Using exactly the same reasoning as in the study of Sub-subcase 2b.ii, we show that $T(q'_0 p_0 \bar{p}) < T(q'_0 q_1 \bar{p})$, so that γ is not time-optimal. This completes the analysis of Case 4.

Finally, we consider Case 5.

CASE 5. $t_0 < t'_0 < t'_1 < t_1$.

Here, again, the points q'_0, q'_1 must belong to one-dimensional strata S_0, S_1 of \mathcal{T}_1 , such that Δ_B vanishes identically on $S_0 \cup S_1$. On the other hand, Δ_A can never vanish on S_0 or on S_1 , because both S_0 and S_1 are turnpikes. If we have $\Delta_A(q'_0)\Delta_A(q'_1) < 0$,

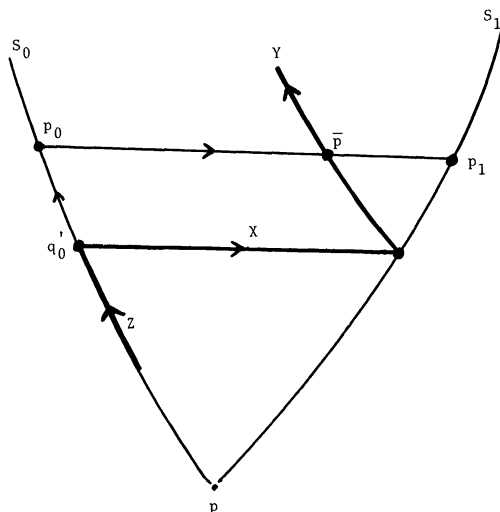


FIG. 4

then the reasoning of Sub-subcase 2b.i gives us a stratum $S \in \mathcal{T}_1$ that crosses γ somewhere between q'_0 and q'_1 , and is such that S^ε is a barrier in $Sq(\varepsilon)^+$.

If $\Delta_A(q'_0)\Delta_A(q'_1) > 0$, then we distinguish the subcases.

SUBCASE 5.i. $\beta(q'_0) > 0$ and $\beta(q'_1) > 0$.

SUBCASE 5.ii. $\beta(q'_0) < 0$ and $\beta(q'_1) < 0$.

(As we shall see below, there is a very minor technical reason why these subcases are slightly different, so that we cannot just limit ourselves to dealing with one of them.)

Consider Subcase 5.i, and suppose that both S_0^ε and S_1^ε fail to be barriers. Then we are in the situation shown in Fig. 5. (Recall that $[t'_0, t'_1]$ was the *leftmost* maximal subinterval of $[t_0, t_1]$ on which γ is an X -trajectory.) Since $(x, y) \sim (x', y)$ whenever $(x, y) \in S_0$ and $(x', y) \in S_1$, we can apply Lemma 3.6 of [A], and conclude that $T(q_0 p_1 q'_1) = T(q_0 q'_0 q'_1)$. So, if we modify γ by replacing the arc from q_0 to q'_0 to q'_1 by the arc from q_0 to p_0 to q'_1 , we see that the new trajectory $\tilde{\gamma}$ is also time-optimal. However, the reasoning of Sub-subcase 2b.i can be applied to $\tilde{\gamma}$, to conclude that $T(\bar{p} r p_1) < T(\bar{p} q_0 p_1)$, and therefore $\tilde{\gamma}$ is not optimal. This contradiction settles Subcase 5.i.

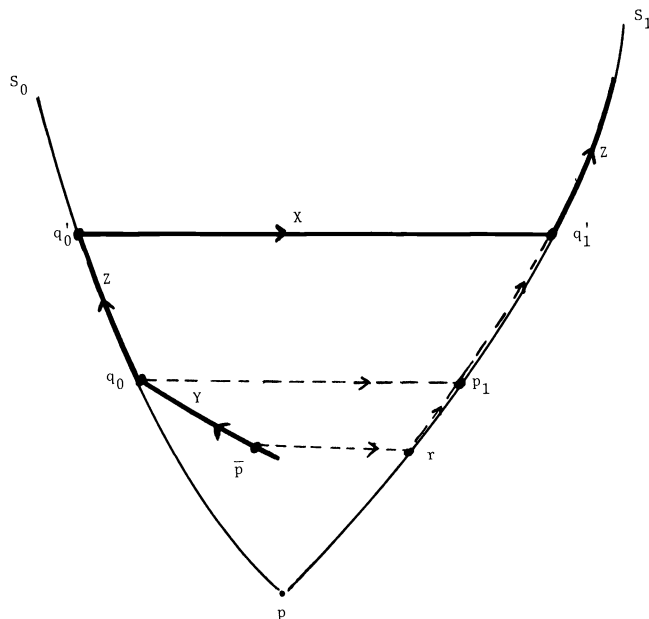


FIG. 5

Note that the preceding reasoning depends strongly on the fact that, for each point m_0 in S_0 , lying between q_0 and q'_0 , the horizontal line through m_0 actually meets S_1 . If we are in Subcase 5.ii, then γ might be as shown in Fig. 6, so that the horizontal line through q_0 does not meet S_1 , because S_1 leaves $Sq(\varepsilon)^+$ through its right vertical edge, at a height below that of q_0 . So, in this subcase, the reasoning has to be modified slightly. Let $[t'_1, t'^*_1]$ be the maximal subinterval of $[t_0, t_1]$ that contains t'_1 and is such that $\gamma|_{[t'_1, t'^*_1]} \in \text{Traj}(Z)$. Let $q'^*_1 = \gamma(t'^*_1)$. Then Lemma 3.6 of [A] implies that $T(q'_0 p_1 q'^*_1) = T(q'_0 q'_1 q'^*_1)$. Therefore we can form a new trajectory γ_2 by substituting the arc $q'_0 \rightarrow p_1 \rightarrow q'^*_1$ for the arc $q'_0 \rightarrow q'_1 \rightarrow q'^*_1$. The new trajectory γ_2 is also optimal. If there is another singular piece to the right of t'^*_1 , let t'_2, t'^*_2 be the endpoints of the

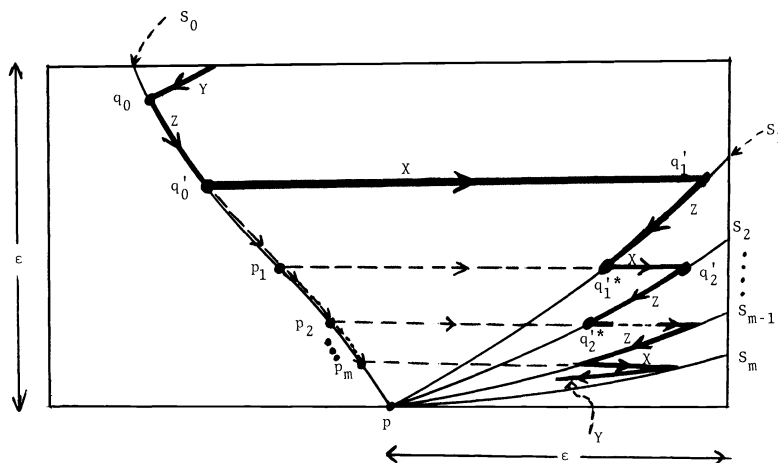


FIG. 6

leftmost such piece, and let S_2 be the stratum containing it. If $\Delta_A > 0$ on S_2 then there would have to be a barrier of the form S^ε , $S \in \mathcal{T}_1$, lying between S_1 and S_2 (by the reasoning of Sub-subcase 2b.i). If $\Delta_A < 0$ on S_2 , then $\gamma \upharpoonright [t'_2, t_2^*]$ goes downwards along S_2 . Let $q'_2 = \gamma(t'_2)$, $q_2^* = \gamma(t_2^*)$. Then Lemma 3.6 of [A] shows that $T(p_1 p_2 q'_2) = T(p_1 q'_2 q_2^*)$. So we may replace the arc $p_1 q'_2 q_2^*$ of γ_2 by the arc $p_1 p_2 q'_2$, and obtain a new optimal trajectory γ_3 . This procedure can be continued until all the singular pieces to the right of $[t'_0, t'_1]$ have been eliminated. The resulting trajectory γ_m is still time-optimal. However, γ_m is a strict $Y * X * Z * Y$ trajectory, and so γ_m is of the type studied in Case 4. So γ_m cannot be time-optimal, as shown in the study of Case 4. This contradiction concludes the study of Case 5, which was the last missing case.

The proof of Lemma 3.12 is now complete. As explained before, we have proved the following.

THEOREM 3.17. *Let $p \in M$ be any point such that (3.I.i, ii, iii) hold. Then p has a neighborhood U such that $\text{Traj}(X \vee Y \vee Z)$ is boundedly sufficient for U . \square*

This result, together with Corollary 2.4, gives the following.

THEOREM 3.18. *Let Σ be a C.A.S. that has the DSAP. Let $p \in M$ be such that either $X(p) \neq 0$ or $Y(p) \neq 0$. Then there exists a neighborhood U of p such that $\text{Traj}(X \vee Y \vee Z)$ is boundedly sufficient for U . \square*

4. The degenerate cases. In this section, we prove the analogues of Theorem 3.17 for analytic systems that do not have the DSAP. First, we consider a C.A.S. that does not have the DAP, i.e., a system for which Δ_A vanishes identically.

LEMMA 4.1. *Suppose that Σ is a C.A.S. for which $\Delta_A \equiv 0$. Let $p \in M$ be such that either $X(p) \neq 0$ or $Y(p) \neq 0$. Then p has a neighborhood U such that $\text{Traj}(X \vee Y)$ is boundedly sufficient for U .*

Proof. Assume that $X(p) \neq 0$. (The case when $Y(p) \neq 0$ is similar.) Then we can choose a square coordinate chart $(U, (\xi, \eta))$, centered at p , by means of which U becomes identified with a square $S_q(\varepsilon)$, and having the property that

$$(4.1) \quad X = \partial_x.$$

Then $Y = \alpha \partial_x + \beta \partial_y$, for some analytic functions α, β on U . The hypothesis that $\Delta_A \equiv 0$ implies that $\beta \equiv 0$, so that

$$(4.2) \quad Y = \alpha \partial_x.$$

From this it is clear that every trajectory γ of $\Sigma \downarrow U$ is contained in a horizontal line. If γ is time-optimal, then clearly γ cannot go twice through the same point. Therefore γ either goes from right to left or from left to right. The former can only happen if $\alpha(\gamma(t)) < 0$ for all $t \in \text{Dom}(\gamma)$. Then the t -derivative of the x -coordinate of $\gamma(t)$ is a convex combination of $\alpha(\gamma(t))$ (which is < 0) and of 1. Clearly, if the coefficient of 1 is nonzero, then one could achieve greater speed leftwards by making it zero, so γ would not be time-optimal. Therefore, any time-optimal trajectory in U which goes from right to left is necessarily in $\text{Traj}(Y)$.

Now suppose that γ is a time-optimal trajectory in U that goes from left to right. It is clear that $\gamma \upharpoonright J \in \text{Traj}(X)$, if J is any interval such that $\alpha(\gamma(t)) < 1$ for $t \in J$, and that $\gamma \upharpoonright J \in \text{Traj}(Y)$ if $\alpha(\gamma(t)) > 1$ for $t \in J$. Therefore, γ is regular bang-bang, and it can only switch at points q such that $\alpha(q) = 1$. (This is because the set of zeros of $\alpha - 1$ on any horizontal segment S is discrete, unless $\alpha \equiv 1$ on S . But, in the latter case, if γ is contained in S , then γ is an X -trajectory.) The fact that $\gamma \in \text{Traj}(X \vee Y)^\nu$ for some fixed ν , independent of γ , now follows easily, if ε is small enough, from well-known properties of analytic functions. (For instance, take a CASA stratification of $\text{Sq}(\varepsilon)$ which is compatible with X and ∂_y , and with the set of zeros of $\alpha - 1$. As shown in Lemma 3.1, every one-dimensional stratum of \mathcal{S} is either a horizontal segment or a set of the form $\{(\psi(y), y) : a < y < b\}$ for some function ψ . So, if S is a horizontal segment in $\text{Sq}(\varepsilon)$, then S meets each one- or zero-dimensional stratum of \mathcal{S} at most once, unless $\alpha \equiv 1$ on S . If $0 < \varepsilon' < \varepsilon$, and if S is a horizontal segment in $\text{Sq}(\varepsilon')$ such that $\alpha - 1$ does not vanish identically on S , then $\alpha - 1$ cannot have more than ν zeros on S , where ν is the number of strata of \mathcal{S} that meet $\text{Sq}(\varepsilon')$ and have dimension zero or one.) \square

We now study the case of systems which have the DAP but fail to have the DSAP. It can be proved that, for such a system, there is a good *weakly sufficient* family of optimal trajectories in a neighborhood of any point p such that $X(p) \neq 0$ or $Y(p) \neq 0$. However, it turns out that the proof is simpler if we make the extra hypothesis that the accessibility property holds at p , and that this case is the only one which is actually needed for the proof of the existence of a regular synthesis. So we will limit ourselves to the simpler case.

The reason why we will only obtain weakly sufficient families, rather than sufficient families, is as follows. If the DSAP fails, then $\Delta_B \equiv 0$ (cf. Lemma 2.2). Then Lemma 3.10 of [A], together with Formula (3.43) of [A], imply that any two trajectories from a point p to a point q take the same time, at least if Δ_A never vanishes. So there are many optimal trajectories. In fact, in a neighborhood of a point where $\Delta_A \neq 0$, every trajectory will be optimal, and so there is no hope of proving that every optimal trajectory is of a particularly simple type. However, one can prove that, whenever a point q can be reached from a point p , then q can be reached from p by a particularly simple trajectory, which then turns out to be optimal.

THEOREM 4.2. *Let Σ be a C.A.S. such that $\Delta_B \equiv 0$. Let $p \in M$ be such that the AP holds at p . Then p has a neighborhood U with the property that $\text{Traj}(X \vee Y)$ is weakly boundedly sufficient for U (i.e. there exists an $N > 0$ such that, whenever $q_1 \in U$ can be reached from $q_0 \in U$ by means of a time-optimal trajectory in U , then q_1 can be reached from q_0 by means of a time-optimal $\gamma \in \text{Traj}(\Sigma \downarrow U)$ such that $\gamma \in [\text{Traj}(X \vee Y)]^N$.)*

Proof. First assume that $X(p)$ and $Y(p)$ are linearly independent. In this case we can use Lemma 3.12 of [A], and find a neighborhood U which is the domain of a square coordinate chart centered at p , relative to which X and Y have components $(\alpha, 0)$, $(0, \beta)$, where α, β are analytic on U , and > 0 . Then it is clear that, whenever a point $q_1 \in U$ can be reached from a point $q_0 \in U$ by means of a trajectory in U , then

q_1 can be reached from q_0 by means of a γ in $\text{Traj}(\Sigma \upharpoonright U) \cap \text{Traj}(X * Y)$. We now show that every trajectory in U is time-optimal. (This will establish that $\text{Traj}(X \vee Y)$ is weakly sufficient for U .)

A simple computation shows that

$$(4.3) \quad [X, Y] = -\beta(\partial_y \alpha) \partial_x + \alpha(\partial_x \beta) \partial_y.$$

Therefore

$$(4.4) \quad \Delta_B = \frac{1}{4}(-\alpha^2(\partial_x \beta) + \beta^2(\partial_y \alpha)).$$

Since $\Delta_B \equiv 0$, we get

$$(4.5) \quad \frac{\partial_x \beta}{\beta^2} = \frac{\partial_y \alpha}{\alpha^2}.$$

Therefore

$$(4.6) \quad \partial_x \left(\frac{1}{\beta} \right) = \partial_y \left(\frac{1}{\alpha} \right).$$

Since U is simply connected, there is an analytic function $\psi: U \rightarrow \mathbb{R}$ such that

$$(4.7) \quad \partial_x \psi = \frac{1}{\alpha}, \quad \partial_y \psi = \frac{1}{\beta}.$$

Therefore

$$(4.8) \quad \langle d\psi, X \rangle \equiv \langle d\psi, Y \rangle \equiv 1.$$

From this it follows easily that, if $\gamma \in \text{Traj}(\Sigma \upharpoonright U)$, and $\text{Dom}(\gamma) = [a, b]$, then

$$(4.9) \quad \int_{\gamma} \psi = b - a,$$

i.e.,

$$(4.10) \quad \psi(\gamma(b)) - \psi(\gamma(a)) = T(\gamma).$$

If $\gamma(a) = q_0$, $\gamma(b) = q_1$, we see that $T(\gamma) = \psi(q_1) - \psi(q_0)$, so that $T(\gamma)$ is independent of γ . Therefore all trajectories from q_0 to q_1 in U take exactly the same time.

We now consider the case when $X(p)$ and $Y(p)$ are linearly dependent. The accessibility hypothesis implies, first of all, that $X(p)$ and $Y(p)$ cannot both vanish. Assume, without loss of generality, that $X(p) \neq 0$. Choose a square coordinate chart, centered at p , with domain U_0 , by means of which U_0 is identified with a square $\text{Sq}(\varepsilon_0)$, and which is such that, on U_0 ,

$$(4.11) \quad X = \partial_x.$$

Let Y have components α, β . Then a simple computation shows that

$$(4.12) \quad \Delta_B = \frac{1}{4}((\alpha - 1)(\partial_x \beta) - (\partial_x \alpha)\beta).$$

So, $\Delta_B \equiv 0$ implies that

$$(4.13) \quad (\partial_x \alpha)\beta \equiv (\alpha - 1)(\partial_x \beta).$$

If $\alpha(0, 0) \neq 1$, then the function $\phi: x \rightarrow \beta(x, 0)$ is a solution of

$$(4.14) \quad \dot{\phi} = \phi\mu,$$

where

$$(4.15) \quad \mu(x) = \frac{(\partial_x \alpha)(x, 0)}{\alpha(x, 0) - 1}$$

for x in some interval $]-\delta, \delta[$. Since $X(p)$ and $Y(p)$ are dependent, we have $\beta(0, 0) = 0$, and so $\beta(x, 0) = 0$ for $|x| < \delta$. Therefore, by analyticity, $\beta(x, 0) = 0$ for $|x| < \varepsilon$. This shows that Y is everywhere tangent, on U_0 , to the integral curve of X through p , and therefore the accessibility property does not hold. This contradiction arose from assuming that $\alpha(0, 0) \neq 1$. So $\alpha(0, 0) = 1$. Since $\beta(0, 0) = 0$, we conclude that

$$(4.16) \quad X(p) = Y(p).$$

By making ε smaller we may assume that the accessibility property holds at every point of U_0 , and that $\alpha > 0$, throughout U_0 . Then the proof that $\beta(p) = 0$ implies $\alpha(p) = 1$ also works at all other points $q \in U_0$, and we conclude that

$$\alpha(q) = 1 \quad \text{whenever } q \in U_0, \quad \beta(q) = 0.$$

We now let \mathcal{S} be a CASA stratification of U_0 which is compatible with $\{p\}$, with the set of zeros of $\beta \upharpoonright U_0$, and with the vector fields X , Y , and ∂_y . Let $0 < \varepsilon < \varepsilon_0$ be such that ε is good for \mathcal{S} (cf. the definition of a "good" ε , before Lemma 3.3). We now let $\mathcal{S}_1(\varepsilon)$ be the set of all intersections $S \cap \text{Sq}(\varepsilon)$, as S ranges over all one-dimensional strata of \mathcal{S} that meet $\text{Sq}(\varepsilon)$. Then $\mathcal{S}_1(\varepsilon)$ is finite, and the zero set of $\beta \upharpoonright \text{Sq}(\varepsilon)$ is exactly the union of $\{p\}$ and of the members of $\mathcal{S}_1(\varepsilon)$. Lemma 3.4 implies that, if $S \in \mathcal{S}_1(\varepsilon)$, then $S \in \mathcal{A}_p^+(\varepsilon) \cup \mathcal{A}_p^-(\varepsilon)$. (The remaining possibility, that $S = \Gamma_\varepsilon^+$ or $S = \Gamma_\varepsilon^-$, cannot arise. Indeed, if $S = \Gamma_\varepsilon^+$ or $S = \Gamma_\varepsilon^-$, it would follow that $\beta(x, 0) = 0$ for $|x| < \varepsilon$, contradicting the accessibility assumption.)

It is easy to see that the open set $\{q: q \in \text{Sq}(\varepsilon), \beta(q) \neq 0\}$ is partitioned into a finite number W_1, \dots, W_m of connected components, which have the following properties:

- (4.I.1) W_i is simply connected;
- (4.I.2) W_i is a union of horizontal segments;
- (4.I.3) The right boundary $\partial_R W_i$ of W_i (i.e. the set of all right endpoints of all the segments whose union is W_i) is either
 - (a) a set $S \in \mathcal{S}_1(\varepsilon)$, of the form $\{(\psi(y), y): 0 < y < \varepsilon\}$, or
 - (b) a union $S_1 \cup S_2$, where $S_1 = \{(\psi(y), y): 0 < y < b\}$, $b < \varepsilon$, and $\psi(b-) = \varepsilon$, and $S_2 = \{(\varepsilon, y): b \leq y < \varepsilon\}$, or
 - (c) a set $S \in \mathcal{S}_1(\varepsilon)$ of the form $\{(\psi(y), y): -\varepsilon < y < 0\}$ or
 - (d) a union $S_1 \cup S_2$, where $S_1 = \{(\psi(y), y): a < y < 0\}$, $-\varepsilon < a$, and $\psi(a+) = \varepsilon$, and $S_2 = \{(\varepsilon, y): -\varepsilon < y \leq a\}$, or
 - (e) the union of $\{p\}$, of a set of the type described in (a) or (b), and of a set of the type described in (c) or (d). In all cases, ψ is a real analytic function which is either constant or strictly monotonic.
- (4.I.4) $\beta \neq 0$ throughout W_i .

The *left boundary* of W_i is defined similarly, and it satisfies a property similar to (4.I.3), which we shall not state explicitly.

Let $B_i^* = (\partial_R W_i) \cap \text{Sq}(\varepsilon)$. If $\partial_R W_i$ satisfies (a) or (b) of (4.I.3), let B_i be the union of B_i^* and of the vertical segment $\{(0, y): -\varepsilon < y \leq 0\}$. If $\partial_R W_i$ satisfies (c) or (d), let B_i^* be the union of B_i and of the vertical segment $\{(0, y): 0 < y < \varepsilon\}$. Finally, if $\partial_R W_i$ satisfies (e), let $B_i = B_i^*$. Then B_i is a barrier in $\text{Sq}(\varepsilon)$. If $\gamma \in \text{Traj}(\Sigma \upharpoonright \text{Sq}(\varepsilon))$, and if γ contains some point in W_i , but eventually leaves W_i , then it must leave through a

point in B_i . Therefore γ never re-enters W_i . So every trajectory of $\Sigma \upharpoonright \text{Sq}(\varepsilon)$ is a concatenation of at most m pieces, each of which is contained in one of the sets $\text{Clos } W_i$. Therefore, our conclusion will be proved if we show that, whenever γ is a time-optimal trajectory which is contained in $\text{Clos } W_i$ for some i , and goes from q_0 to q_1 , then there is a time-optimal $\gamma' \in \text{Traj}(\Sigma \upharpoonright \text{Clos } W_i)$, that goes from q_0 to q_1 , and is in $\text{Traj}(X * Y)$. We will establish this by proving that:

(4.II.a) Whenever γ_1, γ_2 are trajectories in $\text{Clos } W_i$, such that $\text{In}(\gamma_1) = \text{In}(\gamma_2)$ and $\text{Term}(\gamma_1) = \text{Term}(\gamma_2)$, then $T(\gamma_1) = T(\gamma_2)$

and

(4.II.b) Whenever $q_1 \in \text{Clos } W_i$ can be reached from $q_0 \in \text{Clos } W_i$ by means of a trajectory in $\text{Clos } W_i$, then q_1 can be reached from q_0 by means of a $\gamma \in \text{Traj}(\Sigma \upharpoonright U) \cap \text{Traj}(X * Y)$.

To prove (4.II.a), it is clearly sufficient to assume that γ_1 and γ_2 are entirely contained in W_i . Since β never vanishes on W_i , we can define an analytic function ψ and W_i by

$$(4.17) \quad \psi = \frac{1 - \alpha}{\beta}.$$

Then

$$(4.18) \quad \partial_x \psi = \frac{-(\partial_x \alpha)\beta - (1 - \alpha)(\partial_x \beta)}{\beta^2}.$$

So (4.13) implies that $\partial_x \psi \equiv 0$ on W_i . Since $\partial_y 1 \equiv 0$, and W_i is simply connected, there exists an analytic $\phi: W_i \rightarrow \mathbb{R}$ such that

$$\partial_x \phi = 1, \quad \partial_y \phi = \psi.$$

Therefore

$$\langle d\phi, X \rangle = \partial_x \phi \equiv 1$$

and

$$\langle d\phi, Y \rangle = \alpha(\partial_x \phi) + \beta(\partial_y \phi) \equiv 1.$$

Then, if γ is any trajectory of $\Sigma \upharpoonright W_i$, we have

$$T(\gamma) = \int_{\gamma} d\phi = \phi(\text{Term}(\gamma)) - \phi(\text{In}(\gamma)).$$

So $T(\gamma)$ only depends on the endpoints of γ , and (4.II.a) is proved.

We now prove (4.II.b). Assume that $\beta > 0$ on W_i . (The case when $\beta < 0$ on W_i is identical.) Then the Y -trajectories in $\text{Clos } W_i$ go to the right and up. Suppose that $\gamma: [a, b] \rightarrow \text{Clos } W_i$ is a trajectory of Σ , such that $\gamma(a) = q_0$, $\gamma(b) = q_1$. Then q_1 is to the right of and above q_0 . Let $\gamma'(t) = \Phi_t^Y(q_0)$. Then there are t_1, t_2 such that $\gamma'(t_1) \in \partial_L W_i$, $\gamma'(t_2) \in \partial_R W_i$, and $t_1 \leq 0 < t_2$. Since $\gamma' \upharpoonright]t_1, t_2[$ goes up and to the right, and $X = \partial_x$, the curve $\gamma' \upharpoonright]t_1, t_2[$ is a barrier in W_i . The connected components of $W_i - \gamma'$ are W_i^L, W_i^R , where W_i^R is the union of all the segments $(\mathbb{R} \times \{y\}) \cap W_i$, for $y \leq \bar{y}$ (where \bar{y} is the y -coordinate of q_0), and of all the segments $(]x(t), \infty[\times \{y(t)\}) \cap W_i$, where $\gamma'(t) = (x(t), y(t))$, $t \in]t_1, t_2[$. Since $\gamma' \upharpoonright]t_1, t_2[$ is a barrier in W_i , and q_1 is reachable from q_0 in $\text{Clos } W_i$, it follows that $q_1 \in \text{Clos}(W_i^R)$. Since q_1 has a larger y -coordinate than q_0 , we necessarily have $q_1 \in (]x(t), \infty[\times \{y(t)\}) \cap W_i$ for some t , and so q_1 is reachable from q_0 by an $X * Y$ -trajectory in $\text{Clos } W_i$. \square

5. Conclusion. If we combine the results of Theorem 3.18, Lemma 4.1 and Theorem 4.2, we obtain the following conclusions (for which the connectedness of Σ is no longer needed):

THEOREM 5.1. *Let Σ be a real analytic system*

$$(5.1) \quad \dot{x} = f(x) + ug(x), \quad |u| \leq 1$$

on a two-dimensional real-analytic manifold M . Let $p \in M$ be such that at least one of the vectors $f(p)$, $g(p)$ does not vanish. Assume that either (i) Σ has the DSAP, or (ii) Σ has the AP at p , or (iii) Δ_A vanishes identically near 0. Let $X = f - g$, $Y = f + g$, and let Z denote the singular vector field. Then p has a neighborhood U such that, for the minimum time problem, $\text{Traj}(X \vee Y \vee Z)$ is boundedly sufficient for U .

THEOREM 5.2. *Let Σ , M , f , g , X , Y , Z be as in Theorem 5.1. Let $\gamma: [a, b] \rightarrow M$ be a time-optimal trajectory of Σ . Then there exists a time-optimal trajectory $\gamma': [a, b] \rightarrow M$ of Σ such that $\gamma'(a) = \gamma(a)$, $\gamma'(b) = \gamma(b)$, and γ' is a finite concatenation of X -, Y - and Z -trajectories.*

Proof. If γ goes through a point p where both f and g vanish, then necessarily $b = a$ and we can take $\gamma' = \gamma$. Otherwise γ is a finite concatenation of trajectories γ_i that are contained in open sets U_i for which $\text{Traj}(X \vee Y \vee Z)$ is boundedly sufficient, and the desired conclusion follows. \square

Appendix.

A1. Semianalytic and subanalytic sets. We list here the main facts about semianalytic and subanalytic sets. For details, see Hardt [Ha], Lojasiewicz [Loj], Sussmann [Su6], and Tamm [Ta].

A subset S of a C^ω manifold M is *analytic* if every $p \in M$ has a neighborhood U such that $S \cap U$ is the set of zeros of a C^ω function $f: U \rightarrow \mathbb{R}$. We say that S is *semianalytic* if every $p \in M$ has a neighborhood U such that

$$S \cap U = \bigcup_{j=1}^m \bigcup_{i=1}^{n(j)} S_{ij}$$

where each S_{ij} is a subset of U of the form $\{x: f_{ij}(x) = 0\}$ or $\{x: f_{ij}(x) > 0\}$, and $f_{ij}: U \rightarrow \mathbb{R}$ is real analytic.

We shall not repeat the actual definition of the class of subanalytic sets (cf. [Ha], where they are called “semianalytic shadows”, or [Hi1], or [Ta], or [Su]). The main facts are:

- (A.I) Every semianalytic subset of M is subanalytic;
- (A.II) Every locally finite union or intersection of subanalytic sets is subanalytic;
- (A.III) The complement and the closure of a subanalytic set is subanalytic.

Moreover, subanalyticity is local, i.e.:

- (A.IV) If $S \subseteq M$, then S is a subanalytic subset of M if and only if every $p \in M$ has a neighborhood U such that $S \cap U$ is subanalytic in U .

Also:

- (A.V) If $f: M \rightarrow N$ is a C^ω map, and S is subanalytic in N , then $f^{-1}(S)$ is subanalytic in M .

Finally, we have the most important property of all, namely, that one can take proper images. Precisely, if $f: M \rightarrow N$ is a map, and $A \subseteq M$, we say that f is *proper on A* if $f^{-1}(K) \cap A$ is compact for each compact $K \subseteq N$. Then we have:

(A.VI) If $f: M \rightarrow N$ is analytic, S is a subanalytic subset of M , and f is proper on $\text{Clos } S$, then $f(S)$ is a subanalytic subset of N .

The preceding properties make it possible to prove easily that many sets are subanalytic. Consider formulas $F(x_1, \dots, x_n)$ with free variables x_1, \dots, x_n ranging over analytic manifolds M_1, \dots, M_n . If F_1, \dots, F_m are formulas which define subanalytic sets, then any formula obtained from them by taking conjunctions, disjunctions and negations also has this property. Moreover, we can also allow existential quantifications, provided that they are bounded. (In $(\exists x)F(x, y)$, we say that the quantifier $(\exists x)$ is bounded if for every compact $K \subseteq N$ there is a compact $J \subseteq M$ such that, for each $y \in K$, $(\exists x)F(x, y)$ is equivalent to $(\exists x)(x \in J \wedge F(x, y))$.) Finally, since a universal quantifier can be expressed in terms of existential quantifiers and negations, we can also allow universal quantifiers, provided that they are bounded. (The definition of a bounded universal quantifier \forall is the same as that for \exists , except that $(\exists x)(x \in J \wedge F(x, y))$ must be replaced by $(\forall x)(x \in J \Rightarrow F(x, y))$.)

A C^k stratification of a manifold M is a locally finite partition \mathcal{S} of M into connected, embedded submanifolds of M , of class C^k , such that, if $S \in \mathcal{S}$, then the frontier $\text{Fron } S (= (\text{Clos } S) - S)$ is a union of members of \mathcal{S} , all of them of dimension smaller than $\dim S$. A CASA stratification of M is a stratification whose members are analytic submanifolds and subanalytic sets. A stratification \mathcal{S} is *compatible* with a set A if A is a union of members of \mathcal{S} . If \mathcal{A} is a family of sets, then \mathcal{S} is *compatible with* \mathcal{A} if it is compatible with each $A \in \mathcal{A}$. If X is an analytic vector field on M , we say that \mathcal{S} is *compatible with* X if, for each $S \in \mathcal{S}$, X is either everywhere tangent to S or nowhere tangent to S .

(A.VII) If \mathcal{A} is an arbitrary locally finite family of subanalytic subsets of M , and \mathcal{F} a finite family of analytic vector fields on M , then there is a CASA stratification \mathcal{S} of M which is compatible with all the $A \in \mathcal{A}$ and all the $X \in \mathcal{F}$.

If $f: M \rightarrow N$ is a map, L is a subset of M , and \mathcal{S}, \mathcal{T} are stratifications of M, N , we say that $(\mathcal{S}, \mathcal{T})$ is *compatible with f over L* if \mathcal{S} is compatible with L and, for every $S \in \mathcal{S}$, $S \subseteq L$, the image $f(S)$ is in \mathcal{T} , and $f|_S: S \rightarrow f(S)$ is a submersion. We say that $(\mathcal{S}, \mathcal{T})$ is *one-one compatible with f over L* if, in addition, the map $f|_S$ is one-to-one for every $S \in \mathcal{S}$, $S \subseteq L$, such that $\dim S = \dim f(S)$.

(A.VIII) If $f: M \rightarrow N$ is an analytic map, L is a subanalytic subset of M such that f is proper on $\text{Clos } L$, and \mathcal{A}, \mathcal{B} are locally finite families of subanalytic subsets of M, N , respectively, then there exist CASA stratifications \mathcal{S}, \mathcal{T} of M, N , which are compatible with \mathcal{A}, \mathcal{B} , and are such that $(\mathcal{S}, \mathcal{T})$ is one-one compatible with f over L .

It follows from (A.VII) that, if A is a subanalytic subset of M , then A is a union of members S of a stratification \mathcal{S} . In particular, if A is relatively compact, then A only meets finitely many members of \mathcal{S} , and therefore:

(A.IX) A relatively compact subanalytic subset of M has finitely many components.

Finally, we quote a fact from [Loj]:

(A.X) A subanalytic subset of the plane is necessarily semianalytic.

A2. A counterexample. We show that, in the neighborhood of a point p where X and Y both vanish, there need not be bounds on the number of switchings. Let X be an analytic vector field whose trajectories spiral about 0, and converge to 0 as $t \rightarrow \infty$. Let $Y = \phi X$, where ϕ is a positive function which is < 1 above the x axis, and > 1 below the x axis. If U is an arbitrary neighborhood of 0, one can pick $q_1 \in U$, $q_2 \in U$ such that $q_2 = \Phi_t^X(q_1)$ for a $t > 0$, and that the number of times that the X -trajectory γ from q_1 to q_2 winds around the origin is arbitrarily large. To make γ time-optimal, we must reparametrize γ , so that γ is an X -trajectory above the x -axis, but a Y -trajectory below the x -axis. So U contains time-optimal trajectories with an arbitrarily large number of switchings.

With a little bit of work, one can modify the preceding example so as to get X and Y to have the strong accessibility property everywhere, except at 0.

In these examples, the unboundedness in the number of switchings arises as the time becomes unbounded. We do not know whether there are examples where the number of switchings is unbounded even when the time remains bounded.

REFERENCES

- [A] H. J. SUSSMANN, *The structure of time-optimal trajectories for single-input systems in the plane: the C^∞ nonsingular case*, this Journal, 25 (1987), pp. 433–465.
- [Ba] M. BAYTMAN, *The Optimal Synthesis of Trajectories In The Plane*, Zinatne, Riga (USSR), 1971. (In Russian.)
- [Ha] R. M. HARDT, *Stratification of real analytic maps and images*, Invent. Math., 28 (1975), pp. 193–208.
- [Loj] S. LOJASIEWICZ, *Ensembles semi-analytiques*, Lecture notes at I.H.E.S., Bures-sur-Yvette, 1965.
- [Su] H. J. SUSSMANN, *Subanalytic sets and regular synthesis*, to appear.
- [SuJ] H. J. SUSSMANN AND V. JURDJEVIC, *Controllability of nonlinear systems*, J. Differential Equations, 12 (1972), pp. 95–116.
- [Ta] M. TAMM, *Subanalytic sets in the calculus of variation*, Acta Math., 146 (1981), pp. 167–199.
- [Hi1] H. HIRONAKA, *Subanalytic sets*, in Number Theory, Algebraic Geometry and Commutative Algebra, in Honor of Y. Akizuki, Kinokuniya, Tokyo, 1973.

A NONQUADRATIC BOLZA PROBLEM AND A QUASI-RICCATI EQUATION FOR DISTRIBUTED PARAMETER SYSTEMS*

YOU YUN-CHENG†

Abstract. Closed-loop optimal control of a nonquadratic Bolza problem for linear distributed parameter systems and normal solution of an associated quasi-Riccati operator equation are studied by the approach of a nonlinear integral equation.

Key words. Bolza problem, nonquadratic cost function, quasi-Riccati equation, closed-loop optimal control, nonlinear integral equation, distributed parameter system

AMS(MOS) subject classification. 49A27

1. Introduction. In this paper, we consider an optimal control problem for a linear evolution system,

$$(1.1) \quad \frac{dx}{dt} = Ax(t) + Bu(t), \quad x(0) = x_0,$$

$$(1.2) \quad \min_{u \in L^2(0, T; U)} \left\{ J(u) = M(x(T)) + \int_0^T (Q(x(t)) + \frac{1}{2} \langle Ru(t), u(t) \rangle) dt \right\}.$$

Assume that $T > 0$ is finite and fixed, X and U are real Hilbert spaces, $x(t)$ and x_0 take values in X , and the admissible set of control functions is $\mathcal{U} = L^2(0, T; U)$. In (1.1), $A: D(A) (\subset X) \rightarrow X$ is an infinitesimal generator of C_0 -semigroup of bounded linear operators $e^{At} (t \geq 0) \in \mathcal{L}(X)$, and $B \in \mathcal{L}(U; X)$. In (1.2), we assume that

$$(1.3) \quad \begin{aligned} &M(\cdot) \text{ and } Q(\cdot) \text{ are } C^2 \text{ convex mappings from } X \text{ to } \mathbb{R} \\ &\quad \text{(nonquadratic in general),} \\ &R \in \mathcal{L}(U) \text{ is self-adjoint and coercively positive.} \end{aligned}$$

We take the mild solution of (1.1) to be state function, i.e.,

$$(1.4) \quad x(t) = e^{At}x_0 + \int_0^t e^{A(t-s)}Bu(s) ds, \quad t \in [0, T],$$

where the integration is in the Bochner sense.

This optimal control problem will be referred to as (NQBP). The results in this paper generalize the well-known quadratic optimal control [1].

In [2] and [3], the author proved the closed-loop optimal control and global normal solution of the associated quasi-Riccati equations in the cases $Q(\cdot) = 0$ and $M(\cdot) = 0$ respectively. In these two cases, the optimal nonlinear feedbacks are given by

$$(1.5) \quad u(t) = -R^{-1}B^*e^{A^*(T-t)}M'(H(T-t, x(t))) \quad (\text{see [2]}),$$

$$(1.6) \quad u(t) = -R^{-1}B^* \int_t^T e^{A^*(s-t)}Q'(G(s, t, x(t))) ds \quad (\text{see [3]})$$

* Received by the editors July 29, 1985; accepted for publication (in revised form) April 25, 1986.

† Institute of Mathematics, Fudan University, Shanghai, People's Republic of China. Present address, Center for Control Science and Dynamical Systems, University of Minnesota, Minneapolis, Minnesota 55455.

respectively, where $H(T-t, x)$ and $G(s, t, x)$ are solution operators of following nonlinear algebraic equation [2]

$$(1.7) \quad y + \left\{ \int_0^{T-t} e^{As} B R^{-1} B^* e^{A^*s} ds \right\} M'(y) = e^{A(T-t)} x$$

for any given $(t, x) \in [0, T] \times X$,

and nonlinear Fredholm integral equation

$$(1.8) \quad y(s) + \int_t^T \left\{ \int_t^{\min(s, \sigma)} e^{A(s-\eta)} B R^{-1} B^* e^{A^*(\sigma-\eta)} d\eta \right\} Q'(y(\sigma)) d\sigma = e^{A(s-t)} x,$$

$s \in [t, T]$ for any given $(t, x) \in [0, T] \times X$,

respectively. However, these two approaches are not applicable to the more general case of (NQBP) we consider here.

In [4], more general convex control problems for linear evolutionary processes had been studied. By means of subdifferentials and adjoint equations, those authors established a set of optimality conditions which amounts to the open-loop relations of the optimal control process.

In connection with closed-loop syntheses of the related problems, [5] developed a series of results on local and global existence of solution to Hamilton-Jacobi equations mainly by the constructive approximation approach.

Here in this paper, we consider the $\{x(T), x(\cdot), u(\cdot)\}$ -separate nonquadratic criteria with the C^2 -assumption. The obtained results will provide

- (i) A simpler open-loop formula of the optimal control;
- (ii) A new proof of the global existence of the solution of the quasi-Riccati equation as well as the improved regularity of the solution;
- (iii) A new variational formula, which directly leads to the closed-loop relation;
- (iv) A class of nonlinear integral equations whose solution mappings yield both the optimal feedback operators and normal solutions of the quasi-Riccati equations;
- (v) A generalization to closed-loop optimal strategies of nonquadratic differential games.

The contents of this paper are outlined as follows. In § 2, the existence of the optimal control and an open-loop relation are proved. In § 3, we show that a normal solution of quasi-Riccati equation will provide optimal feedback for the closed-loop solution. Section 4 is devoted to the investigation of solution mapping $K(s, t, x)$ of a nonlinear integral equation. We prove in § 5 that $K(t, t, x) = P(t, x)$ turns out to be a normal solution of quasi-Riccati equation. Hence the closed-loop solution to (NQBP) is given. In § 6 we give some remarks.

Here we list some notations. We shall denote a scalar product by $\langle \cdot, \cdot \rangle$. Let E , E_1 and E_2 be Banach spaces; then $\mathcal{L}(E_1; E_2)$ and $\mathcal{L}(E)$ stand for the Banach spaces of bounded linear operators from E_1 to E_2 and from E to itself respectively. Superscript $*$ will be attached to adjoint operators. $e^{A^*t} \in \mathcal{L}(X)$ is the dual semigroup of e^{At} and is generated by A^* , the adjoint of the densely defined and closed operator A . For a self-adjoint operator S , then $S \geq 0$ (resp. $S > 0$) means nonnegative (resp. coercively positive).

If E is a Banach space, $[a, b] \subset \mathbb{R}$, then we shall denote by $L^2(a, b; E)$ the Hilbert space of strongly measurable functions $g(\cdot): [a, b] \rightarrow X$, such that

$$\int_a^b \|g(t)\|_E^2 dt < +\infty.$$

We shall denote by $C([a, b]; E)$ the Banach space of all strongly continuous functions from $[a, b]$ to E .

For a mapping f from a Banach space E_1 to another E_2 , we shall denote its Fréchet derivative by f' or Df . Subscripts will be used to represent partial Fréchet derivatives or related parameters of mappings according to the context.

All the concepts and facts of nonlinear analysis used later appear in [6]. In particular, we shall make use of the following relation of a differentiable mapping f , i.e.,

$$(1.9) \quad f(x+h) - f(x) = \int_0^1 Df(x+\lambda h)h \, d\lambda.$$

2. Existence and open-loop equality.

THEOREM 1. *For any given $x_0 \in X$, there exists a unique optimal control of (NQBP).*

Proof. First we assert that for a C^1 convex mapping $\varphi(\cdot)$ from a Hilbert space E to \mathbb{R} there exist an element $e \in E$ and a constant $\gamma \in \mathbb{R}$ such that

$$(2.1) \quad \varphi(x) \geq \langle e, x \rangle_E + \gamma \quad \forall x \in E.$$

This is a consequence of the monotonicity of $\varphi'(\cdot)$ and (1.9).

From (1.3) and (2.1), there must be constants $\delta > 0$, $\beta > 0$, and $\alpha(x_0) \in \mathbb{R}$ which depends continuously on x_0 , such that

$$(2.2) \quad \begin{aligned} \delta \|u\|_{\mathcal{U}}^2 &\leq \frac{1}{2} \int_0^T \langle Ru(t), u(t) \rangle \, dt \\ &= J(u; x_0) - M \left(e^{AT} x_0 + \int_0^T e^{A(T-s)} Bu(s) \, ds \right) \\ &\quad - \int_0^T Q \left(e^{At} x_0 + \int_0^t e^{A(t-s)} Bu(s) \, ds \right) \, dt \\ &\leq J(u; x_0) + \alpha(x_0) + \beta \|u\|_{\mathcal{U}}, \end{aligned}$$

where we write (1.2) as $J(u; x_0)$ to indicate its dependence on $u(\cdot)$ and x_0 . Thus we have

$$(2.3) \quad 0 \leq \delta \left(\|u\|_{\mathcal{U}} - \frac{\beta}{2\delta} \right)^2 \leq J(u; x_0) + \left(\alpha(x_0) + \frac{\beta^2}{4\delta} \right) \quad \forall u(\cdot) \in \mathcal{U}.$$

It follows from (2.3) that for a given $x_0 \in X$,

$$(2.4) \quad -\infty < J(u; x_0) < +\infty,$$

and that a minimizing sequence is uniformly bounded in \mathcal{U} so that there is a weakly convergent subsequence in \mathcal{U} .

From (1.3) we can deduce that for any $0 \leq \lambda \leq 1$, $u(\cdot)$ and $v(\cdot)$ in \mathcal{U} ,

$$(2.5) \quad \lambda J(u) + (1-\lambda)J(v) \geq J(\lambda u + (1-\lambda)v) + \frac{1}{2}\lambda(1-\lambda) \int_0^T \langle R(u-v), u-v \rangle \, dt.$$

Hence, $J(u)$ is a strictly convex and (easy to see) strongly continuous functional on \mathcal{U} , so that it is weakly lower semi-continuous.

The above facts imply the existence and uniqueness of optimal control of (NQBP). QED.

COROLLARY 1. Let $t_0 \in [0, T]$, the following optimal control problem $(\text{NQBP})_{t_0}$:

$$(2.6) \quad x(t) = e^{A(t-t_0)}x_0 + \int_{t_0}^t e^{A(t-s)}Bu(s) ds,$$

$$(2.7) \quad \min_{u \in L^2(t_0, T; U)} \left\{ J_{t_0}(u) = M(x(T)) + \int_{t_0}^T (Q(x(t)) + \frac{1}{2}\langle Ru(t), u(t) \rangle) dt \right\}$$

has a unique optimal control for any given $x_0 \in X$.

COROLLARY 2. Let Σ be an arbitrarily fixed bounded subset in X , and $u_{(t_0, x_0)}(\cdot)$ be the optimal control of $(\text{NQBP})_{t_0}$ corresponding to the initial state $x_0 \in \Sigma$. Then,

$$1^\circ) \quad \sup_{(t_0, x_0) \in [0, T] \times \Sigma} \{J_{(t_0, x_0)}^* = \inf_{u \in L^2(t_0, T; U)} J_{t_0}(u; x_0)\} < +\infty,$$

$$2^\circ) \quad \sup_{(t_0, x_0) \in [0, T] \times \Sigma} \|u_{(t_0, x_0)}(\cdot)\|_{L^2(t_0, T; U)} < +\infty.$$

Proof. Similar to (2.3), with appropriate choice of constants $\tilde{\alpha}(x_0)$ only depending continuously on x_0 , and $\tilde{\beta}$, we have the following inequality:

$$(2.8) \quad 0 \leq \delta \left(\|u\|_{L^2(t_0, T; U)} - \frac{\tilde{\beta}}{2\delta} \right)^2 \leq J_{t_0}(u; x_0) + \left(\tilde{\alpha}(x_0) + \frac{\tilde{\beta}^2}{4\delta} \right) \quad \forall u \in L^2(t_0, T; U), t_0 \in [0, T], x_0 \in X.$$

Obviously,

$$(2.9) \quad J_{(t_0, x_0)}^* \leq M(e^{A(T-t_0)}x_0) + \int_{t_0}^T Q(e^{A(t-t_0)}x_0) dt \leq \text{const}(\Sigma).$$

From (2.8) and the continuous dependence of $\tilde{\alpha}(x_0)$ on $x_0 \in \Sigma$, we have

$$(2.10) \quad J_{(t_0, x_0)}^* \geq - \left(\tilde{\alpha}(x_0) + \frac{\tilde{\beta}^2}{4\delta} \right) \geq \text{const}(\Sigma).$$

Hence 1° is valid. Moreover 1° and (2.8) show that 2° is valid. QED.

THEOREM 2. For any given $t_0 \in [0, T]$ and $x_0 \in X$, if $u(\cdot)$ is the optimal control and $x(\cdot)$ is the corresponding optimal trajectory of $(\text{NQBP})_{t_0}$, then the following open-loop equality must be satisfied:

$$(2.11) \quad u(t) = -R^{-1}B^*(e^{A^*(T-t)}M'(x(T)) + \int_t^T e^{A^*(\sigma-t)}Q'(x(\sigma)) d\sigma),$$

$$(2.12) \quad x(t) = e^{A(t-t_0)}x_0 - \int_{t_0}^t e^{A(t-s)}BR^{-1}B^* \cdot \left(e^{A^*(T-s)}M'(x(T)) + \int_s^T e^{A^*(\sigma-s)}Q'(x(\sigma)) d\sigma \right) ds, \quad t \in [t_0, T].$$

Proof. For each fixed $v(\cdot) \in \mathcal{U}$, let $\psi(\lambda; v) = J_{t_0}(u + \lambda v; x_0)$, $\lambda \in \mathbb{R}$. As $\psi(0; v) = \min_{\lambda \in \mathbb{R}} \psi(\lambda; v)$, it must be $\psi'_\lambda(0; v) = 0$, $\forall v(\cdot) \in \mathcal{U}$. We can write (2.6) as

$$(2.13) \quad x(\cdot) = h_{t_0}(\cdot) + (\Gamma_{t_0}u)(\cdot), \quad x(T) = h_{t_0}(T) + \Lambda_{t_0}u,$$

where $h_{t_0}(t) = e^{A(t-t_0)}x_0$, $\Gamma_{t_0} \in \mathcal{L}(L^2(t_0, T; U); L^2(t_0, T; X))$ and $\Lambda_{t_0} \in \mathcal{L}(L^2(t_0, T; U); X)$ are defined to be

$$(2.14) \quad \begin{aligned} (\Gamma_{t_0}v)(t) &= \int_{t_0}^t e^{A(t-s)}Bv(s) ds, \quad v \in L^2(t_0, T; U), \\ \Lambda_{t_0}v &= \int_{t_0}^T e^{A(T-s)}Bu(s) ds, \quad v \in L^2(t_0, T; U). \end{aligned}$$

Substitute (2.13) into (2.7) for $u + \lambda v$; it follows that

$$\begin{aligned}
 \psi'_\lambda(0; v) &= \langle M'(h_0(T) + \Lambda_0 u), \Lambda_0 v \rangle_X + \langle Q'(h_0 + \Gamma_0 u), \Gamma_0 v \rangle_{L^2(t_0, T; X)} \\
 &\quad + \langle Ru, v \rangle_{L^2(t_0, T; U)} \\
 (2.15) \quad &= \langle Ru + \Lambda_0^* M'(h_0(T) + \Lambda_0 u) + \Gamma_0^* Q'(h_0 + \Gamma_0 u), v \rangle_{L^2(t_0, T; U)} = 0 \\
 &\quad \forall v(\cdot) \in L^2(t_0, T; U).
 \end{aligned}$$

Thus we obtain (2.11):

$$\begin{aligned}
 u(t) &= -R^{-1} \{ \Lambda_0^* M'(h_0(T) + \Lambda_0 u) + \Gamma_0^* Q'(h_0 + \Gamma_0 u) \} \\
 &= -R^{-1} B^* (e^{A^*(T-t)} M'(x(T)) + \int_t^T e^{A^*(\sigma-t)} Q'(x(\sigma)) d\sigma), \quad t \in [t_0, T].
 \end{aligned}$$

Substitution of (2.11) into (2.6) leads to (2.12). QED.

3. Quasi-Riccati operator equation and normal solution. We consider a quasi-Riccati operator equation associated with (NQBP),

$$\begin{aligned}
 (3.1) \quad &\frac{d}{dt} \langle P(t, x), y \rangle + \langle P_x(t, x) Ax, y \rangle + \langle P(t, x), Ay \rangle + \langle Q'(x), y \rangle \\
 &- \langle P_x(t, x) BR^{-1} B^* P(t, x), y \rangle = 0, \quad (t, x, y) \in [0, T] \times D(A) \times D(A), \\
 &P(T, x) = M'(x), \quad x \in X.
 \end{aligned}$$

DEFINITION 1. If a nonlinear mapping $P(t, x): [0, T] \times X \rightarrow X$ satisfies the following conditions, then it is called a normal solution of the quasi-Riccati equation (3.1),

- 1° $P(t, x): [0, T] \times X \rightarrow X$ is strongly continuous in (t, x) ;
- 2° $\langle P(t, x), y \rangle$ is continuously differentiable in $t \in [0, T]$, for each x and y in $D(A)$;
- 3° $P(t, x)$ is Fréchet differentiable in $x \in X$, for each $t \in [0, T]$;
 - (i) $P_x(t, x): [0, T] \times X \rightarrow \mathcal{L}(X)$ is strongly continuous in (t, x) , i.e., $(t, x) \rightarrow (\hat{t}, \hat{x})$ in $[0, T] \times X$ implies $P_x(t, x)\xi \rightarrow P_x(\hat{t}, \hat{x})\xi$ in X , $\forall \xi \in X$;
 - (ii) $P_x(t, x)$ is bounded in $\mathcal{L}(x)$ -norm for (t, x) in any given bounded subset of $[0, T] \times X$;
- 4° $P(t, x)$ satisfies (3.1);
- 5° $P(t, \cdot): X \rightarrow X$ is a gradient operator, for each $t \in [0, T]$;
- 6° Cauchy problem

$$(3.2) \quad \frac{dx}{dt} = Ax - BR^{-1} B^* P(t, x), \quad x(0) = x_0$$

has a global mild solution [7] $x(\cdot) \in C([0, T]; X)$ for each given $x_0 \in X$.

Assume that $P(t, x)$ is a normal solution of (3.1). By definition of gradient operators [6], for each $t \in [0, T]$, there are anti-derivatives $\Phi(t, x): x \rightarrow \mathbb{R}$ such that

$$(3.3) \quad \Phi_x(t, x) = P(t, x), \quad x \in X.$$

Those $\Phi(t, x)$ satisfying (3.3) may well be different from each other by a constant $c(t)$. We set, without loss of generality, a definition as follows.

DEFINITION 2. The anti-derivative $\Phi(t, x)$ of a normal solution $P(t, x)$ of (3.1) is called canonical, if $\Phi(t, x)$ is such that

$$(3.4) \quad \Phi(t, 0) = M(0), \quad 0 \leq t \leq T.$$

LEMMA 1. Assume that $P(t, x)$ is a normal solution of (3.1) and $\Phi(t, x)$ is the canonical anti-derivative of $P(t, x)$. Then, for each $\{x_0, u(\cdot)\} \in D(A) \times C^1([0, T]; U)$ and corresponding trajectory $x(\cdot)$, the function $\Phi(t, x(t))$ is absolutely continuous on $[0, T]$.

The proof of Lemma 1 is similar to that of [5, Lemma 2], by means of properties 1°–3° of Definition 1, and the following equality, which is obtained from (1.9), (3.3) and (3.4),

$$(3.5) \quad \Phi(t, x(t)) = \int_0^1 \langle P(t, sx(t)), x(t) \rangle ds + M(0), \quad t \in [0, T].$$

The detail is omitted here.

LEMMA 2. The assumptions are the same as in Lemma 1. Then, for each $\{x_0, u(\cdot)\} \in D(A) \times C^1([0, T]; U)$ and corresponding trajectory $x(\cdot)$, the following relation holds:

$$(3.6) \quad \begin{aligned} \frac{d}{dt} \Phi(t, x(t)) &= Q(0) - Q(x(t)) + \langle Bu(t), P(t, x(t)) \rangle \\ &+ \int_0^1 \langle P_x(t, sx(t)) BR^{-1} B^* P(t, sx(t)), x(t) \rangle ds. \end{aligned}$$

The proof of Lemma 2 is straight from (3.5) and similar to that of [5, Lemma 3], so is omitted here.

LEMMA 3. For each $\{x_0, u(\cdot)\} \in X \times \mathcal{U}$ and corresponding trajectory $x(\cdot)$ given by (1.4), there exists a sequence $\{x_n^0, u_n(\cdot)\}_1^\infty \subset D(A) \times C^1([0, T]; U)$ such that

$$\begin{aligned} x_n^0 &\rightarrow x_0 \quad \text{in } X, \\ u_n(\cdot) &\rightarrow u(\cdot) \quad \text{in } \mathcal{U} = L^2(0, T; U), \\ x_n(\cdot) &\rightarrow x(\cdot) \quad \text{in } C([0, T]; X), \end{aligned}$$

where $x_n(\cdot)$ is the trajectory corresponding to x_n^0 and $u_n(\cdot)$.

Proof. This is simply a consequence of the density of $D(A)$ in X , the mollification of L^2 -Bochner integrable functions, and the Hölder inequality.

THEOREM 3. Assume that the quasi-Riccati equation (3.1) has a normal solution $P(t, x)$. Then, for any given $x_0 \in X$, there exists a closed-loop optimal control of (NQBP), given by

$$(3.7) \quad u(t) = -R^{-1} B^* P(t, x(t)), \quad t \in [0, T],$$

where $x(\cdot)$ is the corresponding optimal trajectory.

Proof. Let $\mu(t, x) = \frac{1}{2} \langle R^{-1} B^* P(t, x), B^* P(t, x) \rangle$. It is easy to see that

$$(3.8) \quad \mu'_x(t, x) = P_x(t, x) BR^{-1} B^* P(t, x).$$

For each $\{x_0, u(\cdot)\} \in D(A) \times C^1([0, T]; U)$ and corresponding $x(\cdot)$, by (3.6), (3.8) and (1.9), we have

$$(3.9) \quad \begin{aligned} \frac{d}{dt} \Phi(t, x(t)) &+ Q(x(t)) + \frac{1}{2} \langle Ru(t), u(t) \rangle \\ &= \frac{1}{2} \langle R(u(t) + R^{-1} B^* P(t, x(t))), u(t) + R^{-1} B^* P(t, x(t)) \rangle \\ &+ Q(0) - \mu(t, 0), \end{aligned}$$

where $\Phi(t, x)$ is the canonical anti-derivative of $P(t, x)$.

Integrate (3.9) for $t \in [0, T]$, by Lemma 1, $\Phi_x(T, x) = P(T, x) = M'(x)$ and $\Phi(T, 0) = M(0)$, and it follows that $\Phi(T, x) = M(x)$, $\forall x \in X$, and

$$\begin{aligned}
 J(u) &= M(x(T)) + \int_0^T (Q(x(t)) + \frac{1}{2} \langle Ru(t), u(t) \rangle) dt \\
 &= \left\{ \Phi(0, x_0) + Q(0)T - \int_0^T \mu(t, 0) dt \right\} \\
 &\quad + \frac{1}{2} \int_0^T \langle R(u(t) + R^{-1}B^*P(t, x(t))), u(t) + R^{-1}B^*P(t, x(t)) \rangle dt \\
 &\cong \Phi(0, x_0) + Q(0)T - \int_0^T \mu(t, 0) dt \equiv \rho(x_0),
 \end{aligned}
 \tag{3.10}$$

where $\rho(x_0)$ is a constant determined by x_0 . In view of Lemma 3, (1.3), the continuity of $\Phi(t, x)$ in x , and the property 1° of normal solution $P(t, x)$, we know that (3.10) is also valid for each $\{x_0, u(\cdot)\} \in X \times \mathcal{U}$ and corresponding $x(\cdot)$.

On the other hand, the property 6° of $P(t, x)$ shows that the feedback control (3.7) is admissible, i.e., in \mathcal{U} . Therefore, (3.10) indicates that (3.7) must be optimal, and

$$\min_{u(\cdot) \in \mathcal{U}} J(u; x_0) = \rho(x_0) \quad \forall x_0 \in X.
 \tag{3.11}$$

Thus we have completed the proof. QED.

4. Solution mapping $K(s, t, x)$ of a nonlinear integral equation. In order to explore the existence and possible expression of normal solution of the quasi-Riccati equation (3.1), we consider a nonlinear integral equation:

$$\begin{aligned}
 y(s) &= e^{A^*(T-s)} M' \left(e^{A(T-t)} x - \int_t^T e^{A(T-\eta)} B R^{-1} B^* y(\eta) d\eta \right) \\
 &\quad + \int_s^T e^{A^*(\sigma-s)} Q' \left(e^{A(\sigma-t)} x - \int_t^\sigma e^{A(\sigma-\eta)} B R^{-1} B^* y(\eta) d\eta \right) d\sigma,
 \end{aligned}
 \tag{4.1}$$

$(s, t, x) \in \Omega$,

where

$$\Omega = \{(s, t, x) | 0 \leq t \leq s \leq T, x \in X\}.
 \tag{4.2}$$

LEMMA 4. Let $G_t \in \mathcal{L}(L^2(t, T; X))$ and $H_t \in \mathcal{L}(L^2(t, T; X); X)$ be defined as

$$(G_t \varphi)(s) = \int_t^s e^{A(s-\sigma)} \varphi(\sigma) d\sigma, \quad s \in [t, T], \quad \varphi \in L^2(t, T; X),
 \tag{4.3}$$

$$H_t \varphi = \int_t^T e^{A(T-\sigma)} \varphi(\sigma) d\sigma, \quad \varphi \in L^2(t, T; X).
 \tag{4.4}$$

Suppose that $W \in \mathcal{L}(X)$ and $N \in \mathcal{L}(L^2(t, T; X))$ are nonnegative self-adjoint operators. Then, the following assertions are valid.

1°) The operator

$$V_t = I + (H_t^* W H_t + G_t^* N G_t) B R^{-1} B^* \in \mathcal{L}(L^2(t, T; X))
 \tag{4.5}$$

is bijective and its inverse operator is given by

$$\begin{aligned}
 V_t^{-1} &= I - (H_t^* W H_t + G_t^* N G_t) \sqrt{B R^{-1} B^*} \\
 &\quad \cdot (I + \sqrt{B R^{-1} B^*} (H_t^* W H_t + G_t^* N G_t) \sqrt{B R^{-1} B^*})^{-1} \sqrt{B R^{-1} B^*}.
 \end{aligned}
 \tag{4.6}$$

2°) If $N \in \mathcal{L}(L^2(t, T; X))$ is defined to be

$$(N\varphi)(s) = Y(s)\varphi(s), \quad s \in [t, T],$$

where $Y(\cdot): [t, T] \rightarrow \mathcal{L}(X)$ is a nonnegative self-adjoint operator function and strongly continuous in $s \in [t, T]$, then the above 1° holds and $V_t \in \mathcal{L}(C([t, T]; X))$ has bounded inverse operator $V_t^{-1} \in \mathcal{L}(C([t, T]; X))$ given by (4.6) too.

Proof. Assertion 1° can be verified directly. We see, by transposition,

$$(4.7) \quad \begin{aligned} (G_t^* \varphi)(s) &= \int_s^T e^{A^*(\sigma-s)} \varphi(\sigma) d\sigma, \quad s \in [t, T], \quad \varphi \in L^2(t, T; X), \\ (H_t^* \xi)(s) &= e^{A^*(T-s)} \xi, \quad s \in [t, T], \quad \xi \in X. \end{aligned}$$

By virtue of the facts that both V_t and V_t^{-1} , given by (4.5) and (4.6) with N described as above, map $C([t, T]; X)$ into itself, we obtain 2°. QED.

THEOREM 4. For any given $(t, x) \in [0, T] \times X$, there exists a unique solution $y(\cdot) \in C([t, T]; X)$ of (4.1).

Proof. (1) Existence: According to Corollary 1 and Theorem 2, there exists a unique optimal process $\{\hat{u}_{(t,x)}(\cdot), \hat{x}_{(t,x)}(\cdot)\}$ of (NQB P) $_t$, for the given initial state value x , which satisfies (2.11) and (2.12). Let

$$(4.8) \quad y(s; t, x) = e^{A^*(T-s)} M'(\hat{x}_{(t,x)}(T)) + \int_s^T e^{A^*(\sigma-s)} Q'(\hat{x}_{(t,x)}(\sigma)) d\sigma, \quad s \in [t, T].$$

From (4.8) and (2.12) we can directly verify that $y(\cdot; t, x)$ given by (4.8) is exactly a solution (4.1).

(2) Uniqueness: Let $F: C([t, T]; X) \times [0, T] \times X \rightarrow C([t, T]; X)$ be

$$(4.9) \quad \begin{aligned} F(y(\cdot), t, x)(s) &= y(s) - e^{A^*(T-s)} M' \left(e^{A(T-t)} x - \int_t^T e^{A(T-\eta)} B R^{-1} B^* y(\eta) d\eta \right) \\ &\quad - \int_s^T e^{A^*(\sigma-s)} Q' \left(e^{A(\sigma-t)} x - \int_t^\sigma e^{A(\sigma-\eta)} B R^{-1} B^* y(\eta) d\eta \right) d\sigma, \end{aligned} \quad s \in [t, T].$$

In order to prove the uniqueness, according to the implicit function theorem in Banach spaces (see [6, p. 115]), we only need to show that $D_y F(y(\cdot), t, x) \in \mathcal{L}(C([t, T]; X))$ is boundedly invertible for each $(y(\cdot), x) \in C([t, T]; X) \times X$ and fixed $t \in [0, T]$. In fact,

$$(4.10) \quad \begin{aligned} (D_y F(y(\cdot), t, x) z(\cdot))(s) &= z(s) + \int_t^T \Pi_t(s, \eta) B R^{-1} B^* z(\eta) d\eta \\ &= \{[I + (H_t^* W_t(y, x) H_t + G_t^* N_t(y, x) G_t) B R^{-1} B^*] z(\cdot)\}(s), \quad s \in [t, T] \\ &\quad \forall z(\cdot) \in C([t, T]; X) \end{aligned}$$

where G_t , H_t and their adjoint operators are given by (4.3), (4.4) and (4.7),

$$(4.11) \quad \begin{aligned} \Pi_t(s, \eta) &= e^{A^*(T-s)} M'' \left(e^{A(T-t)} x - \int_t^T e^{A(T-\xi)} B R^{-1} B^* y(\xi) d\xi \right) e^{A(T-\eta)} \\ &\quad + \int_{\max(s, \eta)}^T e^{A^*(\sigma-s)} Q'' \\ &\quad \cdot \left(e^{A(\sigma-t)} x - \int_t^\sigma e^{A(\sigma-\xi)} B R^{-1} B^* y(\xi) d\xi \right) e^{A(\sigma-\eta)} d\sigma, \end{aligned}$$

$$(4.12) \quad W_t(y, x) = M'' \left(e^{A(T-t)} x - \int_t^T e^{A(T-\xi)} B R^{-1} B^* y(\xi) d\xi \right),$$

$$(4.13) \quad (N_t(y, x)\varphi)(s) = Q'' \left(e^{A(s-t)} x - \int_t^s e^{A(s-\xi)} B R^{-1} B^* y(\xi) d\xi \right) \varphi(s),$$

$$s \in [t, T] \quad \forall \varphi \in L^2(t, T; X).$$

The convexity of $M(\cdot)$ and $Q(\cdot)$ implies that $M''(x)$ and $Q''(x) \in \mathcal{L}(X)$ are nonnegative self-adjoint, so that $W_t(y, x) \in \mathcal{L}(X)$ and $N_t(y, x) \in \mathcal{L}(L^2(t, T; X))$ are also non-negative. According to Lemma 4-2° and (4.10), $D_y F(y(\cdot), t, x) \in \mathcal{L}(C([t, T]; X))$ is boundedly invertible. QED.

We denote the solution mapping of (4.1) by

$$(4.14) \quad y(s) = K(s, t, x), \quad (s, t, x) \in \Omega.$$

Let $E(s, t, x)$ be defined as

$$(4.15) \quad E(s, t, x) = e^{A(s-t)} x - \int_t^s e^{A(s-\eta)} B R^{-1} B^* K(\eta, t, x) d\eta.$$

LEMMA 5. For any bounded subset $\Sigma \subset X$, let $\Omega_\Sigma = \{(s, t, x) | 0 \leq t \leq s \leq T, x \in \Sigma\}$. Then,

$$(4.16) \quad \sup_{(s, t, x) \in \Omega_\Sigma} \|K(s, t, x)\| = \text{const}(\Sigma) < +\infty,$$

$$(4.17) \quad \sup_{(s, t, x) \in \Omega_\Sigma} \|E(s, t, x)\| = \text{const}(\Sigma) < +\infty.$$

Proof. These two conclusions are consequences of (4.8), (1.3), 2° of Corollary 2 and (4.15). QED.

LEMMA 6. For any given $x \in X$, the following limit relation is equiconvergent with respect to $t \in [0, T]$,

$$(4.18) \quad \lim_{\delta x \rightarrow 0} \|K(\cdot, t, x + \delta x) - K(\cdot, t, x)\|_{C([t, T]; X)} = 0.$$

Proof. Let $\Delta K(s) = K(s, t, x + \delta x) - K(s, t, x)$ and

$$(4.19) \quad E_{\delta x}(s, t, x) = e^{A(s-t)} \delta x - \int_t^s e^{A(s-\eta)} B R^{-1} B^* \Delta K(\eta) d\eta.$$

By calculation we see that $\Delta K(\cdot)$ satisfies the following equation:

$$(4.20) \quad \begin{aligned} & (1 + (H_t^* \tilde{W}_{\delta x} H_t + G_t^* \tilde{N}_{\delta x} G_t) B R^{-1} B^*) \Delta K(\cdot) \\ & = H_t^* \tilde{W}_{\delta x} e^{A(T-t)} \delta x + G_t^* \tilde{N}_{\delta x} e^{A(\cdot-t)} \delta x, \end{aligned}$$

where G_t , H_t and their adjoint operators are given by (4.3), (4.4) and (4.7),

$$(4.21) \quad \tilde{W}_{\delta x} = \int_0^1 M''(E(T, t, x) + \lambda E_{\delta x}(T, t, x)) d\lambda \in \mathcal{L}(X),$$

and $\tilde{N}_{\delta x} \in \mathcal{L}(L^2(t, T; X)) \cap \mathcal{L}(C([t, T]; X))$ is a multiplication operator:

$$(4.22) \quad (\tilde{N}_{\delta x} \varphi)(s) = \int_0^1 Q''(E(s, t, x) + \lambda E_{\delta x}(s, t, x)) d\lambda \cdot \varphi(s), \quad s \in [t, T].$$

By Lemma 4-2° and (4.6), on account of the following facts,

$$\begin{aligned}
 & \sup_{0 \leq t \leq s \leq T} \|E(s, t, x)\| < +\infty, \quad \sup_{\substack{0 \leq t \leq s \leq T \\ \|\delta x\| \leq 1}} \|E_{\delta x}(s, t, x)\| < +\infty \quad (\text{by Lemma 5}), \\
 & \sup_{t \in [0, T]} \|G_t\|_{\mathcal{L}(L^2(t, T; X); C([t, T]; X))} < +\infty, \quad \sup_{t \in [0, T]} \|G_t^*\|_{\mathcal{L}(C([t, T]; X))} < +\infty \\
 (4.23) \quad & \hspace{15em} (\text{by (4.3) and (4.7)}), \\
 & \sup_{t \in [0, T]} \|H_t\|_{\mathcal{L}(L^2(t, T; X); X)} < +\infty, \quad \sup_{t \in [0, T]} \|H_t^*\|_{\mathcal{L}(X; C(t, T; X))} < +\infty \\
 & \hspace{15em} (\text{by (4.4) and (4.7)}),
 \end{aligned}$$

and for sufficiently small $\|\delta x\|$ (e.g. $\|\delta x\| \leq 1$) and all $t \in [0, T]$,

$$\begin{aligned}
 & \|\tilde{W}_{\delta x}\|_{\mathcal{L}(X)} \leq \text{const}, \quad \|\tilde{N}_{\delta x}\|_{\mathcal{L}(L^2(t, T; X))} \leq \text{const}, \quad \|\tilde{N}_{\delta x}\|_{\mathcal{L}(C([t, T]; X))} \leq \text{const} \\
 (4.24) \quad & \hspace{10em} (\text{from the } C^2 \text{ continuity of } M(\cdot) \text{ and } Q(\cdot)),
 \end{aligned}$$

we obtain, by inversion of (4.20), that

$$(4.25) \quad \|\Delta K(\cdot)\|_{C([t, T]; X)} \leq \text{const} \|\delta x\|,$$

where the constant is independent of $t \in [0, T]$ and δx such that $\|\delta x\| \leq 1$. This amounts to the equiconvergence of (4.18). QED.

THEOREM 5. *The solution mapping $K(s, t, x)$ of (4.1) possesses the following properties:*

1°) $K(s, t, x)$ is strongly continuous in $(s, t, x) \in \Omega$.

2°) $\langle K(s, t, x), y \rangle$ is differentiable in $s \in [t, T]$, for each x and y in $D(A)$, and

$$(4.26) \quad \frac{d}{ds} \langle K(s, t, x), y \rangle = -\langle K(s, t, x), Ay \rangle - \langle Q'(E(s, t, x)), y \rangle,$$

which is continuous in (s, t) such that $0 \leq t \leq s \leq T$.

3°) $K(s, t, x)$ is strongly differentiable in $t \in [0, T]$, for each $s \in [t, T]$ and $x \in D(A)$, and

$$\begin{aligned}
 K_t(s, t, x) = & \left\{ (D_y F(K(\cdot, t, x), t, x))^{-1} \left[e^{A^*(T-\cdot)} M''(E(T, t, x)) e^{A(T-t)} \right. \right. \\
 (4.27) \quad & \left. \left. + \int_t^T e^{A^*(\sigma-\cdot)} Q''(E(\sigma, t, x)) e^{A(\sigma-t)} d\sigma \right] \right\} (s) \\
 & \cdot [-Ax + BR^{-1}B^*K(t, t, x)].
 \end{aligned}$$

Moreover, $K_t(s, t, x)$ is strongly continuous in (s, t) such that $0 \leq t \leq s \leq T$, for each $x \in D(A)$.

4°) $K(s, t, x)$ is Fréchet differentiable in $x \in X$, for each (s, t) such that $0 \leq t \leq s \leq T$, and

$$\begin{aligned}
 K_x(s, t, x) = & \left\{ [D_y F(K(\cdot, t, x), t, x)]^{-1} \left[e^{A^*(T-\cdot)} M''(E(T, t, x)) e^{A(T-t)} \right. \right. \\
 (4.28) \quad & \left. \left. + \int_t^T e^{A^*(\sigma-\cdot)} Q''(E(\sigma, t, x)) e^{A(\sigma-t)} d\sigma \right] \right\} (s)
 \end{aligned}$$

Moreover,

(i) $K_x(s, t, x)\xi$ is strongly continuous in $(s, t, x) \in \Omega$, for each $\xi \in X$;

(ii) $K_x(s, t, x)$ is bounded in $\mathcal{L}(X)$ -norm for $0 \leq t \leq s \leq T$ and $x \in \Sigma$, where $\Sigma \subset X$ is any bounded subset.

We divide the proof of Theorem 5 into following three lemmas.

LEMMA 7. Assertions 1° and 2° of Theorem 5 are valid.

Proof. Assertion 1°: Because the associated space $C([t, T]; X)$ of the mapping F in (4.9) is t -variant, we cannot simply use the abstract implicit function theorem here. Let $\Delta s \rightarrow +0$, $\Delta t \rightarrow +0$ (similarly for other cases), and $\delta x \rightarrow 0$; we have

$$K(s + \Delta s, t + \Delta t, x + \delta x) - K(s, t, x) = I_1 + I_2 + I_3 \rightarrow 0,$$

because of

$$I_1 = K(s + \Delta s, t + \Delta t, x + \delta x) - K(s + \Delta s, t + \Delta t, x) \rightarrow 0 \quad (\text{by Lemma 6}),$$

$$I_2 = K(s + \Delta s, t + \Delta t, x) - K(s + \Delta s, t, x)$$

$$= K(s + \Delta s, t + \Delta t, x) - K(s + \Delta s, t + \Delta t, \hat{x}_{(t,x)}(t + \Delta t)) \rightarrow 0$$

$$(\text{by (4.8), (2.12), Lemma 6 and } \hat{x}_{(t,x)}(t + \Delta t) \rightarrow x \text{ when } \Delta t \rightarrow 0),$$

$$I_3 = K(s + \Delta s, t, x) - K(s, t, x) \rightarrow 0.$$

Assertion 2°: By (4.1), for $x, y \in D(A)$, we can differentiate $\langle K(s, t, x), y \rangle$ straight to achieve (4.26) and the assertion 2°. QED.

LEMMA 8. Assertion 3° of Theorem 5 is valid.

Proof. Let $\Delta > 0$ be sufficiently small and

$$\delta K(s, t, x, \Delta) = \frac{1}{\Delta} (K(s, t + \Delta, x) - K(s, t, x)), \quad s \in [t + \Delta, T].$$

By calculation we see that $\delta K(\cdot, t, x, \Delta)$ satisfies the following equation:

$$\begin{aligned} & (I + (H_{t+\Delta}^* \hat{W}_{t+\Delta} H_{t+\Delta} + G_{t+\Delta}^* \hat{N}_{t+\Delta} G_{t+\Delta}) BR^{-1} B^*) \delta K(\cdot, t, x, \Delta) \\ (4.29) \quad & = [H_{t+\Delta}^* \hat{W}_{t+\Delta} e^{A(T-(t+\Delta))} + G_{t+\Delta}^* \hat{N}_{t+\Delta} e^{A(-(t+\Delta))}] \\ & \quad \cdot \left\{ \frac{1}{\Delta} (I - e^{A\Delta}) x + \frac{1}{\Delta} \int_t^{t+\Delta} e^{A(t+\Delta-\eta)} BR^{-1} B^* K(\eta, t, x) d\eta \right\}, \end{aligned}$$

where G_t , H_t and their adjoint operators are given by (4.3), (4.4) and (4.7),

$$(4.30) \quad \hat{W}_{t+\Delta} = \int_0^1 M''(E(T, t, x) + \lambda E^\Delta(T, t, x)) d\lambda \in \mathcal{L}(X),$$

$$(4.31) \quad (\hat{N}_{t+\Delta} \varphi)(s) = \int_0^1 Q''(E(s, t, x) + \lambda E^\Delta(s, t, x)) d\lambda \cdot \varphi(s), \quad s \in [t + \Delta, T]$$

(a multiplication operator on $L^2(t + \Delta, T; X)$)

in which $E(s, t, x)$ given by (4.15), and

$$\begin{aligned} E^\Delta(s, t, x) &= (e^{A(s-(t+\Delta))} - e^{A(s-t)})x + \int_t^{t+\Delta} e^{A(s-\eta)} BR^{-1} B^* K(\eta, t, x) d\eta \\ (4.32) \quad & - \int_{t+\Delta}^s e^{A(s-\eta)} BR^{-1} B^* (K(\eta, t + \Delta, x) - K(\eta, t, x)) d\eta. \end{aligned}$$

For $0 \leq t \leq s \leq T$ and $x \in D(A)$, in view of (4.3), (4.4), (4.6), (4.7), (4.29)–(4.32), (4.23) and (4.10), we obtain

(4.33)

$$\lim_{\Delta \rightarrow +0} \delta K(s, t, x, \Delta) = \{[D_y F(K(\cdot, t, x), t, x)]^{-1} [H_t^* \hat{W}_t e^{A(T-t)} + G_t^* \hat{N}_t e^{A(\cdot-t)}]\}(s) \\ \cdot (-Ax + BR^{-1}B^*K(t, t, x)).$$

Hence (4.27) holds for the right derivative $K_t^+(s, t, x)$, $0 \leq t < s \leq T$, $x \in D(A)$. Similarly it holds for $K_t^-(s, t, x)$, $0 \leq t \leq s \leq T$, $x \in D(A)$.

The remains can be deduced from (4.27), (4.10), (4.6), (1.3) and Lemmas 6 and 7. QED.

LEMMA 9. *The assertion 4° of Theorem 5 is valid.*

Proof. Now we can fix $t \in [0, T]$ arbitrarily. Apply the abstract implicit function theorem to the mapping F of (4.9); then we have the Fréchet differentiability of $K(s, t, x)$ in $x \in X$; moreover,

$$K_x(\cdot, t, x) = -(D_y F(K(\cdot, t, x), t, x))^{-1} D_x F(K(\cdot, t, x), t, x) \\ = -(D_y F(K(\cdot, t, x), t, x))^{-1} (-e^{A^*(T-\cdot)} M''(E(T, t, x)) e^{A(T-t)} \\ - \int_{\cdot}^T e^{A^*(\sigma-\cdot)} Q''(E(\sigma, t, x)) e^{A(\sigma-t)} d\sigma).$$

Thus (4.28) is true. The remains can be deduced from (4.28), (4.10), (4.6), (1.3) and Lemmas 6 and 7. QED.

Thus we complete the proof of Theorem 5.

5. Feedback operator $P(t, x)$ and closed-loop solution. $K(s, t, x): \Omega \rightarrow X$ is the solution mapping (4.14) of (4.1). Let

$$(5.1) \quad P(t, x) = K(t, t, x): [0, T] \times X \rightarrow X.$$

We shall prove that $P(t, x)$ given by (5.1) is a normal solution of the quasi-Riccati equation (3.1), so that by Theorem 3 it is a feedback operator of closed-loop optimal control of (NQBP).

LEMMA 10. *$P(t, x)$ given by (5.1) possesses the properties 1°, 2° and 3° of Definition 1.*

Proof. By Theorem 5 and (4.26)–(4.28) we have

$$(5.2) \quad \frac{d}{dt} \langle P(t, x), y \rangle = \left\{ \frac{d}{ds} \langle K(s, t, x), y \rangle + \frac{d}{dt} \langle K(s, t, x), y \rangle \right\} \Big|_{s=t} \\ = -\langle P(t, x), Ay \rangle - \langle Q'(x), y \rangle \\ + \left\{ (D_y F(K, t, x))^{-1} \left[e^{A^*(T-\cdot)} M''(E(T, t, x)) e^{A(T-t)} \right. \right. \\ \left. \left. + \int_{\cdot}^T e^{A^*(\sigma-\cdot)} Q''(E(\sigma, t, x)) e^{A(\sigma-t)} d\sigma \right] \right\}(t) \\ \cdot (-Ax + BR^{-1}B^*P(t, x)), \quad (t, x, y) \in [0, T] \times D(A) \times D(A),$$

and

$$(5.3) \quad P_x(t, x) = \left\{ (D_y F(K, t, x))^{-1} \left[e^{A^*(T-\cdot)} M''(E(T, t, x)) e^{A(T-t)} + \int_t^T e^{A^*(\sigma-\cdot)} Q''(E(\sigma, t, x)) e^{A(\sigma-t)} d\sigma \right] \right\} (t),$$

$$(t, x) \in [0, T] \times X.$$

The mentioned properties 1°, 2° and 3° can be verified by means of (5.1)–(5.3) and those properties of $K(s, t, x)$ described in Theorem 5. QED.

LEMMA 11. For each $t \in [0, T]$, $P(t, \cdot): X \rightarrow X$ given by (5.1) is gradient operator.

Proof. According to Theorem 2.5.2 of [6], it is equivalent to show that $P_x(t, x) \in \mathcal{L}(X)$ is self-adjoint for each $t \in [0, T]$ and $x \in X$. In fact, by (5.3), (4.10) and (4.6), we have

$$(5.4) \quad P_x(t, x) = \left\{ e^{A^*(T-t)} M''(E(T, t, x)) e^{A(T-t)} + \int_t^T e^{A^*(\sigma-t)} Q''(E(\sigma, t, x)) e^{A(\sigma-t)} d\sigma \right\} \\ - Z_t(t) \sqrt{BR^{-1}B^*} (I + \sqrt{BR^{-1}B^*} Z_t \sqrt{BR^{-1}B^*})^{-1} \sqrt{BR^{-1}B^*} \\ \cdot \left\{ e^{A^*(T-\cdot)} M''(E(T, t, x)) e^{A(T-t)} + \int_t^T e^{A^*(\sigma-\cdot)} Q''(E(\sigma, t, x)) e^{A(\sigma-t)} d\sigma \right\},$$

where

$$(5.5) \quad Z_t = H_t^* W_E H_t + G_t^* N_E G_t, \\ Z_t(t) \varphi = (Z_t \varphi)(t) \quad \forall \varphi \in L^2(t, T; X),$$

and

$$(5.6) \quad W_E = M''(E(T, t, x)), \\ (N_E \varphi)(\sigma) = Q''(E(\sigma, t, x)) \varphi(\sigma), \quad \sigma \in [t, T], \quad \varphi \in L^2(t, T; X).$$

Note that $Z_t(t) \in \mathcal{L}(L^2(t, T; X); X)$; we can verify that its adjoint operator $Z_t(t)^* \in \mathcal{L}(X; L^2(t, T; X))$ is given by

$$(5.7) \quad Z_t(t)^* = e^{A^*(T-\cdot)} M''(E(T, t, x)) e^{A(T-t)} + \int_t^T e^{A^*(\sigma-\cdot)} Q''(E(\sigma, t, x)) e^{A(\sigma-t)} d\sigma.$$

Substitute (5.7) for the last bracket of (5.4). We see that $P_x(t, x) \in \mathcal{L}(X)$ is self-adjoint for $(t, x) \in [0, T] \times X$. QED.

LEMMA 12. The Cauchy problem (3.2) with $P(t, x)$ given by (5.1) admits a global mild solution $x(\cdot) \in C([0, T]; X)$ for any given $x_0 \in X$.

Proof. We show that (see (4.15))

$$(5.8) \quad x(t) = E(t, 0, x_0), \quad t \in [0, T]$$

is a desired mild solution of (3.2), i.e., it is a strongly continuous solution of following equation:

$$(5.9) \quad x(t) = e^{At} x_0 - \int_0^t e^{A(t-s)} BR^{-1} B^* P(s, x(s)) ds, \quad t \in [0, T].$$

Obviously, $E(\cdot, 0, x_0) \in C([0, T]; X)$. We only need to prove

$$(5.10) \quad K(t, 0, x_0) = K(t, t, E(t, 0, x_0)), \quad t \in [0, T], \quad x_0 \in X,$$

for then (5.9) is satisfied by (5.8):

$$\begin{aligned} E(t, 0, x_0) &= e^{At}x_0 - \int_0^t e^{A(t-s)}BR^{-1}B^*K(s, 0, x_0) ds \\ &= e^{At}x_0 - \int_0^t e^{A(t-s)}BR^{-1}B^*K(s, s, E(s, 0, x_0)) ds \quad (\text{by (5.10)}) \\ &= e^{At}x_0 - \int_0^t e^{A(t-s)}BR^{-1}B^*P(s, E(s, 0, x_0)) ds \quad (\text{by (5.1)}). \end{aligned}$$

Now we prove (5.10). Let $\Delta K_{(t, x_0)}(s) = K(s, t, E(t, 0, x_0)) - K(s, 0, x_0)$, $s \in [t, T]$. By calculation we obtain that $\Delta K_{(t, x_0)}(\cdot)$ satisfies the following equation:

$$(5.11) \quad (I + (H_t^*W_{\Delta K}H_t + G_t^*N_{\Delta K}G_t)BR^{-1}B^*)\Delta K_{(t, x_0)}(\cdot) = 0,$$

where

$$(5.12) \quad W_{\Delta K} = \int_0^1 M'' \left(E(T, 0, x_0) - \lambda \int_t^T e^{A(T-\eta)}BR^{-1}B^*\Delta K_{(t, x_0)}(\eta) d\eta \right) d\lambda,$$

$$(5.13) \quad (N_{\Delta K}\varphi)(s) = \int_0^1 Q'' \left(E(s, 0, x_0) - \lambda \int_t^s e^{A(s-\eta)}BR^{-1}B^*\Delta K_{(t, x_0)}(\eta) d\eta \right) d\lambda \cdot \varphi(s),$$

$s \in [t, T], \quad \varphi \in L^2(t, T; X).$

By Lemma 4, we conclude that $\Delta K_{(t, x_0)}(\cdot) = 0$ in $C([t, T]; X)$. Therefore (5.10) is valid. Besides, the well properties of $P(t, x)$ imply that the strongly continuous solution of (5.9) is unique. QED.

THEOREM 6. $P(t, x)$ given by (5.1) is a normal solution of the quasi-Riccati operator equation (3.1).

Proof. Obviously this $P(t, x)$ is such that (by (4.1))

$$(5.14) \quad P(T, x) = K(T, T, x) = M'(x), \quad x \in X.$$

From (5.2) and (5.3) we can verify directly that (3.1) is satisfied by this $P(t, x)$:

$$\begin{aligned} &\frac{d}{dt} \langle P(t, x), y \rangle + \langle P_x(t, x)Ax, y \rangle + \langle P(t, x), Ay \rangle + \langle Q'(x), y \rangle - \langle P_x(t, x)BR^{-1}B^*P(t, x), y \rangle \\ &= \{ -\langle P(t, x), Ay \rangle - \langle Q'(x), y \rangle + \langle P_x(t, x)(-Ax + BR^{-1}B^*P(t, x)), y \rangle \\ &\quad + \langle P_x(t, x)Ax, y \rangle + \langle P(t, x), Ay \rangle + \langle Q'(x), y \rangle - \langle P_x(t, x)BR^{-1}B^*P(t, x), y \rangle \} = 0. \end{aligned}$$

By Lemmas 10–12 we know that all the properties described in Definition 1 are possessed by this $P(t, x)$. Thus it is a normal solution of (3.1). QED.

THEOREM 7 (Closed-loop theorem). *For any given $x_0 \in X$, there exists a closed-loop optimal control of (NQBP), given by the following state feedback:*

$$(5.15) \quad u(t) = -R^{-1}B^*P(t, x(t)) = -R^{-1}B^*K(t, t, x(t)), \quad t \in [0, T],$$

where $K(s, t, x)$ is the solution operator (4.14) of the nonlinear integral equation (4.1) and $P(t, x)$ is given by (5.1).

Proof. This closed-loop result is obtained by combination of Theorem 3 with Theorem 6. QED.

6. Remarks. We have two remarks about the relation between the closed-loop result of (NQBP) and that of quadratic optimal control, and that of the nonquadratic differential game problem.

Remark 1. If $M(\cdot)$ and $Q(\cdot)$ in (1.2) reduce to quadratic forms, $M(x) = \frac{1}{2}\langle \hat{M}x, x \rangle$ and $Q(x) = \frac{1}{2}\langle \hat{Q}x, x \rangle$, where $\hat{M} \in \mathcal{L}(X)$ and $\hat{Q} \in \mathcal{L}(X)$ are nonnegative self-adjoint, then we can verify that (4.1) becomes following integral equation of Fredholm type [8]:

$$y(s) + \int_t^T \Psi_t(s, \eta) B R^{-1} B^* y(\eta) d\eta = \left[e^{A^*(T-s)} \hat{M} e^{A(T-t)} + \int_s^T e^{A^*(\sigma-s)} \hat{Q} e^{A(\sigma-t)} d\sigma \right] x, \quad (6.1)$$

$(s, t, x) \in \Omega$

where

$$\Psi_t(s, \eta) = e^{A^*(T-s)} \hat{M} e^{A(T-\eta)} + \int_{\max(s, \eta)}^T e^{A^*(\sigma-s)} \hat{Q} e^{A(\sigma-\eta)} d\sigma, \quad (6.2)$$

$(s, \eta) \in [t, T] \times [t, T]$.

The solution operator $K(s, t, x)$ of (6.1) turns out to be a bounded linear operator in x , $K(s, t, x) = \hat{K}(s, t)x$. As a result, (5.1) gives $P(t, x) = \hat{P}(t)x$ where $\hat{P}(t) = \hat{K}(t, t) \in \mathcal{L}(X)$ is the strongly continuous and self-adjoint solution of following Riccati operator equation:

$$\begin{aligned} \frac{d}{dt} \langle \hat{P}(t)x, y \rangle + \langle Ax, \hat{P}(t)y \rangle + \langle \hat{P}(t)x, Ay \rangle \\ + \langle \hat{Q}x, y \rangle - \langle \hat{P}(t) B R^{-1} B^* \hat{P}(t)x, y \rangle = 0, \quad (t, x, y) \in [0, T] \times D(A) \times D(A), \\ \hat{P}(T) = \hat{M}. \end{aligned} \quad (6.3)$$

Thus in this case, the closed-loop result (5.15) coincides with the well-known linear state feedback [1], [9]

$$u(t) = -R^{-1} B^* \hat{P}(t)x(t), \quad t \in [0, T]. \quad (6.4)$$

Remark 2. The method in this paper can be applied to deal with nonquadratic differential game problem of Bolza type on Hilbert spaces:

$$\begin{aligned} \frac{dx}{dt} = Ax(t) + Bu(t) + Cv(t), \quad x(0) = x_0, \\ J(u, v) = M(x(T)) + \int_0^T (Q(x(t)) + \frac{1}{2}\langle R_1 u(t), u(t) \rangle + \frac{1}{2}\langle R_2 v(t), v(t) \rangle) dt, \end{aligned} \quad (6.5)$$

where $C \in \mathcal{L}(V; X)$, $R_1 > 0$ and $R_2 < 0$ coercively, and it is desired to find an optimal closed-loop strategy $\{\hat{u}(\cdot), \hat{v}(\cdot)\} \in L^2(0, T; U) \times L^2(0, T; V)$ such that

$$J(\hat{u}, v) \leq J(\hat{u}, \hat{v}) \leq J(u, \hat{v}) \quad (6.6)$$

for arbitrary admissible feedback control $u(\cdot)$ and $v(\cdot)$. We give a synthesis result whose proof is similar to Theorem 3 and omitted here.

THEOREM 8. *If the following quasi-Riccati operator equation*

$$\begin{aligned} \frac{d}{dt} \langle P(t, x), y \rangle + \langle P_x(t, x) Ax, y \rangle + \langle P(t, x), Ay \rangle + \langle Q'(x), y \rangle \\ - \langle P_x(t, x) (B R_1^{-1} B^* + C R_2^{-1} C^*) P(t, x), y \rangle = 0, \\ P(T, x) = M'(x), \quad x \in X \end{aligned} \quad (6.7)$$

$(t, x, y) \in [0, T] \times D(A) \times D(A)$,

has a normal solution $P(t, x): [0, T] \times X \rightarrow X$ (as in Definition 1), then, for any given $x_0 \in X$, there exists a closed-loop optimal strategy

$$(6.8) \quad \begin{aligned} u(t) &= -R_1^{-1} B^* P(t, x(t)), \\ v(t) &= -R_2^{-1} C^* P(t, x(t)), \end{aligned} \quad t \in [0, T].$$

This result is a generalization of the quadratic differential game [10].

REFERENCES

- [1] R. F. CURTAIN AND A. J. PRITCHARD, *Infinite Dimensional Linear Systems Theory*, Springer-Verlag, Berlin, Heidelberg, New York, 1978.
- [2] Y. C. YOU, *Nonlinear optimal state feedback for distributed parameter systems*, submitted to The 7th International Conference on Analysis and Optimization of Systems, June 1986, Antibes, France.
- [3] ———, *Closed-loop solution to nonquadratic Lagrange problem for distributed parameter systems*, accepted paper of The 4th IFAC Symposium on Control of Distributed Parameter Systems, June 30–July 2, UCLA, Los Angeles, CA.
- [4] V. BARBU AND TH. PRECUPANU, *Convexity and Optimization in Banach Spaces*, Sijthoff and Noordhoff, Bucharest, 1978.
- [5] V. BARBU AND G. DA PRATO, *Hamilton–Jacobi Equations in Hilbert Spaces*, Pitman, Boston, London, 1983.
- [6] M. S. BERGER, *Nonlinearity and Functional Analysis*, Academic Press, New York, 1977.
- [7] A. PAZY, *Semigroups of Linear Operators and Applications to Partial Differential Equations*, Springer-Verlag, Berlin, New York, 1983.
- [8] Y. C. YOU, *On solution of a class of operator Riccati equations*, Chinese Ann. Math. Ser. A, 5 (1984), pp. 219–227.
- [9] ———, *Optimal control for linear systems with quadratic indefinite criterion on Hilbert spaces*, Chinese Ann. Math., Ser. B, 4 (1983), pp. 21–32.
- [10] ———, *Closed-loop syntheses for quadratic differential game of distributed parameter systems*, Chinese Ann. Math., Ser. B, 6 (1985), pp. 325–334.

OPTIMIZATION OF "log x" ENTROPY OVER LINEAR EQUALITY CONSTRAINTS*

YAIR CENSOR† AND ARNOLD LENT‡

Abstract. In this paper we develop a special-purpose iterative algorithm, of the row-action type, for solving the problem of maximizing the "log x" entropy functional over linear equality constraints. The algorithm employs "projections" onto hyperplanes which we call "log x" entropy projections. A complete proof of convergence is given.

Key words. entropy optimization, row-action algorithm, linear equality

AMS(MOS) subject classifications. 49D20, 65K05, 90C25, 94A17

1. Introduction. The " $x \log x$ " entropy functional, $\text{ent } x$, maps the nonnegative orthant \mathbb{R}_+^n of the n -dimensional Euclidean space \mathbb{R}^n into \mathbb{R} according to

$$(1) \quad \text{ent } x := - \sum_{j=1}^n x_j \log x_j,$$

where, by definition, $0 \log 0 := 0$.

Entropy optimization problems which seek to maximize $\text{ent } x$ over linear constraints sets (equality, inequality or interval constraints) arise in various fields of applications. These include (i) *transportation planning* (the gravity model) see, e.g., [31], (ii) *statistics* (adjustment of contingency tables, maximum-likelihood estimation) see, e.g., [12], (iii) *linear numerical analysis* (preconditioning of a matrix prior to calculation of eigenvalues and eigenvectors) see, e.g. [15], (iv) *chemistry* (the chemical equilibrium problem) see, e.g., the remark and references mentioned in [16], (v) *geometric programming* (the dual problem) see, e.g., [40], (vi) *image processing* (image reconstruction from projections, image restoration) see, e.g., [6], [18], [22]. For further general information we mention here [28], [30], [19].

The use of entropy is rigorously founded in several areas, see, e.g., [1], [28], [34], while in other situations entropy optimization is used on a more empirical basis. In image reconstruction from projections (from where our own motivation to study entropy optimization comes), arguments in favor of maximum entropy imaging have been given, typically expressing a conviction that the maximum entropy approach yields the most probable solution agreeing with the available data, i.e., a solution which is most objective or maximally uncommitted with respect to missing information.

Another measure of entropy is the "log x" entropy functional defined on the positive orthant by

$$(2) \quad h(x) := \sum_{j=1}^n \log x_j.$$

* Received by the editors February 19, 1985; accepted for publication (in revised form) May 6, 1986.

† Department of Mathematics and Computer Science, University of Haifa, Mt. Carmel, Haifa 31999, Israel. Present address, Medical Image Processing Group, Department of Radiology, Hospital of the University of Pennsylvania, Philadelphia, Pennsylvania 19104. The work of this author was supported in part by National Science Foundation grant ECS-8117908 and National Institutes of Health grant HL-28438 while the author was visiting the Medical Image Processing Group during the academic year 1982-83 and in February 1985.

‡ Technicare Corporation, Cleveland, Ohio 44139.

This entropy was first proposed by Burg in [3], see [17, §§ 5.17, 5.18] and [28], and has since then provoked a controversy regarding the question of which entropy functional should be used in different situations. This question was discussed in [11], [17], [39], [19, § 10.4.14], and recently in [29].

We do not enter the argument of “ $x \log x$ ” entropy versus “ $\log x$ ” entropy at all but are rather interested in the mathematical question of the construction and study of useful algorithms for solving computationally linearly constrained entropy optimization problems.

Several recent publications discuss algorithms for optimization of the “ $x \log x$ ” entropy functional over various linear constraints. These include [32], where the convergence of the algorithm called “MART” (Multiplicative Algebraic Reconstruction Technique), first proposed in [20], was proved, and [37], where a maximum entropy algorithm called “MENT” was proposed, studied and experimented with. References [23], [24] contain further results about the practical performance of “ $x \log x$ ” entropy optimization algorithms in image reconstruction from projections.

The algorithms presented in [10] are applicable to “ $x \log x$ ” (but *not* to “ $\log x$ ”) entropy optimization over equality, inequality and interval linear constraints. They are based on the method of Bregman [2], see also [5], [8], [31].

In spite of the extensive work done on algorithms for “ $x \log x$ ” entropy optimization problems, little attention has been given to the development of special-purpose algorithms for “ $\log x$ ” entropy optimization. In this paper we develop and study such an algorithm.

Special-purpose algorithms for solving optimization problems with a specific objective function have demonstrated their effectiveness in several fields of applications. For example, in image reconstruction from projections several special-purpose algorithms were proposed, studied and tested [7], [14], [21], [25]. The method of Hildreth [26], further studied in [33], is a special-purpose method for norm minimization over linear inequalities. Reference [4] is a recent review of various such special algorithms.

The algorithm we study here is a special-purpose algorithm, of the row-action type (in the sense of [4]), for solving the “ $\log x$ ” entropy optimization problem over linear equality constraints. We have first proposed this algorithm on a heuristic basis and without mathematical analysis in [9]. It was implemented by Jain and Ranganath [27], who applied it to a problem of two-dimensional spectral estimation. The present paper is, to our knowledge, the first presentation of a thorough study including a proof of convergence of this algorithm. This paper is statedly a mathematical, not an experimental, study and therefore we make no claims whatsoever about the efficacy of the algorithm presented here in practical application. However, the row-action nature of the algorithm (discussed later) points to its potential advantages for handling very large and sparse problems.

Bregman’s theory [2], as applied and extended in [10] is not *immediately* applicable to the case of “ $\log x$ ” entropy because of the singularity of this function at the boundary of its domain. However, by showing that the iterates produced by the algorithm we study “stay away” from that boundary, we are able to analyze convergence along the lines of Bregman’s original theory.

2. Problem formulation and algorithm development. Consider the following “ $\log x$ ” entropy optimization problem:

$$(3) \quad \begin{array}{ll} \text{Max } h(x) \\ \text{subject to } Ax = b \quad \text{and} \quad x > 0 \end{array}$$

where $h(x)$ is given by (2), A is a given real $m \times n$ matrix, $b \in \mathbb{R}^m$ is a given vector and $x > 0$ means $x_j > 0$, $j = 1, 2, \dots, n$. The system $Ax = b$ is alternatively written as $\langle a^i, x \rangle = b_i$, $i = 1, 2, \dots, m$, where $\langle \cdot, \cdot \rangle$ is the (usual) inner product in \mathbb{R}^n and a^i (all vectors are columns) is the i th column of A^T (T stands for transpose).

Define the feasible set of (3) by $F := \{x \in \mathbb{R}^n \mid Ax = b\} \cap \text{int } \mathbb{R}_+^n$ and the null space of A by $N(A) = \{x \in \mathbb{R}^n \mid Ax = 0\}$.

The function $[-h(x)]$ is strictly convex over $\text{int } \mathbb{R}_+^n$ and, for convenience, we rewrite (3) as

$$(4) \quad \begin{aligned} &\text{Min } [-h(x)] \\ &\text{subject to } x \in F. \end{aligned}$$

Introducing dual variables for the equality constraints, we form the Lagrangian associated with (4),

$$(5) \quad L(x, u) = -\sum_{j=1}^n \log x_j + \sum_{i=1}^m u_i (\langle a^i, x \rangle - b_i).$$

Duality (see, e.g., [35, p. 316] permits us to define the dual function

$$(6) \quad \phi(u) = \text{Inf } \{L(x, u) \mid x \in \text{int } \mathbb{R}_+^n\}.$$

The next proposition gives a necessary and sufficient condition for the solvability of (4).

PROPOSITION 1. *If $F \neq \emptyset$ and $A \neq 0$, then a necessary and sufficient condition for the solution of (4) to exist is*

$$(7) \quad N(A) \cap \mathbb{R}_+^n = \{0\}.$$

Proof. Sufficiency. Assume that $\text{Inf } \{-h(x) \mid x \in F\} = -\infty$ and take a sequence $\{x^k\}_{k=0}^\infty$ such that $x^k \in F$ and $-h(x^k) = -k$ for all $k \geq 0$. Such a sequence exists since $\{h(x) \mid x \in F\}$ is an interval. This sequence must be unbounded because

$$\exp h(x) = \prod_{j=1}^n x_j \leq \left(\max_{1 \leq j \leq n} x_j \right)^n$$

and, therefore,

$$\max_{1 \leq j \leq n} x_j^k \geq \exp(k/n) \quad \text{for all } k \geq 0.$$

Take a subsequence of $\{x^k / \|x^k\|\}$ which converges to a $v \in \mathbb{R}^n$. This limit must have the properties $v \geq 0$ and $\|v\| = 1$. However, since $Ax^k = b$, $Ax^k / \|x^k\| = b / \|x^k\|$ and the unboundedness of $\{x^k\}$ implies that $\|x^k\| \rightarrow \infty$, thus, $Av = 0$ which shows that (7) does not hold.

Necessity. Suppose that x^* solves (4) and that there exists

$$v \geq 0, \quad v \neq 0 \quad \text{with } Av = 0.$$

Then, $x^* + tv \in F$ for $t \geq 0$, and $-h(x + tv) \searrow -\infty$, as t gets arbitrarily large, which is a contradiction. \square

In view of Proposition 1, we henceforth make the following assumptions on problem (4)

$$(A1) \quad a^i \neq 0 \text{ for all } i = 1, 2, \dots, m,$$

$$(A2) \quad F \neq \emptyset, \text{ and}$$

$$(A3) \quad N(A) \cap \mathbb{R}_+^n = \{0\}.$$

Next we derive our algorithm for solving (4). The underlying principle is the construction of a primal-dual algorithm, i.e., one in which both primal variables $\{x^k\}$

and dual variables $\{u^k\}$ are iteratively changed in such a way as to increase the dual function $\phi(u)$. However, as sometimes happens with primal-dual algorithms for linear equality constrained problems, the algorithm in its final form does not require that a dual sequence $\{u^k\}$ be actually updated (see, e.g., [32]).

The necessary conditions for the minimization of the Lagrangian (5) are

$$(8) \quad 0 = \frac{\partial L}{\partial x_j} = \frac{-1}{x_j} + \sum_{i=1}^m u_i a_j^i, \quad j = 1, 2, \dots, n.$$

Therefore, we impose the relation

$$(9) \quad x_j^k = \frac{1}{(A^T u^k)_j}, \quad j = 1, 2, \dots, n$$

on the primal and the dual iterates. We use cyclic coordinate ascent for the dual vectors, changing only one coordinate of u^k during each iteration, i.e.

$$(10) \quad u_i^{k+1} = \begin{cases} u_i^k, & i \neq i_k, \\ u_{i_k}^{k+1}, & i = i_k, \end{cases}$$

where $\{i_k\}_{k=0}^\infty$ is the *control sequence* of the algorithm. This is a sequence of indices according to which the algorithm works. In the k th iterative step, when u^{k+1} and x^{k+1} are calculated from u^k and x^k , the index i_k determines which coordinate of u^k will be changed, and which equation $\langle a^{i_k}, x \rangle = b_{i_k}$ will be used to update x^k .

The *cyclic control*, according to which our algorithm is controlled, is given by the sequence

$$(11) \quad i_k = k(\bmod m) + 1,$$

where m is the number of rows in A .

In view of (9) and (10) we write

$$(12) \quad x_j^{k+1} = \frac{1}{\sum_{i=1}^m u_i^{k+1} a_j^i} = \frac{x_j^k \cdot \sum_{i=1}^m u_i^k a_j^i}{\sum_{i=1}^m u_i^k a_j^i - B_k a_j^{i_k}},$$

where B_k is defined by

$$(13) \quad B_k := u_{i_k}^k - u_{i_k}^{k+1}.$$

Thus,

$$x_j^{k+1} = \frac{x_j^k}{1 - B_k a_j^{i_k} x_j^k},$$

or

$$(14) \quad \frac{1}{x_j^{k+1}} = \frac{1}{x_j^k} - B_k a_j^{i_k}, \quad j = 1, 2, \dots, n,$$

which is the final form of the iterative step for the primal iterates.

Equation (13) may be rewritten as

$$(15) \quad u^{k+1} = u^k - B_k e^{i_k}$$

where e^i stands for the i th standard basis-vector in \mathbb{R}^m with 1 in its i th coordinate and zeros elsewhere.

To complete the definition of our algorithm we define B_k by imposing the condition that

$$(16) \quad \langle a^{i_k}, x^{k+1} \rangle = b_{i_k}$$

which means that x^{k+1} has to fulfill the i_k th constraint equation. This leads to the notion of the "log x" *entropy projection* which is studied in the next section.

3. "log x" entropy projections onto hyperplanes. We define the concept of "log x" *entropy projection* as follows. Let $H = \{x \in \mathbb{R}^n \mid \langle a, x \rangle = \beta\}$ be a given hyperplane with $a = (a_j) \neq 0$. Denote

$$(17) \quad H^+ := H \cap \text{int } \mathbb{R}_+^n$$

and assume that $H^+ \neq \emptyset$. Let $y \in \text{int } \mathbb{R}_+^n$ be a given vector, and define, for every $j = 1, 2, \dots, n$,

$$(18) \quad \alpha_j := a_j y_j.$$

Assume that all $\alpha_j \neq 0$. This does not restrict the generality because, as seen below, if $\alpha_j = 0$ it can be deleted beforehand without affecting what we do.

Next define

$$(19) \quad r := \text{Max } \{1/\alpha_j \mid 1 \leq j \leq n, a_j < 0\};$$

if $a_j \geq 0$ for all j , set $r \equiv -\infty$, and

$$(20) \quad t := \text{Min } \{1/\alpha_j \mid 1 \leq j \leq n, a_j > 0\};$$

if $a_j \leq 0$ for all j , set $t \equiv \infty$.

DEFINITION 1. The "log x" *entropy projection* of y onto H^+ is defined as the vector y' obtained from the system

$$(21) \quad y'_j = \frac{y_j}{1 - B y_j a_j}, \quad j = 1, 2, \dots, n,$$

$$(22) \quad \langle a, y' \rangle = \beta,$$

where B obeys the additional restriction

$$(23) \quad r < B < t.$$

PROPOSITION 2. Given $H^+ \neq \emptyset$ as in (17) with $a \neq 0$ and a vector $y > 0$, the system (21)–(23) determines uniquely the vector $y' > 0$, which is the "log x" *entropy projection* of y onto H^+ , and the real number B which is called the projection parameter associated with the "log x" *entropy projection* of y onto H^+ .

Proof. The system (21), (22) gives rise to

$$(24) \quad \sum_{j=1}^n a_j \frac{y_j}{1 - B y_j a_j} = \beta$$

which leads us to consider the behavior of the real-valued function

$$(25) \quad g(s) := \sum_{j=1}^n \frac{\alpha_j}{(1 - s \alpha_j)} = \sum_{j=1}^n \left(\frac{1}{\alpha_j} - s \right)^{-1}$$

of the single variable s , $r < s < t$. Here we see that we may assume $\alpha_j \neq 0$ since $\alpha_j = 0$ contributes nothing to the first sum in (25) anyway. The derivative

$$(26) \quad g'(s) = \sum_{j=1}^n \left(\frac{1}{\alpha_j} - s \right)^{-2}$$

is always nonnegative, thus $g(s)$ is monotonically increasing. At the points $s_j = 1/\alpha_j$, $j = 1, 2, \dots, n$, $g(s)$ is discontinuous in such a manner that

$$(27) \quad g(s) = \beta$$

has exactly one solution s_j^* between each two consecutive s_j 's.

We verify the existence of a solution to (27). There are three cases:

Case (i). $-\infty < r < t < \infty$. In this case $\lim_{s \rightarrow r^+} g(s) = -\infty$ and $\lim_{s \rightarrow t^-} g(s) = +\infty$. Hence, by continuity of g over the open interval (r, t) , there is a solution to (27).

Case (ii). $-\infty = r < t < \infty$. In this case $a_j \geq 0$ for all j , $\lim_{s \rightarrow r^+} g(s) = 0$ and $\lim_{s \rightarrow t^-} g(s) = \infty$. But, by assumption, $H^+ \neq \emptyset$, so β in the definition of H must be positive, hence (27) has a (finite) solution.

Case (iii). $-\infty < r < t = \infty$. This case is handled in the same way as Case (ii) but with $\beta < 0$.

According to (23) we pick precisely that solution s^* such that $r < s^* < t$ and set

$$(28) \quad B = s^*.$$

To conclude the proof we verify that y' which is uniquely determined from (21) and (28) belongs to $\text{int } \mathbb{R}_+^n$. Indeed, if $a_j = 0$ then $y'_j = y_j > 0$. If $a_j > 0$ then, from (18), (20) and (28),

$$(29) \quad 1 - B y_j a_j > 0,$$

which shows, by (21), that $y'_j > 0$. If $a_j < 0$, then from (18), (19) and (28), again (29) holds. \square

4. Convergence of the algorithm. Based on the considerations set forth in the previous sections we study the following algorithm for the solution of the "log x " entropy optimization problem (4).

ALGORITHM.

Initialization: $x^0 > 0$ is a positive vector for which there exists a $u^0 \in \mathbb{R}^m$ such that

$$(30) \quad x_j^0 \cdot (A^T u^0)_j = 1, \quad j = 1, 2, \dots, n.$$

Iterative Step:

$$(31) \quad \frac{1}{x_j^{k+1}} = \frac{1}{x_j^k} - B_k a_j^{i_k}, \quad j = 1, 2, \dots, n$$

where B_k is the projection parameter associated with the "log x " entropy projection of x^k onto the set $H_{i_k}^+$ where

$$(32) \quad H_{i_k} = \{x \in \mathbb{R}^n \mid \langle a^{i_k}, x \rangle = b_{i_k}\},$$

i.e., B_k is the number which is uniquely determined from (31), (16) and

$$(33) \quad r_k < B_k < t_k$$

where

$$(34) \quad r_k := \text{Max} \left\{ \frac{1}{a_j^{i_k} x_j^k} \mid 1 \leq j \leq n, a_j^{i_k} < 0 \right\}$$

or $r_k = -\infty$ if $a_j^{i_k} \geq 0$ for all j , and

$$(35) \quad t_k := \text{Min} \left\{ \frac{1}{a_j^{i_k} x_j^k} \mid 1 \leq j \leq n, a_j^{i_k} > 0 \right\},$$

or $t_k = \infty$ if $a_j^{i_k} \leq 0$ for all j .

Control Sequence: The algorithm is controlled by a cyclic control sequence (11).

Remark 1. The initialization step of this algorithm requires that

$$(36) \quad \{u \in \mathbb{R}^m \mid A^T u > 0\} \neq \emptyset.$$

This follows from assumption (A3) by Gordan's transposition theorem. See, e.g., [36].

The following propositions lead to the conclusion about the convergence of this algorithm to the desired limit.

PROPOSITION 3. *A sequence $\{x^k\}$ produced by the algorithm has, for all $k \geq 0$, the properties (i) $x^k > 0$ and (ii) $x_j^k \cdot (A^T u^k)_j = 1, j = 1, 2, \dots, n$, where u^{k+1} is determined from (15).*

Proof. (i) Follows from Proposition 2. The conditions $H_i^+ \neq \emptyset$ and $a^i \neq 0, i = 1, 2, \dots, m$, necessary for the application of Proposition 2 follow from assumptions (A1) and (A2).

(ii) By induction. Initialization takes care of $k = 0$. Assume the claim is true up to k , then

$$\frac{1}{x_j^{k+1}} = (A^T u^k)_j - B_k a_j^{i_k} = [A^T (u^k - B_k e^{i_k})]_j = (A^T u^{k+1})_j. \quad \square$$

PROPOSITION 4. *The limit*

$$(37) \quad \lim_{k \rightarrow \infty} L(x^k, u^k),$$

where L is the Lagrangian defined in (5), $\{x^k\}$ is any sequence generated by the algorithm, and $\{u^k\}$ is its dual sequence determined from (15), exists and is finite.

Proof. Abbreviating

$$(38) \quad L_k := L(x^k, u^k) \quad \text{and} \quad d_k := L_{k+1} - L_k,$$

we have

$$(39) \quad d_k = -h(x^{k+1}) + \langle u^{k+1}, Ax^{k+1} - b \rangle + h(x^k) - \langle u^k, Ax^k - b \rangle.$$

From Proposition 3(ii) we see that (∇ stands for gradient),

$$(40) \quad A^T u^k = \nabla h(x^k)$$

for all $k \geq 0$. Therefore,

$$(41) \quad d_k = -h(x^{k+1}) + h(x^k) + \langle \nabla h(x^{k+1}), x^{k+1} \rangle - \langle \nabla h(x^k), x^k \rangle - \langle u^{k+1} - u^k, b \rangle.$$

Now, by (31),

$$(42) \quad \begin{aligned} & \langle \nabla h(x^{k+1}), x^{k+1} \rangle - \langle \nabla h(x^k), x^k \rangle \\ &= \langle \nabla h(x^{k+1}) - \nabla h(x^k), x^{k+1} \rangle + \langle \nabla h(x^k), x^{k+1} - x^k \rangle \\ &= -B_k \langle a^{i_k}, x^{k+1} \rangle + \langle \nabla h(x^k), x^{k+1} - x^k \rangle, \end{aligned}$$

and, by (15),

$$(43) \quad \langle u^{k+1} - u^k, b \rangle = \langle -B_k e^{i_k}, b \rangle.$$

Therefore, using also (16), we obtain

$$(44) \quad d_k = -h(x^{k+1}) + h(x^k) + \langle \nabla h(x^k), x^{k+1} - x^k \rangle$$

which is always nonnegative by Theorem 3.4.4 of [38], because $-h(x)$ is a convex function on the open convex set $\text{int } \mathbb{R}_+^n$. This proves that $\{L_k\}$ is monotonically increasing. To conclude the proof we show that it is also bounded from above. Take some fixed $z \in F$, then

$$(45) \quad \langle A^T u^k, z - x^k \rangle = \langle u^k, Az - Ax^k \rangle = \langle u^k, b - Ax^k \rangle.$$

By the same Theorem 3.4.4 of [38],

$$(46) \quad \Gamma := -h(z) + h(x^k) + \langle \nabla h(x^k), z - x^k \rangle \geq 0$$

for all $k \geq 0$. But, by (45), (40), (38) and (5),

$$(47) \quad \Gamma = -h(z) + h(x^k) + \langle u^k, b - Ax^k \rangle = -h(z) - L_k,$$

which proves that, for all $k \geq 0$,

$$(48) \quad L_k \leq -h(z). \quad \square$$

COROLLARY 1. $\lim_{k \rightarrow \infty} d_k = 0$.

Proof. We have the proof by (38) and Proposition 4. \square

Next we show that any sequence $\{x^k\}$ generated by the algorithm is both bounded and bounded away from zero.

DEFINITION 2. A set $S, S \subseteq \text{int } \mathbb{R}_+^n$, is called *bounded away from zero* if there exists a positive vector $q > 0$ such that

$$(49) \quad x \in S \Rightarrow x \geq q.$$

PROPOSITION 5. For any fixed $z \in \text{int } \mathbb{R}_+^n$ and any real α the set

$$(50) \quad \Omega(z, \alpha) := \left\{ x > 0 \left| \sum_{j=1}^n [\log x_j + (z_j/x_j)] \leq \alpha \right. \right\}$$

is (i) *bounded*, and (ii) *bounded away from zero*.

Proof. We prove (i) and (ii) together. Suppose (i) (respectively (ii)) was not true. Then for some α and some $z > 0$ we could have produced a sequence

$$(51) \quad \{x^k\}_{k=0}^\infty \subseteq \Omega(z, \alpha)$$

such that, for at least one component $l, 1 \leq l \leq n, \lim_{k \rightarrow \infty} x_l^k = +\infty$ (respectively, $\lim_{k \rightarrow \infty} x_l^k = 0^+$) would hold. But this would immediately contradict (50) because, for any fixed real positive γ and δ , the following limits hold:

$$(52) \quad \lim_{\xi \rightarrow +\infty} \left(\log \xi + \frac{\gamma}{\xi} \right) = +\infty$$

and

$$(53) \quad \lim_{\eta \rightarrow 0^+} \left(\log \eta + \frac{\delta}{\eta} \right) = +\infty. \quad \square$$

PROPOSITION 6. Any sequence $\{x^k\}$ generated by the algorithm is bounded and bounded away from zero.

Proof. Take a fixed $z \in F$. Then Γ (defined in (46)) may be written in the form

$$(54) \quad \Gamma = \sum_{j=1}^n \left[\log x_j^k + \frac{z_j}{x_j^k} \right] - h(z) - n.$$

On the other hand, by (47) and the monotonicity of $\{L_k\}$, proven in Proposition 4,

$$(55) \quad \Gamma = -h(z) - L_k \leq -h(z) - L_{k-1} \leq \dots \leq -h(z) - L_0.$$

Therefore, for any sequence $\{x^k\}$ generated by the algorithm,

$$(56) \quad \sum_{j=1}^n \left[\log x_j^k + \frac{z_j}{x_j^k} \right] \leq n - L_0 = \alpha,$$

and the proof is complete by Proposition 5. \square

Our next goal is to show that every cluster point x^* of any sequence $\{x^k\}$ generated by the algorithm is feasible for (4).

First, another property of "log x" entropy has to be checked. (It is similar, but not identical, to property (vi) in the definition of Bregman functions in [10, Def. 2.1].)

PROPOSITION 7. *Let $\{y^k\}_{k=0}^\infty$ be a convergent sequence in $\text{int } \mathbb{R}_+^n$ whose limit is $y^* > 0$ and let $\{x^k\}_{k=0}^\infty$ be a bounded sequence in $\text{int } \mathbb{R}_+^n$ which is bounded away from zero. If*

$$(57) \quad \lim_{k \rightarrow \infty} \left[\sum_{j=1}^n \left(\log y_j^k - \log x_j^k + \frac{x_j^k}{y_j^k} - 1 \right) \right] = 0,$$

then

$$(58) \quad \lim_{k \rightarrow \infty} x^k = y^*.$$

Proof. Take a convergent subsequence $x^{k_l} \rightarrow z$, as $l \rightarrow \infty$, $z > 0$. Then the limit in (57) is equal to

$$(59) \quad \sum_{j=1}^n \left[\log \frac{y_j^*}{z_j} + \frac{z_j}{y_j^*} - 1 \right],$$

in which each summand is nonnegative, because $\log x$ is convex, and therefore equal to zero, implying that $z_j = y_j^*$, $j = 1, 2, \dots, n$. \square

PROPOSITION 8. *If x^* is a cluster point of $\{x^k\}$ which is generated by the algorithm then $x^* \in F$.*

Proof. Let $x^{k_l} \rightarrow x^*$ as $l \rightarrow \infty$. By Proposition 6, $x^* > 0$. Use (44) to write

$$(60) \quad d_{k_l} = -h(x^{k_l+1}) + h(x^{k_l}) + \langle \nabla h(x^{k_l}), x^{k_l+1} - x^{k_l} \rangle.$$

Proposition 6 guarantees that $\{x^{k_l+1}\}$ is bounded and Corollary 1 ensures that $\lim_{l \rightarrow \infty} d_{k_l} = 0$. Now use Proposition 7 to deduce that $x^{k_l+1} \rightarrow x^*$. Repeating this argument leads to

$$(61) \quad x^{k_l+i} \xrightarrow{l \rightarrow \infty} x^* \quad \text{for every } i = 0, 1, 2, \dots, m.$$

The remainder of the proof is similar to the proof of [10, Step 5]. Consider the semi-infinite array which has $m+1$ rows, with $\{x^{k_l+i}\}_{l=1}^\infty$ in its i th row. We show that for each i , $1 \leq i \leq m$, some row of the array contains an infinite number of elements belonging to $H_i = \{x \in \mathbb{R}^n \mid \langle a^i, x \rangle = b_i\}$. For each i , at least one element in each column of the array must belong to H_i because $x^{k_l+1} \in H_{i_k}$ and the control sequence is cyclic.

It follows that there must be some row in the array, infinitely many elements of which belong to H_i because otherwise there would be only a finite number of such elements in the whole array. Thus, we can extract from the rows of the array subsequences, all of which converge to x^* , which belong to each of the H_i 's. Therefore, $x^* \in \bigcap_{i=1}^m H_i$ which, together with $x^* > 0$, proves that $x^* \in F$. \square

The convergence theorem now follows.

THEOREM 1. *If assumptions (A1), (A2) and (A3) hold then any sequence $\{x^k\}_{k=0}^\infty$ generated by the algorithm converges to the solution x^* of problem (4).*

Proof. Let x^* be a cluster point of $\{x^k\}$

$$(62) \quad x^{k_l} \xrightarrow{l \rightarrow \infty} x^*,$$

and, by Proposition 8, $x^* \in F$. Using Proposition 3(ii) write

$$(63) \quad \langle u^{k_l}, Ax^{k_l} - b \rangle = \sum_{j=1}^n \frac{1}{x_j^{k_l}} (x_j^{k_l} - x_j^*),$$

therefore, and since $\{x^{k_l}\}$ is bounded away from zero,

$$(64) \quad \lim_{l \rightarrow \infty} \langle u^{k_l}, Ax^{k_l} - b \rangle = 0.$$

This shows, using (62) and continuity of $h(x)$, that

$$(65) \quad \lim_{l \rightarrow \infty} L(x^{k_l}, u^{k_l}) = -h(x^*).$$

From (48), $L_{k_l} \leq -h(z)$ for any $z \in F$; thus, by (65),

$$(66) \quad -h(x^*) \leq -h(z),$$

which shows that x^* is optimal for problem (4). Since $[-h(x)]$ is strictly convex over $\text{int } \mathbb{R}_+^n$, x^* must be unique, which proves that $x^k \rightarrow x^*$ as $k \rightarrow \infty$. \square

Remark 2. An alternative proof of Theorem 1 was suggested by a referee. Proposition 6 guarantees the existence of a fixed real $\delta > 0$, which might depend on x^0, u^0 , such that

$$(67) \quad \inf_{k \geq 0} \min_{1 \leq j \leq n} x_j^k \geq \delta,$$

where $\{x^k\}$ is generated by the algorithm. Define the function $h_\delta: \mathbb{R} \rightarrow \mathbb{R}$ which quadratically extends $\log x$, by

$$(68) \quad h_\delta(x) := \begin{cases} \log x, & \delta/2 \leq x, \\ \log(\delta/2) + \frac{1}{4}(\log(\delta/2))^2 - [x - (\delta/2) - \frac{1}{2}(\log(\delta/2))]^2, & -\infty < x < \delta/2, \end{cases}$$

and form $h^\delta: \mathbb{R}^n \rightarrow \mathbb{R}$ by

$$(69) \quad h^\delta(x) := \sum_{j=1}^n h_\delta(x_j).$$

The function $-h^\delta(x)$ is a strongly zone consistent Bregman function with zone $S = \mathbb{R}^n$, as can be verified according to [10, Defs. 2.1 and 3.1]. By using the auxiliary function $h^\delta(x)$ and by applying Bregman's general method [10, Algorithm 4.1] to the problem

$$(70) \quad \begin{aligned} &\text{Min } [-h^\delta(x)] \\ &\text{subject to } Ax = b. \end{aligned}$$

Theorem 1 can be proved along the following lines. By Proposition 6, the sequence $\{x^k\}$, produced by the algorithm, is bounded. As such, it may be viewed as being generated by Bregman's method [10, Algorithm 4.1] applied to problem (70). Theorem 4.1 of [10] then guarantees that $\lim_{k \rightarrow \infty} x^k = x^*$ where x^* solves (70). The rest follows

as in the proof of Theorem 1. This proof eliminates the need for Propositions 7 and 8 by resorting to Bregman's theory as presented in [10]. Our proof, however, is self-contained.

5. Concluding remarks. We have proposed here a special-purpose algorithm for solving the "log x" entropy optimization problem over linear equality constraints. Although it is a primal-dual algorithm, there is no need in practice to construct and update a dual sequence $\{u^k\}$. All references made here to such a sequence were for the purpose of our theoretical study of this algorithm. As seen from (30)–(35), the new algorithm is of the *row-action* type as defined in [4]. This means that in each iterative step only the immediately previous iterate is needed, and access is required to *only one row of the system* of equations of the feasible set. As discussed at some length in [4] such properties usually make an algorithm potentially advantageous for handling large and sparse systems which arise in various fields of applications.

In image reconstruction from projections, for example, one might encounter such huge and sparse systems (see, e.g., [6]) that would make attempts to apply any general-purpose algorithm, which is not of the row-action type, practically infeasible.

These remarks are based on our familiarity with the results obtained by applying other row-action methods to problems of image reconstruction from projections. We do not have, so far, any computational experience with our new algorithm except for the results of Jain and Ranganath [27], who used it for a spectral estimation problem. Computational experience with this particular algorithm, as well as a study of its efficacy in practical situations, such as in image reconstruction from projections, are open and await investigation.

The significance of the study, presented here, of the convergence of the algorithm, is twofold. From the practical point of view it legitimizes the use of the algorithm as a tool in "log x" entropy optimization problems in fields of applications. On the theoretical side, it demonstrates that the general algorithm of Bregman [10, Algorithm 4.1] is applicable to functions outside the class \mathcal{B} of Bregman functions as described there.

Another recent study of Bregman's method is [13], where relaxation parameters are introduced into the method. In the present paper we do not discuss the possibility of incorporating such parameters. We also leave out some practical considerations such as the construction of x^0 and the calculation of B_k in practice.

Acknowledgments. We are grateful to Ms. Anna Cogan for preparing the manuscript. The constructive remarks of three anonymous referees are greatly appreciated.

REFERENCES

- [1] J. ACZÉL AND Z. DARÓCZY, *On Measures of Information and Their Characterization*, Academic Press, New York, 1975.
- [2] L. M. BREGMAN, *The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming*, U.S.S.R. Comput. Math. and Math. Phys., 7 (1967), pp. 200–217.
- [3] J. P. BURG, *Maximum entropy spectral analysis*, in Proceedings of the 37th Annual Meeting of the Society of Exploration Geophysicists, Oklahoma City, OK, 1967.
- [4] Y. CENSOR, *Row-action methods for huge and sparse systems and their applications*, SIAM Rev., 23 (1981), pp. 444–466.
- [5] ———, *Entropy optimization via entropy projections*, in System Modeling and Optimization, R. F. Drenick and F. Kozin, eds., Lecture Notes in Control and Information Science, 38, Springer-Verlag, Berlin, New York, 1982, pp. 450–454.

- [6] Y. CENSOR, *Finite series expansion reconstruction methods*, Proc. IEEE, 71 (1983), pp. 409–419.
- [7] Y. CENSOR, P. P. B. EGGERMONT AND D. GORDON, *Strong under-relaxation in Kaczmarz's method for inconsistent systems*, Numer. Math., 41 (1983), pp. 83–92.
- [8] Y. CENSOR, T. ELFVING AND G. T. HERMAN, *Methods for entropy maximization with applications in image processing*, in Proc. of the Third Scandinavian Conference on Image Analysis, P. Johansen and P. W. Becker, eds., Chartwell-Bratt, Lund, Sweden, 1983, pp. 296–300.
- [9] Y. CENSOR, A. V. LAKSHMINARAYANAN AND A. LENT, *Relaxational methods for large-scale entropy optimization problems, with applications in image reconstruction*, in Information Linkage Between Applied Mathematics and Industry, P. C. C. Wang et al., eds., Academic Press, New York, 1979, pp. 539–546.
- [10] Y. CENSOR AND A. LENT, *An iterative row-action method for interval convex programming*, J. Optim. Theory Appl., 34 (1981), pp. 321–353.
- [11] L. R. D'ADDARIO, *Maximum entropy imaging: theory and philosophy*, in Image Analysis and Evaluation, R. Shaw, ed., Society of Photographic Scientists and Engineers, Washington, D.C., 1977, pp. 221–225.
- [12] J. N. DARROCH AND D. RATCLIFF, *Generalized iterative scaling for log linear models*, Ann. Math. Statist., 43 (1972), pp. 1470–1480.
- [13] A. R. DE PIERRO AND A. N. IUSEM, *A relaxed version of Bregman's method for convex programming*, J. Optim. Theory Appl., to appear.
- [14] P. P. B. EGGERMONT, G. T. HERMAN AND A. LENT, *Iterative algorithms for large partitioned linear systems with applications to image reconstruction*, Linear Algebra Appl., 40 (1981), pp. 37–67.
- [15] T. ELFVING, *On some methods for entropy maximization and matrix scaling*, Linear Algebra Appl., 34 (1980), pp. 321–339.
- [16] S. ERLANDER, *Entropy in linear programs*, Math. Programming, 21 (1981), pp. 137–151.
- [17] B. R. FRIEDEN, *Image enhancement and restoration*, in Picture Processing and Digital Filtering, T. S. Huang, ed., Topics in Applied Physics, Vol. 6, Springer-Verlag, Berlin, New York, Heidelberg, 1975, pp. 177–248.
- [18] ———, *Statistical models for the image restoration problem*, Comput. Graphics Image Process., 12 (1980), pp. 40–59.
- [19] ———, *Probability, Statistical Optics, and Data Testing*, Springer-Verlag, Berlin, New York, Heidelberg, 1983.
- [20] R. GORDON, R. BENDER AND G. T. HERMAN, *Algebraic reconstruction techniques (ART) for three-dimensional electron microscopy and X-ray photography*, J. Theoret. Biol., 29 (1970), pp. 471–481.
- [21] G. T. HERMAN, *A relaxation method for reconstructing objects from noisy X-rays*, Math. Programming, 8 (1975), pp. 1–19.
- [22] ———, *Image Reconstruction from Projections: The Fundamentals of Computerized Tomography*, Academic Press, New York, 1980.
- [23] ———, *Mathematical optimization versus practical performance: A case study based on the maximum entropy criterion in image reconstruction*, Math. Programming Stud., 20 (1982), p. 96–112.
- [24] ———, *Application of maximum entropy and Bayesian optimization methods to image reconstruction from projections*, in Maximum-Entropy and Bayesian Methods in Inverse Problems, C. R. Smith and W. T. Grandy, Jr., eds., D. Reidel, Dordrecht, Holland, 1985, pp. 319–338.
- [25] G. T. HERMAN AND A. LENT, *A family of iterative quadratic optimization algorithms for pairs of inequalities, with application in diagnostic radiology*, Math. Programming Stud., 9 (1978), pp. 15–29.
- [26] C. HILDRETH, *A quadratic programming procedure*, Naval Res. Logist. Quart., 4 (1957), pp. 79–85, Erratum, *ibid.*, p. 361.
- [27] A. K. JAIN AND S. RANGANATH, *Two-dimensional spectral estimation*, Proceedings of the RADC (Rome Air Development Center) Workshop on Spectral Estimation, held at RADC, Griffis Air Base, Rome, NY, May 1978, pp. 151–157.
- [28] E. T. JAYNES, *On the rationale of maximum-entropy methods*, Proc. IEEE, 70 (1982), pp. 939–952.
- [29] R. JOHNSON AND J. E. SHORE, *Which is the better entropy expression for speech processing: $-S \log S$ or $\log S''$* , IEEE Trans. Acoust. Speech Signal Process. ASSP-32 (1984), pp. 129–136.
- [30] J. N. KAPUR, *Twenty-five years of maximum-entropy principle*, J. Math. Phys. Sci., 17 (1983), pp. 103–156.
- [31] B. LAMOND AND N. P. STEWART, *Bregman's balancing method*, Transportation Res., Part B, 15 (1981), pp. 239–248.
- [32] A. LENT, *A convergent algorithm for maximum entropy image restoration with a medical X-ray application*, in Image Analysis and Evaluation, R. Shaw, ed., Society of Photographic Scientists and Engineers, Washington, D.C., 1977, pp. 249–257.
- [33] A. LENT AND Y. CENSOR, *Extensions of Hildreth's row-action method for quadratic programming*, this Journal, 18 (1980), pp. 444–454.

- [34] R. D. LEVINE AND M. TRIBUS, EDS., *The Maximum Entropy Formalism*, MIT Press, Cambridge, MA, 1978.
- [35] D. G. LUENBERGER, *Introduction to Linear and Nonlinear Programming*, Addison-Wesley, Reading, MA, 1973.
- [36] O. L. MANGASARIAN, *Nonlinear Programming*, McGraw-Hill, New York, 1969.
- [37] G. N. MINERBO, *MENT: A maximum entropy algorithm for reconstructing a source from projection data*, Comput. Graphics Image Process., 10 (1979), pp. 48–68.
- [38] J. M. ORTEGA AND W. C. RHEINHOLDT, *Iterative Solution of Nonlinear Equations in Several Variables*, Academic Press, New York, 1970.
- [39] S. J. WERNECKE, *Maximum entropy techniques for digital image reconstruction*, in Image Analysis and Evaluation, R. Shaw, ed., Society of Photographic Scientists and Engineers, Washington, D.C., 1977, pp. 238–243.
- [40] D. S. WONG, *Maximum likelihood, entropy maximization, and the geometric programming approaches to the calibration of trip distribution models*, Transportation Res., Part B, 15 (1981), pp. 329–343.

A SUPERLINEARLY CONVERGENT FEASIBLE METHOD FOR THE SOLUTION OF INEQUALITY CONSTRAINED OPTIMIZATION PROBLEMS*

ELIANE R. PANIER† AND ANDRÉ L. TITS†

Abstract. When iteratively solving optimization problems arising from engineering design applications, it is sometimes crucial that all iterates satisfy a given set of “hard” inequality constraints, and generally desirable that the objective function value improve at each iteration. In this paper, we propose an algorithm of the successive quadratic programming (SQP) type which, unlike other algorithms of this type, does enjoy such properties. Under mild assumptions, the new algorithm is shown to converge from any initial point, locally superlinearly. Numerically tested, it has proven to be competitive with the most successful currently available nonlinear programming algorithms, while the latter do not exhibit the desired properties.

Key words. constrained optimization, successive quadratic programming, superlinear convergence, engineering design

AMS(MOS) subject classifications. 90C30, 65K10

1. Introduction. While some of the specifications associated with engineering design problems can often be relaxed, others, such as stability or physical realizability, have to be met imperatively (see [13] for a discussion of optimization problems arising from design problems). The former type of specification calls for tradeoff exploration through close interaction between designer and design process. However, this tradeoff exploration can meaningfully take place only once the latter specifications are satisfied. Since each iteration of an optimization algorithm involves one or more function evaluations and since typically, in a design environment, function evaluations call for computationally expensive system simulations, it is essentially required that hard constraints be satisfied *at each iteration*.¹ It is also desirable that the design obtained after each iteration improve on the previous one.

In the simplest case, a design problem can be formulated as

$$(P) \begin{cases} \min f(x) \\ \text{s.t. } x \in X \end{cases}$$

where $X = \{x \text{ s.t. } g_j(x) \leq 0, j = 1, \dots, m\}$ and $f: \mathbb{R}^n \rightarrow \mathbb{R}$ and $g_j: \mathbb{R}^n \rightarrow \mathbb{R}, j = 1, \dots, m$, are smooth functions. For this optimization problem, the stipulations outlined above amount to the requirement that, given $x_0 \in X$, the optimization algorithm construct a sequence $\{x_k\}_{k=0}^\infty$ such that, for all k ,

$$(1.1) \quad x_k \in X$$

and

$$(1.2) \quad f(x_{k+1}) \leq f(x_k).$$

* Received by the editors August 9, 1985; accepted for publication (in revised form) May 13, 1986. This work was supported by National Science Foundation grants DMC-84-20740 and CDR-85-00108, by a grant from the Minta Martin Foundation, College of Engineering, University of Maryland, by a grant from the Engineering Research Center at the University of Maryland, and by a grant from the Westinghouse Corporation.

† Electrical Engineering Department and Systems Research Center, University of Maryland, College Park, Maryland 20742.

¹ In fact some simulation programs (such as SPICE2 [10]) will refuse parameter values which violate some physical realizability constraints.

Methods of feasible directions [23], [14] satisfy these two requirements. They have been extended to handle problems with functional constraints [3], [15] and multiple objectives [13] and enhanced to efficiently handle design problems [21]. They have been used very successfully in solving engineering design problems arising in diverse application areas [12], [1], [11]. However, they suffer from an important shortcoming in that they are generally *slow*, as their rate of convergence is at best linear.

This paper presents an algorithm which enjoys properties (1.1) and (1.2) as well as a superlinear rate of convergence. This algorithm is of the successive quadratic programming (SQP) type. Successive quadratic programming algorithms were first introduced by Wilson [22]. Subsequently, Robinson [20] showed that Wilson's method is locally quadratically convergent and that it can be viewed as a form of Newton's method for solving the first order necessary conditions of optimality for constrained nonlinear programming problems. The question of global convergence and Hessian approximation were then considered by a number of authors (see e.g. [4], [2]). Numerical experiments have shown that these methods (in particular a version due to Powell [17]) often dramatically outperform algorithms of other classes [6]. However, existing SQP type algorithms do not enjoy properties (1.1) and (1.2).

Given an estimate $x \in X$ of the solution x^* to problem (P) and an estimate H of the Hessian of the Lagrangian at x^* , the SQP iteration yields a search direction d^0 given by the solution of the quadratic program

$$(1.3) \quad \begin{aligned} \min \quad & \frac{1}{2} d^T H d + \langle \nabla f(x), d \rangle \\ \text{s.t.} \quad & g_j(x) + \langle \nabla g_j(x), d \rangle \leq 0, \quad j = 1, \dots, m. \end{aligned}$$

Let us assume for the time being that H is positive definite. Then clearly d^0 is a descent direction for f at x , since, using the first order condition of optimality for (1.3), we get

$$(1.4) \quad \begin{aligned} \langle \nabla f(x), d^0 \rangle &= -\langle H d^0, d^0 \rangle - \sum_j \mu_j \langle \nabla g_j(x), d^0 \rangle \\ &= -\langle H d^0, d^0 \rangle + \sum_j \mu_j g_j(x) \\ &\leq -\rho |d^0|^2 \end{aligned}$$

for some positive ρ and some nonnegative multipliers μ_j . However d^0 may not be a feasible direction at x , since the constraints in (1.3) merely imply, for the constraints active at x ,

$$\langle \nabla g_j(x), d^0 \rangle \leq 0$$

and thus property (1.1) may not be satisfied. Feasibility is recovered if one substitutes in the right-hand side of the constraints of (1.3) a negative number $-\varepsilon$. However the new solution d^1 may not be any more a descent direction for f , thus jeopardizing property (1.2). Indeed, (1.4) now becomes

$$\langle \nabla f(x), d^1 \rangle \leq -\rho |d^1|^2 + \varepsilon \sum_j \mu_j.$$

Choosing $\varepsilon = |d^1|^\nu$, with $\nu > 2$, would resolve this difficulty, at least for d^1 small, which is the case if x is close to the solution of (P). Unfortunately the transformed problem would not be a quadratic program any more. Following an idea used by Herskovits in a different context [5], we propose to solve successively two quadratic programs: first (1.3), giving d^0 , then

$$\begin{aligned} \min \quad & \frac{1}{2} d^T H d + \langle \nabla f(x), d \rangle \\ \text{s.t.} \quad & g_j(x) + \langle \nabla g_j(x), d \rangle \leq -|d^0|^\nu, \quad j = 1, \dots, m \end{aligned}$$

yielding a search direction d^1 . The hope is that d^1 will be small enough, and close enough to d^0 , near the solution to (P), for property (1.2) as well as the basic convergence properties of the SQP type algorithms to be preserved. As shown in a later section, such will indeed be the case, even without assuming positive definiteness of H over the entire space (the milder assumption (4.1) will be used instead).

Once a feasible descent direction is obtained, an Armijo type rule may be suitable as a line search procedure. However, in order to preserve a superlinear rate of convergence, it is necessary to avoid any Maratos-like effect [7], by which the step length is truncated even close to the solution. Mayne and Polak [9] solve this problem in a different context—SQP methods using a penalty function for the stepsize calculation—by replacing the line search by a search along a suitably defined arc, tangent to d^1 at x . In our context, a further “bending” towards the feasible region is necessary to avoid truncation of the step due to infeasibility. It turns out that the amount of bending must be closely monitored. Indeed, the bent unit step, say, $d^1 + \tilde{d}$, must be very close to d^1 when d^1 is small (in the neighborhood of a solution of (P)). Otherwise, $d^1 + \tilde{d}$ may not inherit enough descent properties from d^1 , resulting again in a truncated step. Also, if \tilde{d} is too large, even the unit step iteration may not yield superlinear convergence. A suitable correction \tilde{d} will be obtained as the solution of a linear least squares problem.

The last problem to be addressed is that of global convergence. As suggested above, d^1 is guaranteed to be a descent direction for f only in the neighborhood of a solution to (P). Away from a solution, a first order search direction will be used. A suitable mechanism will ensure that our algorithm selects the SQP direction when a solution is approached, so that superlinear convergence can occur.

The resulting algorithm is relatively complex, as it involves the solution of two quadratic programs and of one linear least squares problem at most iterations. Clearly however, the close relationship between these three problems should result, in a clever implementation, in little more computational effort than that required for the solution of a single quadratic program.

The remainder of this paper is organized as follows. The proposed algorithm is stated in § 2. In § 3, it is shown that, under mild assumptions, this algorithm is convergent irrespective of the initial guess. Rate of convergence analysis is the object of § 4, where conditions for superlinear convergence are put forth. Finally, § 5 is devoted to implementation aspects and to numerical experiments.

2. The algorithm. Throughout the paper, the following two hypotheses will be assumed to hold.

H1. The set X is not empty;

H2. The functions $f, g_j, j = 1, \dots, m$ are continuously differentiable.

The algorithm we propose for solving (P) is as follows.

ALGORITHM A.

Parameters.

$$M > 0, \quad \alpha \in (0, \tfrac{1}{2}), \quad \beta \in (0, 1), \quad \nu > 2, \quad \kappa > 2, \quad \tau \in (2, 3).$$

Data.

$$x_0 \in X, \quad H_0 \in \mathbb{R}^{n \times n}.$$

Step 0. Initialization.

Set $k = 0$.

Step. 1. Computation of a search direction.

(i) Solve

$$(QP_0) \begin{cases} \min \frac{1}{2} d^T H_k d + \langle \nabla f(x_k), d \rangle \\ \text{s.t. } g_j(x_k) + \langle \nabla g_j(x_k), d \rangle \leq 0, \quad j = 1, \dots, m \end{cases}$$

to the extent of obtaining a Kuhn-Tucker point d_k^0 of least norm.

If (QP_0) has no Kuhn-Tucker point or if $|d_k^0| > M$ or if $|H_k d_k^0| > |d_k^0|^{1/2}$, go to (iv).

If $|d_k^0| = 0$ stop.

(ii) Solve

$$(QP) \begin{cases} \min \frac{1}{2} d^T H_k d + \langle \nabla f(x_k), d \rangle \\ \text{s.t. } g_j(x_k) + \langle \nabla g_j(x_k), d \rangle \leq -|d_k^0|^\nu, \quad j = 1, \dots, m \end{cases}$$

to the extent of obtaining a Kuhn-Tucker point d_k of least norm.

If d_k exists, set $\theta_k = \langle \nabla f(x_k), d_k \rangle$.

If (QP) has no Kuhn-Tucker point or if $|d_k| > M$ or if $\theta_k > \min(-|d_k^0|^\kappa, -|d_k|^\kappa)$, go to (iv).

(iii) Compute a correction \tilde{d}_k , solution of the linear least squares problem

$$(LS) \begin{cases} \min \frac{1}{2} |d|^2 \\ \text{s.t. } g_j(x_k + d_k) + \langle \nabla g_j(x_k), d \rangle = -|d_k^0|^\tau \quad \forall j \in I_k \end{cases}$$

where $I_k = \{j \text{ s.t. } g_j(x_k) + \langle \nabla g_j(x_k), d_k \rangle = -|d_k^0|^\nu\}$.

If (LS) has no solution or if $|\tilde{d}_k| > |d_k|$, set $\tilde{d}_k = 0$.

Proceed to Step 2.

(iv) Compute a first order feasible descent direction d_k (see remark below).

Set $\theta_k = \langle \nabla f(x_k), d_k \rangle$.

Set $\tilde{d}_k = 0$.

Step 2. Line search.

Compute t_k , the first number t of the sequence $\{1, \beta, \beta^2, \dots\}$ satisfying

$$(2.1) \quad f(x_k + t d_k + t^2 \tilde{d}_k) \leq f(x_k) + \alpha t \theta_k,$$

$$(2.2) \quad g_j(x_k + t d_k + t^2 \tilde{d}_k) \leq 0, \quad j = 1, \dots, m.$$

Step 3. Updates.

Compute a new approximation H_{k+1} of the Hessian matrix.

Set $x_{k+1} = x_k + t_k d_k + t_k^2 \tilde{d}_k$.

Set $k = k + 1$.

Go back to Step 1. □

Remark. The “first order” direction of Step 1(iv) is any direction satisfying a set of conditions that will be stated later, as the need arises. At this time, let us just point out that algorithms do exist that construct directions satisfying these conditions (e.g. the algorithm in [16] using optimality function θ_ε^2 defined by equation (36) in that paper).

3. Global convergence. In this section we prove that, under mild conditions, the algorithm described in § 1 is convergent.

In addition to H1 and H2, we will assume that the following hypothesis holds.

H3. For any $x \in X$, the vectors $\{\nabla g_j(x), j \in I(x)\}$ are linearly independent, where

$$I(x) \triangleq \{j \text{ s.t. } g_j(x) = 0\}.$$

Before analyzing the convergence properties of Algorithm A, we need to verify that the line search of Step 2 is well defined. A first requirement on the first order direction is needed here.

R1. The direction computed at Step 1(iv) of the algorithm is a strict descent direction for f and for the active constraints associated with the current iterate (i.e., $\langle \nabla f(x_k), d_k \rangle < 0$ and $\langle \nabla g_j(x_k), d_k \rangle < 0$ for all $j \in I(x_k)$).

PROPOSITION 3.1. *The line search yields a step $t_k = \beta^j$ for some finite $j = j(k)$.*

Proof. This is a well-known result in the case when the direction is computed at Step 1(iv) and R1 is satisfied. Thus suppose that the direction is computed through Step 1(i)–(iii). We have

$$f(x_k + td_k + t^2 \tilde{d}_k) = f(x_k) + t \langle \nabla f(x_k + \xi d_k + \xi^2 \tilde{d}_k), d_k + 2\xi \tilde{d}_k \rangle$$

for some $\xi \in [0, t]$. Since f is continuously differentiable, $\theta_k = \langle \nabla f(x_k), d_k \rangle < 0$ (from Step 1(ii)), and $\alpha \in (0, \frac{1}{2})$, there exists $\underline{t} > 0$ such that

$$f(x_k + td_k + t^2 \tilde{d}_k) \leq f(x_k) + t\alpha\theta_k \quad \forall t \in [0, \underline{t}].$$

We also have

$$g_j(x_k + td_k + t^2 \tilde{d}_k) = g_j(x_k) + t \langle \nabla g_j(x_k + \xi d_k + \xi^2 \tilde{d}_k), d_k + 2\xi \tilde{d}_k \rangle$$

for some $\xi \in [0, t]$. Moreover, from the inequalities

$$g_j(x_k) + \langle \nabla g_j(x_k), d_k \rangle \leq -|d_k^0|^\nu < 0, \quad j = 1, \dots, m$$

and

$$g_j(x_k) \leq 0, \quad j = 1, \dots, m,$$

we conclude that either $g_j(x_k) < 0$ or $g_j(x_k) = 0$ and $\langle \nabla g_j(x_k), d_k \rangle < 0$. Therefore, for $j = 1, \dots, m$, there exists some \underline{t}_j such that

$$g_j(x_k + td_k + t^2 \tilde{d}_k) \leq 0 \quad \forall t \in [0, \underline{t}_j]. \quad \square$$

It is of interest to note that this result was obtained without making use of any property of \tilde{d}_k .

Our first convergence result has to do with the sequence of intermediate directions $\{d_k^0\}$.

PROPOSITION 3.2. *Suppose that Algorithm A generates an infinite sequence. Let x^* be a cluster point of this sequence, and $\{x_k\}_{k \in K}$ a subsequence converging to x^* . Suppose moreover that the directions at points x_k , for $k \in K$, are computed through Step 1(i)–(iii). Then, the sequence $\{d_k^0\}_{k \in K}$ tends to zero.*

Proof. We assume by contradiction that there exists a cluster point x^* , a number $\underline{d} > 0$ and subsequences $\{x_k\}_{k \in K}$ and $\{d_k^0\}_{k \in K}$ such that

$$x_k \rightarrow x^*, \quad k \in K, \quad k \rightarrow \infty$$

and

$$|d_k^0| \geq \underline{d} \quad \forall k \in K.$$

We first show that, in that case, the step t_k obtained by the line search is bounded away from zero on K , i.e.,

$$(3.1) \quad \exists \underline{t} > 0 \text{ s.t. } t_k \geq \underline{t} \quad \forall k \in K.$$

From Step 1(ii) we have, for $k \in K$,

$$\theta_k = \langle \nabla f(x_k), d_k \rangle \leq -(\underline{d})^\kappa$$

and

$$g_j(x_k) + \langle \nabla g_j(x_k), d_k \rangle \leq -(\underline{d})^\nu.$$

Then, for $k \in K$, k large enough, we obtain,

$$\langle \nabla f(x_k), d_k \rangle \leq -\delta$$

and

$$\begin{aligned} \langle \nabla g_j(x_k), d_k \rangle &\leq -\delta \quad \forall j \in I(x^*), \\ g_j(x_k) &\leq -\delta \quad \forall j \notin I(x^*) \end{aligned}$$

for some $\delta > 0$. From the identity

$$f(x_k + td_k + t^2 \tilde{d}_k) = f(x_k) + \int_0^1 \langle \nabla f(x_k + t\xi d_k + t^2 \xi^2 \tilde{d}_k), td_k + 2t^2 \xi \tilde{d}_k \rangle d\xi,$$

it then follows that, for $k \in K$, k large enough,

$$\begin{aligned} f(x_k + td_k + t^2 \tilde{d}_k) - f(x_k) - \alpha t \theta_k \\ \leq t \left\{ \int_0^1 [\langle \nabla f(x_k + t\xi d_k + t^2 \xi^2 \tilde{d}_k), d_k + 2t\xi \tilde{d}_k \rangle - \langle \nabla f(x_k), d_k \rangle] d\xi \right. \\ \left. + (1 - \alpha) \langle \nabla f(x_k), d_k \rangle \right\} \\ \leq t \left\{ \sup_{\xi \in [0,1]} |\nabla f(x_k + t\xi d_k + t^2 \xi^2 \tilde{d}_k) - \nabla f(x_k)| |d_k| \right. \\ \left. + 2t \sup_{\xi \in [0,1]} |\nabla f(x_k + t\xi d_k + t^2 \xi^2 \tilde{d}_k)| |\tilde{d}_k| - (1 - \alpha) \delta \right\}. \end{aligned}$$

Since d_k and \tilde{d}_k are bounded and $f \in C^1$, this ensures that there exists $t_f > 0$, independent of k , such that for $t \in [0, t_f]$, $k \in K$, k large enough,

$$f(x_k + td_k + t^2 \tilde{d}_k) - f(x_k) - \alpha t \theta_k \leq 0.$$

Similarly, for $k \in K$, k large enough, $t > 0$ and $j \in I(x^*)$, it holds

$$\begin{aligned} g_j(x_k + td_k + t^2 \tilde{d}_k) - g_j(x_k) &\leq t \left\{ \sup_{\xi \in [0,1]} |\nabla g_j(x_k + t\xi d_k + t^2 \xi^2 \tilde{d}_k) - \nabla g_j(x_k)| |d_k| \right. \\ &\quad \left. + 2t \sup_{\xi \in [0,1]} |\nabla g_j(x_k + t\xi d_k + t^2 \xi^2 \tilde{d}_k)| |\tilde{d}_k| - \delta \right\} \end{aligned}$$

so that there exists some $t_j > 0$ independent of k such that, for $t \in [0, t_j]$, $k \in K$, k large enough,

$$g_j(x_k + td_k + t^2 \tilde{d}_k) \leq 0.$$

Also, there exists $t_j > 0$ independent of k such that, for $t \in [0, t_j]$, $k \in K$, k large enough, and $j \notin I(x^*)$,

$$g_j(x_k + td_k + t^2 \tilde{d}_k) \leq g_j(x_k) + \frac{\delta}{2} \leq -\delta + \frac{\delta}{2} \leq 0 \quad \forall t \in [0, t_j].$$

Our claim (3.1) is thus proven, with $\underline{t} = \min \{t_f, t_j, j = 1, \dots, m\}$.

Now, for $k \in K$, k large enough, we have,

$$\begin{aligned} (3.2) \quad f(x_{k+1}) &\leq f(x_k) + \alpha t_k \theta_k \\ &\leq f(x_k) - \alpha \underline{t} \delta. \end{aligned}$$

On the other hand, from (2.1), the sequence $\{f(x_k)\}$ is monotonically decreasing and hence, since f is continuous, $f(x_k) \rightarrow f(x^*)$ as $k \rightarrow \infty$. This contradicts (3.2). \square

In order to prove global convergence of Algorithm A, we need to strengthen the first requirement on the first order direction, replacing it with R1'.

R1'. If a subsequence $\{x_k\}_{k \in K}$ converges to a point x^* which is not a Kuhn-Tucker point for problem (P), then the corresponding first order directions are bounded and satisfy the inequalities

$$\langle \nabla f(x_k), d_k \rangle \leq -\delta,$$

$$\langle \nabla g_j(x_k), d_k \rangle \leq -\delta \quad \forall i \in I(x^*)$$

for all $k \in K$, for some $\delta > 0$.

THEOREM 3.3. *Algorithm A described in § 2 either stops at a Kuhn-Tucker point or generates a sequence $\{x_k\}$ for which each accumulation point is a Kuhn-Tucker point for (P).*

Proof. The first statement is obvious, the only stopping point being in Step 1(i). Thus, suppose that $\{x_k\}_{k \in K} \rightarrow x^*$. If the first order direction is selected infinitely many times, the result follows from an argument identical to that used in the proof of Proposition 3.2, using requirement R1' and the fact that the function f is monotonically decreasing. We then suppose, without loss of generality, that the direction is always computed through Step 1(i)-(iii) on K and that the active set associated with (QP_0) keeps a constant value

$$I = I_k = \{j \text{ s.t. } g_j(x_k) + \langle \nabla g_j(x_k), d_k^0 \rangle = 0\} \quad \forall k \in K.$$

From Proposition 3.2, we have

$$d_k^0 \rightarrow 0, \quad k \in K, \quad k \rightarrow \infty.$$

Therefore, $I \subset I(x^*)$. Also, the vector d_k^0 satisfies the optimality conditions

$$(3.3a) \quad H_k d_k^0 + \nabla f(x_k) + \sum_{j \in I} (\mu_k)_j \nabla g_j(x_k) = 0,$$

$$(3.3b) \quad (\mu_k)_j \geq 0,$$

for some multiplier vector μ_k . Because $I \subset I(x^*)$, for $k \in K$, k large enough, the vectors $\nabla g_j(x_k)$, $j \in I$ are linearly independent. If we denote by $R_I(x_k)$ the $n \times |I|$ matrix

$$R_I(x_k) = (\nabla g_j(x_k) \text{ s.t. } j \in I)$$

we obtain the expression of the unique multiplier vector μ_k as

$$\mu_k = -(R_I^T(x_k) R_I(x_k))^{-1} R_I(x_k)^T (H_k d_k^0 + \nabla f(x_k)).$$

Due to the condition $|H_k d_k^0| \leq |d_k^0|^{1/2}$ we obtain

$$\mu_k \rightarrow \mu^*, \quad k \in K, \quad k \rightarrow \infty$$

with

$$\mu^* = -(R_I^T(x^*) R_I(x^*))^{-1} R_I(x^*)^T \nabla f(x^*).$$

Taking the limit in (3.3) yields

$$\nabla f(x^*) + \sum_{j \in I} \mu_j^* \nabla g_j(x^*) = 0, \quad \mu_j^* \geq 0. \quad \square$$

We conclude this section by showing that the existence of an accumulation point in the sequence generated by Algorithm A induces some regularity properties on this

sequence. This result will be used in § 4. We first need to introduce a second and last requirement on the first order direction.²

R2. The first order direction satisfies the relation $\langle \nabla f(x_k), d_k \rangle \leq -c|d_k|^\alpha$ for some $\alpha \geq 1$ and $c > 0$.³

PROPOSITION 3.4. *Suppose that the sequence $\{x_k\}$ generated by Algorithm A has some accumulation point. Then*

$$\{|x_{k+1} - x_k|\} \rightarrow 0, \quad k \rightarrow \infty.$$

Proof. Since $f(x_k)$ is monotonically decreasing, existence of an accumulation point of $\{x_k\}$ and continuity of f imply that the sequence $\{f(x_k)\}$ is bounded. Also, the line search in Algorithm A yields

$$f(x_{k+1}) \leq f(x_k) + \alpha t_k \theta_k.$$

It follows that

$$(3.4) \quad t_k \theta_k \rightarrow 0, \quad k \rightarrow \infty.$$

Now $t_k \theta_k$ is bounded from above by $-t_k |d_k|^\kappa$ if the direction is computed through Step 1(i)–(iii) and by $-ct_k |d_k|^\alpha$ if the direction is computed through Step 1(iv). Thus, in both cases, (3.4) and the fact that the step t_k is bounded by 1 imply

$$|t_k d_k| \rightarrow 0, \quad k \rightarrow \infty.$$

Since

$$\begin{aligned} |x_{k+1} - x_k| &\leq t_k |d_k| + t_k^2 |\tilde{d}_k| \\ &\leq 2t_k |d_k| \end{aligned}$$

the claim holds. \square

4. Rate of convergence. In order to study the rate of convergence of the algorithm, we need some stronger regularity assumptions on the functions involved in problem (P). We replace H2 by the following hypothesis.

H2'. The functions $f, g_j, j = 1, \dots, m$ are three times continuously differentiable. Hypotheses H1 and H3 are still assumed to hold.

Let x^* be a Kuhn–Tucker point for (P). Denote by μ^* the unique multiplier vector computed at x^* and, for any $x \in \mathbb{R}^n$ and $\mu \in \mathbb{R}^m$, denote by $L(x, \mu)$ the Lagrangian function

$$L(x, \mu) = f(x) + \sum_j (\mu)_j g_j(x).$$

The optimality conditions associated with x^* can then be written

$$\begin{aligned} \nabla_x L(x^*, \mu^*) &= 0, \\ \mu^* &\geq 0, \quad g_j(x^*) \leq 0, \quad j = 1, \dots, m, \\ (\mu^*)_j g_j(x^*) &= 0, \quad j = 1, \dots, m. \end{aligned}$$

The point x^* is said to satisfy *second order sufficiency conditions with strict complementary slackness* if the multipliers satisfy $\mu_j^* > 0$ for all $j \in I(x^*)$ and if the Hessian of the Lagrangian function $\nabla_{xx} L(x^*, \mu^*)$ is positive definite on the subspace $\{p \text{ s.t. } \langle \nabla g_j(x^*), p \rangle = 0 \text{ for all } j \in I(x^*)\}$.

² We could replace R2 by any condition sufficient for Proposition 3.4 to hold.

³ The direction computed by the method described in [16, eq. (36)] satisfies R2 with $\alpha = 2$ and $c = 1$.

PROPOSITION 4.1. *If some accumulation point x^* of the sequence generated by the algorithm satisfies the second order sufficiency conditions with strict complementary slackness, then the entire sequence converges to x^* .*

Proof. Under the stated assumptions, the Kuhn–Tucker point x^* is isolated (see, e.g., [19]), i.e., for some $\varepsilon > 0$, the ball $B(x^*, \varepsilon)$ does not contain any Kuhn–Tucker point other than x^* . From Proposition 3.4, we have

$$|x_{k+1} - x_k| \rightarrow 0, \quad k \rightarrow \infty.$$

Therefore, for k large enough, $|x_{k+1} - x_k| < \varepsilon/4$ and there exists a subsequence $\{x_k\}_{k \in K}$ such that $|x_k - x^*| < \varepsilon/4$ on K . It is then impossible to leave $B(x^*, \varepsilon)$ without creating another cluster point and hence a Kuhn–Tucker point in that ball. \square

In the sequel, we will assume that the sequence generated by the algorithm converges to such a point x^* . We will denote by R^* and P^* the $n \times |I(x^*)|$ and $n \times n$ matrices, respectively, defined by

$$\begin{aligned} R^* &= \{\nabla g_j(x^*), j \in I(x^*)\}, \\ P^* &= I - R^*(R^{*T}R^*)^{-1}R^{*T}. \end{aligned}$$

Given some iterate x_k close enough to x^* , we will similarly define matrices R_k and P_k by

$$\begin{aligned} R_k &= \{\nabla g_j(x_k), j \in I(x^*)\}, \\ P_k &= I - R_k(R_k^T R_k)^{-1}R_k^T. \end{aligned}$$

Without loss of generality, we will suppose that the matrices H_k are symmetric. We will assume moreover that the sequence $\{H_k\}$ converges to a matrix H^* satisfying

$$(4.1) \quad P^*H^*P^* = P^*\nabla_{xx}^2 L(x^*, \mu^*)P^*.$$

This holds, for example, when one uses secant approximations as in [9] or, under suitable conditions, when one uses the BFGS update formula (see [18]). Hypothesis (4.1) and the second order sufficiency condition guarantee the existence of a positive number ρ satisfying⁴

$$(4.2) \quad d^T P_k H_k P_k d \geq \rho |P_k d|^2 \quad \forall d \in \mathbb{R}^n$$

for k large enough.

Propositions 4.2 and 4.3 give important asymptotic properties.

PROPOSITION 4.2. *For k large enough,*

- (i) (QP₀) has a unique Kuhn–Tucker point of least norm,
(QP) has a unique Kuhn–Tucker point of least norm,
 $\{d_k^0\} \rightarrow 0, \quad \{d_k\} \rightarrow 0,$
where d_k^0 and d_k are computed through Step 1(i) and (ii).
- (ii) $\{\mu_k^0\} \rightarrow \mu^*, \quad \{\mu_k\} \rightarrow \mu^*$
where μ_k^0 and μ_k are the multipliers associated with the quadratic problems (QP₀) and (QP).
- (iii) $I_k^0 \triangleq \{j \text{ s.t. } (\mu_k^0)_j > 0\} = \{j \text{ s.t. } g_j(x_k) + \langle \nabla g_j(x_k), d_k^0 \rangle = 0\} = I(x^*),$
 $I_k \triangleq \{j \text{ s.t. } (\mu_k)_j > 0\} = \{j \text{ s.t. } g_j(x_k) + \langle \nabla g_j(x_k), d_k \rangle = -|d_k^0|^\nu\} = I(x^*).$

⁴ In fact if (4.2) holds, a positive matrix H_k^+ can easily be constructed such that $P_k H_k^+ P_k = P_k H_k P_k$ (see [18]).

Proof. x^* is a Kuhn-Tucker point for the problem

$$\begin{aligned} & \min \frac{1}{2}(x - x^*)^T H^*(x - x^*) + \langle \nabla f(x^*), x - x^* \rangle \\ & \text{s.t. } g_j(x^*) + \langle \nabla g_j(x^*), x - x^* \rangle \leq 0 \end{aligned}$$

at which the second order sufficiency conditions are satisfied with strict complementary slackness and linear independence of the gradients of the active constraints.

We can write $d_k^0 = x - x_k$ where x is solution of the problem

$$\begin{aligned} & \min \frac{1}{2}(x - x_k)^T H_k(x - x_k) + \langle \nabla f(x_k), x - x_k \rangle \\ & \text{s.t. } g_j(x_k) + \langle \nabla g_j(x_k), x - x_k \rangle \leq 0. \end{aligned}$$

Since $x_k \rightarrow x^*$ and $H_k \rightarrow H^*$, parts (i) and (ii) for d_k^0 follow from Theorem 2.1 of [20]. We can also write $d_k = x - x_k$ where x is solution of the problem

$$\begin{aligned} & \min \frac{1}{2}(x - x_k)^T H_k(x - x_k) + \langle \nabla f(x_k), x - x_k \rangle \\ & \text{s.t. } g_j(x_k) + \langle \nabla g_j(x_k), x - x_k \rangle \leq -|d_k^0|^\nu \end{aligned}$$

and, as $d_k^0 \rightarrow 0$, parts (i) and (ii) for d_k also follow from Theorem 2.1 of [20]. That (iii) is true follows from the fact that $\mu_k^0 \rightarrow \mu^*$, $\mu_k \rightarrow \mu^*$, and that, from strict complementarity, $I(x^*) = \{j \text{ s.t. } \mu_j^* > 0\}$. \square

PROPOSITION 4.3. *The solutions of (QP₀) and (QP) satisfy*

$$(4.3) \quad \{d_k^0\} \sim \{d_k\},$$

i.e., there exist some constants $C_1 > 0$, $C_2 > 0$ and an integer \hat{k} such that

$$C_1 |d_k^0| \leq |d_k| \leq C_2 |d_k^0| \quad \forall k \geq \hat{k}.$$

Proof. For k large enough, $I_k^0 = I_k = I(x^*)$ and d_k satisfies

$$(4.4) \quad H_k d_k + \nabla f(x_k) + R_k \mu_k = 0.$$

Let us define Δd_k and $\Delta \mu_k$ by

$$d_k = d_k^0 + \Delta d_k, \quad \mu_k = \mu_k^0 + \Delta \mu_k.$$

We have from (4.4)

$$H_k d_k^0 + \nabla f(x_k) + R_k \mu_k^0 + H_k \Delta d_k + R_k \Delta \mu_k = 0$$

which gives

$$(4.5) \quad H_k \Delta d_k + R_k \Delta \mu_k = 0.$$

Now, Δd_k solves

$$R_k^T \Delta d_k = \begin{pmatrix} -|d_k^0|^\nu \\ \vdots \\ -|d_k^0|^\nu \end{pmatrix}$$

and can be decomposed into

$$\Delta d_k = \Delta d_k^1 + \Delta d_k^2$$

with

$$\Delta d_k^1 = P_k \Delta d_k$$

and

$$(4.6) \quad \Delta d_k^2 = -R_k (R_k^T R_k)^{-1} \begin{pmatrix} |d_k^0|^\nu \\ \vdots \\ |d_k^0|^\nu \end{pmatrix}.$$

Thus, since H_k is bounded, it follows from (4.5) that

$$H_k \Delta d_k^1 + R_k \Delta \mu_k = O(|d_k^0|^\nu).$$

This gives

$$\Delta d_k^{1^T} H_k \Delta d_k^1 + \Delta d_k^T P_k^T R_k \Delta \mu_k = O(|d_k^0|^\nu)$$

and, since $P_k^T R_k = 0$, using (4.2), we get

$$(4.7) \quad |\Delta d_k^1|^2 = O(|d_k^0|^\nu).$$

From (4.6) and (4.7) we thus obtain

$$\Delta d_k = o(|d_k^0|).$$

□

In order to establish the main results, we now need three lemmas.

LEMMA 4.4. *There exists some constant $\gamma > 0$ such that, for k large enough,*

$$\sum_{j \in I(x^*)} (\mu_k)_j g_j(x_k) \leq -\gamma \left(\sum_{j \in I(x^*)} g_j(x_k)^2 \right)^{1/2}.$$

Proof. For k large enough we have, from the convergence of the multipliers,

$$\begin{aligned} \sum_{j \in I(x^*)} (\mu_k)_j g_j(x_k) &\leq \frac{1}{2} \min \{ \mu_j^* \text{ s.t. } j \in I(x^*) \} \sum_{j \in I(x^*)} g_j(x_k) \\ &\leq -\frac{1}{2} \min \{ \mu_j^* \text{ s.t. } j \in I(x^*) \} \left(\sum_{j \in I(x^*)} g_j(x_k)^2 \right)^{1/2}. \end{aligned}$$

We set $\gamma = \frac{1}{2} \min \{ \mu_j^* \text{ s.t. } j \in I(x^*) \}$, which is positive due to strict complementarity. □

LEMMA 4.5. *Direction d_k computed through Step 1(i) and (ii) can be decomposed into $d_k = P_k d_k + d_k^1$ with*

$$|d_k^1| \leq C \left(\sum_{j \in I(x^*)} g_j(x_k)^2 \right)^{1/2} + O(|d_k^0|^\nu)$$

for k large enough, for some $C > 0$.

Proof. The direction d_k satisfies

$$R_k^T d_k = -h_k - \begin{pmatrix} |d_k^0|^\nu \\ \vdots \\ |d_k^0|^\nu \end{pmatrix}$$

where h_k is an $|I(x^*)|$ -vector whose components are the values $g_j(x_k)$ for $j \in I(x^*)$. It is thus possible to rewrite d_k as

$$d_k = P_k d_k + d_k^1$$

with

$$d_k^1 = -R_k (R_k^T R_k)^{-1} \left(h_k + \begin{pmatrix} |d_k^0|^\nu \\ \vdots \\ |d_k^0|^\nu \end{pmatrix} \right).$$

This implies that, for k large enough,

$$|d_k^1| \leq C \left(\sum_{j \in I(x^*)} g_j(x_k)^2 \right)^{1/2} + O(|d_k^0|^\nu)$$

for some C . □

LEMMA 4.6. *There exists a positive constant $\bar{\gamma}$ such that, for k large enough, the solution d_k of (QP) satisfies the inequality*

$$\theta_k = \langle \nabla f(x_k), d_k \rangle \leq -\bar{\gamma} |d_k|^2.$$

Proof. Direction d_k computed through Step 1(ii) satisfies the Kuhn–Tucker conditions

$$\nabla f(x_k) + \sum_{j \in I(x^*)} (\mu_k)_j \nabla g_j(x_k) + H_k d_k = 0$$

and, multiplying by d_k ,

$$\theta_k = - \sum_{j \in I(x^*)} (\mu_k)_j \langle \nabla g_j(x_k), d_k \rangle - d_k^T H_k d_k$$

which yields, using the complementarity conditions,

$$\theta_k = \sum_{j \in I(x^*)} (\mu_k)_j g_j(x_k) + \sum_j (\mu_k)_j |d_k^0|^v - d_k^T H_k d_k.$$

Replacing d_k by its decomposition, we obtain

$$\begin{aligned} \theta_k &= \sum_{j \in I(x^*)} (\mu_k)_j g_j(x_k) + \sum_j (\mu_k)_j |d_k^0|^v - d_k^T P_k H_k P_k d_k \\ &\quad - 2d_k^T P_k H_k d_k^1 - d_k^{1^T} H_k d_k^1. \end{aligned}$$

Using (4.2), Lemmas 4.4 and 4.5, and the fact that the matrices H_k and the multipliers μ_k are bounded, we obtain

$$\theta_k \leq -\rho |d_k|^2 + O(|d_k^0|^v).$$

Since $\{d_k\} \sim \{d_k^0\}$, the claim holds. \square

The next proposition shows that, for k large enough, the algorithm never needs to compute a first order direction.

PROPOSITION 4.7. *For k large enough, the solutions of (QP₀) and (QP) satisfy the following inequalities*

- (i) $|d_k^0| \leq M,$
- (ii) $|H_k d_k^0| \leq |d_k^0|^{1/2},$
- (iii) $|d_k| \leq M,$
- (iv) $\theta_k = \langle \nabla f(x_k), d_k \rangle \leq \min \{-|d_k^0|^\kappa, -|d_k|^\kappa\}.$

Proof. Relations (i)–(iii) obviously hold since the sequences $\{d_k^0\}$ and $\{d_k\}$ converge to zero and the matrices H_k are bounded. Inequality (iv) follows from Lemma 4.6. \square

A crucial requirement for achieving superlinear convergence is that a unit stepsize be used in a neighborhood of the solution. The next proposition shows that Algorithm A does achieve this goal.

PROPOSITION 4.8. *For k large enough, the direction is always computed through Step 1(i)–(iii) and the stepsize t_k is one.*

Proof. (The proof is analogous to that of Proposition 15 in [8]; see also [9].)

In all the relations given in this proof, the phrase “for k large enough” is implicit.

The first part of the theorem is obvious in view of Proposition 4.7. In order to prove the second part, we first show that the property

$$\tilde{d}_k = O(|d_k|^2)$$

holds (close to the solution, \tilde{d}_k is always well defined). By definition, \tilde{d}_k is the minimal norm solution of

$$g_j(x_k + d_k) + \langle \nabla g_j(x_k), \tilde{d}_k \rangle = -|d_k^0|^\tau, \quad j \in I_k.$$

Expanding, we obtain

$$\begin{aligned} g_j(x_k) + \langle \nabla g_j(x_k), d_k \rangle + \langle \nabla g_j(x_k), \tilde{d}_k \rangle \\ + \frac{1}{2} \langle d_k, \nabla_{xx} g_j(x_k + \xi d_k) d_k \rangle = -|d_k^0|^\tau, \quad j \in I_k \end{aligned}$$

for some $0 \leq \xi \leq 1$. Hence, using the definition of I_k

$$R_k^T \tilde{d}_k = O(|d_k|^2).$$

Thus, \tilde{d}_k solves the problem

$$\begin{aligned} \min \frac{1}{2} |\tilde{d}_k|^2 \\ \text{s.t. } R_k^T \tilde{d}_k = O(|d_k|^2) \end{aligned}$$

and, since R_k is full column rank, is given by

$$\tilde{d}_k = R_k (R_k^T R_k)^{-1} O(|d_k|^2),$$

which proves our first claim. Now, according to Step 2 in Algorithm A, two conditions are needed for the line search to yield a unit stepsize, namely feasibility of the resulting point (2.2) and sufficient decrease (2.1). Expanding g_j around $x_k + d_k$ we obtain, for $j \in I(x^*)$,

$$\begin{aligned} (4.8) \quad g_j(x_k + d_k + \tilde{d}_k) &= g_j(x_k + d_k) + \langle \nabla g_j(x_k + d_k), \tilde{d}_k \rangle + O(|d_k|^4) \\ &= g_j(x_k + d_k) + \langle \nabla g_j(x_k), \tilde{d}_k \rangle + O(|d_k|^3) \\ &= -|d_k^0|^\tau + O(|d_k^0|^3). \end{aligned}$$

The last term is negative since the sequence $\{d_k^0\}$ converges to zero. Thus the feasibility condition is satisfied.

We also have, since f is three times continuously differentiable,

$$f(x_k + d_k + \tilde{d}_k) = f(x_k) + \langle \nabla f(x_k), d_k \rangle + \langle \nabla f(x_k), \tilde{d}_k \rangle + \frac{1}{2} d_k^T \nabla_{xx} f(x_k) d_k + O(|d_k|^3).$$

The Kuhn-Tucker conditions

$$\nabla f(x_k) + H_k d_k + \sum_j (\mu_k)_j \nabla g_j(x_k) = 0$$

and the complementarity relations imply

$$\frac{1}{2} \langle \nabla f(x_k), d_k \rangle = -\frac{1}{2} d_k^T H_k d_k - \sum_j (\mu_k)_j \langle \nabla g_j(x_k), d_k \rangle - \frac{1}{2} \sum_j (\mu_k)_j g_j(x_k) + O(|d_k^0|^\nu)$$

and

$$\langle \nabla f(x_k), \tilde{d}_k \rangle = O(|d_k|^3) - \sum_j (\mu_k)_j \langle \nabla g_j(x_k), \tilde{d}_k \rangle.$$

We obtain therefore,

$$\begin{aligned} (4.9) \quad f(x_k + d_k + \tilde{d}_k) - f(x_k) &= \frac{1}{2} \theta_k - \sum_j (\mu_k)_j \langle \nabla g_j(x_k), d_k \rangle \\ &\quad - \sum_j (\mu_k)_j \langle \nabla g_j(x_k), \tilde{d}_k \rangle - \frac{1}{2} d_k^T H_k d_k + \frac{1}{2} d_k^T \nabla_{xx} f(x_k) d_k \\ &\quad + O(|d_k|^\nu) + O(|d_k|^3) - \frac{1}{2} \sum_j (\mu_k)_j g_j(x_k). \end{aligned}$$

Now, since the g_j 's are three times continuously differentiable, the relation

$$g_j(x_k + d_k + \tilde{d}_k) = O(|d_k|^\tau), \quad j \in I(x^*)$$

obtained in (4.8) yields, for $j \in I(x^*)$,

$$g_j(x_k) + \langle \nabla g_j(x_k), d_k \rangle + \langle \nabla g_j(x_k), \tilde{d}_k \rangle + \frac{1}{2} d_k^T \nabla_{xx} g_j(x_k) d_k = O(|d_k|^\tau).$$

Hence,

$$\begin{aligned} & -\sum_j (\mu_k)_j \langle \nabla g_j(x_k), d_k \rangle - \sum_j (\mu_k)_j \langle \nabla g_j(x_k), \tilde{d}_k \rangle \\ & = \sum_j (\mu_k)_j g_j(x_k) + \frac{1}{2} \sum_j (\mu_k)_j d_k^T \nabla_{xx} g_j(x_k) d_k + O(|d_k|^\tau). \end{aligned}$$

Substituting those values into (4.9), we obtain

$$\begin{aligned} f(x_k + d_k + \tilde{d}_k) - f(x_k) &= \frac{1}{2} \theta_k + \frac{1}{2} \sum_j (\mu_k)_j g_j(x_k) \\ & \quad + \frac{1}{2} d_k^T \left(\nabla_{xx} f(x_k) + \sum_j (\mu_k)_j \nabla_{xx} g_j(x_k) - H_k \right) d_k \\ & \quad + O(|d_k|^\tau) + O(|d_k|^\nu). \end{aligned}$$

This, together with Lemmas 4.4 and 4.5, gives

$$\begin{aligned} & f(x_k + d_k + \tilde{d}_k) - f(x_k) - \alpha \theta_k \\ & \leq \left(\frac{1}{2} - \alpha \right) \theta_k + \frac{1}{2} d_k^T P_k \left(\nabla_{xx} f(x_k) + \sum_j (\mu_k)_j \nabla_{xx} g_j(x_k) - H_k \right) P_k d_k \\ & \quad + O(|d_k|^\tau) + O(|d_k|^\nu). \end{aligned}$$

Due to the convergence of the projections of the approximate Hessian matrices, we obtain

$$f(x_k + d_k + \tilde{d}_k) - f(x_k) - \alpha \theta_k \leq \left(\frac{1}{2} - \alpha \right) \theta_k + o(|d_k|^2).$$

In view of Lemma 4.6, the right-hand side of the last inequality is nonpositive. Thus the “sufficient decrease” condition is satisfied. \square

THEOREM 4.9. *Under the stated assumptions, the convergence is two-step superlinear, i.e., the following relation holds*

$$\lim_{k \rightarrow \infty} \frac{|x_{k+2} - x^*|}{|x_k - x^*|} = 0.$$

Proof. The proof is similar to the one of [18, Thm. 1]. \square

5. Implementation and computational results. Several implementation issues have to be addressed. First, the sequence $\{H_k\}$ of $n \times n$ matrices is thus far unspecified, subject only to requirement (4.1). While a secant approximation to the Hessian $\nabla_{xx}^2 L(x_k, \mu_k)$ would be suitable, use of an update formula avoids many function evaluations. Under some assumptions, matrices H_k generated by the BFGS formula [18] are shown to satisfy (4.1). The latter option, with $H_0 = I$, was selected for our experiments. Second, the order in which the tests (2.1) and (2.2) are performed needs to be specified. In our implementation, in line with the premise that the objective function may not be defined outside the feasible set, (2.2) was tested first and (2.1) was tested only when (2.2) was satisfied. Third, Algorithm A as stated does not efficiently handle affine constraints. In our experiments, the correction corresponding to such constraints was set to zero in the right-hand sides of the constraints in (QP) and (LS).

TABLE 1
Computational results.

No	Code	NF	NDF	FV	VC	KT
12	VF02AD	12	12	-.30000000E+02	.58E-09	.35E-07
	OPRQP	40	26	-.30000004E+02	.76E-05	.15E-09
	A	7	7	-.30000000E+02	.0	.12E-06
29	VF02AD	13	13	-.22627417E+02	.0	.16E-05
	OPRQP	64	39	-.22627421E+02	.56E-05	.10E-05
	A	14	10	-.22627417E+02	.0	.17E-06
30	VF02AD	14	14	.10000000E+01	.0	.56E-08
	OPRQP	18	18	.10000000E+01	.38E-08	.28E-09
	A	14	13	.10000000E+01	.0	.0
31	VF02AD	10	10	.60000000E+01	.27E-09	.12E-04
	OPRQP	24	22	.59999631E+01	.62E-05	.13E-06
	A	11	8	.60000000E+01	.0	.41E-06
33	VF02AD	5	5	-.40000000E+01	.0	.0
	OPRQP	43	39	-.40000000E+01	.32E-10	.0
	A	4	4	-.40000000E+01	.0	.0
34	VF02AD	8	8	-.83403245E+00	.15E-08	.0
	OPRQP	60	37	-.83403515E+00	.73E-05	.0
	A	9	8	-.83403245E+00	.0	.43E-08
43	VF02AD	12	12	-.44000000E+02	.35E-09	.75E-05
	OPRQP	31	24	-.44000013E+02	.79E-05	.19E-06
	A	9	9	-.44000000E+02	.0	.68E-04
57	VF02AD	4	4	.30646306E-01	.0	.0
	OPRQP	40	24	.28459078E-01	.89E-05	.89E-06
	A	33	19	.28459673E-01	.0	.20E-07
66	VF02AD	7	7	.51816327E+00	.39E-08	.57E-06
	OPRQP	18	17	.51815751E+00	.10E-04	.11E-10
	A	8	8	.51816324E+00	.0	.0
84	VF02AD	6	6	-.52803365E+07	.63E-01	.0
	OPRQP	43	5	-.55883016E+07	.68E+00	.22E+06
	A	4	4	-.52803389E+07	.0	.0
100	VF02AD	20	20	.68063006E+03	.76E-07	.29E-03
	OPRQP	49	31	.68063005E+03	.76E-05	.73E-08
	A	42	14	.68063006E+03	.0	.21E-03
113	VF02AD	15	15	.24306209E+02	.16E-0	.11E-03
	OPRQP	30	28	.24306193E+02	.13E-04	.11E-08
	A	18	14	.24306209E+02	.0	.17E-04
117	VF02AD	17	17	.32348679E+02	.36E-07	.28E-05
	OPRQP	41	40	.32348442E+02	.54E-05	.73E-06
	A	28	16	.32348679E+02	.0	.68E-04

No: number of the test problem in [6].

Code: name of the program.

NF: number of objective function evaluations.

NDF: number of gradient evaluations of the objective function.

FV: objective function value at the final point.

VC: sum of constraint violation, given by $\sum_{j=1}^m \max(0, g_j(x))$, at the final point.

KT: norm of Kuhn-Tucker vector (i.e. norm of the gradient of the Lagrangian function at the final point).

However, in order to avoid potential zigzagging, the right-hand side in the condition defining I_k in (LS) was not set to zero for the affine constraints, but rather the corresponding “=” sign was changed to a “ \geq ”. Finally, *scaling* can be introduced at various places in the algorithm, and values have to be selected for the various parameters. If the right-hand side in the constraints in (QP) is too big, d_k may not be a descent direction for f in the early iterations, while if it is too small, the stepsize may be truncated, due to infeasibility, until a very small neighborhood of the solution is reached. In our experiments, the right-hand side of the constraints in (QP) and of the condition defining I_k in (LS) was replaced by $\max(-|d_k^0|^3, -10^{-2}|d_k^0|)$, which seems to often result in a satisfactory behavior on reasonably well scaled problems. For a similar reason, we replaced the right-hand side of the constraints in (LS) by $\max(-|d_k^0|^{5/2}, -10^{-2}|d_k^0|)$. The right-hand side of the test on θ_k in Step 1(ii) was scaled by a small number. This test was always satisfied throughout our experiments. Finally, we used $\alpha = .3$, $\beta = .8$ and $M = \infty$.

Algorithm A was tested on fourteen of the seventeen problems in [6] which do not involve equality constraints but do include *nonlinear* inequality constraints, and for which a feasible initial point is provided. Problems numbered 67, 70 and 85 were discarded due to some disparity between function values we computed and those given in [6]. When tested on Problem 93, with the chosen values of the algorithm parameters, Algorithm A had to resort to the first order direction (Step 1(ii)) for the initial iterations due to infeasibility of (QP), thus making the performance of Algorithm A dependent on the choice of the first order method. Table 1 shows the results obtained on the thirteen remaining problems. The results obtained with Algorithm A are compared to the best results among those given in [6], i.e., those obtained with algorithms VF02AD and OPRQP. The format of this table is as in [6].

In most cases, Algorithm A is competitive with VF02AD. It always performs better than OPRQP. This is remarkable since neither VF02AD nor OPRQP enjoys properties (1.1) and (1.2). It could be argued that a comparison based only on the number of function evaluations unduly favors Algorithm A, which calls for the solution of up to two quadratic programs and one linear least squares problem at each iteration. However, as pointed out in the introduction, clever implementation should reduce the computational effort needed to solve these three problems to little more than that required for the solution of a single quadratic program. Also, in the context of engineering design problems, function evaluations typically require such extensive computation that time spent in solving quadratic programs can generally be regarded as negligible. Finally, the number of constraint function evaluations is not indicated in Table 1. Typically, the number of such evaluations will be somewhat larger for Algorithm A than for its contenders due to the Maratos effect avoidance scheme.

REFERENCES

- [1] M. K. H. FAN, C. D. WALRATH, C. LEE, A. L. TITS, W. T. NYE, M. RIMER, R. T. GRANT AND W. S. LEVINE, *Two case studies in optimization-based computer-aided design of control systems*, Proc. 24th IEEE Conf. on Decision and Control (December 1985), p. 1794.
- [2] U. M. GARCIA-PALOMARES AND O. L. MANGASARIAN, *Superlinearly convergent quasi-Newton algorithms for nonlinearly constrained optimization problems*, Math. Programming, 11 (1976), pp. 1-13.
- [3] C. GONZAGA, E. POLAK AND R. TRAHAN, *An improved algorithm for optimization problems with functional inequality constraints*, IEEE Trans. Automat. Control, AC-25 (1980), pp. 49-54.
- [4] S. P. HAN, *A globally convergent method for nonlinear programming*, J. Optim. Theory Appl., 22 (1977), pp. 297-309.

- [5] J. HERSKOVITS, *A two-stage feasible direction algorithm for nonlinear constrained optimization*, Math. Programming (1987), to appear.
- [6] W. HOCK AND K. SCHITTKOWSKI, *Test examples for nonlinear programming codes*, in Lecture Notes in Econom. and Math. Systems, 187, Springer-Verlag, Berlin, New York., 1981.
- [7] N. MARATOS, *Exact penalty function algorithms for finite dimensional and optimization problems*, Ph.D. thesis, Imperial College of Science and Technology, London, U.K. (1978).
- [8] D. Q. MAYNE AND E. POLAK, *A superlinearly convergent algorithm for constrained optimization problems*, Computing and Control Publication 78/52, Imperial College of Science and Technology, London (1978).
- [9] ———, *A superlinearly convergent algorithm for constrained optimization problems*, Math. Programming Study, 16 (1982), pp. 45–61.
- [10] L. W. NAGEL, *SPICE2: A computer program to simulate semiconductor circuits*, Memo No. ERL-M520, Electronics Research Laboratory, University of California, Berkeley, CA, May 1975.
- [11] W. T. NYE AND A. L. TITS, *An enhanced methodology for interactive optimal design*, Proc. 1983 IEEE International Symposium on Circuits and Systems (May 1983), pp. 1050–1051.
- [12] W. T. NYE, *DELIGHT: An interactive system for optimization-based engineering design*, Ph.D. thesis, Department EECS, University of California, Berkeley, CA, 1983.
- [13] W. T. NYE AND A. L. TITS, *An application-oriented, optimization-based methodology for interactive design of engineering systems*, Internat. J. Control, 43 (1986), pp. 1693–1721.
- [14] E. POLAK, *Computational Methods in Optimization*, Academic Press, New York, 1971.
- [15] E. POLAK AND D. Q. MAYNE, *An algorithm for optimization problems with functional inequality constraints*, IEEE Trans. Automat. Control, AC-21 (1976), pp. 184–193.
- [16] E. POLAK, R. TRAHAN AND D. Q. MAYNE, *Combined phase I–phase II methods of feasible directions*, Math. Programming, 17 (1979), pp. 32–61.
- [17] M. J. D. POWELL, *A fast algorithm for nonlinearly constrained optimization calculations*, in Numerical Analysis, Dundee, 1977, Lecture Notes in Math., 630, G. A. Watson, ed., Springer-Verlag, Berlin, New York, 1977, pp. 144–157.
- [18] ———, *The convergence of variable metric methods for nonlinearly constrained optimization calculations*, in Nonlinear Programming 3, O. L. Mangasarian, R. R. Meyer and S. M. Robinson, eds., Academic Press, New York, 1978, pp. 27–63.
- [19] S. M. ROBINSON, *A quadratically-convergent algorithm for general nonlinear programming problems*, Math. Programming, 3 (1972), pp. 145–156.
- [20] ———, *Perturbed Kuhn–Tucker points and rates of convergence for a class of nonlinear-programming algorithms*, Math. Programming, 7 (1974), pp. 1–16.
- [21] A. L. TITS, W. T. NYE AND A. SANGIOVANNI-VINCENTELLI, *Enhanced methods of feasible directions for engineering design problems*, J. Optim. Theory Appl., 51 (1986), pp. 475–504.
- [22] R. B. WILSON, *A simplified algorithm for concave programming*, Ph.D. dissertation, Harvard University, Cambridge, MA, 1963.
- [23] G. ZOUTENDIJK, *Methods of Feasible Directions*, Elsevier, Amsterdam, 1960.

STRUCTURE OF EXTREMALS IN OPTIMAL CONTROL PROBLEMS*

JACOB KOGAN†

Abstract. In this paper we study the structure of intersections of neighboring extremals in optimal control. The obtained results extend a known criterion for existence of a field of extremals from the calculus of variations to the optimal control theory.

Key words. calculus of variation, conjugate points, field of extremals, optimal control problem

AMS(MOS) subject classification. 49C05

1. Introduction. The classical theory of the calculus of variations presents necessary conditions for absence of conjugate points, namely the points of intersection of neighboring extremals initiating at the same point. The absence of conjugate points in the calculus of variations enables one to construct a field of extremals and to derive sufficient conditions for a strong extremum (see for example [1], [3], [5]–[9], [13]).

In this paper we accept the necessary conditions from the calculus of variations and investigate the structure of intersections of neighboring extremals in optimal control. It is shown in this work that, in contrast with the classical calculus of variations, neighboring extremals in an optimal control problem can intersect each other. On the other hand we demonstrate in this study that the set of intersections of extremals in optimal control possesses a simple elegant structure. The main result of the paper is the following: *Let $x(t)$ be an extremal trajectory defined on a time interval $[t_1, t_2]$. There exists a finite partition $t_1 = r_1 < r_2 < \dots < r_k = t_2$ of the time interval $[t_1, t_2]$ such that for each neighboring extremal $y(t)$ with $y(t_1) = x(t_1)$ and $y(s) = x(s)$ there exists an element r of the partition and the following condition holds:*

$$x(t) = y(t) \text{ on } [t_1, r] \quad \text{and} \quad x(t) \neq y(t) \quad \text{for each } t \in (r, t_2].$$

We show also that the partition is determined by the attainable set of the control system under consideration. We call the point r a branching point of the extremal $x(t)$ and say that $y(t)$ branches out of $x(t)$ at r .

The paper is organized as follows: In the second section we state the problem and introduce the main assumptions. The third section is devoted to study of extremals in a linear control system with a quadratic cost. The fourth section is the main part of the study. We consider there a nonlinear control problem and derive necessary and sufficient conditions for branching.

2. The problem. We introduce first, for the sake of convenience, the following convention: In order to distinguish between functions of a real variable and elements of the real Euclidean space R^n , we shall denote (for example) a function of a real variable by $x(\cdot)$, in contrast with a point x in R^n . The norm of $x \in R^n$ is denoted by $|x|$, $\langle x, y \rangle$ denotes the scalar product in R^n and $\dot{x}(t)$ denotes differentiation with respect to time, i.e. $\dot{x}(t) = dx(t)/dt$. The norm of an $n \times n$ matrix M is denoted by $|M|$, namely $|M| = \max \{|Mx| : |x| = 1\}$.

Consider an optimal control system

$$(2.1) \quad \dot{x}(t) = F(t, x(t), u(t))$$

* Received by the editors January 27, 1986; accepted for publication (in revised form) May 15, 1986.

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

where $u(\cdot) : [t_1, t_2] \mapsto R^m$ is measurable and $x(\cdot) : [t_1, t_2] \mapsto R^n$ is absolutely continuous, where measurability is understood to be in the Lebesgue sense, and equalities are always "almost everywhere." Following Berkovitz (see [2, p. 22]) we define an admissible pair as follows.

DEFINITION 2.1. Let $x(t)$ be an absolutely continuous function from $[t_1, t_2]$ to R^n and $u(t)$ be a measurable function from $[t_1, t_2]$ to R^m . The pair $(x(t), u(t))$ is admissible if it satisfies (2.1).

Consider a cost functional defined on the set of admissible pairs $(x(t), u(t))$ as follows:

$$(2.2) \quad c(x(\cdot), u(\cdot)) = \int_{t_1}^{t_2} f(t, x(t), u(t)) dt.$$

In order to describe the structure of extremals we need to recall some auxiliary notions.

DEFINITION 2.2. A trajectory $x(t)$ is an extremal trajectory of the optimal control problem (2.1), (2.2) if $(x(t), u(t))$ is an admissible pair and there exists a scalar $\eta_0 \leq 0$ and an absolutely continuous vector valued function $\eta(t)$ which is defined on $[t_1, t_2]$ and is a solution of the following ordinary differential equation:

$$(2.3) \quad \frac{d}{dt} \eta(t) = -\eta_0 \frac{\partial f}{\partial x}(t, x(t), u(t)) - \eta(t) \frac{\partial F}{\partial x}(t, x(t), u(t)),$$

and such that the triple $(x(t), u(t), \eta(t))$ satisfies the Pontryagin Maximum Principle, i.e., the following additional condition holds on $[t_1, t_2]$:

$$(2.4) \quad \begin{aligned} &\eta_0 f(t, x(t), u(t)) + \eta(t) F(t, x(t), u(t)) \\ &= \max_u \{ \eta_0 f(t, x(t), u) + \eta(t) F(t, x(t), u) \} \end{aligned}$$

(see [2, p. 186]). The last definitions show that an extremal trajectory, which is the main object of the study, usually appears together with an admissible control and a corresponding solution of the adjoint equation (2.3). For the sake of the technical convenience we introduce now the notion of an extremal triple.

DEFINITION 2.3. A triple $(x(t), u(t), \eta(t))$ is an extremal triple if $(x(t), u(t))$ is an admissible pair, $\eta(t)$ is a solution of (2.3) and the triple $(x(t), u(t), \eta(t))$ satisfies condition (2.4).

DEFINITION 2.4. The Hamiltonian H of the optimal control problem (2.1), (2.2) is defined as follows:

$$H(t, x, u, \eta) = \eta_0 f(t, x, u) + \eta F(t, x, u).$$

In this paper we consider the structure of regular extremals only. Namely, we suppose throughout that $\eta_0 = -1$.

We present now the definition of a branching point of the extremal trajectory $x(t)$. Suppose that there exists a neighboring extremal $y(t)$ such that $x(t)$ and $y(t)$ coincide over an initial subinterval $[t_1, r]$ of the time interval $[t_1, t_2]$ and differ on the rest of it. Namely, $x(t) = y(t)$ for each $t \in [t_1, r]$ and $x(t) \neq y(t)$ for each $t \in (r, t_2]$. In this case we say that r is a branching point of the extremal trajectory $x(t)$, and $y(t)$ branches out of $x(t)$ at r .

We wish to emphasize that our interest concentrates on the branching points of the extremal trajectory $x(t)$ formed by neighboring extremals initiating at the same point $x(t_1)$, where by a neighboring extremal we mean an extremal $y(t)$ such that the norm $\|x(\cdot), \eta(\cdot) - (y(\cdot), \mu(\cdot))\|_{C[t_1, t_2]}$ is small (how small will be specified later).

Here $\eta(t)$, $\mu(t)$ are the corresponding solutions of the adjoint equation (2.3) and $\sup \{|x(t) - y(t)| + |\eta(t) - \mu(t)| : t \in [t_1, t_2]\}$ is denoted by $\|(x(\cdot), \eta(\cdot)) - (y(\cdot), \mu(\cdot))\|_{C[t_1, t_2]}$.

Let $y(t)$ be an extremal trajectory of the control problem (2.1), (2.2) with $y(t_1) = x(t_1)$. It is shown in the last section (see Theorem 4.2) that there exists a positive ε such that, if $\|(x(\cdot), \eta(\cdot)) - (y(\cdot), \mu(\cdot))\|_{C[t_1, t_2]} < \varepsilon$ and $x(r) \neq y(r)$ for some $r \in [t_1, t_2]$, then $x(t) \neq y(t)$ for each $t \in [r, t_2]$. On the other hand we show by an example (see Example 4.1) that, if the condition $\|(x(\cdot), \eta(\cdot)) - (y(\cdot), \mu(\cdot))\|_{C[t_1, t_2]} < \varepsilon$ is not satisfied, the extremal $y(t)$ can be different from $x(t)$ on a subinterval $[r, r + \delta)$ and intersect $x(t)$ once again on $[r + \delta, t_2]$.

In order to present the formal definition of a branching point we need to define rigorously what is meant by a neighboring extremal in this study. We say, first, that an extremal $y(t)$ is an ε -neighboring extremal if $\|(x(\cdot), \eta(\cdot)) - (y(\cdot), \mu(\cdot))\|_{C[t_1, t_2]} < \varepsilon$. A point r is an ε -branching point of the extremal trajectory $x(t)$ if there exists an ε -neighboring extremal $y(t)$ such that

$$x(t) = y(t) \quad \text{on } [t_1, r], \quad x(t) \neq y(t) \quad \text{on } (r, t_2].$$

It is clear, that if $\varepsilon_1 < \varepsilon_2$, then the set of ε_1 -branching points is a subset of that of ε_2 -branching points. It is shown in the last section (see Definition 4.1) that there exists an $\varepsilon^* > 0$ such that, for each positive ε less than ε^* , the set of ε^* -branching points coincides with that of ε points. From here on this ε^* will define the set of neighboring extremal trajectories as follows.

DEFINITION 2.5. An extremal trajectory $y(t)$ is a neighboring extremal if the inequality $\|(x(\cdot), \eta(\cdot)) - (y(\cdot), \mu(\cdot))\|_{C[t_1, t_2]} < \varepsilon^*$ holds.

This definition enables us to define a branching point as follows.

DEFINITION 2.6. A point r is a branching point of the extremal trajectory $x(t)$ if there exists a neighboring extremal trajectory $y(t)$ such that

$$x(t) = y(t) \quad \text{on } [t_1, r], \quad x(t) \neq y(t) \quad \text{on } (r, t_2].$$

Our main purpose in this work is to determine the conditions under which the branching points of an extremal $x(t)$ exist or do not exist and to characterize the set of branching points of $x(t)$.

Relationship with the calculus of variations. The classical theory of the calculus of variations deals with the linear control system

$$\dot{x}(t) = u(t)$$

with a cost functional

$$\int_{t_1}^{t_2} f(t, x(t), u(t)) dt.$$

Sufficient conditions for **nonexistence** of conjugate points on the interval $[t_1, t_2]$ are the following:

(1) $f_{uu}(t, x(t), u(t))$ is positive definite along the chosen extremal $x(t)$ (the strengthened Legendre condition),

$$(2) \quad \int_{t_1}^{t_2} \begin{pmatrix} h(t) \\ \dot{h}(t) \end{pmatrix}^* \begin{pmatrix} f_{xx}(t, x(t), u(t)) & f_{ux}(t, x(t), u(t)) \\ f_{xu}(t, x(t), u(t)) & f_{uu}(t, x(t), u(t)) \end{pmatrix} \begin{pmatrix} h(t) \\ \dot{h}(t) \end{pmatrix} dt$$

is positive definite for each $h(t)$ such that $h(t_1) = h(t_2) = 0$, where $*$ indicates the transpose, i.e., the second variation of the cost functional along $x(t)$ is positive definite. Namely, if conditions 1 and 2 hold for the extremal $x(t)$, then the interval $[t_1, t_2]$ contains no point conjugate to t_1 (see [8, Thm. 4, p. 123]).

In order to describe the branching points of $x(t)$ we adopt a natural generalization of the sufficient conditions of the calculus of variations namely we assume that the following holds for the extremal triple $(x(t), u(t), \eta(t))$.

Hypothesis 2.1. (1) There exists a positive scalar m such that for each vector $v \in R^m$

$$v^* H_{uu}(t, x(t), u(t), \eta(t)) v \leq -m|v|^2 \quad \text{along } x(t),$$

and for each positive ρ there exists a positive δ such that

$$|H_{uu}(t, x(t), u(t), \eta(t)) - H_{uu}(t, y, w, \xi)| < \rho$$

provided

$$|(t, x(t), u(t), \eta(t)) - (t, y, w, \xi)| < \delta.$$

$$(2) \quad \int_{t_1}^{t_2} \begin{pmatrix} z(t) \\ v(t) \end{pmatrix}^* \begin{pmatrix} H_{xx}(t, x(t), u(t), \eta(t)) & H_{ux}(t, x(t), u(t), \eta(t)) \\ H_{xu}(t, x(t), u(t), \eta(t)) & H_{uu}(t, x(t), u(t), \eta(t)) \end{pmatrix} \begin{pmatrix} z(t) \\ v(t) \end{pmatrix} dt$$

is negative definite for each $v(t)$, $z(t)$ such that

$$z(t) = \int_{t_1}^t \Phi(t, \sigma) B(\sigma) v(\sigma) d\sigma, \quad z(t_2) = 0,$$

where $A(t)$, $B(t)$ are Lebesgue integrable matrices on $[t_1, t_2]$ given by $A(t) = \partial F / \partial x(t, x(t), u(t))$, $B(t) = \partial F / \partial u(t, x(t), u(t))$ and $\Phi(t, t_1)$ is the transition matrix of $d/dt \phi = A(t)\phi$, i.e., $\phi(t) = \Phi(t, t_1)x_1$ is the solution of the equation $d/dt \phi = A(t)\phi$ with $\phi(t_1) = x_1$. In other words: $z(t)$ is a solution of the linear equation

$$\dot{z}(t) = A(t)z(t) + B(t)v(t)$$

with boundary conditions $z(t_1) = z(t_2) = 0$.

We wish to clarify now the connection between Hypothesis 2.1 and the sufficient conditions of the calculus of variations. Note, that in the calculus of variations the cost function $f(t, x, u)$ has continuous first and second partial derivatives with respect to all its arguments. In this case the first condition of Hypothesis 2.1 is just the strengthened Legendre condition. The second condition is well known in the calculus of variations property of the second variation of the cost functional to be positive definite.

We wish to be able to present extremals as solutions of the Hamiltonian system (2.5), (2.6) displayed below. To this end we assume the existence of an optimal feedback as follows.

Hypothesis 2.2. There exists a positive scalar ε_1 such that for each triple (t, x, η) with $|(x, \eta) - (x(t), \eta(t))| < \varepsilon_1$ there exists a unique control function $\bar{u}(t, x, \eta)$ continuously differentiable with respect to (x, η) and measurable in t such that

$$-f(t, x, \bar{u}(t, x, \eta)) + \eta F(t, x, \bar{u}(t, x, \eta)) = \max_u \{-f(t, x, u) + \eta F(t, x, u)\} \quad \text{on } [t_1, t_2].$$

Without any loss of generality (due to Hypothesis 2.1) we shall assume throughout that this ε_1 is sufficiently small in order to guarantee that for each quadruple $(t, y, w, \eta) \in R \times R^n \times R^m \times R^n$ such that $|(y, w, \mu) - (x(t), u(t), \eta(t))| < \varepsilon_1$ the following condition holds:

$$v^* H_{uu}(t, y, w, \mu) v \leq -\frac{m}{2}|v|^2 \quad \text{for each } v \in R^m$$

This assumption enables one to reduce the study of extremals to the study of solutions $(x(t), \eta(t))$ of the following system of ordinary differential equations:

$$(2.5) \quad \dot{x}(t) = H_{\eta}(t, x(t), \bar{u}(t, x(t), \eta(t)), \eta(t)),$$

$$(2.6) \quad \dot{\eta}(t) = -H_x(t, x(t), \bar{u}(t, x(t), \eta(t)), \eta(t)).$$

The presentation of extremals as solutions of the Hamiltonian system (2.5), (2.6) will be extremely useful in this study.

Our first result deals with a linear control system with a quadratic cost functional. We will show in the last section of the paper that, in the case of a nonlinear control system, the branching points of the extremal trajectory $x(t)$ are determined by the linearized system about the extremal $x(t)$. Hence, characterization of the branching points in this simplest case is a significant step forward.

3. Linear control problem. We consider in this section a linear control system

$$(3.1) \quad \dot{x}(t) = A(t)x(t) + B(t)u(t)$$

with a quadratic cost

$$(3.2) \quad c(x(\cdot), u(\cdot)) = \int_{t_1}^{t_2} \begin{pmatrix} x(t) \\ u(t) \end{pmatrix}^* \begin{pmatrix} R_{11}(t) & R_{12}(t) \\ R_{21}(t) & R_{22}(t) \end{pmatrix} \begin{pmatrix} x(t) \\ u(t) \end{pmatrix} dt,$$

where $x \in R^n$, $u \in R^m$ and $A(t)$, $B(t)$, $R_{11}(t)$, $R_{12}(t)$, $R_{21}(t)$, $R_{22}(t)$ are matrices with appropriate dimensionalities whose elements are Lebesgue integrable functions on $[t_1, t_2]$ and $*$ indicates the transpose. For convenience we assume throughout that $R_{12}(t) = R_{21}^*(t)$.

Note that in this case Hypothesis 2.1 turns into the following.

Hypothesis 3.1. (1) $R_{22}(t)$ is uniformly positive definite on $[t_1, t_2]$. Namely, there exists a positive scalar δ such that $u^* R_{22}(t) u \geq \delta |u|^2$ on $[t_1, t_2]$ for each $u \in R^m$.

(2) For each admissible pair $(z(t), v(t))$ with $z(t_1) = z(t_2) = 0$ the cost $c(z(\cdot), v(\cdot)) \geq 0$ and $c(z(\cdot), v(\cdot)) = 0$ if and only if $(z(t), v(t)) = (0, 0)$.

In the case of the linear control problem (3.1), (3.2), Hypothesis 2.2 is fulfilled and one can derive an explicit formula for extremal control as follows:

$$(3.3) \quad \bar{u}(t, x, \eta) = \frac{1}{2} R_{22}^{-1}(t) \{ B^*(t) \eta^* - 2 R_{12}(t) x \}.$$

The adjoint equation is

$$(3.4) \quad \dot{\eta}(t) = 2x^*(t) R_{11}(t) + 2u^*(t) R_{12}(t) - \eta(t) A(t)$$

and the Pontryagin Maximum Principle is

$$(3.5) \quad -2u^*(t) R_{22}(t) - 2x^*(t) R_{12}(t) + \eta(t) B(t) = 0.$$

We wish to show now that branching is a unique possible form of intersection of extremals initiating at the same point. First we mention the following important property of extremals of the linear control problem (3.1), (3.2).

THEOREM 3.1. Let $(x(t), u(t))$ be an extremal pair. Then for each admissible pair $(y(t), w(t))$ with $x(t_1) = y(t_1)$ and $x(t_2) = y(t_2)$

$$c(x(\cdot), u(\cdot)) \leq c(y(\cdot), w(\cdot)) \quad \text{and} \quad c(x(\cdot), u(\cdot)) = c(y(\cdot), w(\cdot))$$

$$\text{if and only if } (x(t), u(t)) = (y(t), w(t)).$$

Proof. We wish to show that $c(y(\cdot), w(\cdot)) - c(x(\cdot), u(\cdot)) \geq 0$. The pair $(x(t), u(t))$ is an extremal pair, hence the first element of the Taylor expansion of the difference vanishes (see e.g. [11, p. 357]). The second (and the last term) of the expansion is $c(y(\cdot) - x(\cdot), w(\cdot) - u(\cdot))$. Due to Hypothesis 3.1, $c(y(\cdot) - x(\cdot), w(\cdot) - u(\cdot)) \geq 0$ and $c(y(\cdot) - x(\cdot), w(\cdot) - u(\cdot)) = 0$ if and only if $x(t) = y(t)$ and $u(t) = w(t)$. This finishes the proof.

Let $(x(t), u(t))$ and $(y(t), w(t))$ be extremal pairs with $x(t_1) = y(t_1)$ and $x(s) = y(s)$ for some $s \in [t_1, t_2]$. Denote

$$\int_{t_1}^s \begin{pmatrix} x(t) \\ u(t) \end{pmatrix}^* \begin{pmatrix} R_{11}(t) & R_{12}(t) \\ R_{21}(t) & R_{22}(t) \end{pmatrix} \begin{pmatrix} x(t) \\ u(t) \end{pmatrix} dt \quad \text{by } c_s(x(\cdot), u(\cdot)).$$

Due to Theorem 3.1, $c_s(x(\cdot), u(\cdot)) \leq c_s(y(\cdot), w(\cdot))$ and $c_s(y(\cdot), w(\cdot)) \leq c_s(x(\cdot), u(\cdot))$, namely $c_s(y(\cdot) - x(\cdot), w(\cdot) - u(\cdot)) = 0$ and $y(t) = x(t)$, $w(t) = u(t)$ on $[t_1, s]$. This remark justifies introduction of the notion of a branching point.

In order to illustrate the phenomenon of branching we wish to consider a simple example. In spite of its simplicity, this example, as will be shown in this section, presents a typical case of branching.

Example 3.1. A linear control system with a nonempty set of branching points. Consider a linear control system

$$\frac{d}{dt}x(t) = B(t)u(t)$$

with a cost functional

$$c(x(\cdot), u(\cdot)) = \int_0^1 u^2(t) dt$$

with $x \in \mathbb{R}^2$, where u is a scalar and $B(t)$ is given by

$$\begin{pmatrix} 1 \\ 0 \end{pmatrix} \quad \text{for } 0 \leq t < \frac{1}{3}, \quad \begin{pmatrix} 0 \\ 0 \end{pmatrix} \quad \text{for } \frac{1}{3} \leq t < \frac{2}{3}, \quad \begin{pmatrix} 0 \\ 1 \end{pmatrix} \quad \text{for } \frac{2}{3} \leq t \leq 1.$$

An extremal control in this case has the following form

$$u(t) = B^*(t)\eta^*(t),$$

where $\eta(t)$ is a solution of the adjoint ordinary differential equation $\dot{\eta}(t) = 0$ (see [11, p. 180]) and $*$ indicates the transpose. Note, that $\eta(t) = \eta(0)$ and we shall denote it simply as η . It then follows that $u(t) = B(t)^*\eta^*$ for each $t \in [0, 1]$. An extremal trajectory $x(t)$ is, therefore, given by

$$x(t) = x(0) + \int_0^t B(s)B^*(s)\eta^* ds.$$

The product $B(t)B^*(t)$ can easily be computed and the derived extremal trajectory $x(t)$ has the following form:

$$x(t) = x(0) + \gamma(t)$$

where the vector $\gamma(t)$ is given by

$$\begin{pmatrix} t\eta_1 \\ 0 \end{pmatrix} \quad \text{for } 0 \leq t < \frac{1}{3}, \quad \begin{pmatrix} \frac{1}{3}\eta_1 \\ 0 \end{pmatrix} \quad \text{for } \frac{1}{3} \leq t < \frac{2}{3}, \quad \begin{pmatrix} \frac{1}{3}\eta_1 \\ (t - \frac{2}{3})\eta_2 \end{pmatrix} \quad \text{for } \frac{2}{3} \leq t \leq 1.$$

Here η_1, η_2 are the components of the vector η .

We would like to study the extremal trajectories initiating at the same initial point $x(0)$.

The definition of $\gamma(t)$ implies that for each pair of different vectors η, μ such that $\eta_1 = \mu_1$ and $\eta_2 \neq \mu_2$, the corresponding extremal trajectories $x(t)$ and $y(t)$ coincide on the time interval $[0, \frac{2}{3}]$ and are different on $(\frac{2}{3}, 1]$. Namely, if, for instance, $x(0) = 0$, $\eta = (1, 0)$, and $\mu = (1, 1)$ then $x(t)$ is given by

$$\begin{pmatrix} t \\ 0 \end{pmatrix} \text{ for } 0 \leq t < \frac{1}{3}, \quad \begin{pmatrix} \frac{1}{3} \\ 0 \end{pmatrix} \text{ for } \frac{1}{3} \leq t < \frac{2}{3}, \quad \begin{pmatrix} \frac{1}{3} \\ 0 \end{pmatrix} \text{ for } \frac{2}{3} \leq t \leq 1,$$

and $y(t)$ is given by

$$\begin{pmatrix} t \\ 0 \end{pmatrix} \text{ for } 0 \leq t < \frac{1}{3}, \quad \begin{pmatrix} \frac{1}{3} \\ 0 \end{pmatrix} \text{ for } \frac{1}{3} \leq t < \frac{2}{3}, \quad \begin{pmatrix} \frac{1}{3} \\ t - \frac{2}{3} \end{pmatrix} \text{ for } \frac{2}{3} \leq t \leq 1,$$

i.e., $x(t) = y(t)$ on $[0, \frac{2}{3}]$, and $x(t) \neq y(t)$ on $(\frac{2}{3}, 1]$.

Suppose that different extremal trajectories $x(t), y(t)$ coincide on a certain subinterval $[0, \tau]$ of $[0, 1]$. We denote by $\eta(t), \mu(t)$ the corresponding solutions of the adjoint differential equation. In this case $\eta_1 = \mu_1$. On the other hand, inasmuch as $x(t)$ and $y(t)$ are different, $\eta_2 \neq \mu_2$. This means that $\tau = \frac{2}{3}$.

The above considerations prove that $\frac{2}{3}$ is the unique branching point of the considered optimal control problem. Hence, the set of the branching points of this control problem is a finite set. As it will be shown in the end of the section, this result is not incidental, but rather holds for a general linear control system with a convex cost. Namely, the set of branching points is a finite set which does not depend on a chosen extremal.

In order to investigate branching points of extremals, introduction of some additional auxiliary definitions is needed. We present now the formal definition of the space of admissible controls of the linear control system (3.1) defined on a subinterval $[t_1, s]$ of the time interval $[t_1, t_2]$.

DEFINITION 3.1. A measurable function $v(t)$ defined on a subinterval $[t_1, s]$ of the time interval $[t_1, t_2]$ with range in R^m is an admissible control if there exists an absolutely continuous function $\phi(t)$ defined on $[t_1, s]$ with range in R^n which is a solution of the differential equation (3.1), i.e.,

$$\dot{\phi}(t) = A(t)\phi(t) + B(t)v(t) \quad \text{a.e. on } [t_1, s].$$

The space of all admissible controls $v(t)$ defined on the interval $[t_1, s]$ is denoted by $U[t_1, s]$.

For each subspace V of R^n we denote by V^\perp its orthogonal complement in R^n . We introduce now a family of vector spaces which determines branching points of extremal trajectories.

DEFINITION 3.2. For each real number s we define two vector spaces $V(t_1, s)$ and $V(t_1, s^+)$ by the following formulas:

$$V(t_1, s) = \left\{ \int_{t_1}^s \Phi(t_1, t)B(t)v(t) dt \mid v(\cdot) \in U[t_1, s] \right\},$$

$$V(t_1, s^+) = \bigcap_{h>0} V(t_1, s+h),$$

where $\Phi(t, \tau)$ is the transition matrix of $d/dt \phi = A(t)\phi$, i.e., $\phi(t) = \Phi(t, \tau)x_\tau$ is the solution of the equation $d/dt \phi(t) = A(t)\phi(t)$ with $\phi(\tau) = x_\tau$.

Denote the dimension of $V(t_1, s)$ by $\dim V(t_1, s)$. Consider $\dim V(t_1, s)$ as a function of s . Choose s_1, s_2 such that $t_1 \leq s_1 \leq s_2$. Definition 3.2 implies that $V(t_1, s_1)$ is a subspace of $V(t_1, s_2)$. On the other hand for each s the space $V(t_1, s)$ is a subspace

of R^n and, therefore, $\dim V(t_1, s) \leq n$. The above consideration shows that $\dim V(t_1, s)$ is a monotone nondecreasing step function with respect to s . This implies that the number of its discontinuity points is finite. Note, that $V(t_1, s) \subset V(t_1, s^+)$; but $V(t_1, s) \neq V(t_1, s^+)$ if and only if s is a point of discontinuity of $\dim V(t_1, s)$. In the remainder of the section we will show that the points of discontinuity of $\dim V(t_1, s)$ are the branching points of the linear control problem (3.1), (3.2).

At this point we wish to mention a relationship between the vector space $V(t_1, s)$ and the attainable set of the control system (3.1). The attainable set at time s for the control system with initial time t_1 and initial state x_1 is the set of all points $x \in R^n$ such that, for some trajectory $\varphi(t)$ satisfying the condition $\varphi(t_1) = x_1$, the relation $\varphi(s) = x$ holds. In the case when the initial state x_1 is chosen to be 0 the attainable set at time s is a linear subspace of R^n which is denoted by $AT(s; t_1, 0)$.

The set $AT(s; t_1, 0)$ is given by the following formula:

$$AT(s; t_1, 0) = \left\{ \int_{t_1}^s \Phi(s, t) B(t) v(t) dt, v(\cdot) \in U[t_1, s] \right\}.$$

Hence, there is an evident relation between the sets $V(t_1, s)$ and $AT(s; t_1, 0)$, namely

$$V(t_1, s) = \Phi(t_1, s) AT(s; t_1, 0).$$

Since $\Phi(t_1, s)$ is nonsingular, it follows that $\dim V(t_1, s) = \dim AT(s; t_1, 0)$. This implies that the points of discontinuity of $\dim V(t_1, s)$ are exactly the points where the dimension of the attainable set changes. The next statements clarify the importance of this set.

Necessary and sufficient conditions for branching.

LEMMA 3.1. *Let $(x(t), u(t), \eta(t))$ be an extremal triple such that $x(t_1) = 0$ and $\eta(t_1) \in V(t_1, r)^\perp$. Then $(x(t), u(t), \eta(t)) = (0, 0, \eta(t_1)\Phi(t_1, t))$ on $[t_1, r]$.*

Proof. Straightforward verification shows that the triple $(0, 0, \eta(t_1)\Phi(t_1, t))$ is an extremal triple of the optimal control problem (3.1), (3.2) on the initial time interval $[t_1, r]$. Namely, $(0, \eta(t_1)\Phi(t_1, t))$ is a solution of the Hamiltonian system (3.7), (3.8) on the time interval $[t_1, r]$. Note that $(x(t), \eta(t))$ is also a solution of the same Hamiltonian system. Moreover, these two solutions have the same initial conditions, hence $(x(t), \eta(t)) = (0, \eta(t_1)\Phi(t_1, t))$ on $[t_1, r]$. In accordance with (3.3) and the condition $\eta(t_1) \in V(t_1, r)^\perp$ the control $u(t)$ is 0 on $[t_1, r]$.

LEMMA 3.2. *Let $(x(t), u(t), \eta(t))$ be an extremal triple such that $x(t) = 0$ on $[t_1, r]$. Then $\eta(t_1) \in V(t_1, r)^\perp$.*

Proof. Note, that if $x(t) = 0$ on $[t_1, r]$ then

$$0 = \dot{x}(t) = \frac{1}{2} B(t) R_{22}^{-1}(t) B^*(t) \eta^*(t) \quad \text{on } [t_1, r],$$

so that

$$0 = \eta(t) B(t) R_{22}^{-1}(t) B^*(t) \eta^*(t) \quad \text{on } [t_1, r].$$

The matrix $R_{22}^{-1}(t)$ is positive definite, hence $\eta(t) B(t) = 0$ on $[t_1, r]$. This implies, due to formula (3.3), that $u(t) = 0$ on $[t_1, r]$. Then, due to (3.4)

$$\eta(t) = \eta(t_1) \Phi(t_1, t) \quad \text{on } [t_1, r],$$

and finally (3.5) yields

$$\eta(t_1) \Phi(t_1, t) B(t) = 0 \quad \text{on } [t_1, r].$$

In other words $\eta(t_1) \in V(t_1, r)^\perp$.

We present now the main result of the section.

THEOREM 3.2. *Let $(x(t), u(t), \eta(t))$ and $(y(t), w(t), \mu(t))$ be two extremal triples with $x(t_1) = y(t_1)$. Then $x(t) = y(t)$ on $[t_1, r]$ if and only if $\eta(t_1) - \mu(t_1) \in V(t_1, r)^\perp$. If $x(r) = y(r)$, then $(x(t), u(t)) = (y(t), w(t))$ on $[t_1, r]$ and $\eta(t) - \mu(t) = (\eta(t_1) - \mu(t_1))\Phi(t_1, t)$ on $[t_1, r]$.*

Proof. Invoke Lemmas 3.1 and 3.2.

Thus it turns out that the set of branching points does not depend on the chosen extremal trajectory and the cost functional. It makes sense, therefore, to speak about the set of branching points of the linear system (3.1) on the time interval $[t_1, t_2]$. This set is exactly the set of discontinuity points of the function $\dim V(t_1, s)$. In particular, in the case where the coefficients of the system are constants, i.e., $A(t) = A$ and $B(t) = B$, there are no branching points at all.

Remark 3.1. Note that in the case of the linear control system (3.1) with the quadratic cost (3.2) the Hamiltonian system (2.5), (2.6) is a system of linear differential equations with respect to (x, η) . Let $(x(t; \xi), \eta(t; \xi))$ be a solution of this system with $x(t_1; \xi) = 0$ and $\eta(t_1, \xi) = \xi$. For each time $t \in [t_1, t_2]$ consider a linear mapping $\phi_t: R^n \rightarrow R^n$ defined as follows:

$$\phi_t(\xi) = x(t; \xi).$$

In accordance with Theorem 3.2 the rank of the matrix $\partial\phi_t(\cdot)/\partial\xi$ is $\dim V(t_1, t)$. Namely

$$(3.6) \quad \text{rank} \frac{\partial\phi_t(\cdot)}{\partial\xi} = \dim V(t_1, t).$$

On the other hand $\dim V(t_1, t) = \dim AT(t; t_1, 0)$. Hence for each $x \in AT(t; t_1, 0)$ there exists an extremal $x(s)$ with $x(t_1) = 0$ and $x(t) = x$. This remark and the relation (3.6) will be extremely useful for future investigation of the branching points in nonlinear control problems.

4. Nonlinear control problem. In this section we shall be concerned with the nonlinear control problem (2.1), (2.2). The main result of the section is the following: *Let $(x(t), u(t), \eta(t))$ be an extremal triple such that Hypothesis 2.1 and Hypothesis 2.2 are satisfied. The set of branching points of $x(t)$ coincides with that of the linear control system*

$$(4.1) \quad \dot{z}(t) = A(t)z(t) + B(t)v(t),$$

where $A(t) = \partial F / \partial x(t, x(t), u(t))$ and $B(t) = \partial F / \partial u(t, x(t), u(t))$.

However, in order to obtain the result we need to impose an additional restriction. Consider a solution $(y(t), \mu(t))$ of the Hamiltonian system (2.5), (2.6) with $y(t_1) = x_1$, $\mu(t_1) = \xi$. Denote this solution by $(x(t; x_1, \xi), \eta(t; x_1, \xi))$. Once selected, $x(t_1)$ will not be changed throughout this section. Hence, this solution can be unambiguously denoted by $(x(t; \xi), \eta(t; \xi))$.

We wish to ensure the smooth (C^1) dependence of solutions $\{x(t; \xi)\}$ on initial conditions $\{\xi\}$. To this end an additional assumption is introduced.

Assumption 4.1. There exists a Lebesgue integrable function $m(t)$ such that for each extremal triple $(y(t), w(t), \mu(t))$ generated by $\mu(t_1)$ close to $\eta(t_1)$, the following condition holds:

$$\begin{aligned} & \left| \begin{pmatrix} H_\eta(t, y(t), w(t), \mu(t)) \\ -H_x(t, y(t), w(t), \mu(t)) \end{pmatrix} \right| \\ & + \left| \begin{pmatrix} H_{\eta x}(t, y(t), w(t), \mu(t)), & H_{\eta\eta}(t, y(t), w(t), \mu(t)) \\ -H_{xx}(t, y(t), w(t), \mu(t)), & -H_{x\eta}(t, y(t), w(t), \mu(t)) \end{pmatrix} \right| \leq m(t). \end{aligned}$$

The mapping $(x(t; \cdot), \eta(t; \cdot)): R^n \mapsto R^n \times R^n (\xi \mapsto (x(t; \xi), \eta(t; \xi)))$ is, therefore, continuously differentiable with respect to ξ . (A proof can be easily constructed with minor changes from that of [4, Thm. 7.2, p. 25].) Our first goal is to derive sufficient conditions for matching of extremals.

Sufficient conditions for matching of extremals. Consider a solution $(y(t), \mu(t))$ of the Hamiltonian system (2.5), (2.6) with $y(t_1) = x(t_1)$. The matrices $A(t)$, $B(t)$ will henceforth be as defined at the beginning of the section and $V(t_1, r)$ and $\Phi(t_1, t)$ be correspondingly defined. Let ε_2 be a positive constant such that the condition $|\eta(t_1) - \mu(t_1)| < \varepsilon_2$ implies that for each $t \in [t_1, t_2]$

1. $|(x(t), \eta(t)) - (y(t), \mu(t))| < \varepsilon_1$,
2. $|\mu(t_1) - \eta(t_1)| |\Phi(t_1, t)| < \varepsilon_1$.

(We remind the reader, that the constant ε_1 was introduced in Hypothesis 2.2.)

LEMMA 4.1. *Let $(y(t), w(t), \mu(t))$ be an extremal triple such that $y(t_1) = x(t_1)$ and $|\mu(t_1) - \eta(t_1)| < \varepsilon_2$. If $\mu(t_1) - \eta(t_1) \in V(t_1, r)^\perp$, then $(y(t), w(t)) = (x(t), u(t))$ on $[t_1, r]$.*

Denote $(\mu(t_1) - \eta(t_1))\Phi(t_1, t)$ by $\Delta(t)$. The proof consists of the following three steps. First we show that the triple $(x(t), u(t), \eta(t) + \Delta(t))$ is an extremal triple of the control problem (2.1), (2.2) on $[t_1, r]$. Second, note that $(x(t), \eta(t) + \Delta(t))$ and $(y(t), \mu(t))$ are the solutions of the Hamiltonian system (2.5), (2.6) over an initial subinterval $[t_1, r]$ with the same initial conditions $(y(t_1), \mu(t_1))$. This yields

$$x(t) = y(t), \quad \eta(t) + \Delta(t) = \mu(t) \quad \text{on } [t_1, r].$$

Third, using the results of the first and second steps we point out the following relation:

$$u(t) = \bar{u}(t, x(t), \eta(t)) = \bar{u}(t, x(t), \eta(t) + \Delta(t)) = \bar{u}(t, y(t), \mu(t)) = w(t) \quad \text{on } [t_1, r].$$

This will finish the proof.

Proof. We wish to show that the triple $(x(t), u(t), \eta(t) + \Delta(t))$ is an extremal triple of the control problem (2.1), (2.2) over $[t_1, r]$. Note that the relation

$$H(t, x(t), u(t), \eta(t) + \Delta(t)) = \max_u H(t, x(t), u, \eta(t) + \Delta(t)) \quad \text{holds on } [t_1, r].$$

Indeed, due to the definition of $\bar{u}(t, x, \eta)$ (see Hypothesis 2.2)

$$H_u(t, x(t), u(t), \eta(t)) = 0 \quad \text{on } [t_1, t_2],$$

and $H_u(t, x(t), u(t), \eta(t) + \Delta(t)) = H_u(t, x(t), u(t), \eta(t)) + \Delta(t)B(t)$ on $[t_1, r]$.

Since $\mu(t_1) - \eta(t_1) \in V(t_1, r)^\perp$ and all bounded measurable $v(\cdot): [t_1, r] \rightarrow R^m$ belong to $U[t_1, r]$, we have

$$\Delta(t)B(t) = 0 \quad \text{on } [t_1, r].$$

It then follows that

$$H_u(t, x(t), u(t), \eta(t) + \Delta(t)) = 0 \quad \text{on } [t_1, r].$$

Suppose that

$$\max_u H(t, x(t), u, \eta(t) + \Delta(t)) = H(t, x(t), v(t), \eta(t) + \Delta(t)) \quad \text{on } [t_1, r].$$

The last relation yields

$$H_u(t, x(t), v(t), \eta(t) + \Delta(t)) = 0 \quad \text{on } [t_1, r].$$

This implies (in light of a lemma that we shall provide later) that $u(t) = v(t)$ on $[t_1, r]$, namely $u(t) = \bar{u}(t, x(t), \eta(t) + \Delta(t))$ and, therefore, $(x(t), u(t), \eta(t) + \Delta(t))$ is indeed an extremal triple. This completes the proof.

Here is the missing link we promised.

LEMMA 4.2. Let $\Delta(t) = \Delta(t_1)\Phi(t_1, t)$. If $|\Delta(t)| < \varepsilon_1$ for each $t \in [t_1, t_2]$, then the condition

$$H_u(t, x(t), u(t), \eta(t) + \Delta(t)) = 0 \quad \text{on } [t_1, r]$$

implies that $\bar{u}(t, x(t), \eta(t) + \Delta(t)) = u(t)$ on $[t_1, r]$.

Proof. Let $t \in [t_1, r]$. Our main goal is to show that $\bar{u}(t, x(t), \eta(t) + \Delta(t)) = u(t)$. Let λ be a scalar between 0 and 1; then

$$H_u(t, x(t), u(t), \eta(t) + \lambda \Delta(t)) = 0$$

and

$$v H_{uu}(t, x(t), u(t), \eta(t) + \lambda \Delta(t)) v \leq -\frac{m}{2} |v|^2 \quad \text{for each } v \in R^m.$$

Hence (due to the Implicit Function Theorem) for each $\lambda \in [0, 1]$ there exist positive scalars ε_λ and δ_λ , and a unique differentiable with respect to (y, μ) function $v_\lambda(t, y, \mu)$ defined on triples (t, y, μ) in an ε_λ neighborhood of $(t, x(t), \eta(t) + \lambda \Delta(t))$ (i.e. $|(y, \mu) - (x(t), \eta(t) + \lambda \Delta(t))| < \varepsilon_\lambda$) such that

$$H_u(t, y, v_\lambda(t, y, \mu), \mu) = 0 \quad \text{and} \quad v_\lambda(t, x(t), \eta(t) + \theta \Delta(t)) = u(t) \\ \text{for } \theta \in [\lambda - \delta_\lambda, \lambda + \delta_\lambda].$$

Consider a vector valued function $\phi(\theta)$ defined by the relation

$$\phi(\theta) = \bar{u}(t, x(t), \eta(t) + \theta \Delta(t)).$$

Note, that $\phi(\theta)$ is a continuous function (due to Hypothesis 2.2) and $\phi(0) = u(t)$. We intend to show that $\phi(\theta) = \phi(0)$ for each $\theta \in [0, 1]$.

Define a scalar $\lambda \in [0, 1]$ by the relation

$$\lambda = \inf_{\theta} \{ \phi(\theta) \neq \phi(0) \mid \theta \in [0, 1] \}.$$

It is clear that $\lambda \geq 0$, $\phi(\lambda) = \phi(0) = u(t)$, and we will show that $\lambda = 1$. Suppose the opposite, namely $\lambda < 1$. In this case

$$H_u(t, x(t), v_\lambda(t, x(t), \eta(t) + \theta \Delta(t)), \eta(t) + \theta \Delta(t)) = 0 \quad \text{for } \theta \in [\lambda - \delta_\lambda, \lambda + \delta_\lambda],$$

and

$$v_\lambda(t, x(t), \eta(t) + \theta \Delta(t)) = u(t) = \phi(\lambda) = \phi(0) \quad \text{for } \theta \in [\lambda - \delta_\lambda, \lambda + \delta_\lambda].$$

On the other hand

$$H_u(t, x(t), \phi(\theta), \eta(t) + \theta \Delta(t)) = 0 \quad \text{for } \theta \in [0, 1].$$

This implies (due to the Implicit Function Theorem) that

$$\phi(\theta) = u(t) = \phi(\lambda) = \phi(0) \quad \text{for } \theta \in [0, \lambda + \delta_\lambda].$$

This contradiction completes the proof.

The next step is a derivation of necessary conditions for matching of extremals.

Necessary conditions for matching of extremals. Our main goal in the remainder of the section is to show that the condition $x(t) = y(t)$ on $[t_1, r]$ implies that $\eta(t_1) - \mu(t_1) \in V(t_1, r)^\perp$. Consider a solution $(y(t), \mu(t))$ of the Hamiltonian system (2.5), (2.6) with $y(t_1) = x_1$, $\mu(t_1) = \xi$. Denote this solution by $(x(t; \xi), \eta(t; \xi))$. The questions of considerable interest are “how many different ξ steer x_1 to the same terminal point $x(t; \eta(t_1))$?” or, to put it another way, “how many solutions does an equation $x(t; \xi) = x(t; \eta(t_1))$ have?” In other words “how many neighboring extremals intersect $x(s; \eta(t_1))$ at time t ?” The answers to these questions are closely related to knowledge of rank $\partial x / \partial \xi(t; \eta(t_1))$.

In order to determine the rank $\partial x/\partial \xi(t; \eta(t_1))$ we consider the auxiliary linear control system

$$(4.2) \quad \dot{z}(t) = A(t)z(t) + B(t)v(t), \quad z(t_1) = 0$$

with a cost functional

$$(4.3) \quad \int_{t_1}^{t_2} h(t, z(t), v(t)) dt,$$

where $h(t, z, v)$ is a quadratic function with respect to (z, v) , namely

$$h(t, z, v) = \begin{pmatrix} z \\ v \end{pmatrix}^* \begin{pmatrix} R_{11}(t) & R_{12}(t) \\ R_{21}(t) & R_{22}(t) \end{pmatrix} \begin{pmatrix} z \\ v \end{pmatrix},$$

such that Hypothesis 3.1 is satisfied. In this case the adjoint equation has the following form:

$$\dot{\theta} = \frac{\partial h}{\partial z}(t, z(t), v(t)) - \theta A(t).$$

Let $(z(t), \theta(t))$ be an extremal trajectory of the control problem (4.2), (4.3), and a corresponding solution of the adjoint equation with $z(t_1) = 0$ and $\theta(t_1) = \xi$. We denote this pair by $(z(t; \xi), \theta(t; \xi))$. Note that $\text{rank } \partial z/\partial \xi(t; 0)$ is the dimension of the attainable set at time t . Specifically,

$$(4.4) \quad \text{rank } \frac{\partial z}{\partial \xi}(t; 0) = \dim V(t_1, t)$$

(see (3.6) and Remark 3.1).

We intend to construct the linear control problem (4.2), (4.3) in such a way that

$$\text{rank } \frac{\partial x}{\partial \xi}(t; \eta(t_1)) = \text{rank } \frac{\partial z}{\partial \xi}(t; 0) \quad \text{for each } t \in [t_1, t_2],$$

where $x(t; \eta(t_1))$ is the extremal trajectory of the control problem (2.1), (2.2), and $z(t; 0)$ is the extremal trajectory of the linear control problem (4.2), (4.3). The matrices $A(t)$ and $B(t)$ have been already defined and all that we have left to construct is a cost functional $h(t, z, v)$. Define $h(t, z, v)$ as follows:

$$h(t, z, v) = -\frac{1}{2} \begin{pmatrix} z \\ v \end{pmatrix}^* \begin{pmatrix} H_{xx}(t, x(t), u(t), \eta(t)) & H_{ux}(t, x(t), u(t), \eta(t)) \\ H_{xu}(t, x(t), u(t), \eta(t)) & H_{uu}(t, x(t), u(t), \eta(t)) \end{pmatrix} \begin{pmatrix} z \\ v \end{pmatrix}.$$

Here by $*$ we indicate the transpose. Note that Hypothesis 2.1 implies the fulfillment of Hypothesis 3.1 for the linear-quadratic control problem (4.2), (4.3).

Denote by the $(n+n) \times (n+n)$ matrix

$$\Psi(t, t_1) = \begin{pmatrix} \Psi_{11}(t, t_1) & \Psi_{12}(t, t_1) \\ \Psi_{21}(t, t_1) & \Psi_{22}(t, t_1) \end{pmatrix} \quad \text{with } \Psi(t_1, t_1) = I,$$

a fundamental solution of the following linear differential equation:

$$\dot{X} = \begin{pmatrix} H_{xx}(t, x(t), u(t), \eta(t)) & H_{\eta\eta}(t, x(t), u(t), \eta(t)) \\ -H_{x\eta}(t, x(t), u(t), \eta(t)) & -H_{\eta\eta}(t, x(t), u(t), \eta(t)) \end{pmatrix} X,$$

where $X \in R^{n+n}$. In this case

$$(4.5) \quad \frac{\partial x}{\partial \xi}(t; \eta(t_1)) = \Psi_{12}(t, t_1)$$

(see Coddington and Levinson [4, p. 25]). On the other hand a direct computation shows that

$$\frac{\partial z}{\partial \xi}(t; 0) = \Psi_{12}(t, t_1) \text{ and, therefore, } \frac{\partial z}{\partial \xi}(t; 0) = \frac{\partial x}{\partial \xi}(t; \eta(t_1)).$$

This implies that $\text{rank } \partial x / \partial \xi(t; \eta(t_1)) = \dim V(t_1, t)$ (see 4.4). Hence, due to the Implicit Function Theorem, for each $t \in [t_1, t_2]$ there exists a constant $\varepsilon(t) > 0$, such that the set

$$\Xi(t) = \{\xi \mid \xi \in R^n, |\xi - \eta(t_1)| < \varepsilon(t), x(t; \xi) = x(t; \eta(t_1))\}$$

is a subset of an n -dim $V(t_1, t)$ -dimensional manifold.

In what follows we describe the family of the sets $\Xi(t)$; in particular we shall show, that $\Xi(t)$ is an n -dim $V(t_1, t)$ -dimensional manifold for each $t \in [t_1, t_2]$. First, choose a positive ε_3 such that the condition $|\Delta \Phi(t_1, s)| < \varepsilon(s)$ holds for each $|\Delta| < \varepsilon_3$, $t_1 \leq s \leq t_2$. Next, choose a vector Δ in such a way that

$$(1) \quad |\Delta| < \varepsilon_3,$$

$$(2) \quad \Delta \in V(t_1, t)^\perp.$$

(The continuity of $\partial x / \partial \xi(t; \eta(t_1))$ in t (see (4.5)) implies existence of such positive ε_3 .) For each vector Δ satisfying conditions 1 and 2 consider an extremal triple

$$(y(s), w(s), \eta(s) + \Delta \Phi(t_1, s)).$$

Due to Lemma 4.1 the corresponding extremal pair $(y(s), w(s))$ coincides with $(x(s), u(s))$ on $[t_1, t]$, in particular $y(t) = x(t)$. The collection of vectors Δ which satisfy the conditions 1 and 2 above is a linear manifold. The dimension of this linear manifold is

$$\dim V(t_1, t)^\perp = n - \dim V(t_1, t).$$

This implies that

$$\Xi(t) = \{\eta(t_1) + \Delta \mid \Delta \text{ satisfies conditions 1 and 2}\}.$$

Hence $\Xi(t)$ itself is an n -dim $V(t_1, t)$ -dimensional linear manifold.

We obtain, therefore, the following.

LEMMA 4.3.

$$\Xi(t) = \{\xi \mid |\xi - \eta(t_1)| < \varepsilon_3, \xi - \eta(t_1) \in V(t_1, t)^\perp\}.$$

COROLLARY 4.1.

$$\Xi(s) \subset \Xi(\sigma) \quad \text{provided } s \geq \sigma.$$

Proof. If $s \geq \sigma$ then

$$V(t_1, \sigma) \subset V(t_1, s);$$

this implies

$$V(t_1, s)^\perp \subset V(t_1, \sigma)^\perp.$$

The result now follows from Lemma 4.3.

At this point we are able to present the definition of the positive scalar ε^* , which defines the family of the neighboring extremals, as follows.

DEFINITION 4.1. Denote $\min(\varepsilon_2, \varepsilon_3)$ by ε^* . (We wish to remind the reader that ε_2 was introduced in Lemma 4.1.) An extremal trajectory $y(t)$ is a neighboring extremal trajectory if

$$\max_{t_1 \leq t \leq t_2} |x(t) - y(t)| + \max_{t_1 \leq t \leq t_2} |\eta(t) - \mu(t)| < \varepsilon^*.$$

In the case of a linear control system with a quadratic cost functional there exists a unique extremal trajectory which steers an initial point x_1 to a terminal point x_2 over the time interval $[t_1, t_2]$. The next theorem shows that the selected trajectory $x(t)$ has this property "in the small."

THEOREM 4.1. Consider an extremal triple $(y(t), w(t), \mu(t))$. If

- (1) $y(t)$ is a neighboring extremal,
- (2) $y(t_1) = x(t_1)$, $y(r) = x(r)$ for some $r \in [t_1, t_2]$, then $y(t) = x(t)$ on $[t_1, r]$.

Proof. The conditions imply that $\mu(t_1) \in \Xi(r)$. On the other hand, due to Corollary 4.1, the relation $\Xi(r) \subset \Xi(t)$ holds for each t such that $t_1 \leq t < r$. Hence, $\mu(t_1) \in \Xi(t)$ for each $t \in [t_1, r]$. This completes the proof.

We show by the following example that the condition $|(x(\cdot), \eta(\cdot)) - (y(\cdot), \mu(\cdot))|_{C[t_1, t_2]} < \varepsilon^*$ cannot be removed. Hence, Theorem 4.1 has a local nature.

Example 4.1. The notion of neighboring extremal plays an important role in the investigation of branching points. In this example we present a bilinear control problem with extremals $x(t)$ and $y(t)$ satisfying the relations

$$x(0) = y(0), \quad x(\pi) = y(\pi), \quad x(t) \neq y(t) \quad \text{for } t \in (0, \pi).$$

Consider the following bilinear control system:

$$(4.6) \quad \dot{x}_1(t) = u(t)x_2(t),$$

$$(4.7) \quad \dot{x}_2(t) = -u(t)x_1(t),$$

where $u(t)$ is a scalar function.

Consider the cost functional

$$(4.8) \quad c(x(\cdot), u(\cdot)) = \frac{1}{2} \int_0^{2\pi} u^2(t) dt.$$

The adjoint equation has the following form:

$$\dot{\eta}_1(t) = u(t)\eta_2(t), \quad \dot{\eta}_2(t) = -u(t)\eta_1(t).$$

The Hamiltonian H of this problem is given as follows:

$$H(t, x, u, \eta) = -\frac{1}{2}u^2 + (\eta_1x_2 - \eta_2x_1)u.$$

Note that $\eta_1(t)x_2(t) - \eta_2(t)x_1(t)$ is a constant and, therefore, $u(t) = u(0)$ for $t \in [0, 2\pi]$.

The extremal trajectories $x_\omega(t)$ of the above optimal control problem are rotations around the origin with constant angular velocities ω . For each real number ω , the extremals $x_\omega(t)$ and $x_{-\omega}(t)$ initiating at the same point $x(0)$ intersect each other at $t = \pi/\omega$. Hypothesis 2.2 is evidently fulfilled and we will choose ω in such a way that the extremal trajectory $x_\omega(t)$ will satisfy Hypothesis 2.1 on $[0, 2\pi]$.

In order to find the suitable ω , choose the following set of initial conditions:

$$x_1(0) = 1, \quad x_2(0) = 0, \quad \eta_1(0) = 0, \quad \eta_2(0) = -1.$$

Straightforward verification would show that in this case $u(t) = 1$ and

$$x_1(t) = \cos t, \quad x_2(t) = -\sin t, \quad \eta_1(t) = -\sin t, \quad \eta_2(t) = -\cos t.$$

The linearized system about this solution is the following:

$$\dot{z}_1(t) = z_2(t) - v(t) \sin t, \quad \dot{z}_2(t) = -z_1(t) - v(t) \cos t.$$

Those solutions $z(t) = (z_1(t), z_2(t))$ of this linear equation that also satisfy the boundary conditions

$$z_1(0) = z_1(2\pi) = z_2(0) = z_2(2\pi) = 0$$

have the form

$$(4.9) \quad z_1(t) = -\sin t \int_0^t v(s) ds, \quad z_2(t) = -\cos t \int_0^t v(s) ds.$$

We now wish to show that the extremal triple $(x(t), u(t), \eta(t))$ satisfies Hypothesis 2.1. A straightforward computation shows that

(1) $H_{uu}(t, x, u, \eta)$ is identically equal to -1 ,

$$(2) \quad \int_0^{2\pi} \begin{pmatrix} z(t) \\ v(t) \end{pmatrix}^* \begin{pmatrix} H_{xx}(t, x(t), u(t), \eta(t)) & H_{ux}(t, x(t), u(t), \eta(t)) \\ H_{xu}(t, x(t), u(t), \eta(t)) & H_{uu}(t, x(t), u(t), \eta(t)) \end{pmatrix} \begin{pmatrix} z(t) \\ v(t) \end{pmatrix} dt \\ = - \int_0^{2\pi} v^2(t) dt$$

is negative definite for each $z(t)$ satisfying condition (4.9).

On the other hand, the trajectory $y(t) = (\cos t, \sin t)$ is an extremal trajectory of the optimal control problem (4.6)–(4.8) and

$$y(\pi) = x(\pi), \quad y(0) = x(0) \quad \text{but} \quad y(t) \neq x(t) \quad \text{for} \quad 0 < t < \pi.$$

Note that $\mu_1(0) = 0$, $\mu_2(0) = 1$ and $\eta_1(0) = 0$, $\eta_2(0) = -1$; hence $|(x(\cdot), \eta(\cdot)) - (y(\cdot), \mu(\cdot))|_{C[t_1, t_2]} \geq 2$. On the other hand, one can show that in this case ε^* , which defines the extremal trajectories, must be less than 1.

We present now the last necessary definition and state the main result of the section.

DEFINITION 4.2. We denote the set of discontinuity points of the function $\dim V(t_1, s)$ on the interval $[t_1, t_2]$ by $\{r_i\}$.

THEOREM 4.2. Each $r \in \{r_i\}$ is a branching point of the extremal trajectory $x(t)$. Let $y(t)$ be a neighboring extremal trajectory. If $y(t_1) = x(t_1)$ and $y(s) \neq x(s)$ for some $s \in [t_1, t_2]$, then there exists $r \in [t_1, s]$ such that the following conditions hold:

- (1) $r \in \{r_i\}$,
- (2) $y(t) = x(t)$ on $[t_1, r]$,
- (3) $y(t) \neq x(t)$ on $(r, t_2]$.

COROLLARY 4.2. The set of branching points of the extremal trajectory $x(t)$ is a finite set. The number of its elements is no more than the dimension of the state space R^n .

Acknowledgment. The author would like to thank the referee whose valuable remarks and corrections much improved the exposition of the results.

REFERENCES

- [1] N. I. AKHIEZER, *The Calculus of Variations*, Blaisdell, Boston, 1962.
- [2] L. D. BERKOVITZ, *Optimal Control Theory*, Springer-Verlag, Berlin, New York, 1974.
- [3] C. CARATHÉODORY, *Calculus of Variations and Partial Differential Equations of the First Order*, two vols., Holden-Day, San Francisco, CA, 1965, 1967.

- [4] E. A. CODDINGTON AND N. LEVINSON, *Theory of Ordinary Differential Equations*, McGraw-Hill, New York, 1955.
- [5] L. CESARI, *Optimization—Theory and Applications*, Springer-Verlag, Berlin, New York, 1983.
- [6] F. H. CLARKE AND V. ZEIDAN, *Sufficiency and the Jacobi condition in the calculus of variations*, Technical report, Université de Montreal, CRM-1273, 1985.
- [7] W. F. FLEMING AND R. V. RISHEL, *Deterministic and Stochastic Optimal Control*, Springer-Verlag, Berlin, New York, 1975.
- [8] I. M. GELFAND AND S. V. FOMIN, *Calculus of Variations*, Prentice-Hall, Englewood Cliffs, NJ, 1963.
- [9] M. R. HESTENES, *Calculus of Variations and Optimal Control Theory*, John Wiley, New York, 1966.
- [10] A. D. IOFFE AND V. M. TIHOMIROV, *Theory of extremal problems*, in *Studies in Mathematics and its Applications*, Vol. 6, North-Holland, Amsterdam, 1979.
- [11] E. B. LEE AND L. MARKUS, *Foundations of Optimal Control Theory*, John Wiley, New York, 1967.
- [12] M. SPIVAK, *Calculus on Manifolds*, W. A. Benjamin, New York, 1965.
- [13] L. C. YOUNG, *Calculus of Variations and Optimal Control Theory*, W. B. Saunders, Philadelphia, PA, 1969.

REFINED STUDY ON STRUCTURAL CONTROLLABILITY OF DESCRIPTOR SYSTEMS BY MEANS OF MATROIDS*

KAZUO MUROTA†

Abstract. For a dynamical system described in the descriptor form $F dx/dt = Ax + Bu$, where the coefficients are classified into generic parameters and fixed constants, the structural controllability is investigated under a physically reasonable assumption that can be justified by the dimensional analysis. A necessary and sufficient condition for the structural controllability is given in matroid-theoretic terms; the condition can be tested by efficient algorithms for the matroid union/intersection problem.

Key words. structural controllability; descriptor form; matroid; algebraic independence; polynomial algorithm

AMS(MOS) subject classifications. 93, 93B05, 93B25, 15, 05C50

1. Introduction. Since the notion of structural controllability was introduced in [21], many papers have appeared for its extensions and refinements [2], [3], [10], [11], [24]–[30], [35], [39]. A dynamical system described in the state-space standard form

$$(1.1) \quad dx/dt = Ax + Bu$$

is called structurally controllable if it is controllable in the ordinary sense [18] when the nonvanishing entries of A and B are replaced by independent free parameters. The structural controllability of (1.1) is known to be expressed by graph-theoretic conditions [10], [21], [24], [35].

Though the notion of structural controllability is quite appealing, it is often unjustifiable to assume that the nonvanishing entries of A and B of (1.1) are independent. In this respect, it is more appropriate to work with the descriptor form [22], [23]

$$(1.2) \quad F dx/dt = Ax + Bu$$

($x \in \mathbb{R}^n$ and $u \in \mathbb{R}^m$), which is more elementary and hence more suitable for the representation of a physical structure. The generic controllability of the descriptor system (1.2) is discussed in graph-theoretic terms in [3], [26], [30], [39] under the assumption that the nonvanishing entries of F , A and B of (1.2) are independent parameters.

As will readily be imagined, however, not all the nonvanishing entries of the matrices in (1.2) can be modeled as independent parameters, but some must be fixed constants, usually simple integers such as ± 1 . With this problem of fixed constants taken into account, the generic controllability condition is derived in [2] for the standard form (1.1) when the coefficients can be expressed in “matrix nets” as

$$A = A_0 + \sum_{i=1}^k \mu_i A_i, \quad B = B_0 + \sum_{i=1}^k \mu_i B_i,$$

where A_i and B_i are fixed matrices and μ_i independent free parameters. The condition given in [2] is, however, purely algebraic and cannot readily be tested by efficient algorithms.

In investigating structural aspects of large-scale systems in general, it would be of fundamental importance to set up a mathematical model which accounts for relevant aspects of a real system with sufficient faith and which admits rigorous mathematical

* Received by the editors July 8, 1985; accepted for publication (in revised form) June 5, 1986.

† Institute of Socio-Economic Planning, University of Tsukuba, Sakura, Ibaraki 305, Japan. Present address, Department of Mathematical Engineering and Instrumentation Physics, University of Tokyo, Bunkyo-ku, Tokyo 113, Japan.

analysis. No less important are algorithmic considerations, especially when combinatorial methods are involved in the analysis of the model. A combinatorial method for systems analysis would not be very powerful for large systems unless it is accompanied by efficient algorithms.

In this respect, the matroid-theoretic point of view can be quite useful both in establishing adequate mathematical models and in constructing efficient algorithms (see, e.g., [14], [15]). In the field of electrical network theory, for example, a number of fundamental problems, such as the problem of the topological degrees of freedom and that of the order of complexity for networks with mutual couplings, have been solved with the aid of matroid-theoretic concepts and algorithms (see, e.g., [14]–[16]). Once results are obtained by means of matroids, it is possible, as a matter of course, to restate all the arguments without reference to matroids. It would, however, make things less clear, hiding the essential simplicity and insight.

In this paper, we will formulate the problem of structural controllability so as to reflect the real situations fairly well and give a solution to it by means of matroids. To be more specific, the generic controllability is considered for a descriptor system (1.2) in which the nonvanishing entries of F , A and B are classified into two groups, i.e., independent free parameters and fixed constants. Following the arguments in [31] on the algebraic implications of the consistency in the system of equations (1.2) with respect to physical dimensions, a physically reasonable assumption is made on the matrix representing the fixed constants. A necessary and sufficient condition for the structural controllability in this refined sense is derived with the aid of the combinatorial canonical form of a layered mixed matrix, the mathematical tool introduced in [32]. The condition can be checked by the efficient combinatorial algorithms for the matroid union/intersection problem or for the independent-flow problem.

The proposed method for testing the structural controllability has several practical advantages. The algorithm is guaranteed to run in $O(n^2(n+m) \log n)$ time in the worst case and runs much faster in many cases. It is free from the numerical difficulty of rounding errors as compared to the technique of parameter variation, which computes the rank of the controllability matrix by substituting several different numerical values to the parameters. The algorithm can be implemented easily as it is composed of the Gaussian elimination on simple rational numbers, the construction of matchings, and the search for paths in graphs. In the special case where no fixed constants are involved, the result of the present paper naturally reduces to the previously known results mentioned above.

2. Preliminaries. This section presents the minimum set of matroid-theoretic concepts to be required in later arguments and gives a brief summary of the combinatorial canonical form of a layered mixed matrix introduced in [32]. See, e.g., [38] for the complete account of matroids.

A matroid is a pair $\mathcal{M} = (S, \mathcal{I})$ of a finite set S and a collection \mathcal{I} of subsets of S such that

- (I1) $\emptyset \in \mathcal{I}$;
- (I2) If $X \in \mathcal{I}$ and $Y \subset X$, then $Y \in \mathcal{I}$;
- (I3) If $X, Y \in \mathcal{I}$ and $|X| = |Y| + 1$, then $Y \cup x \in \mathcal{I}$ for some $x \in X \setminus Y$.

A member of \mathcal{I} is called an *independent set*, and a maximal subset in \mathcal{I} (maximal with respect to set inclusion) a *base*. An element of S is a *coloop* if it is contained in every base. A matroid is called a *free matroid* if every subset is independent in it.

For a matroid \mathcal{M} defined on S and a subset X of S , the *restriction* of \mathcal{M} to X , denoted as $\mathcal{M}|X$, is a matroid on X in which $Y (\subset X)$ is independent iff Y is independent

in \mathcal{M} , and the *contraction* of \mathcal{M} to X , denoted as $\mathcal{M}.X$, is a matroid on X in which $Y (\subset X)$ is independent iff $Y \cup B$ is independent in \mathcal{M} where B is a base of $\mathcal{M}|(S \setminus X)$. A subset X of S consists of coloops of \mathcal{M} iff $\mathcal{M}.X$ is a free matroid.

All bases of a matroid have the same cardinality, which is called the *rank* of the matroid \mathcal{M} and is denoted as $\text{rank}[\mathcal{M}]$. The function $\rho: 2^S \rightarrow \mathbb{Z}$ which assigns the rank of $\mathcal{M}|X$ to $X (\subset S)$ is called the *rank function* of \mathcal{M} ; $\rho(X)$ is the maximum size of an independent set included in X . The rank function ρ^X of $\mathcal{M}.X$ is given by

$$(2.1) \quad \rho^X(Y) = \rho(Y \cup (S \setminus X)) - \rho(S \setminus X) \quad \text{for } Y \subset X.$$

A matroid is also determined by the collection of bases. For a matroid \mathcal{M} on S , its dual, denoted as \mathcal{M}^* , is the matroid on S in which a subset of S is a base iff it is the complement of a base of \mathcal{M} .

Given a matrix A over a field \mathbf{K} , we will denote by $\mathcal{M}(A)$ the matroid defined on the column-set C of A with respect to the ordinary linear dependence among column-vectors of A . A matroid thus obtained is called a linear matroid represented over \mathbf{K} . Let $\mathcal{N}(A)$ designate the (multi)set of nonvanishing entries of a matrix A . In the particular case where $\mathcal{N}(A)$ is algebraically independent [37], the matroid $\mathcal{M}(A)$ is a transversal matroid expressed by the bipartite graph G associated with A ; the vertex-set of G is the union of the row-set R and the column-set C of A , the edge-set of G has the natural one-to-one correspondence with $\mathcal{N}(A)$, and a set $X (\subset C)$ is independent iff X can be matched into R in G .

The relation of algebraic independence over a field also enjoys the matroidal properties. Let \mathbf{K} and $\mathbf{F} (\mathbf{K} \subset \mathbf{F})$ be fields and $S (\subset \mathbf{F})$ a finite set. Then a matroid is defined on S with respect to the algebraic independence over \mathbf{K} . Such a matroid is called an algebraic matroid. The following will be used later as a key property of algebraic independence.

LEMMA 2.1. *Let X and $\{z\}$ be independent in a matroid. If $X \cup z$ is not independent, then $(X \setminus x) \cup z$ is independent for some $x \in X$. \square*

For k matroids $\mathcal{M}_i (i = 1, \dots, k)$ defined on S , their union $\mathcal{M}_1 \vee \dots \vee \mathcal{M}_k$ is a matroid on S in which $X (\subset S)$ is independent iff X can be expressed as $X = X_1 \cup \dots \cup X_k$ with X_i being independent in $\mathcal{M}_i (i = 1, \dots, k)$. The rank function ρ of the union matroid is expressed in terms of the rank functions ρ_i of \mathcal{M}_i as

$$(2.2) \quad \rho(X) = \min \left\{ \sum_{i=1}^k \rho_i(Y) + |X \setminus Y| \mid Y \subset X \right\}, \quad X \subset S.$$

In general, the union of contractions $\mathcal{M}_i.X (X \subset S)$ does not agree with the contraction of the union of \mathcal{M}_i , i.e., $\bigvee_{i=1}^k (\mathcal{M}_i.X) \neq (\bigvee_{i=1}^k \mathcal{M}_i).X$. Still the following holds true.

LEMMA 2.2. *Let $\hat{S} (\subset S)$ be the set of all coloops of $\bigvee_{i=1}^k \mathcal{M}_i$.*

(i) *$\hat{S} = S \setminus X_0$, where X_0 is the (uniquely determined) smallest subset of S that gives the minimum value of*

$$\tilde{\rho}(X) = \sum_{i=1}^k \rho_i(X) - |X|, \quad X \subset S.$$

(ii) *$\bigvee_{i=1}^k (\mathcal{M}_i.\hat{S})$ is the free matroid.*

Proof. (i) Since $\tilde{\rho}$ is submodular, its minimizers constitute a sublattice of 2^S . In particular, there exists the smallest minimizer X_0 of $\tilde{\rho}$. For $x \in S$, $x \in \hat{S}$ iff $\rho(S \setminus x) = \rho(S) - 1$, which is equivalent, by (2.2), to $\min \{\tilde{\rho}(Y) \mid x \notin Y\} = \min \{\tilde{\rho}(Y) \mid Y \subset S\}$, namely to $x \notin X_0$.

$$\text{rank } \bar{A}[R_\infty, C_\infty] = |C_\infty| \quad (< |R_\infty| \text{ if } C_\infty \neq \emptyset).$$

Note that the column-set of \bar{A} has a natural correspondence with that of A .

LEMMA 2.4. $C \setminus C_0$ is the set of all coloops of the matroid $\mathcal{M}(A) = \mathcal{M}(Q) \vee \mathcal{M}(T_1) \vee \mathcal{M}(T_2)$.

Proof. This follows from Lemma 2.2(i), since, by the definition of the combinatorial canonical form and [34, Lemma 5.1(5)], C_0 is the smallest subset of C that gives the minimum of $\tilde{\rho}$ in Lemma 2.2(i), where $k=3$, and ρ_1, ρ_2 and ρ_3 are the rank functions of $\mathcal{M}(Q)$, $\mathcal{M}(T_1)$ and $\mathcal{M}(T_2)$, respectively. \square

When a matrix A over \mathbf{F} is expressed as

$$(2.7) \quad A = Q_A + T_A,$$

where Q_A is a matrix over the subfield \mathbf{K} of \mathbf{F} and $\mathcal{N}(T_A)$ is algebraically independent over \mathbf{K} , it is called a *mixed matrix* with respect to \mathbf{K} . The following is an immediate consequence of Lemma 2.3.

LEMMA 2.5 [33]. Let $A = Q_A + T_A$ be an $m \times n$ mixed matrix. Then we have

$$\text{rank } A = \text{rank } [\mathcal{M}([I|Q_A]) \vee \mathcal{M}([I|T_A])] - m. \quad \square$$

3. Formulation of the structural controllability.

3.1. Controllability of a descriptor system. We consider a linear dynamical system represented by the descriptor form [22], [23]

$$(3.1) \quad F \frac{d\mathbf{x}}{dt} = A\mathbf{x} + B\mathbf{u},$$

where $\mathbf{x} (\in \mathbf{R}^n)$ and $\mathbf{u} (\in \mathbf{R}^m)$ are the descriptor-vector (standing for the internal variables) and the input-vector, respectively, and F , A and B are constant real matrices of sizes $n \times n$, $n \times n$ and $n \times m$, respectively. If we express system (3.1) (with the zero initial state) in terms of the Laplace transform, we have

$$(3.2) \quad [A - sF|B] \begin{pmatrix} \mathbf{x} \\ \mathbf{u} \end{pmatrix} = \mathbf{0},$$

where s is a symbol standing for the differentiation with respect to time. (Throughout this paper, s is treated as an indeterminate.) The coefficient matrix of (3.2) is sometimes called the modal controllability matrix. It should be emphasized that F is not assumed to be nonsingular.

Dynamical behaviors of such a singular differential system have been investigated by many authors [4], [12], [36], [40] and several different definitions for the controllability of the descriptor system have been proposed. In order for system (3.1) to be uniquely solvable for consistent initial conditions, it must satisfy

$$(3.3) \quad \det(A - sF) \neq 0.$$

Following [3], [12], [26], we will adopt the controllability of the exponential modes (or R -controllability of [40]) as the controllability of (3.1). Namely, we will say that (3.1) is controllable if

$$(3.4) \quad \text{rank } [A - zF|B] = n \quad \text{for any complex number } z.$$

This condition is known to be equivalent to the controllability of the state-space system that is derived from (3.1) by the strict equivalence in the sense of [9] for matrix pencils. See [4] for detailed discussions on other possible definitions of the controllability of (3.1).

3.2. Physical observations. When a dynamical system is described in the descriptor form (3.1) in terms of elementary physical variables, it is often justified to assume that the nonvanishing entries of the coefficient matrices F , A and B are classified into two groups, one of generic parameters and the other of fixed constants. In other words, as advocated in [33], we can distinguish two kinds of numbers that characterize physical systems as follows: (i) those numbers representing independent physical parameters such as resistances in electrical networks which, being contaminated by various noises and errors, take inaccurate values independent of one another, so that they can be modeled as algebraically independent generic numbers, and (ii) those numbers accounting for various sorts of conservation laws such as Kirchhoff's, which, stemming from topological incidence relations, are accurate (often ± 1) in value so that no serious numerical difficulty arises in arithmetic operations on them. See [33] for further discussions.

In accordance with this physical consideration, we will assume that the matrices F , A and B in (3.1) are real matrices expressed as

$$(3.5) \quad F = Q_F + T_F, \quad A = Q_A + T_A, \quad B = Q_B + T_B,$$

where Q_F , Q_A and Q_B are matrices with rational entries and $\mathcal{N}(T_F) \cup \mathcal{N}(T_A) \cup \mathcal{N}(T_B)$ (the set of nonvanishing entries of T_F , T_A and T_B) is algebraically independent over the rational number field \mathbf{Q} . This implies that F , A and B are mixed matrices with respect to \mathbf{Q} , of the form (2.7) explained in § 2. According to (3.5), we have

$$(3.6) \quad A - sF = (Q_A - sQ_F) + (T_A - sT_F),$$

which is again a mixed matrix, but with respect to $\mathbf{Q}(s)$, since $\mathcal{N}(T_A - sT_F)$ is algebraically independent over $\mathbf{Q}(s)$. Likewise, the modal controllability matrix $[A - sF|B]$ is a mixed matrix with respect to $\mathbf{Q}(s)$ with the additive decomposition of the form (2.7) given by

$$(3.7) \quad [A - sF|B] = [Q_A - sQ_F|Q_B] + [T_A - sT_F|T_B].$$

It should be clear that from the algebraic point of view, the assumption of the algebraic independence of $\mathcal{N}(T_F) \cup \mathcal{N}(T_A) \cup \mathcal{N}(T_B)$ is tantamount to regarding their members as independent free parameters. As for the fixed constants, the assumption that the entries of Q_F , Q_A and Q_B are rationals is not essential to the subsequent

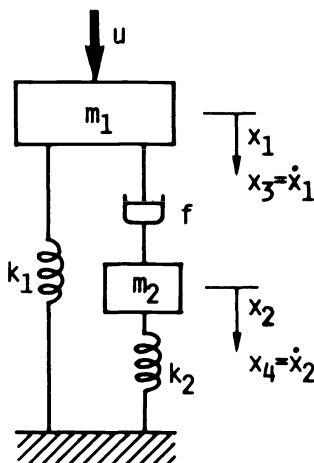


FIG. 3.1. A simple mechanical system of Example 3.1.

algebraic arguments. In the case where nonrational fixed constants are to be taken into account, we may choose any appropriate extension field of \mathbf{Q} as the subfield. The rationality of fixed constants is needed only to reduce the computational complexity for testing the controllability condition by matroid-theoretic algorithms, as will be discussed in § 4.

Example 3.1. Consider the mechanical system in Fig. 3.1 consisting of two masses m_1 and m_2 , two springs k_1 and k_2 , and a damper f ; u is the force exerted from outside. We may choose $\mathbf{x} = (x_1, x_2, x_3, x_4, x_5, x_6)$ as the descriptor-vector, where x_1 (resp. x_2) is the displacement of mass m_1 (resp. m_2) and x_3 (resp. x_4) is its velocity, x_5 is the force by the damper f , and x_6 is the relative velocity of the two masses. Then the system can be expressed in the descriptor form (3.1) with

$$(3.8) \quad F = \begin{pmatrix} 1 & & & & & \\ & 1 & & & & \\ & & m_1 & & & \\ & & & m_2 & & \\ & & & & 0 & \\ 1 & -1 & & & & 0 \end{pmatrix}, \quad A = \begin{pmatrix} 0 & 1 & & & & \\ & 0 & 1 & & & \\ -k_1 & 0 & -1 & & & \\ & -k_2 & 0 & 1 & & \\ & & -1 & f & & \\ & & & & 1 & \end{pmatrix}, \quad B = \begin{pmatrix} 0 \\ 0 \\ 1 \\ 0 \\ 0 \\ 0 \end{pmatrix}.$$

If we regard $\{m_1, m_2, k_1, k_2, f\}$ as independent free parameters, i.e., as being algebraically independent, (3.7) is given by

$$(3.9) \quad [Q_A - sQ_F | Q_B] = \left(\begin{array}{ccccc|c} -s & 1 & & & & 0 \\ & -s & 1 & & & 0 \\ & & 0 & -1 & & 1 \\ & & & 0 & 1 & 0 \\ & & & & -1 & 0 \\ -s & s & & & & 1 & 0 \end{array} \right),$$

$$[T_A - sT_F | T_B] = \left(\begin{array}{ccccc|c} 0 & & & & & 0 \\ & 0 & & & & 0 \\ -k_1 & -sm_1 & & & & 0 \\ & -k_2 & -sm_2 & & & 0 \\ & & & 0 & f & 0 \\ & & & & 0 & 0 \end{array} \right).$$

In actual situations, a small damper might exist in parallel with the spring k_1 . How to deal with such a quantity depends on how we recognize the problem; it may be ignored as above or included as a transcendental element.

The crucial physical observation made in [31] is now introduced. As mentioned above, the fixed constants usually stand for topological/geometrical incidence coefficients, which have no physical dimensions. Thus, it would be natural to expect

that the entries of Q_F , Q_A and Q_B of (3.5) are dimensionless constants. On the other hand, the indeterminate s in (3.2) should have the physical dimension of the inverse of time, since it represents the differentiation with respect to time.

Since the system of equations (3.2) is to represent a physical system, relevant physical dimensions are associated with both the variables (\mathbf{x}, \mathbf{u}) and the equations, or alternatively, with the columns and the rows of the matrix $[A - sF|B]$. Paying particular attention to the dimension of time associated with the matrix $[A - sF|B]$, we denote by c_j and r_i the exponent to the dimension of the inverse of time associated with the j th column and the i th row, respectively. Then, by the principle of dimensional homogeneity [13], [19], the (i, j) entry of $[A - sF|B]$ should have the dimension of the inverse of time with the exponent $r_i - c_j$.

Combining this fact with the nondimensionality of Q_F , Q_A and Q_B , as well as the consideration on the physical dimension of the symbol s , we obtain

$$(3.10) \quad \begin{aligned} r_i - c_j &= 1 && \text{if } (Q_F)_{ij} \neq 0, \\ r_i - c_j &= 0 && \text{if } (Q_A)_{ij} \neq 0, \\ r_i - c_{n+j} &= 0 && \text{if } (Q_B)_{ij} \neq 0. \end{aligned}$$

In the matrix form, this is written as

$$(3.11) \quad [Q_A - sQ_F|Q_B] = \text{diag}(s^{r_1}, \dots, s^{r_n})[Q_A - Q_F|Q_B] \text{diag}(s^{-c_1}, \dots, s^{-c_{n+m}}).$$

This implies that every nonvanishing subdeterminant of $[Q_A - sQ_F|Q_B]$ is a monomial in s with a rational coefficient. Thus, the matrix $[Q_A - sQ_F|Q_B]$ representing the fixed constants enjoys a very simple property.

Based on the above physical observations, we will investigate the controllability of (3.1) in the situation where the coefficient matrices are expressed as (3.5) with the following properties:

- (A1): $\mathcal{N}(T_F) \cup \mathcal{N}(T_A) \cup \mathcal{N}(T_B)$ is algebraically independent over \mathbf{Q} , and
- (A2): Every nonvanishing subdeterminant of $[Q_A - sQ_F|Q_B]$ is of the form αs^p with a rational number α and an integer p .

Whereas a matrix of the form (3.11) obviously satisfies (A2), the converse is also true [31], as stated in Lemma 3.1 below. This characterization of (A2) enables us to check it by an efficient graph-theoretic algorithm. See [31] for the detail.

LEMMA 3.1. $[Q_A - sQ_F|Q_B]$ satisfies (A2) iff it can be expressed as (3.11) for some integers r_i and c_j . \square

It should be noted that in the mechanical system of Example 3.1, the nonvanishing entries of Q_F , Q_A and Q_B are dimensionless and that the assumption (A2) is satisfied indeed.

4. Main results.

4.1. Controllability conditions in terms of matroids. In the first place, the following remarkable consequences of (A1) and (A2) are noted.

LEMMA 4.1.

$$\begin{aligned} \mathcal{M}([I|Q_A - sQ_F|Q_B]) &= \mathcal{M}([I|Q_A - Q_F|Q_B]), \\ \mathcal{M}([I|T_A - sT_F|T_B]) &= \mathcal{M}([I|T_A - T_F|T_B]). \end{aligned}$$

Proof. The former is due to Lemma 3.1, while the latter is immediate from (A1). \square

This lemma asserts that those matroids admit simple representations; the former, defined as a linear matroid over $\mathbf{Q}(s)$, is in reality representable over the subfield \mathbf{Q} ,

and the latter, being a transversal matroid, can be expressed by a bipartite graph. This fact is of practical significance in that it reduces the computational complexity in handling those matroids to a great extent.

We are concerned with combinatorial characterizations of the following conditions:

- (C1): $\det(A - sF) \neq 0$,
- (C2): $\text{rank}[A|B] = n$,
- (C3): $\text{rank}[A - zF|B] = n$ for $z \neq 0$, $z \in \mathbb{C}$.

The first is for the unique solvability given in (3.3). The set of conditions (C2) and (C3) is equivalent to the controllability given in (3.4); (C2) is for the controllability of the zero modes, while (C3) is for the nonzero modes.

The conditions (C1) and (C2) are rephrased as follows.

LEMMA 4.2. $\text{rank}(A - sF) = \text{rank}[\mathcal{M}([I|Q_A - Q_F]) \vee \mathcal{M}([I|T_A - T_F])] - n$. Hence

(C1) is equivalent to

$$(M1): \text{rank}[\mathcal{M}([I|Q_A - Q_F]) \vee \mathcal{M}([I|T_A - T_F])] = 2n.$$

Proof. From (3.6) and Lemma 2.5, we have

$$\text{rank}(A - sF) = \text{rank}[\mathcal{M}([I|Q_A - sQ_F]) \vee \mathcal{M}([I|T_A - sT_F])] - n.$$

Lemma 4.1 then simplifies the right-hand side. \square

LEMMA 4.3.

$$\text{rank}[A|B] = \text{rank}[\mathcal{M}([I|Q_A|Q_B]) \vee \mathcal{M}([I|T_A|T_B])] - n.$$

Hence (C2) is equivalent to

$$(M2): \text{rank}[\mathcal{M}([I|Q_A|Q_B]) \vee \mathcal{M}([I|T_A|T_B])] = 2n.$$

Proof. By the direct application of Lemma 2.5 to (3.7) with $s = 0$. \square

To deal with (C3), we consider three $n \times (3n + m)$ matrices

$$\begin{aligned} Q_D(s) &= [O_n|I_n|Q_A - sQ_F|Q_B], \\ (4.1) \quad T_{D1}(s) &= [I_n|O_n|-sT_F|O_{n,m}], \\ T_{D0} &= [-I_n|-I_n|T_A|T_B] \end{aligned}$$

and a composite $(3n) \times (3n + m)$ matrix

$$(4.2) \quad D(s) = \begin{pmatrix} Q_D(s) \\ T_{D1}(s) \\ T_{D0} \end{pmatrix}.$$

The last $n + m$ columns of $D(s)$ correspond to the variables \mathbf{x} and \mathbf{u} in the natural manner, whereas the first $2n$ columns may be viewed as disjoint copies of the rows of $[A - sF|B]$. Let $\{v_1, \dots, v_n\}$ denote the first n columns and $\{w_1, \dots, w_n\}$ the next n columns of $D(s)$. Then the column-set of $D(s)$, as well as of the three matrices of (4.1), is designated by

$$(4.3) \quad S = \{v_1, \dots, v_n\} \cup \{w_1, \dots, w_n\} \cup \{x_1, \dots, x_n\} \cup \{u_1, \dots, u_m\}.$$

As will be explained in § 5, $\mathcal{N}(T_{D1}(s)) \cup \mathcal{N}(T_{D0})$ may effectively be regarded as being algebraically independent over the subfield $\mathbf{K} = \mathbf{Q}(s)$ by virtually scaling the rows of T_{D1} and T_{D0} as well as the columns of $\{v_1, \dots, v_n\}$ with algebraically independent transcendentals. Hence, by Lemma 2.3, the matroid $\mathcal{M}(D(s))$ is equal to the union $\mathcal{M}(Q_D(s)) \vee \mathcal{M}(T_{D1}(s)) \vee \mathcal{M}(T_{D0})$ of three matroids, each of which is representable without involving symbol s by Lemma 4.1 as follows.

LEMMA 4.4.

$$\mathcal{M}(D(s)) = \mathcal{M}([O|I|Q_A - Q_F|Q_B]) \vee \mathcal{M}([I|O|T_F|O]) \vee \mathcal{M}([I|I|T_A|T_B]).$$

Associated with the matrix $Q_D(s)$ of (4.1) we introduce a weight function ζ_1 on S as follows. By (A2) and Lemma 3.1, there exists a set of integers r_i ($i=1, \dots, n$) and c_j ($j=1, \dots, n+m$) attached to the i th row and the j th column of $[Q_A - sQ_F|Q_B]$ such that (3.11) holds. (Such numbers are not unique from the mathematical point of view.) In case the physical dimensions associated with (3.1) are given, these numbers are readily obtained from the exponents to the inverse time dimension, as discussed in § 3. Even when the dimensions are not known, it is easy to find the numbers r_i and c_j by $O(n(n+m))$ graph operations [31]. The weight function ζ_1 is then defined by

$$(4.4) \quad \zeta_1(v_i) = 0, \quad \zeta_1(w_i) = -r_i, \quad \zeta_1(x_j) = -c_j, \quad \zeta_1(u_j) = -c_{n+j}.$$

For a subset X of S , $\zeta_1(X)$ designates the sum of the weights of the elements of X . Suppose X is independent in $\mathcal{M}([O|I|Q_A - Q_F|Q_B])$ and put $R_X = \{w_1, \dots, w_n\} \setminus X$ and $C_X = (\{x_1, \dots, x_n\} \cup \{u_1, \dots, u_m\}) \cap X$. Then the submatrix of $[Q_A - sQ_F|Q_B]$ corresponding to row-set R_X and column-set C_X is nonsingular and has the determinant of the form αs^p ($\alpha \in \mathbb{Q}$, $\alpha \neq 0$) by Lemma 3.1. The weight $\zeta_1(X)$ represents the exponent p , i.e.,

$$(4.5) \quad p = \zeta_1(X) + r_0, \quad r_0 = \sum_{i=1}^n r_i.$$

Another weight function ζ_2 on S is defined with reference to $\mathcal{M}([I|O|T_F|O])$ by

$$(4.6) \quad \zeta_2(v_i) = 0, \quad \zeta_2(w_i) = 0, \quad \zeta_2(x_j) = 1, \quad \zeta_2(u_j) = 0.$$

Since

$$(4.7) \quad \text{rank } D(z) = \text{rank } [A - zF|B] + 2n \quad \text{for any } z \in \mathbb{C},$$

condition (C3) is equivalent to

$$(4.8) \quad \text{rank } D(z) = 3n \quad \text{for } z \neq 0, \quad z \in \mathbb{C}.$$

Let $\hat{S} (\subset S)$ be the set of coloops of the matroid $\mathcal{M}(D(s))$ that is represented as in Lemma 4.4. Then the following lemma holds true, which will be established later in § 5 by a succession of algebraic arguments by means of the combinatorial canonical form of $D(s)$. Remember that $\mathcal{M}.\hat{S}$ denotes the contraction of a matroid \mathcal{M} to \hat{S} .

LEMMA 4.5. (C3) is equivalent to (M0) and (M3), where

(M0): $\text{rank } [\mathcal{M}([O|I|Q_A - Q_F|Q_B]) \vee \mathcal{M}([I|O|T_F|O]) \vee \mathcal{M}([I|I|T_A|T_B])] = 3n$,
and

(M3): $\zeta_1(X) + \zeta_2(Y)$ is constant for all $X, Y (\subset \hat{S})$ such that $X \cap Y = \emptyset$, X is independent in $\mathcal{M}([O|I|Q_A - Q_F|Q_B]).\hat{S}$, Y is independent in $\mathcal{M}([I|O|T_F|O]).\hat{S}$ and $\hat{S} \setminus (X \cup Y)$ is independent in $\mathcal{M}([I|I|T_A|T_B]).\hat{S}$. \square

Note that (4.5) shows (M3) is unaffected by the choice of ζ_1 and that such subsets X and Y as described in (M3) do exist by Lemma 2.2.

As explained in the proof of Theorem 4.6 below, (M0) is implied by (C1), and therefore (M3) constitutes the essential part. The problem of checking (M3) can be categorized as a version of weighted matroid-partition problem [7]. A special emphasis should be laid on the fact that there exists a well-established efficient combinatorial algorithm for this problem so that (M1), (M2) and (M3) can be checked efficiently by graph manipulations and arithmetic operations on rational numbers without serious numerical difficulty due to rounding errors. The algorithm, adapted to our purpose, will be described in some detail in the latter half of this section.

The main result of the present paper is stated below.

THEOREM 4.6. *Assume (A1) and (A2). The descriptor system (3.1) satisfies (C1), (C2) and (C3) iff*

$$(M1): \text{rank} [\mathcal{M}([I_n|Q_A - Q_F]) \vee \mathcal{M}([I_n|T_A - T_F])] = 2n,$$

$$(M2): \text{rank} [\mathcal{M}([I_n|Q_A|Q_B]) \vee \mathcal{M}([I_n|T_A|T_B])] = 2n,$$

and

$$(M3): \zeta_1(X) + \zeta_2(Y) \text{ is constant for all } X, Y (\subset \hat{S}) \text{ such that } X \cap Y = \emptyset, X \text{ is independent in } \mathcal{M}([O_n|I_n|Q_A - Q_F|Q_B]).\hat{S}, Y \text{ is independent in } \mathcal{M}([I_n|O_n|T_F|O_{n,m}]).\hat{S} \text{ and } \hat{S} \setminus (X \cup Y) \text{ is independent in } \mathcal{M}([I_n|I_n|T_A|T_B]).\hat{S}, \text{ where } \hat{S} \text{ is the set consisting of all coloops of } \mathcal{M}([O|I|Q_A - Q_F|Q_B]) \vee \mathcal{M}([I|O|T_F|O]) \vee \mathcal{M}([I|I|T_A|T_B]).$$

Proof. The theorem follows from Lemmas 4.2, 4.3, and 4.5 if we show that (M0) in Lemma 4.5 is implied by (M1). In fact, by Lemma 4.4, (M0) is equivalent to $\text{rank } D(s) = 3n$, and hence to $\text{rank } [A - sF|B] = n$ by (4.7). This follows obviously from (C1), i.e., from (M1). \square

4.2. Algorithm for testing the controllability condition. The conditions (M1) and (M2) of Theorem 4.6 can be checked efficiently by the matroid union/partition algorithm [7], since the matroids involved are linear matroids represented over the rational numbers and transversal matroids. See [33] for the description of the algorithm for this particular case.

Before presenting a concrete procedure for (M3), we will outline the key idea to cope with the seemingly complicated condition that $\zeta_1(X) + \zeta_2(Y)$ remains constant for all possible choices of (X, Y) with the specified properties. As described in [15], the matroid union/partition problem can be formulated in a natural manner as an independent-flow problem [8]. Then the set of coloops \hat{S} of the union matroid can be identified easily by path-searching on the auxiliary graph associated with a maximum independent flow.

Moreover, by associating an appropriate cost with each arc in the independent-flow problem, the quantity $\zeta_1(X) + \zeta_2(Y)$ can be expressed as the cost of a maximum independent flow. Each arc of the auxiliary graph is then given the “length” that represents the imputed cost. Then (M3) can be shown to be equivalent to the graph-theoretic condition that there exists no directed cycle of nonzero “length” in a certain subgraph, representing \hat{S} , of the auxiliary graph. This condition is known to be further equivalent to the existence of potentials associated with vertices of the subgraph such that the imputed cost, i.e., the “length,” of an arc in a strongly connected component is expressed as the difference of the potentials associated with its two end-vertices.

The concrete procedure for (M3) is as follows. In accordance with [31], [33] we consider the matroid intersection, rather than the union, formulated in the independent-flow problem.

The dual matroid \mathcal{M}_Q^* of $\mathcal{M}_Q = \mathcal{M}([I_n|Q_A - Q_F|Q_B])$ defined on

$$(4.9) \quad \bar{S} = \{w_1, \dots, w_n\} \cup \{x_1, \dots, x_n\} \cup \{u_1, \dots, u_m\}$$

is again a linear matroid over \mathbf{Q} which is expressed by the linear dependence among the row vectors of the matrix

$$(4.10) \quad \begin{array}{l} \mathbf{w}: \\ \mathbf{x}: \\ \mathbf{u}: \end{array} \begin{pmatrix} Q_A - Q_F & Q_B \\ I_n & 0 \\ 0 & I_m \end{pmatrix},$$

the row-set of which is indexed by \bar{S} of (4.9) as indicated.

The underlying graph $G = (V, A^*)$ of the independent-flow problem is defined as follows. It has the union of two disjoint copies of \bar{S} as the vertex-set V :

$$V = V_T \cup V_Q,$$

where

$$V_T = \{w_i^T | i = 1, \dots, n\} \cup \{x_j^T | j = 1, \dots, n\} \cup \{u_j^T | j = 1, \dots, m\},$$

$$V_Q = \{w_i^Q | i = 1, \dots, n\} \cup \{x_j^Q | j = 1, \dots, n\} \cup \{u_j^Q | j = 1, \dots, m\}.$$

In general the copies of $v \in \bar{S}$ in V_T and V_Q are denoted by v^T and v^Q , respectively. The arc-set A^* is given by

$$A^* = A_w \cup A_x \cup A_u \cup A_T,$$

where

$$A_w = \{(w_i^T, w_i^Q) | i = 1, \dots, n\},$$

$$A_x = \{(x_j^T, x_j^Q) | j = 1, \dots, n\},$$

$$A_u = \{(u_j^T, u_j^Q) | j = 1, \dots, m\}$$

and

$$A_T = \mathcal{N}(T_F) \cup \mathcal{N}(T_A) \cup \mathcal{N}(T_B).$$

The arc corresponding to $(T_F)_{ij} (\neq 0)$ (or $(T_A)_{ij} \neq 0$) connects from w_i^T to x_j^T , and the arc $(T_B)_{ij} (\neq 0)$ from w_i^T to u_j^T . Note that G has parallel arcs from w_i^T to x_j^T if $(T_F)_{ij}(T_A)_{ij} \neq 0$.

Each arc is given the unit capacity. The cost function $\gamma: A^* \rightarrow \mathbf{R}$ is defined with reference to (4.4) by

$$(4.11) \quad \gamma(a) = \begin{cases} r_i, & a = (w_i^T, w_i^Q) \in A_w, \\ c_j, & a = (x_j^T, x_j^Q) \in A_x, \\ c_{n+j}, & a = (u_j^T, u_j^Q) \in A_u, \\ 1, & a \in \mathcal{N}(T_F), \\ 0, & a \in \mathcal{N}(T_A) \cup \mathcal{N}(T_B). \end{cases}$$

The entrance-set V^+ of the independent-flow problem is $V^+ = \{w_i^T | i = 1, \dots, n\}$, on which we understand the free matroid is defined, whereas the exit-set V^- is V_Q , to which the matroid \mathcal{M}_Q^* represented by (4.10) is attached (with the obvious correspondence between $V^- = V_Q$ and \bar{S} of (4.9)).

By the integrality, we may assume that an independent flow f from V^+ to V^- in this network is chosen to be integer-valued, i.e.,

$$(4.12) \quad f(a) \in \{0, 1\} \quad \text{for } a \in A^*.$$

By the definition of an independent flow, $f(\delta^+ v) \in \{0, 1\}$ for $v \in V^+$, where $\delta^+ v$ denotes the set of arcs going out of v , and the vector $(f(a) | a \in A_w \cup A_x \cup A_u)$, when identified with a subset of \bar{S} by (4.12), determines a subset, say $J (\subset \bar{S})$, which is independent in the matroid \mathcal{M}_Q^* attached to V^- . By the construction of G , J is independent also in the transversal matroid $\mathcal{M}([I | T_A - T_F | T_B])$.

To define the auxiliary network associated with f , we need to transform the matrix of (4.10), which we denote by P . Since J is independent in \mathcal{M}_Q^* , it can be augmented to a base, say $J \cup J_1$, where $|J \cup J_1| = n + m$. The square submatrix of P with row-set $J \cup J_1$ is then nonsingular, and we define \tilde{P} to be the $(2n + m) \times (n + m)$ matrix obtained from P by post-multiplying the inverse of that submatrix. The row-set of \tilde{P} is still

indexed by \bar{S} and the linear dependence among the row vectors of \tilde{P} represents \mathcal{M}_Q^* . Note that the column-set of \tilde{P} has the natural correspondence with $J \cup J_1$.

The auxiliary network $\bar{N} = (\bar{V}, \bar{A})$ associated with an independent flow f in the present network is defined as follows. (See [8], [15] for the general definition of the auxiliary network for the independent-flow problem.) The vertex-set \bar{V} is identical with V , where the entrance S^+ and the exit S^- is defined as

$$\begin{aligned} S^+ &= \{v \in V^+ | f(\delta^+ v) = 0\}, \\ S^- &= \{v^Q \in V^- | \tilde{P}_{vj} \neq 0 \text{ for some } j \in J_1\}. \end{aligned}$$

The arc-set \bar{A} is given by

$$\bar{A} = B_* \cup B^* \cup A^-,$$

where

$$\begin{aligned} B_* &= \{a | a \in A^*, f(a) = 0\}, \\ B^* &= \{\bar{a} | \bar{a} \text{ is the reorientation of } a \in A^* \text{ such that } f(a) = 1\}, \\ A^- &= \{(u^Q, v^Q) | u \in \bar{S} \setminus J, v \in J, \tilde{P}_{uv} \neq 0, \tilde{P}_{uj} = 0 \text{ for } j \in J_1\}. \end{aligned}$$

The length function $\bar{\gamma}: \bar{A} \rightarrow \mathbf{R}$ is defined by

$$(4.13) \quad \bar{\gamma}(\bar{a}) = \begin{cases} \gamma(\bar{a}) & \text{if } \bar{a} \in B_*, \\ -\gamma(a) & \text{if the reorientation of } \bar{a} \in B^* \text{ is } a \in A^*, \\ 0 & \text{if } \bar{a} \in A^-. \end{cases}$$

Consider the auxiliary network $\bar{N} = (\bar{V}, \bar{A})$ associated with a maximum independent flow, and let $\hat{V} (\subset \bar{V})$ be the set of vertices of \bar{N} which are not reachable to S^- by directed paths in it. The subgraph induced by \hat{V} is denoted as $\hat{G} = (\hat{V}, \hat{A})$ and the subnetwork of \bar{N} , restricted to \hat{G} , as $\hat{N} = (\hat{V}, \hat{A})$. It may be remarked that if (M1) is satisfied, the value of the maximum independent flow is equal to n and therefore S^+ is empty.

The algorithmic characterization of (M3) of Theorem 4.6 is now stated.

THEOREM 4.7. *Suppose (M1) is satisfied. The following three conditions are equivalent.*

- (i) (M3) holds true.
- (ii) The sum of the lengths $\bar{\gamma}(a)$ along any directed cycle in \hat{G} is equal to zero.
- (iii) There exists a "potential" function $\pi: \hat{V} \rightarrow \mathbf{R}$ satisfying

$$\bar{\gamma}(a) = \pi(\partial^- a) - \pi(\partial^+ a)$$

for all $a = (\partial^+ a, \partial^- a) \in \hat{A}$ such that $\partial^+ a$ and $\partial^- a$ belong to the same strongly connected component of \hat{G} .

Proof. The equivalence of (i) and (ii) follows from the well-known facts about the independent-flow problem, while that of (ii) and (iii) is obvious. \square

This theorem provides an efficient way to test (M3). As is well known, a maximum independent flow can be obtained by repeatedly finding a path in auxiliary networks (cf. [15], [33] for the detail of this procedure). The result of [5] implies that the amount of computation needed for finding a maximum independent flow is bounded by $O(n^2(n+m) \log n)$ in the worst case. Then the network $\hat{N} = (\hat{V}, \hat{A})$ can be constructed from the corresponding auxiliary network $\bar{N} = (\bar{V}, \bar{A})$ in $O(|\bar{V}| + |\bar{A}|)$ time in a straightforward manner. Once \hat{N} is constructed, the third condition (iii) above can be checked in $O(|\hat{V}| + |\hat{A}|)$ time with arithmetic operations (subtractions) on rational numbers by finding a spanning forest as well as the decomposition into strongly connected components [1]. Noting that $|\hat{V}| = O(n+m)$ and $|\hat{A}| = O(n(n+m))$, we see that the total amount of computation for testing (M3) is bounded by $O(n^2(n+m) \log n)$ in the worst case.

The other two conditions (M1) and (M2) of Theorem 4.6 can also be checked in $O(n^2(n+m)\log n)$ time in the worst case [5] by the established algorithm for the matroid union/intersection problem. Therefore the proposed algorithm for testing the structural controllability based on Theorem 4.6 is guaranteed to run in $O(n^2(n+m)\log n)$ time even in the worst case, and the actual running time is much less than the bound in many cases, depending on how many fixed constants are involved.

It should also be emphasized that the proposed method is practicable and stable from the numerical point of view; it requires arithmetic operations only on simple rational numbers, typically simple integers such as those representing the underlying topological relations, so that it is free from serious numerical difficulty of rounding errors.

Example 4.1. Recall the mechanical system of Example 3.1. If we choose time $[T]$, length $[L]$ and mass $[M]$ as the fundamental physical dimensions, the dimensions associated with x_j ($j = 1, \cdots, 6$) and u are given by

$$L, L, T^{-1}L, T^{-1}L, T^{-2}LM, T^{-1}L, \text{ and } T^{-2}LM,$$

and those with w_i ($i = 1, \cdots, 6$) by

$$T^{-1}L, T^{-1}L, T^{-2}LM, T^{-2}LM, T^{-2}LM, T^{-1}L.$$

Therefore, $[Q_A - sQ_F|Q_B]$ of (3.9) admits the expression of the form (3.11) with

(4.14)
$$\begin{aligned} c_1 = c_2 = 0, \quad c_3 = c_4 = 1, \quad c_5 = 2, \quad c_6 = 1, \quad c_7 = 2, \\ r_1 = r_2 = 1, \quad r_3 = r_4 = r_5 = 2, \quad r_6 = 1. \end{aligned}$$

The conditions (M1) and (M2) are found to be satisfied, as will be partly mentioned below. The independent-flow problem for (M3) is depicted in Fig. 4.1, in which $V^+ = \{w_1^T, \cdots, w_6^T\}$, $V^- = \{x_1^Q, \cdots, x_6^Q, u^Q, w_1^Q, \cdots, w_6^Q\}$, and the cost $\gamma(a)$ of each arc a is given in parentheses.

As indicated by bold lines, there exists a maximum independent flow f of value 6; it corresponds to $J = \{x_1, x_2, w_1, w_2, w_5, w_6\}$, which is independent both in $\mathcal{M}_Q^* = \mathcal{M}([I|Q_A - Q_F|Q_B])^*$ and in $\mathcal{M}([I|T_A - T_F|T_B])$. This shows that

$$\text{rank} [\mathcal{M}([I|Q_A - Q_F|Q_B]) \vee \mathcal{M}([I|T_A - T_F|T_B])] = 12,$$

or $\text{rank} [A - sF|B] = 6$. Moreover, the flow f of Fig. 4.1 reveals, at the same time, that (M1) is satisfied, since $f((u^T, u^Q)) = 0$.

If J is augmented by $J_1 = \{u\}$ to a base of \mathcal{M}_Q^* , the matrix \tilde{P} is given by

(4.15)
$$\tilde{P} = \begin{array}{c} \begin{array}{cccccc} x_1 & x_2 & w_1 & w_2 & w_5 & w_6 & u \end{array} \\ \begin{array}{l} w_1: \\ w_2: \\ w_3: \\ w_4: \\ w_5: \\ w_6: \\ \hline x_1: \\ x_2: \\ x_3: \\ x_4: \\ x_5: \\ x_6: \\ \hline u: \end{array} \end{array} \begin{array}{cccccc} & & 1 & & & & \\ & & & 1 & & & \\ & & & & 1 & & 1 \\ & & & & -1 & & \\ & & & & 1 & & \\ & & & & & 1 & \\ \hline 1 & & & & & & \\ & 1 & & & & & \\ 1 & & 1 & & & & \\ & 1 & & 1 & & & \\ & & & & -1 & & \\ 1 & -1 & & & & 1 & \\ \hline & & & & & & 1 \end{array}$$

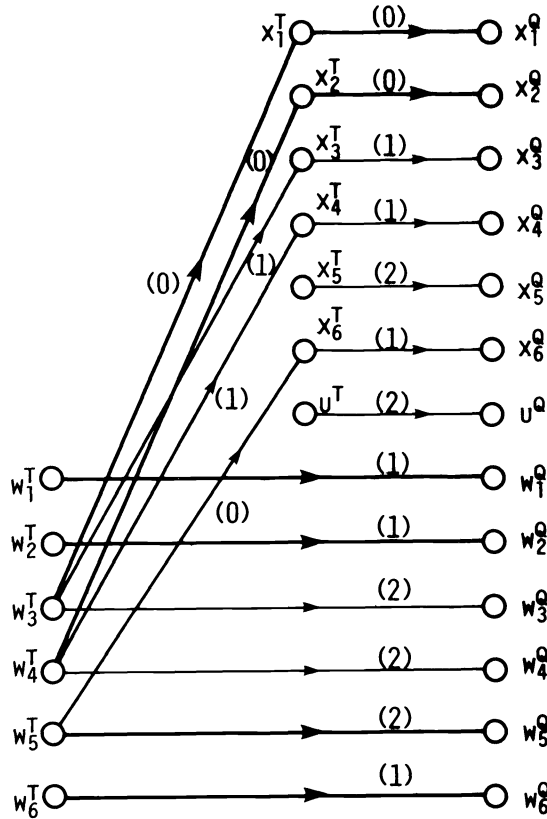


FIG 4.1. Independent-flow problem for Example 4.1. (A maximum independent-flow is drawn in bold lines.)

The associated auxiliary network $\bar{N} = (\bar{V}, \bar{A})$ is drawn in Fig. 4.2, where the “length” $\bar{\gamma}(a)$ is attached in parentheses to each arc a of $B_* \cup B^*$. The entrance S^+ is empty, while the exit $S^- = \{u^Q, w_3^Q\}$. All the vertices except those in $\hat{V} = \{w_1^T, w_1^Q, w_2^T, w_2^Q, w_6^T, w_6^Q\}$ are reachable to S^- , and the subnetwork $\hat{N} = (\hat{V}, \hat{A})$ consists of three disconnected arcs $\hat{A} = \{(w_1^Q, w_1^T), (w_2^Q, w_2^T), (w_6^Q, w_6^T)\}$. Then condition (ii) of Theorem 4.7 is trivially met, and this mechanical system is found to be generically controllable.

Example 4.2. Consider a hypothetical descriptor system (3.1) with $\mathbf{x} = (x_1, x_2, x_3)$, $\mathbf{u} = (u)$, and

$$(4.16) \quad F = \begin{pmatrix} 0 & 0 & p_1 \\ 1 & 1 & p_2 \\ 0 & 0 & 0 \end{pmatrix}, \quad A = \begin{pmatrix} 1 & p_3 & 0 \\ 0 & 0 & 1 \\ -1 & -1 & p_4 \end{pmatrix}, \quad B = \begin{pmatrix} p_5 \\ 0 \\ 0 \end{pmatrix},$$

where $\{p_i | i = 1, \dots, 5\}$ is to be understood as independent parameters. The matrix $[A - sF | B]$ is then a mixed matrix of (3.7) with

$$(4.17) \quad [Q_A - sQ_F | Q_B] = \left(\begin{array}{ccc|c} 1 & 0 & 0 & 0 \\ -s & -s & 1 & 0 \\ -1 & -1 & 0 & 0 \end{array} \right),$$

$$[T_A - sT_F | T_B] = \left(\begin{array}{ccc|c} 0 & p_3 & -sp_1 & p_5 \\ 0 & 0 & -sp_2 & 0 \\ 0 & 0 & p_4 & 0 \end{array} \right).$$

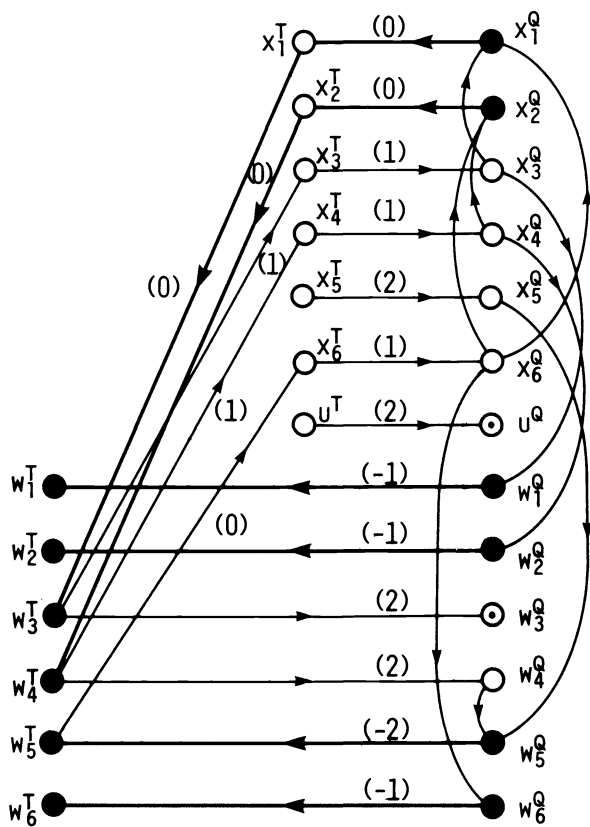


FIG. 4.2. Auxiliary network for Example 4.1. ($S^+ = \emptyset$, $S^- = \{u^Q, w_3^Q\}$.)

Note that the matrix $[Q_A - sQ_F | Q_B]$ above enjoys the property (A2) and it has the expression (3.11) with, e.g.,

$$r_1 = 0, \quad r_2 = 1, \quad r_3 = 0, \quad c_1 = c_2 = 0, \quad c_3 = 1, \quad c_4 = 0.$$

It is easy to see by inspection that (C1) and (C2), or equivalently (M1) and (M2), are satisfied. In Fig. 4.3, the independent-flow problem for (M3) is illustrated, where $V^+ = \{w_1^T, w_2^T, w_3^T\}$, $V^- = \{x_1^Q, x_2^Q, x_3^Q, u^Q, w_1^Q, w_2^Q, w_3^Q\}$, and the cost γ is given in parentheses as before.

The auxiliary network associated with a maximum independent flow f is given in Fig. 4.4 with $\tilde{\gamma}$ in parentheses. The flow f corresponds to the common independent set $J = \{w_3, x_2, x_3\}$. The entrance S^+ is empty and the exit $S^- = \{u^Q\}$, to which the vertices in $\hat{V} = \{x_3^T, x_3^Q, w_2^T, w_2^Q, w_3^T, w_3^Q\}$ are not reachable.

The subnetwork $\hat{N} = (\hat{V}, \hat{A})$, given in Fig. 4.5 with $\tilde{\gamma}$ in parentheses contains two simple directed cycles; the length, relative to $\tilde{\gamma}$, of the cycle consisting of $\{w_2^T, w_2^Q, w_3^Q, w_3^T, x_3^T\}$ vanishes, whereas that of $\{w_2^T, w_2^Q, x_3^Q, x_3^T\}$ does not. Theorem 4.7 reveals that this system does not satisfy (M3), i.e., there exists a nonzero mode that is not controllable. The graph-theoretic arguments of [3], [26], [30], which treat the nonvanishing entries of F , A and B of (4.16) as if they were independent, would fail to detect this fact.

5. Deriving the controllability condition of nonzero modes. This section is devoted to the proof of Lemma 4.5. Recall the matrix $D(s)$ of (4.2) consisting of three matrices

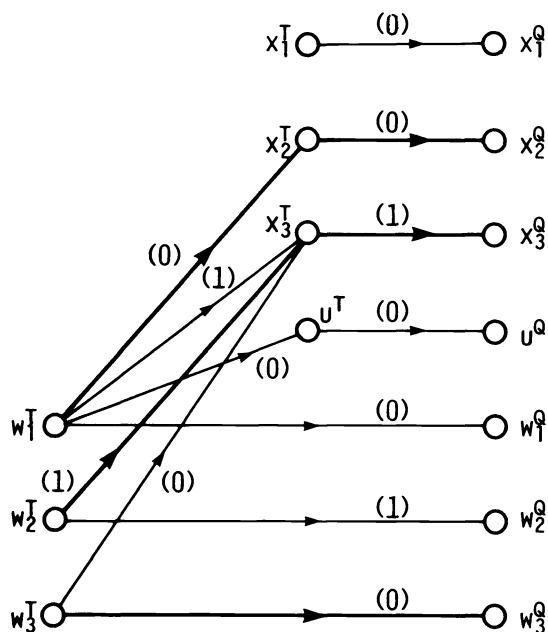


FIG. 4.3. Independent-flow problem for Example 4.2. (A maximum independent-flow is drawn in bold lines.)

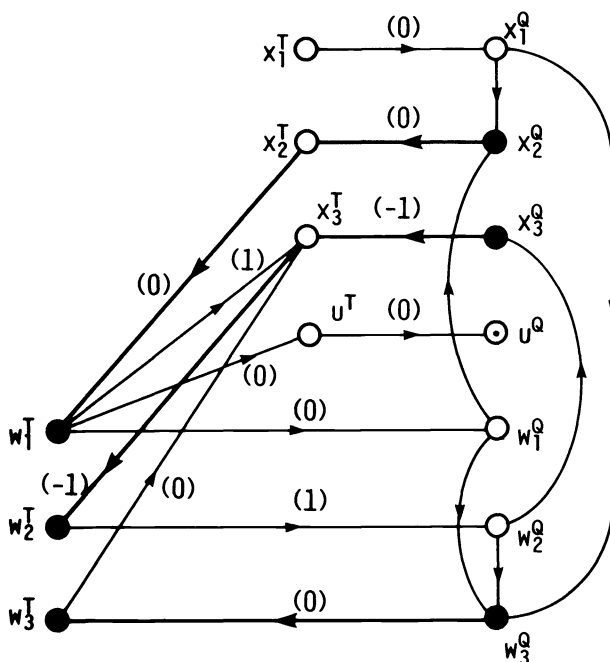


FIG. 4.4. Auxiliary network for Example 4.2.

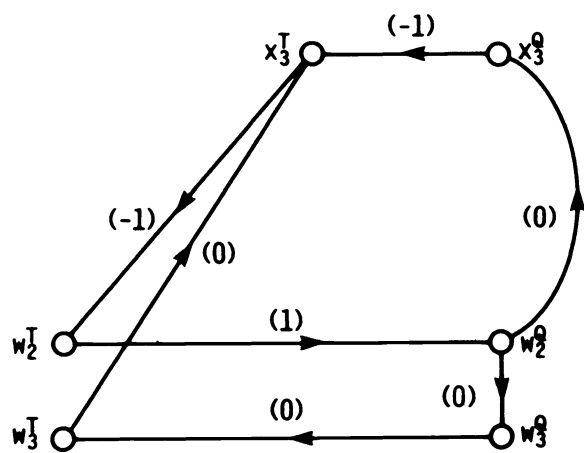


FIG. 4.5. Subnetwork \hat{N} for Example 4.2.

$Q_D(s)$, $T_{D1}(s)$ and T_{D0} of (4.1). By replacing the $3n$ occurrences of unity in $T_{D1}(s)$ and T_{D0} with algebraically independent real numbers, say, t_1, \dots, t_{3n} , we consider another composite $(3n) \times (3n + m)$ matrix $\tilde{D}(s)$, i.e.,

(5.1)
$$\tilde{D}(s) = \begin{pmatrix} Q_D(s) \\ \tilde{T}_{D1}(s) \\ \tilde{T}_{D0} \end{pmatrix},$$

where

(5.2)
$$\begin{aligned} \tilde{T}_{D1}(s) &= [\text{diag}(t_1, \dots, t_n) | O_n | -sT_F | O_{n,m}], \\ \tilde{T}_{D0} &= [-\text{diag}(t_{n+1}, \dots, t_{2n}) | -\text{diag}(t_{2n+1}, \dots, t_{3n}) | T_A | T_B] \end{aligned}$$

and

$$\mathcal{N}(\tilde{T}_{D1}(s)) \cup \mathcal{N}(\tilde{T}_{D0}) \text{ is algebraically independent over } \mathbf{Q}(s).$$

Then $\tilde{D}(s)$ is a matrix of the form (2.3), to which Lemma 2.3 applies with $\mathbf{K} = \mathbf{Q}(s)$.

It has been shown in § 4 that (C3) is equivalent to (4.8). Since the transformation from $D(s)$ to $\tilde{D}(s)$ can be interpreted as scaling of the rows of T_{D1} and T_{D0} as well as the columns of $\{v_1, \dots, v_n\}$ with algebraically independent transcendentals, we see that (4.8) is further equivalent to

(5.3)
$$\text{rank } \tilde{D}(z) = 3n \quad \text{for } z \neq 0, \quad z \in \mathbf{C}.$$

Let $\bar{D}(s)$ be the combinatorial canonical form of $\tilde{D}(s)$, which is obtained by (2.4) with a nonsingular matrix U over $\mathbf{Q}(s)$ (see § 2). It is a block-triangular matrix of the form (2.5) with the column-set C and the row-set R partitioned as

$$\begin{aligned} C &= C_0 \cup C_1 \cup \dots \cup C_r \cup C_\infty, \\ R &= R_0 \cup R_1 \cup \dots \cup R_r \cup R_\infty. \end{aligned}$$

An important consequence of (A2) is that, in the transformation of the form (2.4) connecting $\tilde{D}(s)$ to $\bar{D}(s)$, the matrix $U = U(s)$, which is nonsingular in $\mathbf{Q}(s)$, can be chosen so that each entry is of the form αs^p ($\alpha \in \mathbf{Q}, p \in \mathbf{Z}$) and that

(5.4)
$$\det U(s) = \alpha s^p, \quad \alpha \in \mathbf{Q} \setminus \{0\}, \quad p \in \mathbf{Z}.$$

For instance, $\tilde{D}(s)$ for the system of Example 4.2 is given by

$$(5.5) \quad \tilde{D}(s) = \begin{array}{c|cccccccc|c} v_1 & v_2 & v_3 & w_1 & w_2 & w_3 & x_1 & x_2 & x_3 & u \\ \hline 0 & & & 1 & & & 1 & 0 & 0 & 0 \\ & 0 & & & 1 & & -s & -s & 1 & 0 \\ & & 0 & & & 1 & -1 & -1 & 0 & 0 \\ \hline t_1 & & & 0 & & & 0 & 0 & -sp_1 & 0 \\ & t_2 & & & 0 & & 0 & 0 & -sp_2 & 0 \\ & & t_3 & & & 0 & 0 & 0 & 0 & 0 \\ \hline -t_4 & & & -t_7 & & & 0 & p_3 & 0 & p_5 \\ & -t_5 & & & -t_8 & & 0 & 0 & 0 & 0 \\ & & -t_6 & & & -t_9 & 0 & 0 & p_4 & 0 \end{array}$$

The transformation (2.4) with

$$(5.6) \quad U = \begin{pmatrix} 1 & 0 & 1 \\ 0 & 0 & -1 \\ 0 & 1 & -s \end{pmatrix}$$

makes it into the combinatorial canonical form

$$(5.7) \quad \bar{D}(s) = \begin{array}{c|cccc|ccccc|c} u & w_1 & x_1 & x_2 & v_1 & v_2 & w_2 & w_3 & x_3 & v_3 \\ \hline 0 & 1 & 0 & -1 & & & & 1 & & \\ 0 & 0 & 1 & 1 & & & & -1 & & \\ p_5 & -t_7 & 0 & p_3 & -t_4 & & & & & \\ \hline & & & & t_1 & & & & -sp_1 & \\ & & & & & 0 & 1 & -s & 1 & \\ & & & & & t_2 & 0 & 0 & -sp_2 & \\ & & & & & -t_5 & -t_8 & 0 & 0 & \\ & & & & & 0 & 0 & -t_9 & p_4 & -t_6 \\ \hline & & & & & & & & & t_3 \end{array}$$

with $C_0 = \{u, w_1, x_1, x_2\} (|R_0| = 3)$, $C_1 = \{v_1\}$, $C_2 = \{v_2, w_2, w_3, x_3\}$, $C_3 = \{v_3\}$, and $C_\infty = \emptyset (|R_\infty| = 0)$.

By (5.4), $U(z)$ is nonsingular for any particular complex number $z \neq 0$. Therefore $\tilde{D}(z)$ and $\bar{D}(z)$, as matrices over \mathbb{C} , have a common rank, and (5.3) is expressed with reference to the rank of the diagonal blocks of $\bar{D}(z)$, as follows.

LEMMA 5.1. (C3) is equivalent to the following:

- (i) $R_\infty = \emptyset$,
- (ii) $\text{rank } \bar{D}(z)[R_0, C_0] = |R_0|$ for any $z \neq 0$, $z \in \mathbb{C}$, and
- (iii) $\text{rank } \bar{D}(z)[R_i, C_i] = |R_i|$ for any $z \neq 0$, $z \in \mathbb{C}$; $i = 1, \dots, r$.

Proof. This is immediate from (2.6) and the facts that (C3) is equivalent to (5.3) and that $\tilde{D}(z)$ and $\bar{D}(z)$ have the same rank for each value of $z (\neq 0)$. \square

The following is obvious.

LEMMA 5.2. $R_\infty = \emptyset$ iff (M0) holds true.

The third lemma asserts that condition (ii) of Lemma 5.1 above is always satisfied.

LEMMA 5.3. $\text{rank } \bar{D}(z)[R_0, C_0] = |R_0|$ for any $z \neq 0$, $z \in \mathbb{C}$.

Proof. Let $\bar{D}(s)[R_0, C_0] = \bar{D}_0(s)$ be expressed as

$$(5.8) \quad \bar{D}_0(s) = \begin{pmatrix} Q_0(s) \\ T_1(s) \\ T_0 \end{pmatrix},$$

where $Q_0(s)$ is the matrix over $\mathbf{Q}(s)$ obtained from $Q_D(s)$, and $T_1(s)$ and T_0 are the submatrices, respectively, of $\tilde{T}_{D1}(s)$ and \tilde{T}_{D0} of (5.2). We have

$$(5.9) \quad \text{rank} [\mathcal{M}(Q_0(s)) \vee \mathcal{M}(T_1(s)) \vee \mathcal{M}(T_0)] = |R_0|.$$

Firstly suppose that $z (\neq 0)$ is an algebraic number (over \mathbf{Q}). Then $\mathcal{N}(T_1(z)) \cup \mathcal{N}(T_0)$ is algebraically independent over $\mathbf{Q}(z)$, and therefore Lemma 2.3 yields

$$(5.10) \quad \text{rank } \bar{D}_0(z) = \text{rank} [\mathcal{M}(Q_0(z)) \vee \mathcal{M}(T_1(z)) \vee \mathcal{M}(T_0)].$$

As a consequence of (A2), we have $\mathcal{M}(Q_0(s)) = \mathcal{M}(Q_0(1)) = \mathcal{M}(Q_0(z))$ if $z \neq 0$, whereas $\mathcal{M}(T_1(s)) = \mathcal{M}(T_1(z))$ is obvious. This fact combines (5.9) with (5.10), establishing

$$\text{rank } \bar{D}_0(z) = |R_0| \quad \text{for } z (\neq 0) \text{ algebraic.}$$

Next, consider the case where z is transcendental over \mathbf{Q} . Since $\text{rank } \bar{D}_0(s) = |R_0|$, there exists a nonvanishing minor (subdeterminant) $g(s; \mathcal{N}_T)$ of order $|R_0|$, which may be regarded as a polynomial over \mathbf{Q} in $\{s\} \cup \mathcal{N}_T$, where $\mathcal{N}_T = \mathcal{N}(T_1(1)) \cup \mathcal{N}(T_0)$. It suffices to consider such a $z (\in \mathbf{C} \setminus \{0\})$ that is a root of g , i.e., $g(z; \mathcal{N}_T) = 0$.

This means that $\{z\} \cup \mathcal{N}_T$ is algebraically dependent over \mathbf{Q} , whereas \mathcal{N}_T , as well as $\{z\}$, is algebraically independent over \mathbf{Q} . By Lemma 2.1, there exists $t \in \mathcal{N}_T$ such that $\mathcal{N}'_T = (\mathcal{N}_T \setminus \{t\}) \cup \{z\}$ is algebraically independent.

Since none of the columns of $\bar{D}_0(s)$ is a coloop of $\mathcal{M}(\bar{D}_0(s))$ (see Lemma 2.4), the matrix $\bar{D}_0(s)$ with one column deleted remains of rank $|R_0|$. In particular, there exists a square nonsingular submatrix of $\bar{D}_0(s)$ of order $|R_0|$ that does not contain the entry t . Let $h(s) = h(s; \mathcal{N}_T \setminus \{t\})$ denote its determinant, which is a nontrivial polynomial over \mathbf{Q} in $\{s\} \cup (\mathcal{N}_T \setminus \{t\})$. By the algebraic independence of \mathcal{N}'_T , $h(z)$ does not vanish. This implies that $\bar{D}_0(z)$ is of rank $|R_0|$. \square

Now we turn to condition (iii) of Lemma 5.1. Put $R^* = R \setminus R_0$ and $C^* = C \setminus C_0$.

LEMMA 5.4. Assume (M0) (or equivalently $R_\infty = \emptyset$). Then condition (iii) of Lemma 5.1 is satisfied iff $\det \bar{D}(s)[R^*, C^*]$ can be written as

$$(5.11) \quad \det \bar{D}(s)[R^*, C^*] = \beta s^p,$$

where β is a nonvanishing polynomial in $\mathcal{N}(T_F) \cup \mathcal{N}(T_A) \cup \mathcal{N}(T_B)$ over \mathbf{Q} , and p an integer.

Proof. First note that (M0) implies $|R^*| = |C^*|$. Condition (iii) of Lemma 5.1 is obviously equivalent to

$$\det \bar{D}(z)[R^*, C^*] \neq 0 \quad \text{for any } z \neq 0, \quad z \in \mathbf{C}.$$

Then the assertion above follows immediately. \square

Put

$$(5.12) \quad \bar{D}(s)[R^*, C^*] = \begin{pmatrix} Q^*(s) \\ T_1^*(s) \\ T_0^* \end{pmatrix},$$

where $Q^*(s)$ is the matrix over $\mathbf{Q}(s)$ obtained from $Q_D(s)$, and $T_1^*(s)$ and T_0^* are the submatrices, respectively, of $\tilde{T}_{D1}(s)$ and \tilde{T}_{D0} of (5.2). In the following, $Q^*(s)[X]$ ($X \subset C^*$) means the submatrix of $Q^*(s)$ consisting of the columns of X and all the rows,

and similarly for $T_1^*(s)$ and T_0^* . Recall that $Q^*(s)$, $T_1^*(s)$ and T_0^* all have the full row rank. By the generalized Laplace expansion, we obtain the following.

LEMMA 5.5. *Suppose (M0) holds. Then*

$$(5.13) \quad \det \bar{D}(s)[R^*, C^*] = \sum_{X, Y} \det Q^*(s)[X] \cdot \det T_1^*(s)[Y] \cdot \det T_0^*[C^* \setminus (X \cup Y)],$$

where the summation is taken over all $X, Y (\subset C^*)$ such that X is a base of $\mathcal{M}(Q^*(s))$, Y is a base of $\mathcal{M}(T_1^*(s))$, and $C^* \setminus (X \cup Y)$ is a base of $\mathcal{M}(T_0^*)$.

The matrix $Q^*(s)$ inherits the property that every subdeterminant is of the form αs^p ($\alpha \in \mathbb{Q}, p \in \mathbb{Z}$). For a base $X (\subset C^*)$ of $\mathcal{M}(Q^*(s))$, we have

$$(5.14) \quad \begin{aligned} \det Q^*(s)[X] &= \alpha s^p, \\ \alpha &\in \mathbb{Q} \setminus \{0\}, \quad p = \zeta_1(X) + p_0, \end{aligned}$$

where ζ_1 is defined in (4.4) associated with $Q_D(s)$ and p_0 is an integer independent of X .

For a base $Y (\subset C^*)$ of $\mathcal{M}(T_1^*(s))$, we have similarly

$$(5.15) \quad \begin{aligned} \det T_1^*(s)[Y] &= \beta s^p, \\ \beta &\in \mathbb{Q}[\mathcal{N}(T_F)] \setminus \{0\}, \quad p = \zeta_2(Y). \end{aligned}$$

LEMMA 5.6. *Suppose (M0) holds. Then condition (iii) of Lemma 5.1 is equivalent to (M3) of Lemma 4.5.*

Proof. First notice the relations $\mathcal{M}(Q^*(s)) = \mathcal{M}([O|I|Q_A - Q_F|Q_B]).\hat{S}$, $\mathcal{M}(T_1^*(s)) = \mathcal{M}([I|O|T_F|O]).\hat{S}$, and $\mathcal{M}(T_0^*) = \mathcal{M}([I|I|T_A|T_B]).\hat{S}$, where it should be noted that \hat{S} , the set of coloops of $\mathcal{M}(D(s))$, is identical with C^* if (M0) holds.

Then from Lemma 5.5, (5.14) and (5.15), as well as from the fact that the nonvanishing terms in the right-hand side of (5.13) do not cancel one another (by virtue of the algebraic independence of $\mathcal{N}(T_F) \cup \mathcal{N}(T_A) \cup \mathcal{N}(T_B)$), it follows that $\det \bar{D}(s)[R^*, C^*]$ takes the form of (5.11) iff (M3) holds true. The proof is completed by Lemma 5.4. \square

Finally, Lemma 4.5 follows from Lemmas 5.1, 5.2, 5.3 and 5.6.

6. Conclusion. The result of the present paper includes many previously known results on the structural controllability as special cases. In particular, it is a direct generalization of [30] and [31]; in [30] the structural controllability condition for a descriptor system (without fixed constants) has been expressed in terms of graph-theoretic conditions using the Dulmage–Mendelsohn decomposition [6] of bipartite graphs, and in [29], [31] the structural controllability of a descriptor system is investigated under the same setting as in this paper with an additional assumption of the nonsingularity of F .

There seems to be several different definitions of the controllability of a descriptor system, to which the present approach can readily be adapted.

As has been partly demonstrated in this paper, some of the matroid-theoretic concepts should be useful for the analysis of dynamical systems, just as for electrical networks [14], [16]. See [15], [31] for matroid-theoretic approaches to other dynamical systems problems; the former describes the result of [17] on the controllability condition under some combinatorial constraints, and the latter deals with the problem of determining the dynamical degree [12] of a dynamical system.

Acknowledgments. The author expresses his hearty thanks to Professor Masao Iri of the University of Tokyo for constant discussions and encouragements. Professor

Satoru Fujishige of the University of Tsukuba kindly read the manuscript and suggested improvements. Some comments by the anonymous referees were helpful in revision.

REFERENCES

- [1] A. V. AHO, J. E. HOPCROFT AND J. D. ULLMAN, *The Design and Analysis of Computer Algorithms*, Addison-Wesley, Reading, MA, 1974.
- [2] B. D. O. ANDERSON AND H.-M. HONG, *Structural controllability and matrix nets*, Internat. J. Control, 35 (1982), pp. 397–416.
- [3] T. AOKI, S. HOSOE AND Y. HAYAKAWA, *Structural controllability for linear systems in descriptor form*, Trans. Soc. Instr. Control Engrg., 19 (1983), pp. 628–635. (In Japanese.)
- [4] D. COBB, *Controllability, observability, and duality in singular systems*, IEEE Trans. Automat. Control, AC-29 (1984), pp. 1076–1082.
- [5] W. H. CUNNINGHAM, *Matroid partition and intersection algorithms*, Dept. Math. Stat., Carleton Univ., Ottawa, Ontario, Canada, 1984.
- [6] A. L. DULMAGE AND N. S. MENDELSON, *A structure theory of bipartite graphs of finite exterior dimension*, Trans. Royal Soc. Canada, Section III, 53 (1959), pp. 1–13.
- [7] J. EDMONDS, *Minimum partition of a matroid into independent subsets*, J. Nat. Bur. Stand., 69B (1965), pp. 67–72.
- [8] S. FUJISHIGE, *Algorithms for solving the independent-flow problems*, J. Oper. Res. Soc. Japan, 21 (1978), pp. 189–204.
- [9] F. R. GANTMACHER, *The Theory of Matrices*, Chelsea, New York, 1959.
- [10] K. GLOVER AND L. M. SILVERMAN, *Characterization of structural controllability*, IEEE Trans. Automat. Control, AC-21 (1976), pp. 534–537.
- [11] Y. HAYAKAWA, S. HOSOE AND M. ITO, *Structural controllability of linear time-invariant compartmental systems*, Trans. Inst. Electr. Comm. Engrg. Japan, J65A (1982), pp. 371–378. (In Japanese.)
- [12] ———, *Dynamical degree and controllability for linear systems with intermediate standard form*, Trans. Inst. Electr. Comm. Engrg. Japan, J64A (1981), pp. 752–759. (In Japanese.)
- [13] H. E. HUNTLEY, *Dimensional Analysis*, Macdonald and Co., London, 1952.
- [14] M. IRI, *Applications of matroid theory*, in Mathematical Programming—The State of the Art, A. Bachem, M. Grötschel and B. Korte, eds., Springer, Berlin, 1983, pp. 158–201.
- [15] M. IRI AND S. FUJISHIGE, *Use of matroid theory in operations research, circuits and systems theory*, Internat. J. Syst. Sci., 12 (1981), pp. 27–54.
- [16] M. IRI AND N. TOMIZAWA, *A unifying approach to fundamental problems in network theory by means of matroids*, Electr. Comm. Japan, 58A (1975), pp. 28–35.
- [17] M. IRI, N. TOMIZAWA AND S. FUJISHIGE, *On the controllability and observability of a linear system with combinatorial constraints*, Trans. Soc. Instr. Control Engrg., 17 (1977), pp. 225–242. (In Japanese.)
- [18] T. KAILATH, *Linear Systems*, Prentice-Hall, Englewood Cliffs, NJ, 1980.
- [19] H. L. LANGHAAR, *Dimensional Analysis and Theory of Models*, John Wiley, New York, 1951.
- [20] E. L. LAWLER, *Combinatorial Optimization: Networks and Matroids*, Holt, Reinhart and Winston, New York, 1976.
- [21] C.-T. LIN, *Structural controllability*, IEEE Trans. Automat. Control, AC-19 (1974), pp. 201–208.
- [22] D. G. LUENBERGER, *Dynamic equations in descriptor form*, IEEE Trans. Automat. Control, AC-22 (1977), pp. 312–321.
- [23] ———, *Time-invariant descriptor systems*, Automatica, 14 (1978), pp. 473–480.
- [24] H. MAEDA, *On structural controllability theorem*, IEEE Trans. Automat. Control, AC-26 (1981), pp. 795–798.
- [25] H. MAEDA AND T. YAMADA, *Strong structural controllability*, this Journal, 17 (1979), pp. 123–138.
- [26] T. MATSUMOTO AND M. IKEDA, *Structural controllability based on intermediate standard forms*, Trans. Soc. Instr. Control Engrg., 19 (1983), pp. 601–606. (In Japanese.)
- [27] K. MUROTA, *Structural controllability of a linear time-invariant system with auxiliary variables*, Trans. Soc. Instr. Control Engrg., 19 (1983), pp. 104–109. (In Japanese.)
- [28] ———, *Structural solvability and controllability of systems*, Doctoral dissertation, Dept. Math. Engrg. Instr. Phys., Univ. Tokyo, Japan, 1983; revised version, Springer, New York–Berlin, to appear.
- [29] ———, *Structural controllability of a system with some fixed coefficients*, Trans. Soc. Instr. Control Engrg., 19 (1983), pp. 683–690. (In Japanese.)
- [30] ———, *Structural controllability of a system in descriptor form expressed in terms of bipartite graphs*, Trans. Soc. Instr. Control Engrg., 20 (1984), pp. 272–274. (In Japanese.)

- [31] K. MUROTA, *Use of the concept of physical dimensions in the structural approach to systems analysis*, Japan J. Appl. Math., 2 (1985), pp. 471–494.
- [32] ———, *Combinatorial canonical form of layered mixed matrices and block-triangularization of large-scale systems of linear/nonlinear equations*, Discussion Paper Series 257, Inst. Socio-Economic Planning, Univ. Tsukuba, 1985.
- [33] K. MUROTA AND M. IRI, *Structural solvability of systems of equations—A mathematical formulation for distinguishing accurate and inaccurate numbers in structural analysis of systems*, Japan J. Appl. Math., 2 (1985), pp. 247–271.
- [34] K. MUROTA, M. IRI AND M. NAKAMURA, *Combinatorial canonical form of layered mixed matrices and its application to block-triangularization of systems of linear/nonlinear equations*, SIAM J. Alg. Disc. Meth., 8 (1987), pp. 123–149.
- [35] R. W. SHIELDS AND J. B. PEARSON, *Structural controllability of multiinput linear systems*, IEEE Trans. Automat. Control, AC-21 (1976), pp. 203–212.
- [36] G. C. VERGHESE, B. C. LÉVY AND T. KAILATH, *A generalized state-space for singular systems*, IEEE Trans. Automat. Control, AC-26 (1981), pp. 811–831.
- [37] B. L. VAN DER WAERDEN, *Algebra*, Springer, Berlin, 1955.
- [38] D. J. A. WELSH, *Matroid Theory*, Academic Press, London, 1976.
- [39] T. YAMADA AND D. G. LUENBERGER, *Generic controllability theorem for descriptor systems*, IEEE Trans. Automat. Control, AC-30 (1985), pp. 144–152.
- [40] E. L. YIP AND R. F. SINCOVEC, *Solvability, controllability, and observability of continuous descriptor systems*, IEEE Trans. Automat. Control, AC-26 (1981), pp. 702–707.

ABSTRACT DYNAMIC PROGRAMMING MODELS UNDER COMMUTATIVITY CONDITIONS*

SERGIO VERDU[†] AND H. VINCENT POOR[‡]

Abstract. The unifying purpose of the abstract dynamic programming models is to find sufficient conditions on the recursive definition of the objective function that guarantee the validity of the dynamic programming iteration. This paper presents backward, forward, and backward-forward models that weaken previous sufficient conditions and that include, but are not restricted to, optimization problems. The backward-forward model is devoted to the simultaneous solution of a collection of interrelated sequential problems based on the independent computation of a cost-to-arrive function and a cost-to-go function. Several extremization and nonextremization problems illustrate the applicability of the proposed models.

Key words. backward and forward dynamic programming operator models, finite and infinite horizon sequential optimization, discrete-time Markov processes, fixed-interval detection and smoothing

AMS(MOS) subject classifications. 90C39, 90C48, 93E20

1. Introduction. The fact that dynamic programming has found application in a wide variety of sequential optimization problems has led several researchers to investigate what class of objective functions can be optimized by dynamic programming. The unifying purpose of the abstract dynamic programming models is not to facilitate the solution of specific problems, but to extract the essential features that guarantee the solvability of a sequential problem by dynamic programming. In the models proposed by Mitten [1], Denardo [2], Nemhauser [3], Karp and Held [4] and Bertsekas [5] it is assumed that the real-valued objective function can be defined recursively by a generating operator or local income function [6] which maps a set of functions of states into itself. If this operator is monotone, then further restrictions such as the contraction-mapping assumption of [2], the continuity conditions of [5] or the finiteness of the state-space [4] suffice to validate the dynamic programming solution in various finite and infinite horizon settings. Also, Brown and Strauch [7] have proposed a related abstract model where the return space is not necessarily the extended real line but a multiplicative lattice.

In this paper we propose an abstract discrete-time dynamic programming model that includes, but is not restricted to, optimization problems. Any functional satisfying a certain commutativity condition with the generating operator (which, unlike previous models, is not restricted to be monotone) of the objective function results in a sequential problem solvable by a dynamic programming iteration. Examples of sequential nonextremization problems fitting this framework are the derivation of marginal probability distributions from decomposable joint distributions, iterative computation of stage-separated functions taking values on additive commutative semigroups with distributive products, generation of symbolic transfer functions, and the computation of unconditional transition probabilities of a Markov process.

Another feature of the framework of this paper is the ability to formulate forward models completely symmetric to backward ones. This enables the analysis of open-loop problems by either approach under the same kind of restrictions (§ 3) and, more significantly, it allows for the formulation of the backward-forward dynamic programming operator model.

* Received by the editors July 29, 1985; accepted for publication (in revised form) June 4, 1986. This work was supported in part by the U.S. Office of Naval Research under contract N00014-81-K-0014 and the National Science Foundation under grants ECS-85-12314 and ECS-85-04752.

[†] Department of Electrical Engineering, Princeton University, Princeton, New Jersey 08544.

[‡] Coordinated Science Laboratory, University of Illinois at Urbana-Champaign, Urbana, Illinois 61801.

The backward-forward model presented in this paper is devoted to the simultaneous solution of a collection of interrelated sequential problems based on the independent computation of a cost-to-arrive function and a cost-to-go function. To fix ideas, consider the following simple example of this method. In a layered network, the shortest path containing a particular arc can obviously be obtained by deleting all other arcs in the same layer and solving for the shortest route by either forward or backward dynamic programming; however, if the problem must be solved for each arc in the network, then rather than repeating the above process, it is more efficient to simply compute the distances of each node from the source and to the destination. Other problems such as fixed-interval minimum error probability detection in data communications and fixed-interval smoothing are shown to fit into this framework.

In § 2, we present an operator model for discrete-time finite-horizon backward and forward problems, and a pair of sufficient conditions, viz., the decomposability of the objective function and the commutativity of operators, are shown to ensure the validity of the dynamic programming iteration. When applied to infimization problems, the commutativity condition is weaker and not more difficult to check in specific problems than the sufficient conditions imposed by previous models. The use of the proposed model is briefly illustrated in § 2 for optimum stochastic control problems, and in § 3 for a variety of deterministic problems with extremization and nonextremization operators. Familiarity with previous abstract dynamic programming models, in particular with the model due to Bertsekas [5] which encompasses most other previous settings, may be advantageous in reading §§ 2 and 3. The formulation and some applications of the backward-forward model are presented in § 4. Finally, § 5 briefly discusses the issues arising in the infinite-horizon problem and sufficient conditions on the commutativity of operators and interchangeability of limits are shown to ensure the validity of the dynamic programming iteration and the fixed-point property of the sought-after function of states.

2. Abstract finite-horizon dynamic programming operator model.

Glossary of notation. The following notation is used throughout the paper.¹

S :	state space.
A :	action space.
μ :	function mapping S to A ; $\mu: S \rightarrow A$.
M :	set of admissible policies from S to A ; $\mu \in M$.
Q :	return space.
L :	Q -valued function of states; $L: S \rightarrow Q$.
H :	operator mapping Q -valued functions of states to Q -valued functions of actions; $H: Q^S \rightarrow Q^A$.
$HL(a)$:	the function of actions $HL: A \rightarrow Q$ evaluated at the point $a \in A$.
V :	operator (functional) mapping Q -valued functions of M to Q ; $V: Q^M \rightarrow Q$.
$V\{q(\mu)\}$:	image in Q of a function $q: M \rightarrow Q$; note that μ is a dummy argument: $V\{q(\mu)\}$ depends on the values of $q(\mu)$ for all $\mu \in M$.
$V\{L_\mu\}$:	Q -valued function of states whose value at $x \in S$ is $V\{L_\mu(x)\}$, where L_μ is a Q -valued function of states parametrized by $\mu \in M$. Analogous notation is used for functions of actions $V\{HL_\mu\}$.
$M_{i,j} = M_i \times \cdots \times M_j$,	
$\mu_{i,j} = (\mu_i, \cdots, \mu_j) \in M_{i,j}$.	

¹ Stage indices are omitted here.

$$\begin{aligned} \mathbf{V}_{i,j}\{q(\mu_{i,j})\} &= \mathbf{V}_i\{\cdots \mathbf{V}_j\{q(\mu_{i,j})\}\cdots\}. \\ p_k: \text{projection of a Cartesian product: } p_k(x_1, \cdots, x_n) &= x_k; \\ p_k: X_1 \times \cdots \times X_n &\rightarrow X_k. \end{aligned}$$

The foregoing return, state and action spaces and operators H and \mathbf{V} take on various meanings depending on the specific problem to which the model is applied. In optimization problems Q is identified with the extended real line and \mathbf{V} is the infimization functional; i.e., for any function $q: M \rightarrow Q$

$$\mathbf{V}\{q(\mu)\} = \inf_{\mu \in M} q(\mu).$$

To fix ideas in a first reading of the sequel, it may be helpful to identify Q and \mathbf{V} with these particular cases.

For $i = 0, \cdots, N$, let M_i be the set of admissible policies mapping S_i to A_i (the state and action spaces at stage i , respectively). Suppose that for each stage $n \in \{0, \cdots, N\}$ we are given a function of states $L_{\mu_n, N}^n: S_n \rightarrow Q$ which is parametrized by the string of policies $\mu_n, N \in M_{n, N}$. ($L_{\mu_n, N}^n(x)$ is the cost incurred by using the policies μ_n, \cdots, μ_N if x is the state at the n th stage.) Then the objective of this section is to investigate under what conditions dynamic programming can be used to find the function of initial states:

$$(1) \quad \mathbf{V}_{0, N}\{L_{\mu_{0, N}}^0\}$$

where $\mathbf{V}_i: Q^{M_i} \rightarrow Q$, $i = 0, \cdots, N$.

The first requisite in order to solve (1) by dynamic programming is the recursive formulation of the sequential dependence of the function of states $L_{\mu_{0, N}}^0$ on the policies $\mu_{0, N}$:

B1. Backward decomposability of the objective function. There exists a collection of operators

$$(2) \quad \begin{aligned} H^i: Q^{S_i} &\rightarrow Q^{A_{i-1}}, \quad i = 1, \cdots, N, \quad \text{such that for all } \mu_{0, N} \in M_{0, N}, \\ L_{\mu_{i-1, N}}^{i-1}(x) &= H^i L_{\mu_{i, N}}^i(\mu_{i-1}(x)), \quad x \in S_{i-1}, \quad i = 1, \cdots, N. \end{aligned}$$

Example. Suppose the objective is to control a deterministic system $x_{k+1} = f(x_k, u_k) \in S$, $u_k \in U$, so as to minimize the cost $\sum_{k=0}^N g(x_k, u_k)$ as a function of $x_0 \in S$. This objective function satisfies property B1. To see this, define

$$(3) \quad A = S \times U, \quad HL(a) = g(a) + L(f(a))$$

and restrict the admissible policies to satisfy $p_1(\mu(x)) = x$. If \mathbf{V} is the infimization operator, then (1) coincides with the sought-after minimum cost. Note that the only difference between the operator H (local income function in optimization problems) and the one used in previous abstract dynamic programming models (e.g., [2] and [5]) is that in those works, the mapping is defined on cartesian products of state and control spaces in lieu of action spaces. While both coincide in this example, the action spaces take on different roles in other problems in the sequel. As we know, we can indeed solve the problem in the present example using dynamic programming; however this is because it satisfies other properties in addition to B1, which are characterized next.

DEFINITION. A Q -valued function of states $R^i: S_i \rightarrow Q$ is a *cost-to-go* function² if

$$R^i(x) = \mathbf{V}_{i, N}\{L_{\mu_{i, N}}^i(x)\} \quad \text{for all } x \in S_i.$$

² This terminology is maintained for nonextremization problems.

THEOREM 2.1. Suppose that property B1 holds and define the following sequence of functions of states $B^i: S_i \rightarrow Q$, $i = 0, \dots, N$:

$$(4a) \quad B^N(x) = V_N\{L_{\mu_N}^N(x)\},$$

and

$$(4b) \quad B^{i-1}(x) = V_{i-1}\{H^i B^i(\mu_{i-1}(x))\}.$$

Then B^i , $i = 0, \dots, N$ are cost-to-go functions if and only if

$$(5) \quad V_{i-1,N}\{H^i L_{\mu_{i,N}}^i(\mu_{i-1}(x))\} = V_{i-1}\{H^i V_{i,N}\{L_{\mu_{i,N}}^i(\mu_{i-1}(x))\}\}$$

for all $x \in S_{i-1}$, $i = 1, \dots, N$.

Proof. Because of (4a) it suffices to show that, for each $i = 1, \dots, N$, if $B^i = V_{i,N}\{L_{\mu_{i,N}}^i\}$, then $B^{i-1} = V_{i-1,N}\{L_{\mu_{i-1,N}}^{i-1}\}$ is equivalent to (5); but this follows immediately from (2) and (4b). \square

COROLLARY. Suppose that B1 and the following condition are satisfied:

B2. Backward commutativity of operators.

$$(6) \quad V_{i,N}\{H^i L_{\mu_{i,N}}^i\} = H^i V_{i,N}\{L_{\mu_{i,N}}^i\}, \quad i = 1, \dots, N.$$

Then, B^i , $i = 0, \dots, N$ are cost-to-go functions and in particular $V_{0,N}\{L_{\mu_{0,N}}^0\} = B^0$.

Property B2 can be represented in the following commutativity diagram:

$$\begin{array}{ccc} Q^{S_i \times M_{i,N}} & \xrightarrow{H^i} & Q^{A_{i-1} \times M_{i,N}} \\ V_{i,N} \downarrow & & \downarrow V_{i,N} \\ Q^{S_i} & \xrightarrow{H^i} & Q^{A_{i-1}} \end{array}$$

where the set of functions of states (resp. actions) parametrized by the elements of $M_{i,N}$ is denoted by $Q^{S_i \times M_{i,N}}$ (resp. $Q^{A_{i-1} \times M_{i,N}}$). The mappings $Q^{S_i \times M_{i,N}} \rightarrow Q^{S_i}$, $Q^{A_{i-1} \times M_{i,N}} \rightarrow Q^{A_{i-1}}$ and $Q^{S_i \times M_{i,N}} \rightarrow Q^{A_{i-1} \times M_{i,N}}$ induced by the pointwise application of the operators $V_{i,N}$ and H^i are denoted with the same symbol. In the special case of optimization problems, property B2 can be viewed as a formalization of a general optimality principle: in order to find the optimal partial return for each action a , it suffices to compute the cost-to-go function at the next stage and evaluate the local income function at a .

The fact that the recursions (2) and (4) are defined backwards is only due to the ordering of the operators V in (1). Mere reversal of the stage indices results in a forward solution of $V_{N,0}\{L_{\mu_{0,N}}^N(x)\}$, $x \in S_N$. The corresponding decomposability and commutativity assumptions are, in this case,

F1. Forward decomposability of the objective function.

$$(2') \quad L_{\mu_{0,i+1}}^{i+1}(x) = H^i L_{\mu_{0,i}}^i(\mu_{i+1}(x)), \quad x \in S_{i+1}, \quad i = 0, \dots, N-1,$$

where

$$L_{\mu_{0,i}}^i: S_i \rightarrow Q; \quad i = 0, \dots, N, \quad H^i: Q^{S_i} \rightarrow Q^{A_{i+1}}, \quad i = 0, \dots, N-1.$$

F2. Forward commutativity of operators.

$$(6') \quad V_{i,0}\{H^i L_{\mu_{0,i}}^i\} = H^i V_{i,0}\{L_{\mu_{0,i}}^i\}, \quad i = 0, \dots, N-1.$$

We then have

THEOREM 2.2. Define the functions $F^i: S_i \rightarrow Q$, $i = 0, \dots, N$ through the recursion:

$$(4a') \quad F^0(x) = V_0\{L_{\mu_0}^0(x)\}$$

and

$$(4b') \quad F^{i+1}(x) = \mathbf{V}_{i+1}\{H^i F^i(\mu_{i+1}(x))\}.$$

Then, under assumptions F1 and F2, F^i are cost-to-arrive functions, i.e., $F^i(x) = \mathbf{V}_{i,0}\{L_{\mu_{0,i}}^i(x)\}$.

If either pair of assumptions is satisfied for a particular problem, then the other one is trivially satisfied for the time-reversed problem. Hence, it is only meaningful to distinguish between the forward and backward versions with respect to the state evolution of the original problem. Rather than presenting only one of the versions and fitting every particular example by possibly reversing the stage indices, we choose to maintain always the original indices and present both the forward and the backward formulations. This is due both to the fact that some problems are solved by forward and backward recursions concurrently (§ 4), and because recursions which evolve in the direction of the system are of interest in some applications (e.g., Viterbi's forward dynamic programming algorithm for real-time decision problems [8]).

When this general framework is applied to specific operators H and \mathbf{V} , the verification of the commutativity property (5), or the stronger version (6), often requires an inductive proof which is common to most problems. Based on such an induction, the next result provides a sufficient condition for B2 (analogously for F2) that entails the verification of the commutativity of H with a single operator \mathbf{V} .

THEOREM 2.3. *Suppose that condition B1 is satisfied, and define the functions $L_{\mu_{i,j}}^{i,j}: S_i \rightarrow Q$, $\mu_{i,j} \in M_{i,j}$, $1 \leq i \leq j \leq N$ through the recursions:*

$$(7a) \quad L_{\mu_j}^{j,j}(x) = \mathbf{V}_{j+1,N}\{L_{\mu_{j,N}}^j(x)\}$$

and

$$(7b) \quad L_{\mu_{i-1,j}}^{i-1,j}(x) = H^i L_{\mu_{i,j}}^{i,j}(\mu_{i-1}(x)).$$

Suppose that, for $1 \leq i \leq j \leq N$, we have

$$(8) \quad \mathbf{V}_j\{H^i L_{\mu_{i,j}}^{i,j}\} = H^i \mathbf{V}_j\{L_{\mu_{i,j}}^{i,j}\}, \quad \text{for all } \mu_{i,j-1} \in M_{i,j-1}.$$

Then condition B2 is satisfied.

Proof. Since $L_{\mu_N}^{N,N} = L_{\mu_N}^N$, particularizing (8) for $i = j = N$ results in $\mathbf{V}_N\{H^N L_{\mu_N}^N\} = H^N \mathbf{V}_N\{L_{\mu_N}^N\}$ (condition B2 for $i = N$). Now fix k and suppose that (4) is satisfied for $k+1, \dots, N$. We will show that under condition (8), we have $\mathbf{V}_{k,N}\{H^k L_{\mu_{k,N}}^k\} = H^k \mathbf{V}_{k,N}\{L_{\mu_{k,N}}^k\}$. The proof will be divided in two stages:

(a) If $\mathbf{V}_{j,N}\{H^j L_{\mu_{j,N}}^j\} = H^j \mathbf{V}_{j,N}\{L_{\mu_{j,N}}^j\}$ and condition (8) holds, then $\mathbf{V}_j\{L_{\mu_{i,j}}^{i,j}\} = L_{\mu_{i,j-1}}^{i,j-1}$ for $1 \leq i < j$ and for all $\mu_{i,j-1} \in M_{i,j-1}$.

(b) If (8) holds and $\mathbf{V}_j\{L_{\mu_{k,j}}^{k,j}\} = L_{\mu_{k,j-1}}^{k,j-1}$ for $k < j \leq N$ and for all $\mu_{k,j-1} \in M_{k,j-1}$, then $\mathbf{V}_{k,N}\{H^k L_{\mu_{k,N}}^k\} = H^k \mathbf{V}_{k,N}\{L_{\mu_{k,N}}^k\}$.

Relationship (a) can be proved by induction: Let $i = j - 1$, then for all $x \in S_i$ and $\mu_i \in M_i$ we have

$$\begin{aligned} L_{\mu_i}^{i,i}(x) &= \mathbf{V}_{j,N}\{L_{\mu_{i,N}}^i(x)\} = \mathbf{V}_{j,N}\{H^j L_{\mu_{j,N}}^j(\mu_i(x))\} \\ &= H^j \mathbf{V}_{j,N}\{L_{\mu_{j,N}}^j\}(\mu_i(x)) = H^j \mathbf{V}_j\{L_{\mu_j}^{j,j}\}(\mu_i(x)) = \mathbf{V}_j\{H^j L_{\mu_j}^{j,j}(\mu_i(x))\} \\ &= \mathbf{V}_j\{L_{\mu_{i,j}}^{i,j}(x)\}, \end{aligned}$$

where the second through sixth equalities follow from (2), (4), (7a), (8), and (7b), respectively. Now suppose that $\mathbf{V}_j\{L_{\mu_{i+1,j}}^{i+1,j}\} = L_{\mu_{i+1,j-1}}^{i+1,j-1}$ for all $\mu_{i+1,j-1} \in M_{i+1,j-1}$ and $i < j - 1$, then, for all $x \in S_i$ and $\mu_i \in M_i$,

$$\begin{aligned} \mathbf{V}_j\{L_{\mu_i}^{i,j}(x)\} &= \mathbf{V}_j\{H^{i+1} L_{\mu_{i+1,j}}^{i+1,j}(\mu_i(x))\} \\ &= H^{i+1} \mathbf{V}_j\{L_{\mu_{i+1,j}}^{i+1,j}\}(\mu_i(x)) \\ &= H^{i+1} L_{\mu_{i+1,j-1}}^{i+1,j-1}(\mu_i(x)) = L_{\mu_{i,j-1}}^{i,j-1}(x), \end{aligned}$$

where the second equation follows from (8). In order to prove (b) it is enough to show that

$$(9) \quad \mathbf{V}_N\{H^k L_{\mu_{k,N}}^k\} = H^k \mathbf{V}_N\{L_{\mu_{k,N}}^k\} \quad \text{for all } \mu_{k,N-1} \in M_{k,N-1},$$

and

$$(10) \quad \mathbf{V}_j\{H^k \mathbf{V}_{j+1,N}\{L_{\mu_{k,N}}^k\}\} = H^k \mathbf{V}_{j,N}\{L_{\mu_{k,N}}^k\} \quad \text{for all } \mu_{k,j-1} \in M_{k,j-1} \text{ and } j = k, \dots, N.$$

Equation (9) follows directly from condition (8). To show (10), note that the assumption in (b) implies that

$$\mathbf{V}_{j+1,N}\{L_{\mu_{k,N}}^k\} = L_{\mu_{k,j}}^{k,j} \quad \text{for all } \mu_{k,j} \in M_{k,j}.$$

So, it suffices to show that

$$\mathbf{V}_j\{H^k L_{\mu_{k,j}}^{k,j}\} = H^k \mathbf{V}_{j,N}\{L_{\mu_{k,N}}^k\},$$

but this readily follows from (7) and (8). \square

Three main differences between the above framework and previous abstract dynamic programming models can be underlined, namely,

(i) Attention is not restricted to extremization operators: $\mathbf{V}_i\{q(\mu)\} = \inf_{\mu \in M} q(\mu)$ or $\sup_{\mu \in M} q(\mu)$. Although, of course, this is the most important case, it is both useful (as illustrated below by several applications) and interesting from a conceptual viewpoint to consider nonextremization operators. Note also that since we do not require the operators \mathbf{V}_i to coincide at each stage we can deal, for example, with extremization problems where inf and sup operators occur successively (e.g., dynamic two-person zero-sum games, where computational schemes based on dynamic programming principles are ubiquitous (cf. [9], [10])).

(ii) In the present model the generating operator of the objective recursive function maps functions of states into functions of actions (as in Dynkin and Yushkevich [11]), rather than into functions of states. In contrast to [11] the duality between states and actions is carried one step further by defining the admissible policies as mappings from the state space to the action space rather than to the underlying control space. The stochastic control formulation of [5], [12, Part I] is equivalent to the special case in which the action space is $A_i \subset S_i \times U_i$, where U_i is a control space and the admissible policies $\mu \in M_i$ are such that the image of each state belongs to its fiber, i.e.,

$$p_1(\mu(x)) = x \quad \text{for all } x \in S_i.$$

Besides its notational convenience, the versatility of the use of general action spaces and action-valued policy functions affords a nice parallelism (§ 3) between forward and backward problems.

(iii) In optimization problems, the commutativity conditions of the present model are weaker than the sufficient conditions imposed on the generating operator by previous models (typically, monotonicity and continuity). In addition, it appears that the strong commutativity condition of Theorem 2.3 (between \mathbf{V}_j and H^i , $i \leq j$) is not more difficult to check directly than previous conditions. Although it is natural to impose the monotonicity condition in order to satisfy commutativity with extremization operators, it is not necessary to do so. For example, consider the problem

$$(11) \quad \min_{x_{0,N} \in S_{0,N}} g_0(x_0, x_1) + [g_1(x_1, x_2) + [\dots + g_{N-1}^2(x_{N-1}, x_N)]^2 \dots]^2; \quad g_i(x_i, x_{i+1}) \in \mathbb{R}.$$

The generating operator $H^i L(x_{i-1}, x_i) = g_{i-1}(x_{i-1}, x_i) + L^2(x_i)$ is not monotone in the function of states, yet it satisfies the above commutativity conditions and, therefore, (11) admits a dynamic programming solution. In order to illustrate how previous conditions imply commutativity, consider the setting due to Bertsekas [5], [12, Part I], which encompasses previous abstract dynamic programming models with real-valued return functions. The finite-horizon assumptions in [12, Prop. 6.1] imply the following conditions:

C1. H^i is monotone and for every $\varepsilon > 0$, there exists $\mu^\varepsilon \in M_j$ such that

$$H^i L_{\mu_{i,j-1}\mu^\varepsilon}^{i,j} - H^i V_j \{L_{\mu_{i,j}}^{i,j}\} \leq \varepsilon.$$

C2. There exists a sequence of policies $\mu_j^k \in M_j$, $k = 0, 1, \dots$, such that

$$L_{\mu_{i,j-1}\mu_j^k}^{i,j} \downarrow \inf_{\mu_j \in M_j} L_{\mu_{i,j}}^{i,j} \quad \text{and} \quad H^i L_{\mu_{i,j-1}\mu_j^k}^{i,j} \downarrow H^i \inf_{\mu_j \in M_j} L_{\mu_{i,j}}^{i,j}.$$

Then, it is straightforward to show that (8) follows from either condition.

In stochastic control problems the key to the existence of μ^ε and $\{\mu_j^k\}$ with the above properties is the fact that the dependence of $L_{\mu_{i,j}}^{i,j}(x)$ on $\mu_{i,j}$ is only through $\mu_{i,j}(x)$. Therefore, for every $\delta > 0$ there exists a uniformly δ -optimum policy $\mu^\delta \in A_j^{S_j}$ (not necessarily in M_j) such that for all $x \in S_j$

$$(12) \quad L_{\mu^\delta}^{i,j}(x) \leq \begin{cases} V_j \{L_{\mu_j}^{i,j}(x)\} + \delta & \text{if } V_j \{L_{\mu_j}^{i,j}(x)\} > -\infty, \\ -1/\delta & \text{if } V_j \{L_{\mu_j}^{i,j}(x)\} = -\infty. \end{cases}$$

In particular cases such μ^δ can be shown to belong to M_j and mild restrictions on the cost-per-stages guarantee that C1 and C2 are satisfied. For example, consider a controlled Markov process problem with additive cost-per-stages:

$$H^i L(a) = g_{i-1}(a) + \int_{S_i} L(\omega) P_{i-1}(d\omega|a), \quad a \in A_{i-1},$$

$$L^{N+1}(x) = r(x), \quad x \in S_{N+1}.$$

If $A_i \subset S_i \times U_i$ and all admissible policies $\mu_i \in M_i$ satisfy $p_1(\mu_i(x)) = x$, for all $x \in S_i$, then it is easy to see that $L_{\mu_{0,N}}^0(x)$ is the expected value of the cost of using $\mu_{0,N} \in M_{0,N}$ when the initial state is $x \in S_0$. To show the existence of $\mu^\delta \in M_j$ satisfying (12), it is enough to assume that the state and control spaces are Borel; A_k , $k = 0, \dots, N$ are analytic sets; the costs-per-stage $g_k: A_k \rightarrow \mathbb{R}$, $k = 0, \dots, N$ and the terminal cost $r: S_{N+1} \rightarrow \mathbb{R}$ are lower semi-analytic; the transition functions are Borel measurable and $M_j = \{\mu \in A_j^{S_j} \text{ such that } p_1(\mu(x)) = x, x \in S_j \text{ and } \mu \text{ is universally measurable}\}$ (see [13], [12, Prop. 7.50]). Now, if $V_{j,N} \{L_{\mu_{j,N}}^j(x)\} > -\infty$ for all $x \in S_j$, then for every $\varepsilon > 0$ we can choose $\mu^\varepsilon \in M_j$ such that C1 is satisfied. On the other hand, if there exists $\mu_j \in M_j$ such that $H^i L_{\mu_{i,j}}^{i,j}(a) < +\infty$ for all $a \in A_{i-1}$, then using (12) we can select $\{\mu_j^k\}$ such that C2 is satisfied. Other stochastic control problems such as those with worst-case, rather than average, objective function and multiplicative nonnegative, rather than additive, costs-per-stage can be shown to satisfy commutativity via conditions C1-C2.

3. Application to classes of backward and forward problems. Once we have illustrated briefly the application of the backward dynamic programming framework and associated commutativity conditions to a class of stochastic control problems, in this section we show the application of the framework of § 2 to classes of backward and forward problems. Because of the causality relation among the state spaces, the utility of the forward formulation is restricted to open-loop problems. The main purpose of the first specific model to be presented (a deterministic optimum control problem) is

to illustrate the symmetry achievable between the backward and forward formulations thanks to the versatility of the action space and action-valued policies. In § 3.2 we present applications of nonextremization operators in a general algebraic setting which includes the conventional finite state-space dynamic programming models.

3.1. Additive-cost deterministic optimum control. It is easy to fit the problem of additive-cost optimum control of a deterministic system $x_{k+1} = f_k(x_k, u_k) \in S_{k+1}$, $u_k \in U_k$, $k = 0, \dots, N$ into the framework of § 2. It is noteworthy that the dynamic programming formulation presented here allows a duality between the forward and backward formulations that unlike previous works, [14], [15] which require the system to be codeterministic, does not impose any restrictions to define the forward dynamic programming recursion. For the backward formulation we make the following identifications:

$$A_i = S_i \times U_i,$$

$$H^i L(a) = g_{i-1}(a) + L(f_{i-1}(a))$$

and

$$M_i = \{\mu \in A_i^{S_i} \text{ such that } p_i(\mu(x)) = x \text{ for all } x \in S_i\}.$$

Note that there is a bijection between the set of admissible policies and the set of admissible controls. On defining L^i through the recursion (2)—with $L^{N+1} = 0$ —it is clear that

$$\inf_{\mu_{0,N} \in M_{0,N}} L_{\mu_{0,N}}^0(x) = \inf_{\substack{u_{0,N} \in U_{0,N} \\ \text{s.t.} \\ x_{k+1} = f_k(x_k, u_k); x_0 = x}} \sum_{k=0}^N g_k(x_k, u_k).$$

In the forward case we define

$$A_i = S_{i-1} \times U_{i-1},$$

$$H^i L(a) = g_i(a) + L(p_i(a))$$

and

$$M_i = \{\mu \in A_i^{S_i} \text{ such that } f_{i-1}(\mu(x)) = x \text{ for all } x \in S_i\}.$$

In this case, there is no one-to-one mapping between the set of admissible policies and the set of admissible controls (had we defined policies as mappings from states to controls, then here we would be able to define the forward recursion only for codeterministic systems); however, there is indeed a bijection between $M_{1,N+1} \times S_{N+1}$ and the subset of $S_0 \times U_{0,N} \times S_{N+1}$ that represents the trajectories of the system. Hence, if L^i is defined through (3')—with $L^0 = 0$ —then it can be checked that

$$\inf_{\mu_{1,N+1} \in M_{1,N+1}} L_{\mu_{1,N+1}}^{N+1}(x) = \inf_{\substack{x_0 \in S_0, u_{0,N} \in U_{0,N} \\ \text{s.t.} \\ x_{k+1} = f_k(x_k, u_k); x_{N+1} = x}} \sum_{k=0}^N g_k(x_k, u_k),$$

and in both cases the strong commutativity property of Theorem 2.3 is obviously satisfied.

3.2. Recursive computation of stage-separated functions. Consider an algebraic system $(Q, +, \cdot)$, where $(Q, +)$ is a commutative semigroup and the internal binary

operation \cdot is left-distributive with respect to additions. Suppose that $S_i, i = 0, \dots, N+1$ are finite sets, $g_i: S_i \times S_{i+1} \rightarrow Q$, and the goal is to compute

$$(13) \quad L_0(x) = \sum_{x_1 \in S_1} \cdots \sum_{x_{N+1} \in S_{N+1}} g_0(x, x_1) \cdot [g_1(x_1, x_2) \cdots [g_N(x_N, x_{N+1})] \cdots]$$

for every $x \in S_0$. Some examples of problems of this type are:

- (i) Shortest path in a layered network with $(\mathbb{R}, \min, +)$.
- (ii) Satisfiability of Boolean clauses with $(\{0, 1\}, \text{OR}, \text{AND})$.
- (iii) Computation of marginal probability distributions, with $(R^+, +, \cdot)$ (§ 4).
- (iv) Symbolic transfer function problems, with $(S, \cup, *)$, where S is the family of sets of finite-length strings of symbols drawn from a finite alphabet, and $*$ denotes string concatenation (see [16] for a generalization of the McNaughton–Yamada algorithm [17] for the computation of regular expressions given arbitrary state graphs).
- (v) Dynamic programming for optimization problems where the return space is only partially ordered. If (G, \geq, \circ) is a conditionally complete associative lattice [7], then the algebraic system $(2^A, \max, \circ)$ satisfies the above properties, where $\max(A, B)$ is the set of maximal elements of $A \cup B$ and \circ is only defined $(a \circ B = \{a \circ b, b \in B\})$ when the left operand is a singleton. Interestingly, the generalized version of the optimality principle given by Brown and Strauch [7] (see also [18]) is a special case of the commutativity condition B2.

The function in (13) can be computed by a backward dynamic programming recursion. Define

$$A_i = S_i \times S_{i+1},$$

$$\mathbf{V}_i\{q(\mu)\} = \sum_{\mu \in M} q(\mu),$$

$$H^i L(a) = g_{i-1}(a) \cdot L(p_2(a)),$$

$$M_i = \{\mu \in A_i^{S_i}, \text{ there exists } x^* \in S_{i+1} \text{ such that } \mu(x) = (x, x^*) \text{ for all } x \in S_i\}.$$

If the recursion (2) is used to define $L_{\mu_i, N}^i$, with $L_{\mu}^N(x) = g_N(\mu(x))$ then it is easy to see that $L_0(x) = \mathbf{V}_{0, N} L_{\mu_{0, N}}^0(x)$. The commutativity condition of Theorem 2.3 follows from the left-distributivity of \cdot with respect to $+$. Analogously, if we identify $A_i = S_{i-1} \times S_i$, $M_i = \{\mu \in A_i^{S_i}; \text{ there exists } x^* \in S_{i-1} \text{ such that } \mu(x) = (x^*, x) \text{ for all } x \in S_i\}$ and $H^i L(a) = L(p_1(a)) \cdot g_i(a)$, then recursion (3') can be used to define $L_{\mu_1, i}^i$, with $L_{\mu}^1(x) = g_0(\mu(x))$, and right-distributivity of \cdot with respect to $+$ implies that forward dynamic programming can be used to recursively compute

$$\sum_{x_0 \in S_0} \cdots \sum_{x_N \in S_N} [\cdots [g_0(x_0, x_1)] \cdot g_1(x_1, x_2)] \cdots] \cdot g_N(x_N, x)$$

for all $x \in S_{N+1}$.

Note that if $(Q, +)$ has an identity element 0 which is also an annihilator of \cdot , i.e., $q + 0 = q$, $q \cdot 0 = 0 \cdot q = 0$ for all $q \in Q$, then a special case of the above formulation is

$$\sum_{u_0 \in U_0} \cdots \sum_{u_N \in U_N} \lambda_0(x, u_0) \cdot [\lambda_1(x_1, u_1) \cdots [\lambda_N(x_N, u_N)] \cdots]$$

subject to $x_{k+1} = f_k(x_k, u_k)$, $x_0 = x$. To see this, let

$$T_i(x_i, x_{i+1}) = \{u_i \in U_i \text{ such that } f_i(x_i, u_i) = x_{i+1}\}$$

and

$$g_i(x_i, x_{i+1}) = \begin{cases} \sum_{u_i \in T_i(x_i, x_{i+1})} \lambda_i(x_i, u_i) & \text{if } T_i(x_i, x_{i+1}) \neq \phi, \\ 0 & \text{if } T_i(x_i, x_{i+1}) = \phi, \end{cases}$$

and use the associative and distributive properties of the algebraic system.

4. Backward-forward finite-horizon dynamic programming.

4.1. General setting and sufficient conditions. Given a collection of sets D_i , operators $V_i: Q^{d_i} \rightarrow Q$, $i = 1, \dots, N$ and a function $J: D_{1,N} \rightarrow Q$, suppose that the goal is to find $V_{1,i-1}\{V_{i+1,N}\{J(d_1, \dots, d_i, \dots, d_N)\}\}$ for all $d_i \in D_i$ and $i = 1, \dots, N$. Assume that these sets, operators and function J satisfy the following conditions:

BF1. There exists backward and forward decomposable functions (i.e., satisfying properties B1 and F1, respectively) $\{\tilde{L}_{\tilde{\mu}_0, N-1}^0: \tilde{S}_0 \rightarrow Q, \tilde{\mu}_0, N-1 \in \tilde{M}_{0, N-1}\}$, $\{\tilde{L}_{\tilde{\mu}_2, N+1}^{N+1}: \tilde{S}_{N+1} \rightarrow Q, \tilde{\mu}_2, N+1 \in \tilde{M}_{2, N+1}\}$, and mappings $\tilde{\psi}_i: D_i \rightarrow \tilde{M}_{i-1}$, $\tilde{\psi}_i: D_i \rightarrow \tilde{M}_{i+1}$, $i = 1, \dots, N$, such that

$$(14) \quad J(d_1, \dots, d_i, \dots, d_N) = \tilde{L}_{\tilde{\psi}_1, N(d_{1,N})}^0(x_0) = \tilde{L}_{\tilde{\psi}_1, N}^{N+1}(d_{1,N})(x_{N+1})$$

for all $x_0 \in \tilde{S}_0$ and $x_{N+1} \in \tilde{S}_{N+1}$.

BF2. Denote by $\tilde{\Psi}_i: Q^{\tilde{M}_{i-1}} \rightarrow Q^{D_i}$ the mapping induced by $\tilde{\psi}_i$, i.e., $\tilde{\Psi}_i(f) = f \circ \tilde{\psi}_i$ (e.g., [19]). The operators $V_i \circ \tilde{\Psi}_i = \tilde{V}_{i-1}: Q^{\tilde{M}_{i-1}} \rightarrow Q$, $i = 1, \dots, N$, and \tilde{H} , the generating operator of the backward function in BF1 satisfy the strong commutativity condition (8). (Analogously, with $\tilde{V}_{i+1} = V_i \circ \tilde{\Psi}_i$, $\tilde{\Psi}_i: Q^{\tilde{M}_{i+1}} \rightarrow Q^{D_i}$ induced by $\tilde{\psi}_i$.)

Fix an index $1 \leq i \leq N$ and an element $d_i \in D_i$. Because of assumptions BF1 and BF2 we can solve $V_{1,i-1}\{V_{i+1,N}\{J(d_1, \dots, d_i, \dots, d_N)\}\}$ by either a backward or a forward dynamic programming iteration. However, a separate iteration has to be carried for each index and each element of D_i . If the operators V_i commute, then a backward-forward solution where $V_{1,i-1}\{V_{i+1,N}\{J(d_1, \dots, d_i, \dots, d_N)\}\}$ is solved simultaneously for all elements and stages is given by the following result.

THEOREM 4.1. *Suppose that conditions BF1 and BF2 are satisfied and that $V_i V_j = V_j V_i$, $1 \leq i < j \leq N$. Define the functions:*

$$B^{i-1}(x) = \tilde{V}_{i-1}\{\tilde{H}^i B^i(\tilde{\mu}_{i-1}(x))\}, \quad B^N(x) = \text{constant},$$

$$F^{i+1}(x) = \tilde{V}_{i+1}\{\tilde{H}^i F^i(\tilde{\mu}_{i+1}(x))\}, \quad F^1(x) = \text{constant}.$$

Then there exist functions $W_i: Q^{\tilde{S}_i} \times D_i \times Q^{\tilde{S}_i}$, $i = 1, \dots, N$ such that

$$(15) \quad V_{1,i-1}\{V_{i+1,N}\{J(d_1, \dots, d_i, \dots, d_N)\}\} = W_i(F^i, d_i, B^i), \quad d_i \in D_i, i = 1, \dots, N.$$

Proof. The quantity $V_{1,i-1}\{V_{i+1,N}\{J(d_1, \dots, d_i, \dots, d_N)\}\}$ is a function of $D_{1,i-1}$, d_i , and $D_{i+1,N}$. Theorem 4.1 will follow by showing that the dependence of $V_{1,i-1}\{V_{i+1,N}\{J(d_1, \dots, d_i, \dots, d_N)\}\}$ on $D_{1,i-1}$ is only through F^i , ($i = 2, \dots, N$), and the dependence on $D_{i+1,N}$ is only through B^i , ($i = 1, \dots, N-1$).

Fix a stage $1 \leq i \leq N$ and an element $d_i \in D_i$. Property BF1 states that for all $x_0 \in \tilde{S}_0$ we have

$$(16) \quad \begin{aligned} & V_{1,i-1}\{V_{i+1,N}\{J(d_1, \dots, d_i, \dots, d_N)\}\} \\ &= V_{1,i-1}\{\tilde{V}_{i,N-1}\{\tilde{L}_{\tilde{\psi}_1, i-1(d_{1,i-1})\tilde{\psi}_i(d_i)\tilde{\mu}_{i,N-1}}^0(x_0)\}\}. \end{aligned}$$

Now, it will be shown that for any $\tilde{\mu}_{0,i-1} \in \tilde{M}_{0,i-1}$, $\tilde{V}_{i,N-1}\{\tilde{L}_{\tilde{\mu}_{0,N-1}}^0(x_0)\}$ depends on $\tilde{M}_{i,N-1}$ only through B^i , and hence the same is true for the right-hand side of (16). Using the

notation introduced in Theorem 2.3, we have (note that here the horizon for the backward problem is $N-1$)

$$\tilde{L}_{\mu_{0,N-1}}^0(x_0) = \tilde{L}_{\mu_{0,N-1}}^{0,N-1}(x_0).$$

Furthermore, the result in part (a) of the proof of Theorem 2.3 (the conditions for its validity are guaranteed by properties BF1 and BF2) implies that

$$(17) \quad \tilde{\mathbf{V}}_{i,N-1}\{\tilde{L}_{\mu_{0,N-1}}^{0,N-1}\} = \tilde{\mathbf{V}}_{i,N-2}\{\tilde{L}_{\mu_{0,N-2}}^{0,N-2}\} = \cdots = \tilde{L}_{\mu_{0,i-1}}^{0,i-1}.$$

But according to (7b), the right-hand side of (17) depends on $\tilde{M}_{i,N-1}$ only through the function $\tilde{L}_{\mu_{i-1}}^{i-1,i-1}$ which because of (7a) can be written as

$$\begin{aligned} \tilde{L}_{\mu_{i-1}}^{i-1,i-1}(x) &= \tilde{\mathbf{V}}_{i,N-1}\{\tilde{L}_{\mu_{i-1,N-1}}^{i-1}(x)\} \\ &= \tilde{\mathbf{V}}_{i,N-1}\{\tilde{H}^i \tilde{L}_{\mu_{i,N-1}}^i(\tilde{\mu}_{i-1}(x))\} \\ &= \tilde{H}^i \tilde{\mathbf{V}}_{i,N-1}\{\tilde{L}_{\mu_{i,N-1}}^i\}(\tilde{\mu}_{i-1}(x)) \\ &= \tilde{H}^i B^i(\tilde{\mu}_{i-1}(x)) \end{aligned}$$

where the second equation follows from (2), and the third and fourth equations follow from property B2 (which is satisfied because of the strong commutativity condition (8)) and the corollary to Theorem 2.1, respectively.

Using the fact that $\mathbf{V}_i \mathbf{V}_j = \mathbf{V}_j \mathbf{V}_i$, we can write

$$\mathbf{V}_{1,i-1}\{\mathbf{V}_{i+1,N}\{J(d_1, \dots, d_i, \dots, d_N)\}\} = \mathbf{V}_{i+1,N}\{\mathbf{V}_{1,i-1}\{J(d_1, \dots, d_i, \dots, d_N)\}\}$$

and an entirely analogous reasoning shows that the dependence of $\mathbf{V}_{1,i-1}\{\mathbf{V}_{i+1,N}\{J(d_1, \dots, d_i, \dots, d_N)\}\}$ on $D_{1,i-1}$ is through the cost-to-arrive function F^i . \square

From the above proof it is easy to check that the conditions of Theorem 4.1 guarantee the validity of a backward-forward recursion for problems where several consecutive intermediate elements are fixed, i.e., (15) can be generalized to

$$(18) \quad \mathbf{V}_{1,i-1}\{\mathbf{V}_{j+1,N}\{J(d_1, \dots, d_i, \dots, d_j, \dots, d_N)\}\} = W_{i,j}(F^i, d_{i,j}, B^j), \quad d_{i,j} \in D_{i,j}.$$

Perhaps the simplest example of the backward-forward model is the problem mentioned in the introduction: given a layered network, find for each arc in the network the shortest path from source to destination that contains that arc. The straightforward approach is to run a forward or backward iteration for each arc (deleting all other arcs in the same layer); however, even if we take advantage of the obvious commonality of some of the computations, the number of steps required by this approach is quadratic in the number of layers. In contrast, if the result of Theorem 4.1 is employed (note that the minimization operators commute), then the solution to the shortest path problem requires only two independent (one forward and one backward) dynamic programming recursions, which simply compute the distance of each node from the source and to the destination. Once the cost-to-arrive and cost-to-go are computed for all nodes in the network, the solution is given by $W_{i,i+1}(F^i, d_{i,i+1}, B^{i+1})$ which is simply the sum of the length of each arc and the distances of its head and tail to the destination and from the source, respectively. The next subsection illustrates the application of the backward-forward dynamic programming setting to the problem of finding the sequence of joint distributions of consecutive states of a discrete-time Markov process, and to the problems of fixed-interval detection and smoothing.

4.2. Applications. Consider a Markov process $\{\mathbf{X}_t: (\Omega, \mathcal{F}) \rightarrow (\Omega_t, \mathcal{F}_t), t = 0, \dots, N\}$ whose finite-dimensional distributions are determined by P_0 , an arbitrary

initial probability measure on (Ω_0, F_0) , and by the transition functions P_i , $i = 1, \dots, N$ such that $P_i(x, \cdot)$ is a probability measure on (Ω_i, F_i) , for each $x \in \Omega_{i-1}$ and $P_i(\cdot, B)$ is measurable for each $B \in F_i$. Suppose that the objective is to obtain the sequence of joint distributions of consecutive states, namely,

$$(19) \quad P[X_k \in B_k, X_{k+1} \in B_{k+1}] = \int_{\Omega_{0,k-1} \times B_{k,k+1} \times \Omega_{k+2,N}} \cdots \int P_0(d\omega_0) \prod_{i=1}^N P_i(\omega_{i-1}, d\omega_i).$$

In order to put this problem in the backward-forward framework, let $D_i = \Omega_i \times F_i$, $\mathbf{V}_i\{q(d_i)\} = \int_{\Omega_i} f(\omega) \nu(d\omega)$ for $q(\omega, B) = f(\omega) \nu(B)$, and select $d_k = (x_k, C_k) \in D_k$ and $d_{k+1} = (x_{k+1}, C_{k+1}) \in D_{k+1}$. We make the following identifications for the state, action, policies and generating operators of the backward and forward formulations:

$$\begin{aligned} \tilde{S}_i &= \Omega_i, & \tilde{A}_i &= \Omega_i \times D_{i+1}, \\ \tilde{S}_i &= F_i, & \tilde{A}_i &= D_{i-1} \times \Omega_i, \end{aligned}$$

$$\tilde{H}^j L(a) = P_j(p_1(a), p_3(a)) L(p_2(a)), \quad \tilde{H}^j L(a) = P_{j+1}(p_1(a), p_3(a)) L(p_2(a)).$$

$$\begin{aligned} \tilde{M}_i &= \{\mu \in \tilde{A}_i^{\tilde{S}_i}, \text{ there exists } (\omega, B) \in \Omega_{i+1} \times F_{i+1} \\ &\quad \text{such that } \mu(x) = (x, \omega, B) \text{ for all } x \in \tilde{S}_i\}, \end{aligned}$$

$$\begin{aligned} \tilde{M}_i &= \{\mu \in \tilde{A}_i^{\tilde{S}_i}, \text{ there exists } (\omega, B) \in \Omega_{i-1} \times F_{i-1} \\ &\quad \text{such that } \mu(E) = (\omega, B, E), \text{ for all } E \in \tilde{S}_i\}. \end{aligned}$$

Note that the cost-to-arrive function is now a probability measure and that there is a one-to-one correspondence between D_i and \tilde{M}_{i-1} and between D_i and \tilde{M}_{i+1} . The commutativity conditions follow in this case from the linearity of the integral, and the problem can be solved by either a forward or a backward recursion with respective value functions:

$$F^{i+1}(E) = \int_{\Omega_i} P_{i+1}^G(\omega, E) F^i(d\omega), \quad F^0(E) = P_0^G(E)$$

and

$$B^{i-1}(x) = \int_{\Omega_i} P_i^G(x, d\omega) B^i(\omega), \quad B^{N+1} = 1.$$

Moreover, because of Fubini's theorem, a backward-forward solution given by (18) with $j = i + 1$ is also possible:

$$\mathbf{V}_{0,k-1}\{\mathbf{V}_{k+2,N}\{J(d_0, \dots, d_k, d_{k+1}, \dots, d_N)\}\} = F^k(C_k) P_{k+1}(x_k, C_{k+1}) B^{k+1}(x_{k+1}),$$

so the sought-after relationship is

$$(20) \quad P[X_k \in B_k, X_{k+1} \in B_{k+1}] = \int_{B_k} F^k(dx_k) \int_{B_{k+1}} P_{k+1}(x_k, dx_{k+1}) B^{k+1}(x_{k+1}).$$

Next we examine another application of the backward-forward finite horizon model with nonextremization operators, namely, the problem of fixed-interval minimum probability-of-error detection. Let (Ω, F, P) be a probability space and let $G \subset F$ be the sub- σ -algebra generated by the observation of an F -measurable transformation of a sequence of transmitted symbols $\{u_t \in U_t, t = 0, \dots, N\}$ where U_t , $t = 0, \dots, N$ are finite sets. Optimum decisions based on the a posteriori distribution $P^G[u_0, \dots, u_N]$

can be made according to various optimality criteria; for example, the receiver may select the sequence in $U_{0,N}$ that maximizes $P^G[u_0, \dots, u_N]$ (maximum likelihood sequence detection), or the sequence of arguments that maximizes the marginals $P^G[u_i]$, $i = 0, \dots, N$ (minimum error probability detection). In data-transmission problems such as asynchronous multiuser problems, transmission of convolutionally encoded data and intersymbol interference problems, the a posteriori distribution can be decomposed in product form:

$$P^G[u_0, \dots, u_N] = \prod_{k=0}^N \lambda_k(x_k, u_k) \quad \text{where } x_{k+1} = f_k(x_k, u_k) \text{ and } x_0 \text{ is } G\text{-measurable.}$$

The maximization of the joint distribution (maximum likelihood sequence detection) is a deterministic optimum control problem which fits into the framework presented in § 3.2, and hence can be solved by either a backward or a forward recursion (in real-time applications, the latter is employed in a near-optimum version where decisions are made after a fixed lag—the Viterbi algorithm [8]). If, instead, the optimality criterion is minimum probability-of-error, then the central task of the detector is to compute the marginal a posteriori distribution of each transmitted symbol, i.e.,

$$(21) \quad P^G[u_i] = \sum_{u_{0,i-1} \in U_{0,i-1}} \sum_{u_{i+1,N} \in U_{i+1,N}} \prod_{k=0}^N \lambda_k(x_k, u_k), \quad i = 0, \dots, N.$$

This problem fits also in the framework of § 3.3 and can be solved also by backward or forward dynamic programming. The forward recursion is simplified by noting that in the foregoing data-transmission problems the following condition holds:

S1. For $k = 0, \dots, N$, if there exists $x \in \Omega_k$, $u \in U_k$ and $u' \in U_k$ such that $f_k(x, u) = f_k(x, u')$ then $u = u'$.

Then the corresponding value functions are as follows:

$$(22) \quad F^{k+1}(x) = \sum_{\substack{x_k \in \Omega_k \\ \text{s.t. there exists} \\ u \in U_k, f_k(x_k, u) = x}} F^k(x_k) \lambda_k(x_k, u),$$

and

$$(23) \quad B^k(x) = \sum_{u_k \in U_k} B^{k+1}(f_k(x, u_k)) \lambda_k(x, u_k).$$

In the problem of intersymbol interference a forward dynamic programming solution to the problem of computing the marginal distributions has been reported by Hayes, Cover and Riera [20]. The main shortcoming of this algorithm is that it requires a separate recursion for each value of each transmitted symbol. A more efficient solution is possible by realizing that (21) fits the backward-forward framework of this section, because it can be solved by either backward or forward recursions and its operators (summations) commute. It is easy to check that in this case (15) takes the form

$$(24) \quad \begin{aligned} P^G[u_i] &= \sum_{u_{0,i-1} \in U_{0,i-1}} \sum_{u_{i+1,N} \in U_{i+1,N}} \prod_{k=0}^N \lambda_k(x_k, u_k) \\ &= \sum_{\substack{(x_i, x_{i+1}) \in \Omega_i \times \Omega_{i+1} \\ \text{s.t.} \\ x_{i+1} = f_i(x_i, u_i)}} F^i(x_i) \lambda_i(x_i, u_i) B^{i+1}(x_{i+1}) \end{aligned}$$

which further reduces to

$$(25) \quad P^G[u_i] = \sum_{\substack{x_{i+1} \in \Omega_{i+1} \\ \text{s.t. there exists} \\ x \in \Omega_i, x_{i+1} = f_i(x, u_i)}} F^{i+1}(x_{i+1}) B^{i+1}(x_{i+1})$$

if the following condition is satisfied (e.g., frequently $f_k(\cdot, \cdot)$, $k=0, \dots, N$ is a shift register system).

S2. For $k=0, \dots, N$, if there exists $x \in \Omega_k$, $x' \in \Omega_k$, $u \in U_k$ and $u' \in U_k$ such that $f_k(x, u) = f_k(x', u')$ then $u = u'$.

Thus, as in the problem of finding the shortest path through every arc, the backward-forward solution of the fixed-interval minimum probability-of-error detection problem exhibits linear complexity in the number of transmitted symbols in contrast to the quadratic complexity of the Hayer-Cover-Riera algorithm [20].

Another illustration of the applicability of the backward-forward framework is the problem of fixed-interval maximum a posteriori sequence smoothing of a discrete-time Markov process, i.e., find

$$\arg \max_{x_0 \in \Omega_0} \cdots \max_{x_N \in \Omega_N} p(x_0, \dots, x_N | y_0, \dots, y_N)$$

assuming that conditional probability density functions exist and that $p[y_0, \dots, y_N | x_0, \dots, x_N] = \prod_{k=0}^N p_k(y_k | x_k)$. If $p_k(x_k)$ and $q_k(x_k | x_{k-1})$ denote the unconditional density and the transition density of the Markov process respectively, then we have

$$\begin{aligned} & \arg \max_{x_0 \in \Omega_0} \cdots \max_{x_N \in \Omega_N} p(x_0, \dots, x_N | y_0, \dots, y_N) \\ (26) \quad & = \arg \max_{x_0 \in \Omega_0} \cdots \max_{x_N \in \Omega_N} p_0(x_0) \prod_{k=1}^N q_k(x_k | x_{k-1}) \prod_{k=0}^N p_k(y_k | x_k). \end{aligned}$$

Identifying $\vec{S}_k = \Omega_k = \vec{S}_k$, $\vec{A}_k = \Omega_{k-1} \times \Omega_k = \vec{A}_{k-1}$ and $\vec{M}_k = \{\mu: \text{there exists } x^* \in \vec{S}_{k-1}, \mu(x) = (x^*, x)\}$ and $\vec{M}_k = \{\mu: \text{there exists } x^* \in \vec{S}_{k+1}, \mu(x) = (x, x^*)\}$, we define the recursions:

$$\begin{aligned} & F_0(x_0) = p_0(x_0)p_0(y_0 | x_0), \\ (27) \quad & F_{i+1}(x_{i+1}) = p_{i+1}(y_{i+1} | x_{i+1}) \max_{x_i \in \Omega_i} q_{i+1}(x_{i+1} | x_i) F_i(x_i), \quad i=0, \dots, N-1 \end{aligned}$$

and

$$\begin{aligned} & B_N = 1, \\ (28) \quad & B_{i-1}(x_{i-1}) = \max_{x_i \in \Omega_i} q_i(x_i | x_{i-1}) p_i(y_i | x_i) B_i(x_i), \quad i=1, \dots, N. \end{aligned}$$

Then, the optimization in (26) can be carried out by either a backward or a forward recursion because

$$\begin{aligned} & \max_{x_0 \in \Omega_0} \cdots \max_{x_N \in \Omega_N} p_0(x_0) \prod_{k=1}^N q_k(x_k | x_{k-1}) \prod_{k=0}^N p_k(y_k | x_k) \\ (29) \quad & = \max_{x_N \in \Omega_N} F_N(x_N) = \max_{x_0 \in \Omega_0} p_0(x_0) p_0(y_0 | x_0) B_0(x_0). \end{aligned}$$

Once the optimum terminal state is obtained through (29), the maximizing arguments in (26) can be recovered by backtracking the optimum transitions resulting from the forward or backward recursion. The alternative to this method is the backward-forward recursion, in which one computes both the cost-to-arrive and the cost-to-go through (27) and (28), respectively, and then solves for (cf. Theorem 4.1)

$$\begin{aligned} & \arg \max_{x_i \in S_i} [\max_{x_0 \in S_0} \cdots \max_{x_{i-1} \in S_{i-1}} \max_{x_{i+1} \in S_{i+1}} \cdots \max_{x_N \in S_N} p(x_0, \dots, x_N | y_0, \dots, y_N)] \\ (30) \quad & = \arg \max_{x_i \in S_i} F_i(x_i) B_i(x_i), \quad i=0, \dots, N. \end{aligned}$$

Notice that³

$$(31) \quad F_i(x_i) = K \max_{x_0 \in S_0} \cdots \max_{x_{i-1} \in S_{i-1}} p[x_0, \cdots, x_i | y_0, \cdots, y_i]$$

and

$$(32) \quad B_i(x_i)p_i(x_i) = K \max_{x_{i+1} \in S_{i+1}} \cdots \max_{x_N \in S_N} p[x_i, \cdots, x_N | y_{i+1}, \cdots, y_N],$$

so in some cases the maximizations in the recursions (27) and (28) admit closed-form solutions. For example, in the case of a finite-dimensional linear Gaussian system (where a fixed-interval smoother in terms of the estimates produced by two filters running backwards and forwards, respectively, is well known [21], [22]), the states are (conditionally) jointly Gaussian and unnormalized marginal distributions can be obtained by maximizing with respect to the unwanted variables; hence

$$(33) \quad \begin{aligned} F_i(x_i) &= Kp(x_i | y_0, \cdots, y_i) \\ &= K \exp(-\tfrac{1}{2}\|x_i - \tilde{x}_{i/i}\|_{\tilde{\Sigma}_{i/i}}^2), \end{aligned}$$

$$(34) \quad \begin{aligned} B_i(x_i)p_i(x_i) &= Kp(x_i | y_{i+1}, \cdots, y_N) \\ &= K \exp(-\tfrac{1}{2}\|x_i - \tilde{x}_{i/i+1}\|_{\tilde{\Sigma}_{i/i+1}}^2) \end{aligned}$$

and

$$(35) \quad p_i(x_i) = K \exp(-\tfrac{1}{2}\|x_i - \bar{x}_i\|_{\Sigma_i}^2),$$

where $\tilde{x}_{i/i}$ and $\tilde{\Sigma}_{i/i}$ are the estimate and covariance of a (forward) Kalman filter, $\tilde{x}_{i/i+1}$ and $\tilde{\Sigma}_{i/i+1}$ coincide with the estimate and covariance of a Kalman filter for a derived linear-Gaussian system running backwards, [22], and \bar{x}_i and Σ_i are the unconditional mean and covariance of x_i . Substituting equations (33)–(35) in (30), the sequence of smoothed estimates is given by (cf. [21], [22])

$$(36) \quad \arg \max_{x_i \in R^i} F_i(x_i)B_i(x_i) = [\tilde{\Sigma}_{i/i}^{-1} + \tilde{\Sigma}_{i/i+1}^{-1} - \Sigma_i^{-1}]^{-1}[\tilde{\Sigma}_{i/i}^{-1} \tilde{x}_{i/i} + \tilde{\Sigma}_{i/i+1}^{-1} \tilde{x}_{i/i+1} - \Sigma_i^{-1} \bar{x}_i],$$

$i = 0, \cdots, N.$

5. A glimpse at infinite-horizon models. The finite-horizon commutativity conditions of § 2 are not sufficient to ensure the validity of the dynamic programming recursion in infinite-horizon operator models. In this section we address this problem in the general nonstationary case for backward models. Forward counterparts of all results can be obtained following the approach of § 2. Furthermore, the fixed-point property of the sought-after function of states is studied in the stationary case.

Suppose that sequences of operators $\{V_i: Q^{M_i} \rightarrow Q, i = 0, 1, \cdots\}$, $\{H^i: Q^{S_i} \rightarrow Q^{A_{i-1}}, i = 1, 2, \cdots\}$ and a sequence of functions of states $\{J^i: S_i \rightarrow Q, i = 1, 2, \cdots\}$ are given. Define $L_{\mu_{i,k}}^{i,k}$ $0 \leq i \leq k$ via (7b) and $L_{\mu_k}^{k,k}(x) = H^{k+1}J^{k+1}(\mu_k(x))$. The first goal is to impose conditions to guarantee that the function $V_{0,\infty}\{\lim_{N \rightarrow \infty} L_{\mu_{0,N}}^{0,N}\}: S_0 \rightarrow Q$ can be obtained via a (backward) infinite-horizon dynamic programming recursion, or more generally that $\lim_{N \rightarrow \infty} T_{i,N}J^{N+1}$ is a cost-to-go function, i.e.,

$$(37) \quad V_{i,\infty}\{\lim_{N \rightarrow \infty} L_{\mu_{i,N}}^{i,N}\} = \lim_{N \rightarrow \infty} T_{i,N}J^{N+1}, \quad i = 0, 1, \cdots$$

where the operator $T_i: Q^{S_{i+1}} \rightarrow Q^{S_i}$ is defined by

$$(38) \quad T_i L(x) = V_i\{H^{i+1}L(\mu_i(x))\}.$$

³ K denotes a generic state-independent term which need not coincide in different equations.

It can be shown that (37) holds if the commutativity condition B2 is fulfilled for $L^{i,N}$, $i \leq N = 1, 2, \dots$ (i.e. commutativity holds for finite-horizons) and the following condition is satisfied.

B3. $\lim_{N \rightarrow \infty} V_{i,N}\{L_{\mu_{i,N}}^{i,N}(x)\}$ and $V_{i,\infty}\{\lim_{N \rightarrow \infty} L_{\mu_{i,N}}^{i,N}(x)\}$ exist and are equal for all $x \in S_i$ and $i = 0, 1, \dots$.

As illustrated by the following example, the equality of the functions in B3 is nontrivial. Consider the algebraic system (see § 3.2) $(Q, +, \cdot) = (\{0, 1\}, \text{OR}, \text{AND})$, let $S_i = \{0, 1, \dots\}$ and

$$g_i(j, k) = \begin{cases} 1 & \text{if } k \neq 0 \text{ and } j = 0 \text{ or } k + 1, \\ 0 & \text{otherwise.} \end{cases}$$

In this case any infinite sequence of states $\{\bar{x}_i\}$ results in $\prod_{i=0}^{\infty} g_i(\bar{x}_i, \bar{x}_{i+1}) = 0$; however if $k > N$ then $g_0(0, k) \cdot g_1(k, k-1) \cdot \dots \cdot g_N(k-N+1, k-N) = 1$. Therefore we have that $V_{i,\infty}\{\lim_N L_{\mu_{i,N}}^{i,N}(j)\} = 0$ for $j = 0, 1, \dots$, while $\lim_N V_{i,N}\{L_{\mu_{i,N}}^{i,N}(0)\} = 1$.

The second question of interest in connection with infinite-horizon models is whether the sought-after function $V_{i,\infty}\{\lim_{N \rightarrow \infty} L_{\mu_{i,N}}^{i,N}\}$ is a fixed point of T in stationary problems (i.e., S_i, A_i, H^i, V_i, J^i do not depend on the stage-index). If (18) holds this is obviously the case, because in the stationary case if $V_{i,\infty}\{\lim_{N \rightarrow \infty} L_{\mu_{i,N}}^{i,N}\}$ exists then it does not depend on the stage-index. Nevertheless, even if B2 or B3 fail to be true for a particular problem the following condition is sufficient for $V_{i,\infty}\{\lim_{N \rightarrow \infty} L_{\mu_{i,N}}^{i,N}\}$ to be a fixed point of T in the stationary case.

B4. $HV_{i,\infty}\{\lim_{N \rightarrow \infty} L_{\mu_{i,N}}^{i,N}\}(a)$ and $V_{i,\infty}\{\lim_{N \rightarrow \infty} HL_{M_{i,N}}^{i,N}(a)\}$ exist and coincide for all $a \in A$.

(Note that the above counterexample to B3 satisfies B4.) It can be checked that the contraction assumptions of Denardo [2] and the continuity, uniform growth and linearity conditions of Bertsekas [5] along with the monotonicity of H imply that B3 and B4 are satisfied in stationary infimization problems. In connection with these problems, another question of interest is the existence of (ε) optimal stationary policies; obviously this problem has no counterpart in our formulation with more general operators V .

REFERENCES

- [1] L. G. MITTEN, *Composition principles for synthesis of optimum multi-stage processes*, Operations Research, 12 (1964), pp. 610-619.
- [2] E. V. DENARDO, *Contraction mappings in the theory underlying dynamic programming*, SIAM Rev., 9 (1967), pp. 165-177.
- [3] G. L. NEMHAUSER, *Introduction to Dynamic Programming*, John Wiley, New York, 1966.
- [4] R. M. KARP AND M. HELD, *Finite-state processes and dynamic programming*, SIAM J. Appl. Math., 15 (1967), pp. 693-718.
- [5] D. P. BERTSEKAS, *Monotone mappings with application in dynamic programming*, this Journal, 15 (1977), pp. 438-464.
- [6] E. PORTEUS, *Conditions for characterizing the structure of optimal strategies in infinite-horizon dynamic programs*, J. Optim. Theory Appl., 36 (1982), pp. 419-432.
- [7] T. A. BROWN AND R. E. STRAUCH, *Dynamic programming in multiplicative lattices*, J. Math. Anal. Appl., 12 (1965), pp. 364-370.
- [8] G. D. FORNEY, *The Viterbi algorithm*, Proc. IEEE, 61 (March 1973), pp. 268-278.
- [9] R. BELLMAN, *Dynamic Programming*, Princeton University Press, Princeton, NJ, 1957.
- [10] T. BASAR AND G. J. OLSDER, *Dynamic Noncooperative Game Theory*, Academic Press, New York, 1982.
- [11] E. B. DYNKIN AND A. A. YUSHKEVICH, *Controlled Markov Processes*, Springer-Verlag, New York, 1979.
- [12] D. P. BERTSEKAS AND S. E. SHREVE, *Stochastic Optimal Control: The Discrete Time Case*, Academic Press, New York, 1978.

- [13] D. BLACKWELL, D. FREEDMAN AND M. ORKIN, *The optimal reward operator in dynamic programming*, Ann. Prob., 2 (1974), pp. 926-941.
- [14] R. E. LARSON, *State Increment Dynamic Programming*, Elsevier, New York, 1968.
- [15] R. E. LARSON AND J. L. CASTI, *Principles of Dynamic Programming, Part I*, Marcel Dekker, New York, 1978.
- [16] A. V. AHO, J. E. HOPCROFT AND J. D. ULLMAN, *The Design and Analysis of Computer Algorithms*, Addison-Wesley, Reading, MA, 1974.
- [17] R. MCNAUGHTON AND H. YAMADA, *Regular expressions and state graphs for automata*, IRE Trans. Computers, 9 (1960), pp. 39-47.
- [18] L. G. MITTEN, *Preference order dynamic programming*, Management Sci., 21 (1974), pp. 43-46.
- [19] J. DUGUNDJI, *Topology*, Allyn and Bacon, Boston, 1966.
- [20] J. F. HAYES, T. M. COVER AND J. B. RIERA, *Optimal sequence detection and optimal symbol-by-symbol detection: Similar algorithms*, IEEE Trans. Comm., COM-30 (January 1982), pp. 152-157.
- [21] D. Q. MAYNE, *A solution of the smoothing problem for linear dynamic systems*, Automatica, 4 (1966), pp. 73-92.
- [22] J. E. WALL, JR., A. S. WILLSKY AND N. R. SANDELL, JR., *On the fixed-interval smoothing problem*, Stochastics, 5 (1981), pp. 1-41.
- [23] M. SNIEDOVICH, *Dynamic programming and principles of optimality*, J. Math. Anal. Appl., 65 (1978), pp. 586-606.
- [24] T. L. MORIN, *Monotonicity and the principle of optimality*, J. Math. Anal. Appl., 86 (1982), pp. 665-674.

OPTIMAL CONTROL OF LINEAR SYSTEMS WITH ALMOST PERIODIC INPUTS*

G. DA PRATO† AND A. ICHIKAWA‡

Abstract. In this paper we consider linear time-invariant and periodic systems with periodic forcing terms. We propose new quadratic control problems, both deterministic and stochastic. We also consider stochastic control with partial observation and show that the separation principle holds. Our mathematical models cover both finite and infinite dimensional systems.

Key words. optimal control, filtering, linear periodic systems

AMS(MOS) subject classifications. 93C, 49B

1. Introduction. Recently much effort has been devoted to the study of periodic systems and periodic optimization problems [6], [10], [11], [16], [22], [25], [27]. An obvious reason for this is that there are many periodic systems in nature [16], [25], [27]. But another important aspect is that periodic controls are easy to implement compared with general time-varying controls and that they sometimes even produce better performances. In [10] one of the authors has considered the quadratic control problem for a periodic system in infinite dimension and shows that under some stabilizability condition the optimal control is given by a feedback control which involves the periodic solution of a Riccati equation. We have then considered similar problems for stochastic differential equations [12]. We have shown that under partial observation the separation principle holds.

Periodic functions are easy to handle, but they lack some important properties. As we can see from simple examples [15], [16], sums of periodic functions are not periodic in general but almost periodic. Almost periodic functions are generalizations of periodic functions in some sense and were introduced by H. Bohr in 1920s. Since then almost periodic functions and differential equations related to them have been extensively studied [1], [15], [16]. It is known that almost periodic functions naturally appear in many physical systems for example in celestial mechanics or in stable electronic circuits [15], [16], [27]. So it is important to study systems with almost periodic functions.

In this paper we consider linear infinite dimensional time-invariant and periodic systems with almost periodic forcing terms. We consider both deterministic and stochastic cases and propose new quadratic control problems which are natural for almost periodic functions. With slightly different formulation, we also consider stochastic control with partial observation. We shall show that the separation principle holds.

2. The semigroup model. In this section we consider linear systems described by a strongly continuous semigroup and solve quadratic control problems.

2.1. Almost periodic solutions of a differential equation. Let Y be a real separable Hilbert space with inner product $\langle \cdot, \cdot \rangle$ and norm $|\cdot|$. Let $f(t)$, $t \in \mathbb{R}$ be a continuous function in Y . It is said to be almost periodic if from every sequence a_n we can extract a subsequence a'_n such that $\lim_{n \rightarrow \infty} f(t + a'_n)$ exists uniformly on the real line \mathbb{R} . We

* Received by the editors December 11, 1985; accepted for publication June 5, 1986.

† Scuola Normale Superiore, 56100 Pisa, Italy.

‡ Faculty of Engineering, Shizuoka University, Hamamatsu, 432 Japan. This work was done while this author was a visiting professor of C.N.R. at the Scuola Normale Superiore.

denote by $AP(Y)$ the Banach space of all continuous almost periodic functions in Y with sup norm. Periodic functions are almost periodic but the converse is not true; see for example $\cos t + \cos \sqrt{2}t$. Note that almost periodic functions are bounded. It is also easy to see that $AP(\mathbb{R})$ (scalar functions) forms an algebra. Let $f, g \in AP(Y)$. Then the mean value $\lim_{T \rightarrow \infty} 1/T \int_0^T |f(t)|^2 dt$ exists. Thus $\lim_{T \rightarrow \infty} 1/T \int_0^T \langle f(t), g(t) \rangle dt$ defines an inner product on $AP(Y)$ which we denote by $\langle f, g \rangle_{ap}$. The corresponding norm is denoted by $\|\cdot\|_{ap}$. Let $L_{ap}^2(Y)$ be the completion of $AP(Y)$ with respect to this inner product. See for details of almost periodic functions [1], [15], [16].

Now we consider the differential equation

$$(2.1) \quad y' = Ay + f,$$

where A is the infinitesimal generator of a strongly continuous semigroup e^{tA} on Y [8], [23], [26] and $f \in AP(Y)$. If $Y = \mathbb{R}^n$ and $A \in \mathbb{R}^{n \times n}$, then e^{tA} is the usual matrix exponential function. If A is stable i.e., e^{tA} is exponentially stable, then

$$(2.2) \quad y(t) = \int_{-\infty}^t e^{(t-s)A} f(s) ds$$

is well defined and is almost periodic. In fact, let a_n be an arbitrary sequence. Then

$$\begin{aligned} y(t + a_n) &= \int_{-\infty}^{t+a_n} e^{(t+a_n-s)A} f(s) ds \\ &= \int_{-\infty}^t e^{(t-\tau)A} f(\tau + a_n) d\tau. \end{aligned}$$

Now we can easily obtain the uniform convergence of a subsequence of $y(t + a_n)$ since f is almost periodic. In general it does not satisfy (2.1) but we can find a sequence $f_n \in AP(Y)$ such that $f_n \rightarrow f$ in $AP(Y)$ and

$$y_n(t) = \int_{-\infty}^t e^{(t-s)A} f_n(s) ds$$

is the solution of (3.1) with $f = f_n$ converging to y in $AP(Y)$.

A continuous function y on \mathbb{R} is called a mild solution of (2.1) if

$$y(t) = e^{(t-s)A} y(s) + \int_s^t e^{(t-r)A} f(r) dr$$

for any $t \geq s$. Then $y(t)$ given by (2.2) is a unique mild solution of (2.1) in $AP(Y)$. In fact if z is another solution, then $z(t) - y(t) = e^{(t-s)A} [z(s) - y(s)]$. Letting $s \rightarrow -\infty$ and noting that z, y are bounded, we obtain $z(t) - y(t) = 0$ for any t . A more general condition for the existence of an almost periodic mild solution to (2.1) is that e^{tA} satisfies an exponential dichotomy [15], [16] i.e., there exists a projection operator Π such that

(i) $Y_1 \triangleq \Pi Y \subset D(A)$ and $A_1 \triangleq A\Pi$ is a bounded operator on Y_1 with

$$\|e^{-tA_1}\| \leq M_1 e^{-a_1 t}, \quad t \geq 0 \quad \text{for some } M_1 > 0, \quad a_1 > 0.$$

(ii) $A: D(A) \cap Y_2 \rightarrow Y_2$, where $Y_2 = (I - \Pi)Y$ and $A_2 \triangleq A(I - \Pi)$ generates an exponentially stable semigroup on Y_2 . Then

$$y(t) = \int_{-\infty}^t e^{(t-s)A_2} (I - \Pi) f(s) ds - \int_t^\infty e^{(t-s)A_1} \Pi f(s) ds$$

is a unique mild solution of (2.1) in $AP(Y)$.

The conditions (i), (ii) are fulfilled if A satisfies (a), (b) below.

(a) The spectrum decomposition assumption of Kato [8] of the following type: the spectrum $\sigma(A)$ of A has a decomposition

$$\sigma(A) = \sigma_1(A) \cup \sigma_2(A)$$

such that $\sigma_1(A) \subset \{\lambda \in \mathbb{C} : \operatorname{Re} \lambda \geq \delta\}$, $\sigma_2(A) \subset \{\lambda \in \mathbb{C} : \operatorname{Re} \lambda < -\delta\}$ for some $\delta > 0$ and there is a rectifiable simple closed curve Γ that encloses an open set containing $\sigma_1(A)$ in its interior and $\sigma_2(A)$ in its exterior.

Now define

$$\Pi = \frac{1}{2\pi i} \int_{\Gamma} R(\lambda, A) d\lambda,$$

where $R(\lambda, A)$ is the resolvent of A . Then A satisfies the properties (i), (ii) except the exponential stability of e^{tA} . To assure this, it is sufficient to assume;

(b) e^{tA_2} satisfies the spectrum determined growth assumption [8] i.e.,

$$\sup \operatorname{Re} \sigma(A_2) = \lim_{t \rightarrow \infty} \frac{\log |e^{tA_2}|}{t}.$$

Then we have

$$|e^{tA_2}| \leq M_2 e^{-\delta t}, \quad t \geq 0 \quad \text{for some } M_2 > 0.$$

Note that (b) is satisfied for analytic semigroups or compact semigroups. Note also that if $Y = \mathbb{R}^n$, then (i), (ii) holds if A has no pure imaginary eigenvalue.

2.2. Quadratic control: the deterministic case. Now we consider the system

$$(2.3) \quad y' = Ay + Bu + f,$$

where $u \in L_{\text{ap}}^2(U)$, U is a real separable Hilbert space, $B \in L(U, Y)$ and $f \in L_{\text{ap}}^2(Y)$. Let U_{ad} be the class of all controls $u \in L_{\text{ap}}^2(U)$ such that (2.3) has an almost periodic mild solution. We wish to minimize over U_{ad} the cost functional

$$(2.4) \quad J(u) = \lim_{T \rightarrow \infty} \frac{1}{T} \int_0^T [|My|^2 + \langle Nu, u \rangle] dt,$$

where $M \in L(Y)$ and $0 < N \in L(U)$ has bounded inverse N^{-1} . We may write (2.4) as

$$(2.5) \quad J(u) = |My|_{\text{ap}}^2 + |N^{1/2}u|_{\text{ap}}^2.$$

This setup guarantees that admissible controls and their responses are necessarily bounded, which is often a requirement in practice. We may relax this as in Remark 2.1 but then we need a formulation as in § 2.6.

It is not clear if there exist any admissible controls. However, if $A - BF$ is stable for some $F \in L(Y, U)$, then the feedback law

$$u = -Fy + v, \quad v \in L_{\text{ap}}^2(U)$$

is admissible and in fact

$$y(t) = \int_{-\infty}^t e^{(t-s)(A-BF)} [Bv(s) + f(s)] ds$$

is the unique mild solution of (2.3) in $AP(Y)$. Because of the presence of f it is reasonable to assume the existence of such F . Here we shall assume slightly more:

$$(2.6) \quad \begin{array}{ll} \text{(i)} & (A, B) \text{ is stabilizable,} \\ \text{(ii)} & (M, A) \text{ is detectable.} \end{array}$$

See [21], [30], for these definitions in finite dimension and [24], [31], for the infinite dimensional case.

PROPOSITION 2.1 ([31], [24]). *Suppose (2.6) holds. Then there exists a unique $0 \leq Q \in L(Y)$ satisfying the algebraic Riccati equation*

$$(2.7) \quad QA + A^*Q + M^*M - QBN^{-1}B^*Q = 0$$

*in the inner product sense [24]. Moreover, $A - BN^{-1}B^*Q$ is stable.*

The immediate consequence of this proposition is that $r(t)$ given by

$$(2.8) \quad r(t) = \int_t^\infty e^{(s-t)(A^* - QBN^{-1}B^*)} Qf(s) ds$$

is in $AP(Y)$ and is the unique mild solution of

$$(2.9) \quad r' + (A^* - QBN^{-1}B^*)r + Qf = 0.$$

The solution to our control problem is given by the following theorem.

THEOREM 2.1. *Assume (2.6). Then there is a unique optimal control for (2.3), (2.4), and it is given by the feedback law*

$$(2.10) \quad \bar{u} = -N^{-1}B^*(Qy + r)$$

and

$$(2.11) \quad J(\bar{u}) = 2\langle r, f \rangle_{\text{ap}} - |N^{-(1/2)}B^*r|_{\text{ap}}^2,$$

where $0 \leq Q$ and r are the unique solutions of (2.7) and (2.8) respectively.

Proof. Note that \bar{u} is admissible since $A - BN^{-1}B^*Q$ is stable. Let u be an arbitrary admissible control and y its response. We differentiate $\langle Qy(t), y(t) \rangle + 2\langle r(t), y(t) \rangle$ formally and remove the terms involving A using (2.7). Then, integrating from 0 to T , we obtain

$$\begin{aligned} & \langle Qy(T), y(T) \rangle + 2\langle r(T), y(T) \rangle - \langle Qy(0), y(0) \rangle - 2\langle r(0), y(0) \rangle \\ &= - \int_0^T [|My|^2 + \langle Nu, u \rangle] dt + \int_0^T |N^{1/2}[u + N^{-1}B^*(Qy + r)]|^2 dt \\ & \quad + \int_0^T [2\langle r, f \rangle - \langle N^{-1}B^*r, B^*r \rangle] dt. \end{aligned}$$

Dividing by T and letting $T \rightarrow \infty$, we obtain

$$(2.12) \quad J(u) = |N^{1/2}[u + N^{-1}B^*(Qy + r)]|_{\text{ap}}^2 + 2\langle r, f \rangle_{\text{ap}} - |N^{-1/2}B^*r|_{\text{ap}}^2$$

where we have used the boundedness of $y(T)$ and $r(T)$. As is well known, this formal procedure can be justified by introducing approximating systems of (2.1), (2.9) with strict solutions [3] and then by passing to the limit. See [3], [9] for arguments based on Yosida approximations of A and [18], [19] for arguments using resolvent operator of A . Now the optimality of \bar{u} and (2.11) follows easily from (2.12). Since $A - BN^{-1}B^*Q$ is stable, the uniqueness of \bar{u} also follows.

2.3. Almost periodic processes. Let (Ω, F, P) be a probability space. Let $y(t)$, $t \in \mathbb{R}$ be a measurable stochastic process in Y . We say that $y(t)$ is weakly almost periodic if $y(t) \in L^2(\Omega, F, P; Y)$ (square integrable) and $Ey(t)$ and $\text{cov}[y(t)]h$, for any $h \in Y$ (mean and covariance) are almost periodic.

Now we consider

$$(2.13) \quad dy = (Ay + f) dt + G dw,$$

where A and f are taken as in § 2.1, H is a real separable Hilbert space, $w(t)$, $t \geq 0$ is an H -valued Wiener process with $\text{cov}[w(t)] = tW$, $W \geq 0$, a nuclear operator on H , and $G \in L(H, Y)$ is strongly continuous and $G(t)h$ for any $h \in H$ is almost periodic. We extend $w(t)$ on \mathbb{R} by setting

$$w(t) = w_1(-t), \quad t < 0$$

where $w_1(t)$, $t \geq 0$ is a Wiener process in H with $\text{cov}[w_1(t)] = tW$ but is independent of $w(t)$, $t \geq 0$. With this convention we are able to consider almost periodic solutions of (2.13). Let $F_t = \sigma\{w(s), s \leq t\}$, $t \in \mathbb{R}$. If A is stable then

$$(2.14) \quad y(t) = \int_{-\infty}^t e^{(t-s)A} f(s) ds + \int_{-\infty}^t e^{(t-s)A} G(s) dw(s)$$

is quadratic mean continuous, F_t -adapted and almost periodic. We define a mild solution of (2.13) as in (2.1). Then (2.14) is the unique mild solution of (2.13). It has a property similar to that of (2.7). We denote by $M_{\text{ap}}^2(Y)$ the space of F_t -adapted almost periodic processes in Y . We define $M_{\text{ap}}^2(U)$ in a similar manner.

2.4. Quadratic control under complete observation. We consider a stochastic version of the control system (2.3):

$$(2.15) \quad dy = (Ay + Bu + f) dt + G(t) dw.$$

Let U_{ad} be the set of all controls $u \in M_{\text{ap}}^2(U)$ for which (2.15) has mild solutions in $M_{\text{ap}}^2(Y)$. We wish to minimize over U_{ad} the cost functional

$$(2.16) \quad J(u) = \lim_{T \rightarrow \infty} \frac{1}{T} E \int_0^T [|My|^2 + \langle Nu, u \rangle] dt.$$

If $A - BK$ is stable for some $K \in L(Y, U)$, then the feedback law

$$u = -Ky + v, \quad v \in M_{\text{ap}}^2(U)$$

is admissible and in fact

$$y(t) = \int_{-\infty}^t e^{(t-s)(A-BK)} [Bv(s) + f(s)] ds + \int_{-\infty}^t e^{(t-s)(A-BK)} G(s) dw(s)$$

is the unique mild solution in $M_{\text{ap}}^2(Y)$. We have a result analogous to Theorem 2.1.

THEOREM 2.2. Assume (2.6). Then there is a unique optimal control for (2.15), (2.16), and it is given by the feedback law

$$(2.17) \quad \bar{u} = -N^{-1}B^*(Qy + r)$$

and

$$(2.18) \quad J(\bar{u}) = 2\langle r, f \rangle_{\text{ap}} - |N^{-(1/2)}B^*r|_{\text{ap}}^2 + \langle \text{tr } GWG^*Q, 1 \rangle_{\text{ap}}$$

where $0 \leq Q$ and r are unique solutions of (2.7), (2.8) respectively, tr denotes the trace of nuclear operators and the last term in (2.18) is the scalar product of real valued almost periodic functions.

Proof. Let y be the mild solution corresponding to an admissible control u . As in the proof of Theorem 2.1 it is possible to justify the formal application of Ito's formula to $\langle Qy(t), y(t) \rangle + 2\langle r(t), y(t) \rangle$ [3], [18], [19]. Then we obtain

$$\begin{aligned} & \langle Qy(T), y(T) \rangle + 2\langle r(T), y(T) \rangle - \langle Qy(0), y(0) \rangle - 2\langle r(0), y(0) \rangle \\ &= \int_0^T \{ |N^{1/2}[u + N^{-1}B^*(Qy + r)]|^2 - |My|^2 - \langle Nu, u \rangle \\ & \quad + 2\langle r, f \rangle - \langle N^{-1}B^*r, r \rangle + \text{tr } GWG^*Q \} dt + 2 \int_0^T \langle Qy + r, Gdw \rangle. \end{aligned}$$

Now, taking expectations, dividing by T and letting $T \rightarrow \infty$, we obtain

$$J(u) = \|N^{1/2}[u + N^{-1}B^*(Qy + r)]\|_{\text{ap}}^2 + 2\langle r, f \rangle_{\text{ap}} - |N^{-(1/2)}B^*r|_{\text{ap}}^2 + \langle \text{tr } GWG^*Q, 1 \rangle_{\text{ap}},$$

where $\|\cdot\|_{\text{ap}}$ denotes the norm in $M_{\text{ap}}^2(U)$. The optimality of \bar{u} and (2.18) follow immediately.

2.5. Quadratic control under partial observation. This subsection is devoted to a more general situation where the system is nondirectly observable and hence feedback controls are not feasible. We first recall usual quadratic problems under partial observations. Given signal and observation processes

$$(2.19) \quad dy = (Ay + Bu + f) dt + G(t) dw, \quad y(0) = y_0,$$

$$(2.20) \quad dz = Cy dt + V dv, \quad z(0) = 0,$$

one wishes to minimize

$$(2.21) \quad J_0(u) = E \int_0^T [|My|^2 + \langle Nu, u \rangle] dt$$

over all controls $u \in L^2((0, T) \times \Omega; U)$ such that $u(t)$ is adapted to $\sigma\{z(s), 0 \leq s \leq t\}$ and (2.19), (2.20) have solutions where $C \in L(Y, R^m)$, $V \in \mathbb{R}^{m \times m}$ nonsingular, v is an m -dimensional Wiener process, $y_0 \in L^2(\Omega, F, P; Y)$ is Gaussian with mean \bar{y}_s and covariance P_0 and $y_0, w(t), v(t)$ are independent. To solve this problem, one needs to consider the filtering problem

$$(2.22) \quad dy = Ay dt + G(t) dw, \quad y(0) = y_0,$$

$$(2.23) \quad dz = Cy dt + V dv, \quad z(0) = 0.$$

The optimal filter $\hat{y}(t)$ of $y(t)$ given $\{z(s), 0 \leq s \leq t\}$ is defined in terms of projections [11], [20] and is given by [4], [8], [11], [20]

$$(2.24) \quad d\hat{y} = A\hat{y} dt + P(t)C^*(VV^*)^{-1}d\eta, \quad \hat{y}(0) = \bar{y}_0,$$

where η is the innovation process [11] given by

$$(2.25) \quad d\eta = dz - C\hat{y} dt$$

and P is the solution of the Riccati equation

$$(2.26) \quad P' - AP - PA^* - GWG^* + PC^*(VV^*)^{-1}CP = 0, \quad P(0) = P_0.$$

Admissible controls for (2.19)–(2.21) are all controls $u \in L^2((0, T) \times \Omega; U)$ such that $u(t) \in L^2(\Omega, H_t, P; U) \cap L^2(\Omega, Z_t, P; U)$ a.e. t , where $H_t = \sigma\{\eta(s), 0 \leq s \leq t\}$ and $Z_t = \sigma\{z(s), 0 \leq s \leq t\}$ see [5], [7], [20]. Define \hat{y} by

$$(2.27) \quad d\hat{y} = (A\hat{y} + Bu + f) dt + P(t)C^*(VV^*)^{-1}d\eta, \quad \hat{y}(0) = \bar{y}_0;$$

then for each admissible control u we have

$$(2.28) \quad J_0(u) = \int_0^T \text{tr } MP(t)M^* dt + E \int_0^T [|M\hat{y}|^2 + \langle Nu, u \rangle] dt.$$

If we denote by $\hat{J}_0(u)$ the second term on the right-hand side of (2.28), then the original problem (2.19)–(2.21) is essentially reduced to the problem of complete observation, (2.27) and $\hat{J}(u)$.

Unfortunately following these steps it is not clear how we can formulate a quadratic control problem for (2.19) and (2.20) in terms of almost periodic functions. So we shall consider a quadratic problem which is slightly different from the previous ones but is nevertheless a natural modification of them.

We take the set of admissible controls

$$(2.29) \quad U_{ad} = \{u \in L^\infty(0, T; L^2(\Omega, F, P; U)) : u(t) \in L^2(\Omega, H_t, P; U) \cap L^2(\Omega, Z_t, P; U) \text{ a.e. } t \text{ such that its response } y \in C_B(0, \infty; L^2(\Omega, F, P; Y))\},$$

where C_B denotes the space of bounded continuous functions. We then wish to minimize

$$(2.30) \quad J(u) = \overline{\lim}_{T \rightarrow \infty} \frac{1}{T} E \int_0^T [|My|^2 + \langle Nu, u \rangle] dt.$$

The requirement in (2.29) is in the spirit of almost periodic functions. In (2.30) we take $\overline{\lim}$ since the limit no longer exists in general.

Remark 2.1. We may replace quadratic problems in §§ 2.2 and 2.4 by problems of the type above. We may also drop boundedness conditions for u and y . Such a problem was considered in [17] for the complete observation case (see also Wonham [28]).

In the sequel we take C , V and G constant and assume the following:

$$(2.31) \quad \begin{aligned} (i) & \quad (A^*, C) \text{ is stabilizable;} \\ (ii) & \quad (W^{1/2}G^*, A^*) \text{ is detectable.} \end{aligned}$$

Then by [8], [21] the solution of the Riccati equation (2.26) converges strongly to $0 \leq P_\infty \in L(Y)$ (see Corollary 3.1). Thus, in this case we have

$$J(u) = \text{tr } MP_\infty M^* + \overline{\lim}_{T \rightarrow \infty} \frac{1}{T} \hat{J}_0(u).$$

We denote by $\hat{J}(u)$ the second term above. Now consider an auxiliary control problem of minimizing $\hat{J}(u)$ over all \hat{U}_{ad} subject to (2.27), where

$$(2.32) \quad \hat{U}_{ad} = \{u \in L^\infty(0, T; L^2(\Omega, F, P; U)) : u(t) \in L^2(\Omega, H_t, P; U) \text{ a.e. } t \text{ such that } \hat{y} \in C_B(0, \infty; L^2(\Omega, F, P; Y))\}.$$

If (2.6) is satisfied, then as Theorem 2.2 holds we can show that the optimal control is given by

$$(2.33) \quad \bar{u} = -N^{-1}B^*(Q\hat{y} + r)$$

and

$$(2.34) \quad \hat{J}(\bar{u}) = 2\langle r, f \rangle_{ap} - |N^{-1/2}B^*r|_{ap}^2 + \text{tr } P_\infty C^*(VV^*)^{-1}CP_\infty Q^*,$$

where Q and r are given as in Theorem 2.1. It is well known [2], [5], [7], [14] that the control \bar{u} is also in U_{ad} i.e., admissible for the control problem defined by (2.19), (2.20) and (2.30). Summing up, we have the separation principles as follows.

THEOREM 2.3. Assume (2.6) and (2.31). Then there is a unique optimal control for the problem defined by (2.19), (2.20), (2.29) and (2.30) and it is given by (2.33). Moreover

$$(2.35) \quad J(\bar{u}) = 2\langle r, f \rangle_{ap} - |N^{-1/2}B^*r|_{ap}^2 + \text{tr } MP_\infty M^* + \text{tr } P_\infty C^*(VV^*)^{-1}CP_\infty Q^*.$$

If we set $f = 0$, then the control problem (2.19), (2.20), (2.30) is the infinite horizon problem with average cost. In this case we have the following corollary.

COROLLARY 2.1. The unique optimal control is given by the feedback law on the filter

$$\bar{u} = -N^{-1}B^*Q\hat{y}$$

and

$$J(\bar{u}) = \text{tr } MP_{\infty}M^* + \text{tr } PC^*(VV^*)^{-1}CPQ.$$

This is the separation principle on infinite horizon. See [2], [5], [7], [14], [20], [30] for the usual separation principle.

2.6. Examples. We give two simple examples.

Example 2.1. A deterministic problem. In § 2.2 we take $Y = U = \mathbb{R}$ and set $A = 3$, $B = 4$, $M = N = 1$ and $f(t) = \sin t$. Then (2.3) is

$$y' = 3y + 4u + \sin t.$$

Then the solution of (2.7) which is nonnegative is $Q = \frac{1}{2}$. Then

$$r(t) = \frac{1}{52}[\cos t + 5 \sin t].$$

It is easy to obtain

$$2\langle r, f \rangle_{\text{ap}} = \frac{5}{52}, \quad |B^*r|_{\text{ap}}^2 = \frac{4}{52}.$$

Thus the optimal control is given by

$$\bar{u} = -2y - \frac{1}{13}(\cos t + 5 \sin t)$$

and

$$J(\bar{u}) = \frac{1}{52}.$$

Here f is periodic, but we may add, for example, $\sin \sqrt{2}t$. Then it becomes almost periodic and we can compute \bar{u} and $J(\bar{u})$ in a similar manner.

Example 2.2. Consider the stochastic parabolic equation

$$dy = \left(\frac{\partial^2}{\partial x^2} y + u \right) dt + \sin x \sin at dw, \quad a \in \mathbb{R}, \quad \text{constant},$$

$$y(t, 0) = y(t, \pi) = 0.$$

In (2.15) we take

$$Y = U = L^2(0, \pi), \quad Ay = \frac{d^2}{dx^2} y, \quad D(A) = H^2(0, \pi) \cap H_0^1(0, \pi), \quad H = \mathbb{R}$$

and $W(t)$ a real standard Wiener process. We take

$$J(u) = \lim_{T \rightarrow \infty} \frac{1}{T} E \int_0^T [|y|^2 + |u|^2] dt.$$

Then the algebraic Riccati equation (2.7) becomes

$$QA + AQ - Q^2 + I = 0,$$

whose solution is $Q = \sqrt{A^2 + I} + A$ i.e.,

$$Qy = \sum_{n=1}^{\infty} (\sqrt{n^4 + 1} - n^2) \langle y, e_n \rangle e_n, \quad e_n = \sqrt{\frac{2}{\pi}} \sin nx.$$

The optimal control is

$$\bar{u} = -(\sqrt{A^2 + I} + A)y$$

and its response \bar{y} is given by

$$\begin{aligned}\bar{y}(t) &= \int_{-\infty}^t e^{-\sqrt{A^2 + I}(t-s)} \sin x \sin as \, dw(s) \\ &= \int_{-\infty}^t e^{-\sqrt{2}(t-s)} \sin as \, dw(s) \sin x.\end{aligned}$$

Thus

$$\bar{u}(t) = -(\sqrt{2} - 1) \int_{-\infty}^t e^{-\sqrt{2}(t-s)} \sin as \, dw(s) \sin x$$

and

$$J(\bar{u}) = \lim_{T \rightarrow \infty} \frac{1}{T} \int_0^T \sin^2 at \langle Q \sin x, \sin x \rangle \, dt = \frac{\pi}{2}(\sqrt{2} - 1).$$

3. The evolution operator model. In § 2 we have taken A time-invariant, but here we replace it by $A(t)$, $t \in \mathbb{R}$ with the following properties:

- (i) $A(t) = A(t + \theta)$, $t \in \mathbb{R}$ for some $\theta > 0$.
- (ii) There exists an evolution operator $U(t, s)$, $t \geq s \geq 0$ such that for the initial value problem

$$y' = A(t)y + g(t), \quad y(0) = y_0, \quad g \in L^2(0, \theta; Y)$$

has a unique mild solution

$$(3.1) \quad y(t) = U(t, s)y_0 + \int_0^t Y(t, s)g(s) \, ds.$$

(iii) If n is large, then $n \in \rho(A(t))$ and $A_n(t) = n^2[n - A(t)]^{-1} - nI$ is well defined. Moreover, $y_n(t) \rightarrow y(t)$ in $C([0, \theta], Y)$ where y_n is the strict solution of the approximating systems

$$y'_n = A_n(t)y_n + g, \quad y_n(0) = y_0.$$

(iv) $A^*(t)$ has properties similar to (i)-(iii).

The conditions (3.1) (i)-(iii) are fulfilled if the usual hypotheses of Tanabe and Kato-Tanabe [23], [26] are satisfied. Note that

$$(3.2) \quad U(t + \theta, s + \theta) = U(t, s) \quad \text{for any } t > s.$$

We replace (2.1) by

$$(3.3) \quad y' = A(t)y + f.$$

If $U(t, s)$ is exponentially stable i.e., $|U(t, s)| \leq C_1 e^{-a(t-s)}$, $t \geq s$ for some $C_1 > 0$, $a > 0$, then using (3.2) we can show that

$$y(t) = \int_{-\infty}^t U(t, s)f(s) \, ds$$

is almost periodic and it is the unique mild solution of (3.3). We can construct strict solutions of approximating systems of (3.3) which converge to $y(t)$.

Below we shall consider quadratic problems as in § 2. Since most of the arguments are similar, we omit details of proofs.

3.1. Quadratic control: the deterministic case. We follow § 2.2 and consider

$$(3.4) \quad y' = A(t)y + B(t)u + f(t),$$

$$(3.5) \quad J(u) = \lim_{T \rightarrow \infty} \frac{1}{T} \int_0^T [|My|^2 + \langle Nu, u \rangle] dt,$$

where B , M and N are strongly continuous θ -periodic operators and we assume that $0 < N$ has bounded inverse N^{-1} . We take admissible controls as in § 2.2. If $A(t) - B(t)K(t)$, with $K \in L(Y, U)$ θ -periodic strongly continuous, generates an exponentially stable evolution operator (we shall write this as $U_{A-BK}(t-s)$), then the feedback law

$$u = -K(t)y + v(t), \quad v \in L_{ap}^2(U)$$

is admissible. As with (2.6) we assume

- (3.6) (i) (A, B) is stabilizable, i.e., there exists a θ -periodic strongly continuous operator $K \in L(Y, U)$ such that $U_{A-BK}(t, s)$ is exponentially stable;
 (ii) (M, A) is detectable, i.e., there exists a θ -periodic strongly continuous operator $L \in L(Y)$ such that $U_{A^*-M^*L^*}(t, s)$ is exponentially stable.

PROPOSITION 3.1 [10]. *Assume (3.6). Then there exists a unique θ -periodic solution to the Riccati equation*

$$(3.7) \quad Q' + QA + A^*Q + M^*M - QBN^{-1}B^*Q = 0.$$

Moreover, $A - BN^{-1}B^*Q$ generates an exponentially stable evolution operator $U_{A^*-BN^{-1}B^*Q}(t, s)$.

If, further, $U_{A^*-QBN^{-1}B^*}(t, s)$ is exponentially stable, then

$$(3.8) \quad r(t) = \int_t^\infty U_{A^*-QBN^{-1}B^*}(s, t)Q(s)f(s) ds$$

is a unique almost periodic solution of

$$(3.9) \quad r' + [A^*(t) - Q(t)B(t)N^{-1}(t)B^*(t)]r + Q(t)f(t) = 0.$$

THEOREM 3.1. *Assume (3.6) and that $U_{A^*-QBN^{-1}B^*}(t-s)$ is exponentially stable. Then the unique optimal control for (3.4), (3.5) is given by the feedback law*

$$(3.10) \quad \bar{u} = N^{-1}B^*(Qy + r)$$

and

$$(3.11) \quad J(\bar{u}) = 2\langle r, f \rangle_{ap} - |N^{-1/2}B^*r|_{ap}^2$$

where Q and r are given in Proposition 3.1 and in (3.8) respectively.

Proof. Similar to the proof of Theorem 2.1. But we approximate (3.4), (3.9) by systems involving $A_n(t)$ and then employ limit arguments [3], [10].

3.2. Quadratic control with complete observation. We replace (2.15), (2.16) by the following:

$$(3.12) \quad dy = [A(t)y + B(t)u + f(t)] dt + G(t) dw,$$

$$(3.13) \quad J(u) = \lim_{T \rightarrow \infty} \frac{1}{T} E \int_0^T [|My|^2 + \langle Nu, u \rangle] dt,$$

where A , B , M and N are given as in § 3.1 and G as in § 2.4. We define admissible controls as in § 2.4.

THEOREM 3.2. *Assume (3.6) and that $U_{A^*-QBN^{-1}B^*}(t, s)$ is exponentially stable. Then the feedback control*

$$(3.14) \quad \bar{u} = -N^{-1}B^*(Qy + r)$$

is the unique optimal control and

$$(3.15) \quad J(\bar{u}) = 2\langle r, f \rangle_{\text{ap}} - |N^{-1/2}B^*r|_{\text{ap}}^2 + \langle \text{tr } GWG^*Q, 1 \rangle_{\text{ap}},$$

where Q and r are unique solutions of (3.7) and (3.8).

Proof. Similar to the proof of Theorem 2.2.

3.3. Quadratic control with partial observation. The signal and observation processes are respectively

$$(3.16) \quad dy = [A(t)y + B(t)u + f(t)] dt + G(t) dw, \quad y(0) = y_0,$$

$$(3.17) \quad dz = C(t)y dt + V(t) dv, \quad z(0) = 0,$$

and the cost functional is

$$(3.18) \quad J(u) = \overline{\lim}_{T \rightarrow \infty} \frac{1}{T} E \int_0^T [|My|^2 + \langle Nu, u \rangle] dt,$$

where A , B , M and N are given as in § 3.1 and G , C and V are θ -periodic strongly continuous operators.

We consider the filtering problem as in § 2.5 and define the innovations process similarly. Then, taking admissible controls as in (2.29), we parallel the developments to 2.5. We assume (2.31) in the periodic sense, i.e., in the sense of (3.6). Then from [10] there exists a unique θ -periodic solution to the Riccati equation

$$(3.19) \quad P' - AP - PA^* - GWG^* + PC^*(VV^*)^{-1}CP = 0.$$

Moreover, by Lemma 3.1 below the solution of (3.19) with $P(0) = P_0$ converges orbitally to the periodic solution as $t \rightarrow \infty$. Thus we have

$$J(u) = \frac{1}{\theta} \int_0^\theta \text{tr } M(t)\bar{P}(t)M^*(t) dt + \overline{\lim}_{T \rightarrow \infty} \frac{1}{T} \hat{J}_0(u),$$

where $\hat{J}_0(u)$ is defined as (2.28) and \hat{y} is now given by

$$d\hat{y} = [A(t)\hat{y} + B(t)u + f(t)] dt + P(t)C^*(VV^*)^{-1}d\eta, \quad \hat{y}(0) = \bar{y}_0.$$

Thus we have the following.

THEOREM 3.3. *Assume (3.6) and that $U_{A^*-QBN^{-1}B^*}(t, s)$ is exponentially stable. Assume further that (2.31), in the periodic sense, holds. Then*

$$\bar{u} = -N^{-1}B^*(Q\hat{y} + r)$$

is the unique optimal control for (3.16)–(3.18) and

$$J(\bar{u}) = 2\langle r, f \rangle_{\text{ap}} - |N^{-1/2}B^*r|_{\text{ap}}^2 + \frac{1}{\theta} \int_0^\theta \text{tr } [M(t)\bar{P}(t)M^*(t) + \bar{P}(t)C^*(t)(V(t)V^*(t))^{-1}C(t)\bar{P}(t)Q(t)] dt,$$

where \bar{P} , Q are the unique θ -periodic solutions of (3.7) and (3.19) respectively and r is given by (3.8).

Now we shall prove the following lemma.

LEMMA 3.1. Let P be the solution of (3.19) with $P(0) = P_0$ and let \bar{P} be the unique θ -periodic solution of (3.19) with $\bar{P}(0) = \bar{P}_0$. Then

$$(3.20) \quad P(t + n\theta) \rightarrow \bar{P}(t) \text{ strongly for any } t \geq 0 \text{ as } n \rightarrow \infty.$$

Proof. We shall show this in three steps. We denote by $\tilde{P}(t)$ the solution of (3.19) with $\tilde{P}(0) = 0$.

(i) $P_0 \leq \bar{P}_0$. By a comparison theorem, see for instance [10], we know that

$$\tilde{P}(t) \leq P(t) \leq \bar{P}(t).$$

By [10] $\tilde{P}(t + n\theta) \rightarrow \bar{P}(t)$ for any $t \geq 0$ as $n \rightarrow \infty$. Thus $P(t + n\theta) \rightarrow \bar{P}(t)$ as $n \rightarrow \infty$.

(ii) $P_0 \geq \bar{P}_0$. Note that $P(t) \geq \bar{P}(t)$ for any $t \geq 0$. Set $Q(t) = P(t) - \bar{P}(t)$, then

$$Q' - (A - BN^{-1}B^*\bar{P})^*Q - Q(A - BN^{-1}B^*\bar{P}) + QBN^{-1}B^*Q = 0.$$

Let $\bar{U}(t, s)$ be the evolution operator generated by $A - BN^{-1}B^*\bar{P}$. By differentiating we obtain

$$\begin{aligned} \frac{d}{ds} \langle Q(t-s)\bar{U}(s, 0)y, \bar{U}(s, 0)y \rangle &= \langle Q(t-s)BN^{-1}B^*Q(t-s)\bar{U}(s, 0)y, \bar{U}(s, 0)y \rangle \\ &= |N^{-1/2}B^*Q(t-s)\bar{U}(s, 0)y|^2, \quad y \in Y. \end{aligned}$$

Integrating this from 0 to t , we obtain

$$\langle Q(0)\bar{U}(t, 0), \bar{U}(t, 0)y \rangle - \langle Q(t)y, y \rangle = \int_0^t |N^{-1/2}BQ(t-s)\bar{U}(s, 0)y|^2 ds,$$

which implies

$$0 \leq \langle Q(t)y, y \rangle \leq \langle Q(0)\bar{U}(t, 0)y, \bar{U}(t, 0)y \rangle \rightarrow 0 \quad \text{as } t \rightarrow \infty \text{ [10].}$$

Hence $P(t) - \bar{P}(t) \rightarrow 0$ strongly as $t \rightarrow \infty$.

(iii) General case. Choose n large enough so that $P_0 \leq nI$ and $\bar{P}_0 \leq nI$.

Let $R(t)$ be the solution of (3.19) with $R(0) = nI$. Then by (ii) $R(t) - \bar{P}(t) \rightarrow 0$ as $t \rightarrow \infty$. By the comparison theorem we have

$$\tilde{P}(t) \leq P(t) \leq R(t).$$

Now

$$\tilde{P}(t) - \bar{P}(t) \leq P(t) - \bar{P}(t) \leq R(t) - \bar{P}(t).$$

Hence

$$P(t + n\theta) - \bar{P}(t + n\theta) \rightarrow 0 \quad \text{strongly as } n \rightarrow \infty.$$

Remark 3.1. Note that (3.20) is the global asymptotic orbital stability of $\bar{P}(t)$, i.e., the trajectory of $P(t)$ approaches to that of $\bar{P}(t)$ asymptotically.

COROLLARY 3.1. Let A , G , C and V be constant in (2.26). Then under condition (2.31), $P(t) \rightarrow P_\infty$, where $0 \leq P_\infty$ is the unique solution of

$$AP + PA^* + GWG^* - PC^*(VV^*)^{-1}CP = 0$$

and $A^* - P_\infty C^*(VV^*)^{-1}C$ is stable.

REFERENCES

- [1] L. AMERIO AND G. PROUSE, *Almost Periodic Functions and Functional Equations*, Van Nostrand, New York, 1971.
- [2] A. V. BALAKRISHNAN, *Stochastic Differential Systems I, LN in Economics Math. Systems*, Springer-Verlag, Berlin, New York, 1973.
- [3] V. BARBU AND G. DA PRATO, *Hamilton-Jacobi Equations in Hilbert Spaces*, Pitman, London, 1983.
- [4] A. BENSOUSSAN, *Filtrage Optimal des Systemes Linéaires*, Dunod, Paris, 1971.
- [5] A. BENSOUSSAN AND M. VIOT, *Optimal control of stochastic linear distributed parameter systems*, this Journal, 13 (1975), pp. 904-926.
- [6] S. BITTANTI, A. LOCATELLI AND C. MAFFEZZONI, *Periodic optimization under small perturbations*, in *Periodic Optimization*, Vol. II, A. Marzollo, ed., Springer-Verlag, New York, 1972, pp. 183-231.
- [7] F. CURTAIN AND A. ICHIKAWA, *The separation principle for stochastic evolution equation*, this Journal, 15 (1977), pp. 367-383.
- [8] R. F. CURTAIN AND A. J. PRITCHARD, *Infinite Dimensional Linear Systems Theory*, Lecture Notes in Control and Information Sci., 8, 1978, Springer-Verlag, New York.
- [9] G. DA PRATO, *Quelques résultats d'existence, unicité et régularité pour un problème de la théorie du contrôle*, J. Math. Pures Appl., 52 (1973), pp. 353-375.
- [10] ———, *Synthesis of optimal control for an infinite dimensional periodic problem*, this Journal, 25 (1987), pp. 706-714.
- [11] ———, *Periodic solutions of an infinite dimensional Riccati equation*, 12th IFIP Conference on Systems Modelling and Optimizations, Budapest, 1985.
- [12] G. DA PRATO AND A. ICHIKAWA, *Filtering and control of linear periodic systems*, submitted for publication.
- [13] R. DATKO, *Some nonautonomous control problems with quadratic cost*, J. Differential Equations, 21 (1976), pp. 231-262.
- [14] M. H. A. DAVIS, *Linear Estimation and Stochastic Control*, Chapman and Halls, London, 1977.
- [15] A. M. FINK, *Almost Periodic Differential Equations*, Lecture Notes, 377, Springer-Verlag, Berlin, New York, 1974.
- [16] C. J. HARRIS AND J. F. MILES, *Stability of Linear Systems: Some Aspects of Kinematic Similarity*, Academic Press, New York, London, 1980.
- [17] A. ICHIKAWA, *Optimal control of a linear stochastic evolution equation with state and control dependent noise*, Proc. IMA Conference on Recent Theoretical Developments in Control, Leicester, U.K., Academic Press, New York, London, 1978, pp. 383-401.
- [18] ———, *Dynamic programming approach to stochastic evolution equations*, this Journal, 17 (1979), pp. 152-174.
- [19] ———, *Semilinear stochastic evolution equations: Boundedness, stability and invariant measures*, Stochastics, 12 (1984), pp. 1-39.
- [20] ———, *Filtering and control of stochastic differential equations with unbounded coefficients*, Stochastic Anal. Appl., to appear.
- [21] H. KWAKERNAK AND R. SIVAN, *Linear Optimal Control Systems*, Wiley-Interscience, New York, 1972.
- [22] T. MOROZAN, *Periodic solutions of affine stochastic differential equations*, Preprint Series in Math., N. 36, INCREST, Bucharest, 1985.
- [23] A. PAZY, *Semigroups of Linear Operators and Applications to Partial Differential Equations*, Springer-Verlag, Berlin, New York, 1983.
- [24] A. J. PRITCHARD AND J. ZABCZYK, *Stability and stabilizability of infinite dimensional systems*, SIAM Rev., 23 (1981), pp. 25-52.
- [25] J. L. SPEYER AND R. T. EVANS, *A second variational theory for optimal periodic processes*, IEEE Trans. Automat. Control, AC-29 (1984), pp. 138-147.
- [26] H. TANABE, *Equations of Evolution*, Pitman, London, 1979.
- [27] Y. V. VENKATESH, *Energy Methods in Time-Varying System Stability and Instability Analysis*, Lecture Notes in Physics, 68, Springer-Verlag, Berlin, New York, 1977.
- [28] W. M. WONHAM, *Optimal stationary control of a linear system with state-dependent noise*, this Journal, 5 (1967), pp. 486-500.
- [29] ———, *On a matrix Riccati equation of stochastic control*, this Journal, 6 (1968), pp. 681-697.
- [30] ———, *Random differential equations in control theory*, in *Probabilistic Methods in Applied Mathematics*, Vol. 2, A. Bharucha-Reid, ed., Academic Press, New York, 1970, pp. 131-212.
- [31] J. ZABCZYK, *Remarks on the algebraic Riccati equation in Hilbert space*, Appl. Math. Optim., 2 (1976), pp. 251-258.

FINE MODULI SPACES OF INFINITE DIMENSIONAL LINEAR SYSTEMS*

SHIN KAWASE† AND NIRO YANAGIHARA‡

Abstract. This paper describes the classification problem of continuous families of infinite dimensional linear systems. We are concerned with a family of controllable and observable linear systems $(F(s), G(s), H(s))$ in Banach spaces, which depend continuously on a parameter s in a Hausdorff topological space S . Our problem is to classify such families of systems by isomorphism relation and parametrize them by "moduli." It is known that there exist "moduli" for finite dimensional controllable and observable systems. We show the result fails in the infinite dimensional case, and derive some conditions under which there exist "moduli," i.e., a fine moduli theorem for infinite dimensional systems. The theory of Banach bundles plays an important role in our study.

Key words. moduli space, infinite dimensional system, Banach bundle

AMS(MOS) subject classifications. Primary 93B99; secondary 14D20, 93C25, 93C45, 93C60

1. Introduction. In this paper we consider the problem of classification of continuous families of infinite dimensional linear systems.

Let S be a Hausdorff topological space. We deal with families of systems parametrized by $s \in S$:

$$\frac{dx(t)}{dt} = F(s)x(t) + G(s)u(t),$$

$$y(t) = H(s)x(t),$$

where $x(t) \in X_s$, X_s being a Banach space depending on s , $u(t) \in \mathbb{C}^n$, $y(t) \in \mathbb{C}^m$. $F(s)$ is the infinitesimal generator of a contraction semigroup of class (C_0) on X_s for each $s \in S$, $G(s): \mathbb{C}^n \rightarrow X_s$ and $H(s): X_s \rightarrow \mathbb{C}^m$ are bounded linear operators for each $s \in S$. If $F(s)$, $G(s)$ and $H(s)$ depend continuously on s , then we have a continuous family of systems $\{(F(s), G(s), H(s)); s \in S\}$ (for the precise definition, see Definition 2). Such a family of systems appears, e.g., when dealing with perturbations in distributed systems.

The problem we will consider here is to classify continuous families of systems under the equivalence relation of *isomorphism* (for isomorphism, see Definition 3). For general motivation and background as well as difficulty for the classification problem we refer to [3].

Our problem consists of two parts: (a) find the equivalence classes of *similar* systems (for similarity, see § 2), and (b) parametrize the equivalence classes by the topological space S . The problem is formulated in terms of representability of the functor as follows [6, p. 54].

Let $\text{Top}_H :=$ category of Hausdorff spaces, and $\text{Sets} :=$ category of sets. To each $S \in \text{Top}_H$, there corresponds $\mathcal{F}(S) := \{\text{isomorphism classes of continuous families of systems parametrized by } s \in S\}$. Then the correspondence $\mathcal{F}: S \mapsto \mathcal{F}(S)$ is considered as a contravariant functor from Top_H to Sets . Suppose $\mathcal{M} \in \text{Top}_H$. Let $h_{\mathcal{M}}: \text{Top}_H \rightarrow \text{Sets}$ be the contravariant functor defined by $h_{\mathcal{M}}(S) = \text{Hom}(S, \mathcal{M})$, $S \in \text{Top}_H$, (where $\text{Hom}(S, \mathcal{M}) :=$ the set of all continuous maps from S to \mathcal{M}). If \mathcal{M} represents \mathcal{F} , i.e., there is a functorial isomorphism between \mathcal{F} and $h_{\mathcal{M}}$, then \mathcal{M} is called a *fine moduli space* for \mathcal{F} . (For moduli spaces, see also [3].)

* Received by the editors February 3, 1986; accepted for publication (in revised form) July 4, 1986.

† Railway Technical Research Institute, Japanese National Railways, Kunitachi, Tokyo, 186, Japan.

‡ Department of Mathematics, Faculty of Science, Chiba University, Chiba City, 260, Japan.

Unfortunately, fine moduli spaces need not exist for \mathcal{F} in general [6]. So, an important subject in this direction is to find subfunctors which can be represented. Hazewinkel showed that the subfunctor $\mathcal{F}_{c,o}(S) := \{\text{isomorphism classes of continuous families of controllable and observable systems of finite dimension } k \text{ parametrized by } s \in S\}$ is representable, i.e., there exists a fine moduli space [3, p. 151]. The theorem is called the *fine moduli theorem* for $\mathcal{F}_{c,o}$. In the infinite dimensional case, however, the conditions of controllability and observability do not guarantee the existence of a fine moduli space, as shown by a counter example in § 3.

Our aim is to show fine moduli theorems for the infinite dimensional case. Our motivation for this paper is as follows: As shown by Hazewinkel and others, the fine moduli theorem is very powerful in studying finite dimensional systems parametrized by topological spaces. Hence, it would be natural to seek the infinite dimensional version of it. However, to get an infinite dimensional version is not a straightforward generalization of the finite dimensional case. Since, as we have described above, the result of Hazewinkel fails in the infinite dimensional case, we need further considerations on subfunctors and another approach. This leads us to the present paper.

For example, take Theorem 1 in § 3, which says that two continuous families of controllable and observable systems are isomorphic if they are pointwise isomorphic. This theorem plays an important role in deriving the fine moduli theorem, and is apparently the same as Hazewinkel's result for the finite dimensional case [3, p. 152]. However, his method is not applicable to the proof for the infinite dimensional case and some other techniques are required to get the proof.

Section 2 contains some necessary mathematical and system theoretical background. In § 3, we state definitions and results, Theorems 1 and 2. Theorem 1 is the main tool of this paper. Theorem 2 is the fine moduli theorem for infinite dimensional systems, which is our main result. Section 4 contains some lemmas. Sections 5–6 are devoted to the proofs of our theorems.

2. Mathematical and system theoretical preliminaries. In this section, we sketch some results of Banach bundle and system theory which are essential to this paper. For details of Banach bundles we refer to [1] and [2].

Let S be a Hausdorff topological space. By a *fiber set* we mean a triple (π, E, S) in which E is a set and $\pi: E \rightarrow S$ is a map. E is termed the *total set*, π the *projection* of E , S the *base space*, and for $s \in S$, $\pi^{-1}(s)$ is termed the *fiber* over s .

By a *bundle* we mean a fiber set (π, E, S) in which E is a topological space and $\pi: E \rightarrow S$ is a continuous map. By a *Banach family* we mean a fiber set with surjective projection in which each fiber has a given Banach space structure.

DEFINITION 1. A *Banach bundle* over S is a bundle (π, E, S) such that:

- (a) π is open and surjective.
- (b) For each $s \in S$, $\pi^{-1}(s)$ has a Banach space structure.
- (c) $x \mapsto \|x\|$ is continuous on E to \mathbb{R} .
- (d) The operation $+$ is continuous on $\{(x, y) \in E \times E; \pi(x) = \pi(y)\}$.
- (e) For each $\lambda \in \mathbb{C}$, the map $x \mapsto \lambda x$ is continuous on E to E .
- (f) If $s \in S$ and $\{x_i\}$ is any net of elements of E such that $\|x_i\| \rightarrow 0$ and $\pi(x_i) \rightarrow s$, then $x_i \rightarrow 0_s$, where 0_s stands for the zero element of the Banach space $\pi^{-1}(s)$.

We shall often write merely E instead of (π, E, S) .

If E is a fiber set over S , then $\prod E$ denotes the product of all the fibers of E . The members of $\prod E$ are called *selections* of E . If E is a Banach family, then $\prod E$ is a topological vector space. If E is a bundle over S , then the space of selections $\prod E$ contains the set $\Gamma(E)$ of continuous selections as a subspace. We refer to continuous

selections as *sections* of E . If $\gamma \in \Gamma(E)$ and $\phi(s)$, $s \in S$, is a continuous complex function on S , then $\phi\gamma \in \Gamma(E)$ (where $\phi\gamma$ is defined by $(\phi\gamma)(s) = \phi(s)\gamma(s)$). Suppose E is a Banach bundle. Then we denote by $\Gamma_b(E)$ the Banach space of bounded sections γ with the norm given by

$$(2.1) \quad \|\gamma\| = \sup \{\|\gamma(s)\|; s \in S\} < \infty.$$

Let E be a Banach family over S and $\Gamma \subset \prod E$; then we set $\Gamma_s = \{\gamma(s); \gamma \in \Gamma\}$ for $s \in S$. We say $\Gamma \subset \prod E$ is *total* for E if Γ_s spans a dense linear subspace of $\pi^{-1}(s)$ for each $s \in S$.

Let $s_0 \in S$ and U be an open set of S such that $s_0 \in U$. A *halo function* for pair $(U, \{s_0\})$ is a continuous function $\phi: S \rightarrow [0, 1]$ with $\phi(s_0) = 1$ and $\text{supp } \phi \subset U$. Let $s \in U$ and H_s denote the set of all halo functions for $(U, \{s\})$. Then for $\gamma \in \Gamma(E)$

$$\|\gamma(s)\| = \inf \{\|\phi\gamma\|; \phi \in H_s\} \quad [1, \text{p. 14}].$$

Let (π, E, S) and (π_1, E_1, S) be Banach families. If $\Phi: E \rightarrow E_1$ is a map which carries fibers into fibers such that $\Phi(s) := \Phi|_{\pi^{-1}(s)}: \pi^{-1}(s) \rightarrow \pi_1^{-1}(s)$ is a bounded linear map for each $s \in S$, then Φ is called a *Banach family map*. If E and E_1 are Banach bundles and $\Phi: E \rightarrow E_1$ is a continuous map which carries fibers into fibers such that $\Phi(s)$ is linear on each fiber of E , then Φ is called a *Banach bundle map*. Obviously a Banach bundle map is a Banach family map. Let E and E_1 be Banach bundles over S . If $\Phi: E \rightarrow E_1$ is a Banach bundle map, then we define a linear map $\Phi_*: \Gamma(E) \rightarrow \Gamma(E_1)$ by

$$(\Phi_*\gamma)(s) = \Phi(s)\gamma(s) \quad \text{for each } s \in S, \gamma \in \Gamma(E).$$

If $\Phi: E \rightarrow E_1$ is a bounded Banach bundle map (i.e., $\|\Phi(s)\|$ is bounded on S), then $\Phi_*: \Gamma_b(E) \rightarrow \Gamma_b(E_1)$ is a bounded linear operator and

$$(2.2) \quad \|\Phi_*\| \leq \sup \{\|\Phi(s)\|; s \in S\} \quad [1, \text{p. 18}].$$

Let (π, E, S) be a Banach bundle. Let U be a subspace of S , $E_U := \pi^{-1}(U)$ and $\pi_U := \pi|_{\pi^{-1}(U)}$. Then (π_U, E_U, U) is a Banach bundle, called the *reduction* of E to U and denoted by E_U . If E and E_1 are Banach bundles over S and if $\Phi: E \rightarrow E_1$ is a Banach bundle map, then so is $\Phi_U := \Phi|_{E_U}: E_U \rightarrow E_{1U}$.

Here are a few elementary properties of Banach bundles.

PROPOSITION 1 [2, p. 12]. *Let $x \in E$. Suppose that $\{x_i\} (i \in I)$ is a net of elements of E and $\pi(x_i) \rightarrow \pi(x)$ in S . Suppose further that for each $\varepsilon > 0$ we can find a net $\{y_i\}$ of elements of E and an element y of E such that:*

- (a) $y_i \rightarrow y$ in E ,
- (b) $\pi(x_i) = \pi(y_i)$ for each i and $\pi(x) = \pi(y)$,
- (c) $\|x - y\| < \varepsilon$,
- (d) $\|x_i - y_i\| < \varepsilon$ for all large i .

Then $x_i \rightarrow x$ in E .

PROPOSITION 2 [1, p. 20]. *Let E and E_1 be Banach bundles over S , and let $\Phi: E \rightarrow E_1$ be a Banach bundle map. Then Φ is locally bounded, i.e., Φ_U is bounded for each member U of an open covering of S .*

PROPOSITION 3 [2, p. 14]. *Let (π, E, S) be a Banach family and let Γ be a vector subspace of $\prod E$ such that:*

- (a) Γ is total for E ,
- (b) for any $\gamma \in \Gamma$, $s \mapsto \|\gamma(s)\|$ is continuous on S to \mathbb{R} .

Then E has a unique topology so that E is a Banach bundle with $\Gamma \subset \Gamma(E)$.

Let X be a Banach space. An *infinite dimensional linear system* (with n inputs and m outputs) with the *state space* X is a triple of operators (F, G, H) , where F is

the infinitesimal generator of a contraction semigroup $\{e^{Ft}; t \geq 0\}$ of class (C_0) on X , and $G: \mathbb{C}^n \rightarrow X$ and $H: X \rightarrow \mathbb{C}^m$ are bounded linear operators.

$L^1([0, \infty); \mathbb{C}^n)$ is the Banach space of all \mathbb{C}^n -valued summable functions on $[0, \infty)$. $C_B([0, \infty); \mathbb{C}^m)$ is the Banach space of all \mathbb{C}^m -valued continuous bounded functions on $[0, \infty)$ with the sup-norm. $C_B([0, \infty); M(m, n))$ is the Banach space of all $m \times n$ matrix-valued functions f on $[0, \infty)$ such that all column vectors belong to $C_B([0, \infty); \mathbb{C}^m)$, with the norm $\|f\| = \sup \{|f(t)|; t \in [0, \infty)\}$, where $|\cdot|$ denotes the matrix norm.

By the *controllability operator* of a system (F, G, H) , we mean the bounded linear operator $\mathcal{C}: L^1([0, \infty); \mathbb{C}^n) \rightarrow X$ defined by

$$\mathcal{C}u = \int_0^\infty e^{tF} G u(t) dt, \quad u \in L^1([0, \infty); \mathbb{C}^n).$$

By the *observability operator*, we mean the bounded linear operator $\mathcal{O}: X \rightarrow C_B([0, \infty); \mathbb{C}^m)$ defined by

$$(\mathcal{O}x)(t) = H e^{tF} x, \quad x \in X, \quad t \in [0, \infty).$$

If $X =$ the closure of $\text{Range } \mathcal{C}$, the system is *controllable*. If \mathcal{O} is injective, the system is *observable*. We say the system is *exactly controllable* if $\text{Range } \mathcal{C} = X$, and *exactly observable* if $\text{Range } \mathcal{O}$ is closed in $C_B([0, \infty); \mathbb{C}^m)$ [4].

We say two systems (F, G, H) and (F_1, G_1, H_1) with the respective state spaces X and X_1 are *similar* if there exists a bounded and boundedly invertible operator $P: X \rightarrow X_1$ which intertwines the two systems, that is, there hold

$$(2.3) \quad P e^{tF} = e^{tF_1} P, \quad PG = G_1 \quad \text{and} \quad H = H_1 P.$$

In this case, we write $(F, G, H) \sim (F_1, G_1, H_1)$. Let \mathcal{C} and \mathcal{C}_1 be the respective controllability operators, and \mathcal{O} and \mathcal{O}_1 be the respective observability operators. Then we have [4]

$$(2.4) \quad P\mathcal{C} = \mathcal{C}_1 \quad \text{and} \quad \mathcal{O}_1 P = \mathcal{O}.$$

The *weighting pattern* W of a system (F, G, H) is an $m \times n$ matrix-valued function $W(t) = H e^{tF} G$ on $[0, \infty)$. Then $W \in C_B([0, \infty), M(m, n))$. Conversely, if $W \in C_B([0, \infty); M(m, n))$, then W is the weighting pattern of a system, which is called a *realization* of W . If $(F, G, H) \sim (F_1, G_1, H_1)$, then they have the same weighting pattern. The converse is not generally true. However we have the following:

PROPOSITION 4 [4]. *Let (F, G, H) and (F_1, G_1, H_1) be two exactly controllable (resp. exactly observable) systems in Banach spaces with the same weighting pattern. In addition, let both systems be observable (resp. controllable). Then the two systems are similar.*

The *Hankel operator* $H_W: L^1([0, \infty); \mathbb{C}^n) \rightarrow C_B([0, \infty); \mathbb{C}^m)$ associated with a weighting pattern W is a bounded operator defined by

$$(H_W u)(t) = \int_0^\infty W(t+\tau) u(\tau) d\tau, \quad u \in L^1([0, \infty); \mathbb{C}^n).$$

Let \mathcal{C} and \mathcal{O} be the controllability and observability operators of (F, G, H) , respectively. Suppose (F, G, H) is controllable and observable. Then it is exactly observable if and only if

$$(2.5) \quad \text{Range } \mathcal{O} = \text{the closure of Range } H_W,$$

and exactly controllable if and only if

$$(2.6) \quad \text{Range } \mathcal{O} = \text{Range } H_W \quad (\text{see [4]}).$$

PROPOSITION 5 [4]. (a) Let $\mathcal{M}_{c,eo}$ be the set of all weighting patterns of controllable and exactly observable systems. Then,

$$\mathcal{M}_{c,eo} = C_B([0, \infty); M(m, n)).$$

(b) Let $\mathcal{M}_{ec,eo}$ be the set of all weighting patterns of exactly controllable and exactly observable systems. Then,

$$\mathcal{M}_{ec,eo} = \{W \in C_B([0, \infty); M(m, n)) \text{ such that Range } H_W \text{ is closed}\}.$$

3. Definitions and statements of our results.

DEFINITION 2. Let S be a Hausdorff space. A continuous family of linear systems (with n inputs and m outputs) over S (or parametrized by $s \in S$) is a quadruple $(E; \bar{F}, \bar{G}, \bar{H})$ of a bundle and maps defined as follows;

- (a) $E = (\pi, E, S)$ is a Banach bundle.
- (b) \bar{F} is a map from E into E such that:
 - (i) For each $s \in S$, $F(s) (= \bar{F}|_{\pi^{-1}(s)})$ is a linear map from $\pi^{-1}(s)$ to $\pi^{-1}(s)$ and generates a contraction semigroup $\{e^{tF(s)}; t \geq 0\}$ of class (C_0) on $\pi^{-1}(s)$.
 - (ii) Let $(\lambda - F(s))^{-1}$, $\lambda \in \mathbb{C}$, be the resolvent of $F(s)$. Then the map:

$$x \mapsto (\lambda - F(s))^{-1}x, \quad x \in E, \quad s = \pi(x),$$

is a Banach bundle map on E into E .

- (c) \bar{G} is a Banach bundle map from the trivial Banach bundle $S \times \mathbb{C}^n$ to E .
- (d) \bar{H} is a Banach bundle map from E to the trivial Banach bundle $S \times \mathbb{C}^m$.

Remark. Given a continuous family of systems $(E; \bar{F}, \bar{G}, \bar{H})$, for each $s \in S$ we have a system $(F(s), G(s), H(s))$ with the state space $\pi^{-1}(s)$, where $F(s) = \bar{F}|_{\pi^{-1}(s)}$, $G(s) = \bar{G}|(\{s\} \times \mathbb{C}^n)$ and $H(s) = \bar{H}|_{\pi^{-1}(s)}$.

DEFINITION 3. Let $(E; \bar{F}, \bar{G}, \bar{H})$ and $(E_1; \bar{F}_1, \bar{G}_1, \bar{H}_1)$ be continuous families of controllable and observable systems over S . If there exists a bijective Banach bundle map $\bar{P}: E \rightarrow E_1$ such that

- (a) \bar{P}^{-1} is a Banach bundle map,
- (b) there hold $\bar{P}\bar{G} = \bar{G}_1$, $\bar{H} = \bar{H}_1\bar{P}$ and $P(s)e^{tF(s)} = e^{tF_1(s)}P(s)$ for each $s \in S$, where $P(s) = \bar{P}|_{\pi^{-1}(s)}$, then the two systems are said to be *isomorphic*, denoted by $(E; \bar{F}, \bar{G}, \bar{H}) \cong (E_1; \bar{F}_1, \bar{G}_1, \bar{H}_1)$.

If $(E; \bar{F}, \bar{G}, \bar{H}) \cong (E_1; \bar{F}_1, \bar{G}_1, \bar{H}_1)$, then clearly $(F(s), G(s), H(s)) \sim (F_1(s), G_1(s), H_1(s))$ for each $s \in S$. The converse is not generally true. For a counterexample, see [3, p. 136]. However we have the following:

THEOREM 1. Let S be a Hausdorff space. Let $(E; \bar{F}, \bar{G}, \bar{H})$ and $(E_1; \bar{F}_1, \bar{G}_1, \bar{H}_1)$ be continuous families of controllable and observable systems over S . If for each $s \in S$, we have $(F(s), G(s), H(s)) \sim (F_1(s), G_1(s), H_1(s))$, then

$$(E; \bar{F}, \bar{G}, \bar{H}) \cong (E_1; \bar{F}_1, \bar{G}_1, \bar{H}_1).$$

Now, we will show a fine moduli space does not exist for $\mathcal{F}_{c,o}(S) := \{\text{isomorphism classes of continuous families of controllable and observable systems over } S\}$. We consider the case $n = m = 1$. Let $(E; \bar{F}, \bar{G}, \bar{H})$ be the continuous family of systems over $[0, 1]$ defined by:

$$E = [0, 1] \times L^2(-\infty, \infty),$$

$$\pi: E \rightarrow [0, 1]; \text{ the projection of } E,$$

$$e^{tF(s)}: L^2(-\infty, \infty) \rightarrow L^2(-\infty, \infty) \quad \text{for } s \in [0, 1] \text{ and } x \in L^2(-\infty, \infty)$$

$$(e^{tF(s)}x)(\omega) = e^{i\omega t}x(\omega),$$

$$G(s): \mathbb{C} \rightarrow L^2(-\infty, \infty) \quad \text{for } s \in [0, 1] \text{ and } c \in \mathbb{C}$$

$$G(s)c = ((1 + \omega^2)/(1 + s\omega^2))g(\omega)c,$$

$$H(s): L^2(-\infty, \infty) \rightarrow \mathbb{C} \quad \text{for } s \in [0, 1] \text{ and } x \in L^2(-\infty, \infty)$$

$$H(s)x = (\pi/2) \int_{-\infty}^{\infty} ((1 + s\omega^2)/(1 + \omega^2)) \overline{h(\omega)} x(\omega) d\omega,$$

where $(1 + \omega^2)g(\omega)$ and $h(\omega)$ are functions in $L^2(-\infty, \infty)$. If $g(\omega) \neq 0$ and $h(\omega) \neq 0$ for all $\omega \in (-\infty, \infty)$, then clearly $(E; \bar{F}, \bar{G}, \bar{H})$ is a continuous family of controllable and observable systems.

Suppose \mathcal{M} is a fine moduli space for $\mathcal{F}_{c,o}$. Then we have a one-to-one map

$$\Psi([0, 1]): \mathcal{F}_{c,o}([0, 1]) \rightarrow \text{Hom}([0, 1], \mathcal{M}).$$

Let $f \in \text{Hom}([0, 1], \mathcal{M})$ be a morphism corresponding to $\{(E; \bar{F}, \bar{G}, \bar{H})\}$. We note that for $s \neq 0$, each system $(F(s), G(s), H(s))$ is similar to $(F(1), G(1), H(1))$. In fact, the bounded and boundedly invertible operator:

$$P(s): \pi^{-1}(s) \rightarrow \pi^{-1}(1), \quad (P(s)x_s)(\omega) = ((1 + s\omega^2)/(1 + \omega^2))x_s(\omega),$$

where $x_s \in \pi^{-1}(s)$, intertwines $(F(s), G(s), H(s))$ and $(F(1), G(1), H(1))$. This means that f is constant on $(0, 1]$. (Here we note that for each $s \in S$, $\Psi(\{s\})\{(F(s), G(s), H(s))\} = f(s)$.) Since f is continuous, f must be constant on $[0, 1]$. But $(F(0), G(0), H(0))$ is not similar to $(F(1), G(1), H(1))$, because $P(0): \pi^{-1}(0) \rightarrow \pi^{-1}(1)$,

$$(P(0)x_0)(\omega) = x_0(\omega)/(1 + \omega^2), \quad x_0 \in \pi^{-1}(0),$$

does not have any bounded inverse.

From the above example we see that in order to guarantee the existence of fine moduli spaces, we must impose some additional conditions on systems. The following fine moduli theorem shows such conditions.

THEOREM 2 (Fine moduli theorem). *Let S be a Hausdorff space.*

(a) *Let $\mathcal{F}_{c,eo}(S) := \{\text{isomorphism classes of continuous families of controllable and exactly observable systems over } S\}$. Then, there exists a fine moduli space for the contravariant functor $\mathcal{F}_{c,eo}: S \mapsto \mathcal{F}_{c,eo}(S)$.*

(b) *Let $\mathcal{F}_{ec,eo}(S) := \{\text{isomorphism classes of continuous families of exactly controllable and exactly observable systems over } S\}$. Then, there exists a fine moduli space for the contravariant functor $\mathcal{F}_{ec,eo}: S \mapsto \mathcal{F}_{ec,eo}(S)$.*

Remark. From the proof of the theorem in § 6, we get fine moduli space for $\mathcal{F}_{c,eo}$ and $\mathcal{F}_{ec,eo}$, explicitly.

4. Lemmas for the proofs of theorems.

LEMMA 1. *Let $(E; \bar{F}, \bar{G}, \bar{H})$ be a continuous family of systems. Let $\{x_i\} (i \in I)$ be a net in E such that $x_i \rightarrow x$ in E . For each $\varepsilon > 0$ there are $y_i \in D(F(\pi(y_i)))$ and $y \in D(F(\pi(y)))$ (D stands for the domain) in E such that:*

- (a) $y_i \rightarrow y$ in E ,
- (b) $\pi(x_i) = \pi(y_i)$ and $\pi(x) = \pi(y)$,
- (c) $\|x - y\| < \varepsilon$,
- (d) $\|x_i - y_i\| < \varepsilon$ for all large i ,
- (e) $\{\|F(\pi(y_i))y_i\|\}_{i \geq i_0}$ is bounded for sufficiently large i_0 .

Proof. First, we note for any $x \in \pi^{-1}(s)$ and any $\lambda > 0$, $(I - \lambda^{-1}F(s))^{-1}x \in D(F(s))$ and $(I - \lambda^{-1}F(s))^{-1}x \rightarrow x$ as $\lambda \rightarrow \infty$ [5, p. 58].

Now, put $s_i = \pi(x_i)$ and $s = \pi(x)$. Given $\varepsilon > 0$, choose $\lambda > 0$ so that $\|(I - \lambda^{-1}F(s))^{-1}x - x\| < \varepsilon$, and put $y = (I - \lambda^{-1}F(s))^{-1}x$ and $y_i = (I - \lambda^{-1}F(s_i))^{-1}x_i$. Then $y \in D(F(s))$ and $y_i \in D(F(s_i))$. We will show y_i and y satisfy (a)–(e).

By the assumption, $\|x - y\| < \varepsilon$. Since the map $x \mapsto (I - \lambda^{-1}F(s))^{-1}x$, $s = \pi(x)$, is a Banach bundle map on E to E (see Definition 2), it follows that $y_i \rightarrow y$ and hence $x_i - y_i \rightarrow x - y$. Thus $\|x_i - y_i\| \rightarrow \|x - y\| < \varepsilon$. So,

$$\|x_i - y_i\| < \varepsilon \quad \text{for all large } i.$$

Since $\|(I - \lambda^{-1}F(s_i))^{-1}\| \leq 1$ (recall $F(s_i)$ is the infinitesimal generator of a contraction semigroup), we have

$$\|F(s_i)y_i\| = \|\lambda((I - \lambda^{-1}F(s_i))^{-1} - I)x_i\| \leq 2\lambda \cdot \|x_i\|.$$

So, we know $\{\|F(\pi(y_i))y_i\|\}_{i \geq i_0}$ is bounded sufficiently large i_0 . Q.E.D.

COROLLARY. In Lemma 1 we can choose y and y_i as follows:

(a) $y \in D(F(\pi(x))^2)$ and $y_i \in D(F(\pi(x_i))^2)$.

(b) $\{\|F(\pi(x_i))^2y_i\|\}_{i \geq i_0}$ is bounded for sufficiently large i_0 .

Proof. Put $s_i = \pi(x_i)$ and $s = \pi(x)$. Given $\varepsilon > 0$, choose $\lambda > 0$ so that

$$\|(I - \lambda^{-1}F(s))^{-2}x - x\| < \varepsilon,$$

and put $y = (I - \lambda^{-1}F(s))^{-2}x$ and $y_i = (I - \lambda^{-1}F(s_i))^{-2}x_i$. Then $y \in D(F(s)^2)$ and $y_i \in D(F(s_i)^2)$, and in a similar way to the proof of Lemma 1, we can show y_i and y satisfy (a)–(d) in Lemma 1. Further,

$$\|F(s_i)^2y_i\| = \|\lambda^2((I - \lambda^{-1}F(s_i))^{-1} - I)^2x_i\| \leq 4\lambda^2\|x_i\|.$$

Thus, $\{\|F(\pi(x_i))^2y_i\|\}_{i \geq i_0}$ is bounded for sufficiently large i_0 . Q.E.D.

LEMMA 2. Let $(E; \bar{F}, \bar{G}, \bar{H})$ be a continuous family of systems. The Banach family map from E to E defined by

$$x \mapsto e^{tF(s)}x, \quad x \in \pi^{-1}(s),$$

is a Banach bundle map for each $t \geq 0$.

Proof. Suppose that $\{x_i\}$ is a net of elements of E and $x_i \rightarrow x$ in E . Put $s_i = \pi(x_i)$ and $s = \pi(x)$. Given $\varepsilon > 0$, take y_i and $y \in E$ as in the above corollary.

Here, we note for $y \in D(F(s)^2)$ and $n \in \mathbb{N}$ there holds

$$\|e^{tF(s)}y - (I - tn^{-1}F(s))^{-n}y\| \leq t^2(2n)^{-1}\|F(s)^2y\|$$

(see [5, p. 58]). Hence, for given $\varepsilon > 0$, we have for large n ,

$$\begin{aligned} & \|e^{tF(s)}x - (I - tn^{-1}F(s))^{-n}y\| \\ (4.1) \quad & \leq \|e^{tF(s)}x - e^{tF(s)}y\| + \|e^{tF(s)}y - (I - tn^{-1}F(s))^{-n}y\| \\ & \leq \|x - y\| + t^2(2n)^{-1}\|F(s)^2y\| < \varepsilon, \end{aligned}$$

and similarly for sufficiently large i and large n

$$(4.2) \quad \|e^{tF(s_i)}x_i - (I - tn^{-1}F(s_i))^{-n}y_i\| \leq \|x_i - y_i\| + t^2(2n)^{-1}\|F(s_i)^2y_i\| < \varepsilon,$$

since, by the corollary, $\{\|F(\pi(x_i))^2y_i\|\}_{i \geq i_0}$ is bounded for sufficiently large i_0 .

By Definition 2, the map $x \mapsto (\lambda - F(s))^{-1}x$, $s = \pi(x)$, is a Banach bundle map, so for any $n \in \mathbb{N}$ the map $x \mapsto (I - tn^{-1}F(s))^{-n}x$ is a Banach bundle map on E . Hence

$$(4.3) \quad (I - tn^{-1}F(s_i))^{-n}y_i \rightarrow (I - tn^{-1}F(s))^{-n}y.$$

From (4.1), (4.2) and (4.3) it follows, applying Proposition 1,

$$e^{tF(s_i)}x_i \rightarrow e^{tF(s)}x.$$

This means the Banach family map $x \mapsto e^{tF(s)}x$, $x \in \pi^{-1}(s)$, is continuous, and hence it is a Banach bundle map. Q.E.D.

LEMMA 3. Let S be a Hausdorff space and $(E; \bar{F}, \bar{G}, \bar{H})$ be a continuous family of systems over S . Let $\mathcal{C}(s)$ and $\mathcal{O}(s)$ be the controllability and observability operators of $(F(s), G(s), H(s))$, respectively. Then:

(a) The Banach family map $\bar{\mathcal{C}}: S \times L^1([0, \infty); \mathbb{C}^n) \rightarrow E$ (associated with $\mathcal{C}(s)$) defined by

$$\bar{\mathcal{C}}(s, u) = \mathcal{C}(s)u, \quad (s, u) \in S \times L^1([0, \infty); \mathbb{C}^n),$$

is a Banach bundle map.

(b) The Banach family map $\bar{\mathcal{O}}: E \rightarrow S \times C_B([0, \infty); \mathbb{C}^m)$ (associated with $\mathcal{O}(s)$) defined by

$$\bar{\mathcal{O}}x = (s, \mathcal{O}(s)x), \quad x \in \pi^{-1}(s),$$

is a Banach bundle map.

Proof. (a) Let $\{(s_i, u_i)\}$ be a net of elements of $S \times L^1([0, \infty); \mathbb{C}^n)$ such that $(s_i, u_i) \rightarrow (s, 0)$ in $S \times L^1([0, \infty); \mathbb{C}^n)$. Then

$$\|\bar{\mathcal{C}}(s_i, u_i)\| = \|\mathcal{C}(s_i)u_i\| = \left\| \int_0^\infty e^{tF(s_i)} G(s_i) u_i(t) dt \right\| \leq \|G(s_i)\| \cdot \|u_i\|.$$

By Proposition 2, $\bar{\mathcal{C}}$ is locally bounded. Hence $\bar{\mathcal{C}}(s_i, u_i) \rightarrow 0$. This means $\bar{\mathcal{C}}$ is continuous, and hence $\bar{\mathcal{C}}$ is a Banach bundle map.

(b) Suppose that $\{x_i\} (i \in I)$ is a net of elements of E and $x_i \rightarrow x$ in E . We put $\pi(x_i) = s_i$ and $\pi(x) = s$. We will prove $\mathcal{O}(s_i)x_i \rightarrow \mathcal{O}(s)x$.

Given $\varepsilon > 0$, choose y_i and $y \in E$ as in Lemma 1. Then since by Proposition 2, \bar{H} is locally bounded, we have

$$(4.4) \quad \begin{aligned} \|\mathcal{O}(s_i)y_i\| &= \|H(s_i)e^{tF(s_i)}y_i\| \\ &\leq \|H(s_i)\| \cdot \|y_i\| < K, \end{aligned}$$

for all large i , where $K > 0$ is a constant. Next, for $h > 0$, as $\{\|F(\pi(x_i))^2 y_i\|\}_{i \geq i_0}$ is bounded for sufficiently large i_0 , we have

$$(4.5) \quad \begin{aligned} \|(\mathcal{O}(s_i)y_i)(t+h) - (\mathcal{O}(s_i)y_i)(t)\| &= \|H(s_i)e^{(t+h)F(s_i)}y_i - H(s_i)e^{tF(s_i)}y_i\| \\ &\leq \|H(s_i)\| \cdot \|(e^{hF(s_i)} - I)y_i\| \\ &\leq \|H(s_i)\| \cdot h \cdot \|F(s_i)y_i\| < K_1, \end{aligned}$$

for sufficiently large i , where $K_1 > 0$ is a constant. Hence, put $M := \{\mathcal{O}(s_i)y_i \text{ for sufficiently large } i \in I \text{ such that (4.4) and (4.5) hold}\}$. Then M is a subset of $C_B([0, \infty); \mathbb{C}^m)$ consisting of uniformly bounded and equi-continuous functions on $[0, \infty)$. By the Ascoli-Arzelà theorem, we see M is relatively compact. On the other hand, since by Lemma 2 the map: $x \mapsto e^{tF(s)}x$ is a Banach bundle map, we have

$$(\mathcal{O}(s_i)y_i)(t) = H(s_i)e^{tF(s_i)}y_i \rightarrow (\mathcal{O}(s)y)(t) = H(s)e^{tF(s)}y$$

for each $t \geq 0$. From this, it follows that

$$(4.6) \quad \mathcal{O}(s_i)y_i \rightarrow \mathcal{O}(s)y \text{ in } C_B([0, \infty); \mathbb{C}^m).$$

Next,

$$(4.7) \quad \begin{aligned} \|\mathcal{O}(s_i)y_i - \mathcal{O}(s_i)x_i\| &= \|H(s_i)e^{tF(s_i)}y_i - H(s_i)e^{tF(s_i)}x_i\| \\ &\leq \|H(s_i)\| \cdot \|y_i - x_i\| < \varepsilon \|H(s_i)\|, \end{aligned}$$

and similarly

$$(4.8) \quad \|\mathcal{O}(s)y - \mathcal{O}(s)x\| < \varepsilon \|H(s)\|.$$

From (4.6), (4.7) and (4.8), it follows, by applying Proposition 1,

$$\mathcal{O}(s_i)x_i \rightarrow \mathcal{O}(s)x \quad \text{in } C_B([0, \infty); \mathbb{C}^m).$$

Thus $\bar{\mathcal{O}}$ is continuous, and hence $\bar{\mathcal{O}}$ is a Banach bundle map. Q.E.D.

5. Proof of Theorem 1.

First step. Let $P(s)$ be the bounded and boundedly invertible operator intertwining the systems $(F(s), G(s), H(s))$ and $(F_1(s), G_1(s), H_1(s))$. We define a Banach family map $\bar{P}: E \rightarrow E_1$ by

$$\bar{P}x = P(s)x, \quad x \in \pi^{-1}(s).$$

Since for each $s \in S$, $P(s)$ is invertible, so is \bar{P} , and by (2.3) there hold $\bar{P}\bar{G} = \bar{G}_1$, $\bar{H} = \bar{H}_1\bar{P}$ and $P(s)e^{tF(s)} = e^{tF_1(s)}P(s)$ for each $s \in S$. Thus it suffices to prove \bar{P} is a Banach bundle map, that is, \bar{P} is continuous.

Second step. We will show \bar{P} is locally bounded. Let $\mathcal{O}(s)$ and $\mathcal{O}_1(s)$ be the observability operators of $(F(s), G(s), H(s))$ and $(F_1(s), G_1(s), H_1(s))$, respectively. Then, by Lemma 3, the operators $\bar{\mathcal{O}}: E \rightarrow S \times C_B([0, \infty); \mathbb{C}^m)$ and $\bar{\mathcal{O}}_1: E_1 \rightarrow S \times C_B([0, \infty); \mathbb{C}^m)$ associated with $\mathcal{O}(s)$ and $\mathcal{O}_1(s)$, respectively, are Banach bundle maps. Hence, by Proposition 2, for each member U of an open covering of S , $\bar{\mathcal{O}}_U: E_U \rightarrow U \times C_B([0, \infty); \mathbb{C}^m)$ and $\bar{\mathcal{O}}_{1U}: E_{1U} \rightarrow U \times C_B([0, \infty); \mathbb{C}^m)$ are bounded Banach bundle maps on E_U and E_{1U} , the reductions of E and E_1 to U , respectively. Thus

$$\bar{\mathcal{O}}_{U*}: \Gamma_b(E_U) \rightarrow \Gamma_b(U \times C_B([0, \infty); \mathbb{C}^m))$$

and

$$\bar{\mathcal{O}}_{1U*}: \Gamma_b(E_{1U}) \rightarrow \Gamma_b(U \times C_B([0, \infty); \mathbb{C}^m))$$

are bounded operators (see § 2). Further, as both of the systems are observable, $\mathcal{O}(s)$ and $\mathcal{O}_1(s)$ are injective, so are $\bar{\mathcal{O}}_U$ and $\bar{\mathcal{O}}_{1U}$. Hence $\bar{\mathcal{O}}_{U*}$ and $\bar{\mathcal{O}}_{1U*}$ are also injective.

Now, by (2.4) there holds $\mathcal{O}_1(s)P(x) = \mathcal{O}(s)$. Hence $\bar{\mathcal{O}}_{1U}\bar{P}_U = \bar{\mathcal{O}}_U$, and thus $\bar{\mathcal{O}}_{1U*}\bar{P}_{U*} = \bar{\mathcal{O}}_{U*}$. So we have

$$\bar{P}_{U*} = (\bar{\mathcal{O}}_{1U*})^{-1}\bar{\mathcal{O}}_{U*}.$$

This means that \bar{P}_{U*} is a closed operator on a Banach space $\Gamma_b(E_U)$ to $\Gamma_b(E_{1U})$. Applying the closed graph theorem, we see \bar{P}_{U*} is a bounded operator.

Let $\mathcal{C}(s)$ and $\mathcal{C}_1(s)$ be the controllability operators of $(F(s), G(s), H(s))$ and $(F_1(s), G_1(s), H_1(s))$, respectively. Let

$$\Gamma := \{\gamma; \gamma(s) = \mathcal{C}(s)u, u \in L^1([0, \infty); \mathbb{C}^n)\}$$

and

$$\Gamma_1 := \{\gamma_1; \gamma_1(s) = \mathcal{C}_1(s)u, u \in L^1([0, \infty); \mathbb{C}^n)\}.$$

Then by Lemma 3, $\Gamma \subset \Gamma(E)$ and $\Gamma_1 \subset \Gamma(E_1)$. Since $(F(s), G(s), H(s))$ and $(F_1(s), G_1(s), H_1(s))$ are controllable, Γ and Γ_1 are total for E and E_1 , respectively (see § 2). Further, by (2.4) $P(s)\mathcal{C}(s) = \mathcal{C}_1(s)$. Hence

$$(5.1) \quad \bar{P}_*\Gamma = \Gamma_1.$$

Take an $s \in S$. Let U be a member of an open covering of S such that $s \in U$. Let H_s be the set of all halo functions for $(U, \{s\})$. Then we have for all $\gamma \in \Gamma(E)$

$$\begin{aligned} \|P(s)\gamma(s)\| &= \|(\bar{P}_* \gamma)(s)\| \\ &= \inf \{\|\phi \bar{P}_* \gamma\|; \phi \in H_s\} \\ &= \inf \{\|\bar{P}_* \phi \gamma\|; \phi \in H_s\} \\ &= \inf \{\|\bar{P}_{U*} \phi \gamma\|; \phi \in H_s\} \\ &\leq \|\bar{P}_{U*}\| \cdot \inf \{\|\phi \gamma\|; \phi \in H_s\} \\ &= \|\bar{P}_{U*}\| \cdot \|\gamma(s)\|. \end{aligned}$$

Here, we note that (see § 2)

$$(\phi \bar{P}_* \gamma)(s) = \phi(s) P(s) \gamma(s) = P(s) \phi(s) \gamma(s) = (\bar{P}_* \phi \gamma)(s).$$

Since $\Gamma(E)$ is total, $\|P(s)\| \leq \|\bar{P}_{U*}\|$ for $s \in U$, and thus \bar{P} is locally bounded.

Third step. We will show $\bar{P}: E \rightarrow E_1$ is continuous. Let $x_i \rightarrow x$ in E and let U be a member of an open covering of S such that $\pi(x) \in U$. As Γ is total, for $\varepsilon > 0$ we can choose $\gamma \in \Gamma$ such that

$$\|x - \gamma(\pi(x))\| < \varepsilon.$$

Since $\gamma \in \Gamma(E)$ and $\pi(x_i) \rightarrow \pi(x)$, there holds $x_i - \gamma(\pi(x_i)) \rightarrow x - \gamma(\pi(x))$. Hence $\|x_i - \gamma(\pi(x_i))\| \rightarrow \|x - \gamma(\pi(x))\|$. Thus

$$\|x_i - \gamma(\pi(x_i))\| < \varepsilon \quad \text{for all large } i.$$

So, we have

$$\begin{aligned} (5.2) \quad \|\bar{P}x - (\bar{P}_* \gamma)(\pi(x))\| &= \|P(\pi(x))x - P(\pi(x))\gamma(\pi(x))\| \\ &\leq \|P(\pi(x))\| \cdot \|x - \gamma(\pi(x))\| < \varepsilon \cdot \|\bar{P}_{U*}\|, \end{aligned}$$

and similarly for all large i ,

$$(5.3) \quad \|\bar{P}x_i - (\bar{P}_* \gamma)(\pi(x_i))\| < \varepsilon \cdot \|\bar{P}_{U*}\|.$$

Next, by (5.1), $\bar{P}_* \gamma \in \Gamma(E_1)$. So

$$(5.4) \quad (\bar{P}_* \gamma)(\pi(x_i)) \rightarrow (\bar{P}_* \gamma)(\pi(x)).$$

Hence, applying Proposition 1, it follows from (5.2), (5.3) and (5.4) that $\bar{P}x_i \rightarrow \bar{P}x$. Thus \bar{P} is continuous. Q.E.D.

6. Proof of Theorem 2.

Proof of (a). First step. We use the notations of Proposition 5. We prove $\mathcal{M}_{c, eo}$ is a fine moduli space for $\mathcal{F}_{c, eo}$.

Define a map $\Psi(S): \mathcal{F}_{c, eo}(S) \rightarrow \text{Hom}(S, \mathcal{M}_{c, eo})$ by

$$(6.1) \quad \{(E; \bar{F}, \bar{G}, \bar{H})\} \mapsto \bar{\mathcal{O}}\bar{G}$$

for $\{(E; \bar{F}, \bar{G}, \bar{H})\} \in \mathcal{F}_{c, eo}(S)$, where $\bar{\mathcal{O}}$ is the Banach bundle map associated with the observability operator $\mathcal{O}(s)$ of $(F(s), G(s), H(s))$. We note that the Banach bundle map $\bar{\mathcal{O}}G: S \times \mathbb{C}^n \rightarrow S \times C_B([0, \infty); \mathbb{C}^m)$ is identified with an element in $\text{Hom}(S, C_B([0, \infty); M(m, n)))$ ($:=$ the set of all continuous maps from S to $C_B([0, \infty); M(m, n))$), and hence in $\text{Hom}(S, \mathcal{M}_{c, eo})$ (see Proposition 5). So, for each $s \in S$, $(\bar{\mathcal{O}}\bar{G})(s)$ is the weighting pattern of $(F(s), G(s), H(s))$.

Further, from this we see $\bar{\mathcal{O}}\bar{G}$ does not depend on the choice of representatives. Indeed, suppose $(E_1; \bar{F}_1, \bar{G}_1, \bar{H}_1) \in \{(E; \bar{F}, \bar{G}, \bar{H})\}$. Then $(E_1; \bar{F}_1, \bar{G}_1, \bar{H}_1) \cong (E; \bar{F}, \bar{G}, \bar{H})$. Hence for each $s \in S$, $(F_1(s), G_1(s), H_1(s))$ and $(F(s), G(s), H(s))$ have the same weighting pattern (see § 3). So for each $s \in S$, $(\bar{\mathcal{O}}\bar{G})(s) = (\bar{\mathcal{O}}_1\bar{G}_1)(s)$, where $\bar{\mathcal{O}}_1$ is the Banach bundle map associated with the observability operator $\mathcal{O}_1(s)$ of $(F_1(s), G_1(s), H_1(s))$.

Second step. We will show that $\Psi(S)$ is a surjection. To see this, given $\bar{W} \in \text{Hom}(S, \mathcal{M}_{c,eo})$ we will show that there exists a continuous family $(E_0; \bar{F}_0, \bar{G}_0, \bar{H}_0)$ of controllable and exactly observable systems such that

$$(6.2) \quad \Psi(S)\{(E_0; \bar{F}_0, \bar{G}_0, \bar{H}_0)\} = \bar{W}.$$

For each $s \in S$ we have a realization $(F_0(s), G_0(s), H_0(s))$ of $\bar{W}(s)$ as follows [4]:

$$(6.3) \quad \begin{aligned} &\text{State space } X_{0s} = \text{the closure of Range } H_{\bar{W}(s)}, \\ &\{e^{tF_0(s)}\} = \text{the restriction on } X_{0s} \text{ of the left translation} \\ &\text{semigroup on } C_B([0, \infty); \mathbb{C}^m), \\ &G_0(s): \mathbb{C}^n \rightarrow X_{0s}, \quad G_0(s)c = \bar{W}(s)c \quad \text{for } c \in \mathbb{C}^n, \\ &H_0(s): X_{0s} \rightarrow \mathbb{C}^m, \quad H_0(s)x = x(0) \quad \text{for } x \in X_{0s}, \end{aligned}$$

where $H_{\bar{W}(s)}$ is the Hankel operator of $\bar{W}(s)$. Let $\mathcal{C}_0(s)$ and $\mathcal{O}_0(s)$ be the controllability and observability operators of $(F_0(s), G_0(s), H_0(s))$, respectively. Since there hold [4]:

$$(6.4) \quad \begin{aligned} &\mathcal{C}_0(s) = H_{\bar{W}(s)} \text{ and} \\ &\mathcal{O}_0(s) = \text{embedding of the closure of Range } H_{\bar{W}(s)} \text{ into } C_B([0, \infty); \mathbb{C}^m), \end{aligned}$$

$(F_0(s), G_0(s), H_0(s))$ is controllable and exactly observable (see § 2).

Let (π, E_0, S) be the Banach family defined by

$$\pi^{-1}(s) := X_{0s},$$

and let \bar{F}_0, \bar{G}_0 and \bar{H}_0 be the Banach family maps defined by

$$\begin{aligned} \bar{F}_0: E_0 &\rightarrow E_0; & \bar{F}_0|_{\pi^{-1}(s)} &:= F_0(s), \\ \bar{G}_0: S \times \mathbb{C}^n &\rightarrow E_0; & \bar{G}_0|_{(\{s\} \times \mathbb{C}^n)} &:= G_0(s), \\ \bar{H}_0: E_0 &\rightarrow S \times \mathbb{C}^m; & \bar{H}_0|_{\pi^{-1}(s)} &:= H_0(s), \end{aligned}$$

respectively. Thus we have obtained a family of controllable and exactly observable systems $(E_0; \bar{F}_0, \bar{G}_0, \bar{H}_0)$ such that (6.2) holds.

Now, it remains to prove it is a continuous family. Put

$$\Gamma_0 := \{\gamma; \gamma(s) = \mathcal{C}_0(s)u, u \in L^1([0, \infty); \mathbb{C}^n)\}.$$

Then, as $(F_0(s), G_0(s), H_0(s))$ is controllable, Γ_0 is total for E_0 . Moreover, as $\bar{W} \in \text{Hom}(S, \mathcal{M}_{c,eo})$, $H_{\bar{W}(s)}: L^1([0, \infty); \mathbb{C}^n) \rightarrow C_B([0, \infty); \mathbb{C}^m)$ is strongly continuous in $s \in S$. Hence by (6.4)

$$\|\gamma(s)\| = \|\mathcal{C}_0(s)u\| = \|H_{\bar{W}(s)}u\| \quad (\gamma \in \Gamma_0)$$

is a continuous function on S to \mathbb{R} . From Proposition 3, it follows there is a unique topology so that E_0 is a Banach bundle over S . Then, since clearly \bar{F}_0, \bar{G}_0 and \bar{H}_0 are maps which satisfy the conditions of Definition 2, we see $\Psi(S)$ is a map of $\mathcal{F}_{c,eo}(S)$ onto $\text{Hom}(S, \mathcal{M}_{c,eo})$.

Third step. We will show that $\Psi(S)$ is an injection. Suppose for $\{(E; \bar{F}, \bar{G}, \bar{H})\}$ and $\{(E_1; \bar{F}_1, \bar{G}_1, \bar{H}_1)\} \in \mathcal{F}_{c, eo}(S)$,

$$\Psi(S)\{(E; \bar{F}, \bar{G}, \bar{H})\} = \Psi(S)\{(E_1; \bar{F}_1, \bar{G}_1, \bar{H}_1)\}.$$

Then for each $s \in S$, $(F(s), G(s), H(s))$ and $(F_1(s), G_1(s), H_1(s))$ are controllable and exactly observable systems with the same weighting pattern. Hence by Proposition 4, for each $s \in S$ $(F(s), G(s), H(s)) \sim (F_1(s), G_1(s), H_1(s))$, and from Theorem 1, it follows that

$$(E; \bar{F}, \bar{G}, \bar{H}) \cong (E_1; \bar{F}_1, \bar{G}_1, \bar{H}_1).$$

So $\{(E; \bar{F}, \bar{G}, \bar{H})\} = \{(E_1; \bar{F}_1, \bar{G}_1, \bar{H}_1)\}$. This means $\Psi(S)$ is an injection. Thus, we see $\text{Hom}(S, \mathcal{M}_{c, eo})$ is homomorphic to $\mathcal{F}_{c, eo}(S)$.

Proof of (b). We prove $\mathcal{M}_{ec, eo}$ is a fine moduli space for $\mathcal{F}_{ec, eo}$.

Define a map $\Psi(S): \mathcal{F}_{ec, eo}(S) \rightarrow \text{Hom}(S, \mathcal{M}_{ec, eo})$ by

$$(6.1') \quad \{(E; \bar{F}, \bar{G}, \bar{H})\} \mapsto \bar{O}G \quad \text{for } \{(E; \bar{F}, \bar{G}, \bar{H})\} \in \mathcal{F}_{ec, eo}(S).$$

Then, in a similar way to the third step of the above proof, we can show $\Psi(S)$ is an injection. Hence it remains to prove $\Psi(S)$ is a surjection.

Given $\bar{W} \in \text{Hom}(S, \mathcal{M}_{ec, eo})$, let $(E_0; \bar{F}_0, \bar{G}_0, \bar{H}_0)$ be the family of controllable and exactly observable systems defined by (6.3) (note $\mathcal{M}_{ec, eo} \subset \mathcal{M}_{c, eo}$). Since for each $s \in S$ the range of the Hankel operator $H_{\bar{W}(s)}$ associated with $\bar{W}(s)$ is closed (see Proposition 5), it follows from (6.4) that

$$\text{Range } \mathcal{O}_0(s) = \text{the closure of Range } H_{\bar{W}(s)} = \text{Range } H_{\bar{W}(s)}.$$

Hence by (2.6), $(F_0(s), G_0(s), H_0(s))$ is exactly controllable. So we have a continuous family $(E_0; \bar{F}_0, \bar{G}_0, \bar{H}_0)$ of exactly controllable and exactly observable systems such that (6.2) holds. Thus we see $\Psi(S)$ is a surjection. Q.E.D.

REFERENCES

- [1] M. J. DUPRE AND R. M. GILLETTE, *Banach Bundles, Banach Modules and Automorphisms of C^* -Algebras*, Pitman, Boston, 1983.
- [2] J. M. G. FELL, *An extension of Mackey's method to Banach *-algebraic bundles*, Mem. Amer. Math. Soc., 90 (1969).
- [3] M. HAZEWINKEL, *(Fine) moduli (spaces) for linear systems: What are they and what are they good for*, in Geometrical Methods for the Theory of Linear Systems, C. I. Byrnes and C. F. Martin, eds., NATO Advanced Study Institutes Series C 62, (1979), pp. 125-193.
- [4] S. KAWASE AND N. YANAGIHARA, *State space isomorphism theorems for infinite dimensional linear systems*, Technical Reports of Math. Sci., Chiba Univ., No. 10 (1985).
- [5] H. TANABE, *Equations of Evolution*, Pitman, Boston, 1979.
- [6] A. TANNENBAUM, *Invariance and System Theory: Algebraic and Geometric Aspects*, Lecture Notes in Mathematics 845, Springer-Verlag, Heidelberg, 1981.

A CLASS OF OPTIMIZATION PROBLEMS WITH NONCOMPACT CONSTRAINTS: GENERAL RESULTS AND APPLICATIONS*

KJELL HOLMÅKER†‡ AND DAVID STEWART†

Abstract. Optimization problems recently studied in liver kinetics feature noncompact constraints and highly nonunique solutions. Here a wider class of problems with these properties is considered. It is shown that it is sufficient to solve the problem with an additional, compact constraint. From the solutions of this restricted problem all solutions of the original problem can be obtained. By means of this result and Pontryagin's maximum principle, solutions are found for some problems of liver-kinetic interest.

Key words. existence of optimal controls, noncompact constraints, Pontryagin's maximum principle, liver enzyme kinetics

AMS(MOS) subject classifications. 49A10, 49B10, 92A09, 80A32

1. Introduction. In some recent papers ([1], [3] and [4]) the following optimization (optimal control) problem from liver kinetics has been studied: Let P and M satisfy the system of differential equations

$$(1.1) \quad \begin{aligned} \frac{dP}{dx} &= -\alpha(P)f(x) \\ \frac{dM}{dx} &= \alpha(P)f(x) - \beta(M)g(x) \end{aligned} \quad (0 \leq x \leq L),$$

with nonnegative α and β and with boundary conditions

$$(1.2) \quad P(0) = P_0 > 0, \quad M(0) = 0.$$

Find the minimum of $M(L)$ with respect to all control functions f and g satisfying

$$(1.3) \quad f(x) \geq 0, \quad g(x) \geq 0 \quad \text{for } 0 \leq x \leq L,$$

$$(1.4) \quad \int_0^L f(x) dx = \int_0^L g(x) dx = 1.$$

Here P and M represent the concentrations of two substances in the blood (precursor and metabolite), and the system (1.1) describes a two-stage enzymatic process in the liver. The quantities α and β are proportional to the rates of conversion in the two stages, and f and g describe the distributions of two corresponding kinds of enzymes along a liver capillary of length L . A detailed presentation of the problem can be found in [1].

Because of the noncompactness of the constraints (1.3) and (1.4) it is not clear that the problem has a solution. In [1], [3] and [4] existence was established by various methods under various assumptions on β ([3] is most general in this respect), but these methods are of an ad hoc nature and seem difficult to generalize. In this paper we shall prove an existence theorem for a much more general problem, where (1.1) may be replaced by an arbitrary system which is linear in the control variables and where the constraints on the controls are of the same type as in (1.3) and (1.4). This

* Received by the editors February 3, 1986; accepted for publication (in revised form) July 24, 1986. This work was supported in part by the Australian Research Grants Scheme.

† Department of Mathematics, University of Queensland, Brisbane, Australia.

‡ The author is on leave from the Department of Mathematics, Chalmers University of Technology, Göteborg, Sweden.

is done in § 2, where we show that it is sufficient to look only for solutions where the control vector is constrained to belong to a certain fixed compact set. The problem with this additional constraint has an optimal solution, which is also a solution of the original problem. Furthermore, from the solutions of the restricted problem we can construct all of the infinitely many solutions of the original problem.

In §§ 3–5 we shall apply the results from § 2 to a problem of liver-kinetic interest, but we believe that they have wider applicability. The problem we consider in §§ 3–5 is an extension of problem (1.1)–(1.4), where there is a third kind of enzyme transforming the precursor directly to the end product, thereby by-passing the metabolite. Instead of (1.1) we then have the system

$$\begin{aligned}\frac{dP}{dx} &= -\alpha(P)f(x) - \gamma(P)k(x), \\ \frac{dM}{dx} &= \alpha(P)f(x) - \beta(M)g(x)\end{aligned}$$

where γ is of the same type as α , and k is of the same type as f and g . The addition of the term $-\gamma(P)k(x)$ in the first equation makes the problem considerably more difficult to handle, especially by previously published methods ([1], [3] and [4]). However, we shall show how the results from § 2 enable us to find the solutions. In §§ 3 and 4 we work with the restricted version of the problem, where there is an additional compact constraint, and we apply Pontryagin's maximum principle to obtain (or at least characterize) the solutions. In § 3 we assume that α and γ are strictly monotonic and that β is unimodal (i.e. has a single maximum point), and we find that the problem (in the restricted version) has a unique solution, which can be explicitly calculated.

In § 4 we treat the case where β is strictly monotonic, but α and γ general. A characterization of the optimal controls is given.

In § 5 we treat the case where α , β and γ are general, and we show that certain optimal solutions can be obtained by means of a simpler problem not involving M , β and g . The results of § 4 characterize the solutions of this simpler problem.

2. An existence theorem. We shall consider a class of control systems containing the systems mentioned in § 1 as special cases. Using the standard notation in control theory with t as independent variable, $x = (x_1, \dots, x_n)^T \in \mathbb{R}^n$ as state vector, and $u = (u_1, \dots, u_m)^T \in \mathbb{R}^m$ as control vector, we consider the system

$$\begin{aligned}(2.1) \quad \dot{x} &= \frac{dx}{dt} = A(x)u, \quad 0 \leq t \leq T, \\ x(0) &\in K, \\ u_i(t) &\geq 0, \quad 0 \leq t \leq T, \quad 1 \leq i \leq m, \\ \int_0^T u(t) dt &= a,\end{aligned}$$

and the problem (\mathcal{P}) of minimizing the cost

$$C(u, x) = g(x(T)) + \int_0^T b(x(t))u(t) dt.$$

Here $T > 0$ is fixed, K is a compact set in \mathbb{R}^n , $a = (a_1, \dots, a_m)^T \in \mathbb{R}^m$, $a_i \geq 0$, $a \neq 0$, $A(x)$ is an $n \times m$ matrix and $b(x)$ is an m -row-vector, both continuous as functions of $x \in \mathbb{R}^n$, and g is a real-valued continuous function. The controls u are (Lebesgue)

measurable functions, and the differential equations are understood in the Carathéodory sense.

DEFINITION 2.1. Let v be measurable and nonnegative (i.e., have nonnegative components) and let y be continuous on $[0, T]$. We say that (v, y) is transformed into (u, x) by a rescaling of time, or briefly that (u, x) is a *time transform* of (v, y) , if u is nonnegative, and there exists a $\theta \in L_1(0, T)$ such that $\theta(t) \geq 0$ for $t \in [0, T]$, the function

$$(2.2) \quad \phi(t) = \int_0^t \theta(\tau) d\tau$$

satisfies $\phi(T) = T$, and

$$(2.3) \quad u(t) = v(\phi(t))\theta(t) \quad \text{for almost all } t \in [0, T],$$

$$(2.4) \quad x(t) = y(\phi(t)) \quad \text{for all } t \in [0, T].$$

LEMMA 2.1. If (v, y) is a solution of (2.1), then any (u, x) which is a time-transform of (v, y) is a solution of (2.1) with the same cost, that is

$$C(u, x) = C(v, y).$$

Proof. Let θ and ϕ be as in Definition 2.1 and let u and x be given by (2.3) and (2.4). Note that $\phi: [0, T] \rightarrow [0, T]$ is absolutely continuous with $\phi' = \theta \geq 0$ a.e. We shall make use of the formula

$$(2.5) \quad \int_0^{\phi(t)} F(\sigma) d\sigma = \int_0^t F(\phi(\tau))\theta(\tau) d\tau, \quad t \in [0, T],$$

which holds for any $F \in L_1(0, T)$. A proof of (2.5) can be found, e.g., in [7, p. 377]. From (2.3) and (2.5) it follows that $u_i \in L_1(0, T)$, and we also see that $u_i(t) \geq 0$ on $[0, T]$.

Since (v, y) satisfies (2.1) we have

$$y(s) = y(0) + \int_0^s A(y(\sigma))v(\sigma) d\sigma, \quad 0 \leq s \leq T.$$

Using (2.3)–(2.5), we then get

$$\begin{aligned} x(t) &= y(\phi(t)) = y(0) + \int_0^{\phi(t)} A(y(\sigma))v(\sigma) d\sigma \\ (2.6) \quad &= x(0) + \int_0^t A(y(\phi(\tau)))v(\phi(\tau))\theta(\tau) d\tau \\ &= x(0) + \int_0^t A(x(\tau))u(\tau) d\tau. \end{aligned}$$

Thus x satisfies $dx/dt = A(x)u(t)$ and $x(0) = y(0) \in K$. In the same way we get

$$(2.7) \quad a = \int_0^T v(s) ds = \int_0^T v(\phi(t))\theta(t) dt = \int_0^T u(t) dt,$$

and

$$\begin{aligned} (2.8) \quad \int_0^T b(y(s))v(s) ds &= \int_0^T b(y(\phi(t)))v(\phi(t))\theta(t) dt \\ &= \int_0^T b(x(t))u(t) dt. \end{aligned}$$

Also $x(T) = y(\phi(T)) = y(T)$.

It follows that (u, x) satisfies (2.1) and that

$$C(u, x) = C(v, y). \quad \square$$

LEMMA 2.2. *Let (u, x) be a solution of (2.1). Then there exists a solution (v, y) of (2.1) such that (u, x) is a time-transform of (v, y) , and*

$$\sum_{i=1}^m v_i(s) = T^{-1} \sum_{i=1}^m a_i \quad \text{for all } s \in [0, T].$$

The costs of (u, x) and (v, y) are the same.

Proof. Let

$$\theta(t) = T \left(\sum_{i=1}^m a_i \right)^{-1} \sum_{i=1}^m u_i(t), \quad 0 \leq t \leq T.$$

We define ϕ by (2.2) and ψ by

$$\psi(s) = \min \{t: 0 \leq t \leq T, \phi(t) = s\}, \quad 0 \leq s \leq T.$$

There may be countably many intervals $[\alpha_j, \beta_j]$ where ϕ is constant. We may assume that $\theta(t) = 0$ everywhere in these intervals (if necessary, redefine $u(t)$ on a set of measure zero). Define

$$(2.9) \quad \begin{aligned} y(s) &= x(\psi(s)), \\ v_i(s) &= \begin{cases} \frac{u_i(\psi(s))}{\theta(\psi(s))} & \text{if } \theta(\psi(s)) > 0, \\ \frac{1}{mT} \sum_{i=1}^m a_i & \text{if } \theta(\psi(s)) = 0, \end{cases} \end{aligned}$$

for $s \in [0, T]$ and $1 \leq i \leq m$. Let us show that v_i is measurable. Let f be a measurable real-valued function on $[0, T]$, and let α be any real number. Then

$$E = \{t \in [0, T]: f(t) < \alpha\}$$

is measurable, and

$$\begin{aligned} \{s \in [0, T]: f(\psi(s)) < \alpha\} &= \{s: \psi(s) \in E\} \\ &= \phi \left(E \setminus \bigcup_j (\alpha_j, \beta_j] \right). \end{aligned}$$

But $E \setminus \bigcup_j (\alpha_j, \beta_j]$ is measurable, and ϕ , being absolutely continuous, maps measurable sets onto measurable sets (see, e.g., [6, pp. 248–250]). Thus $s \mapsto f(\psi(s))$ is measurable, and it follows that v_i is measurable. We see that $v_i(s) \geq 0$ and $\sum_{i=1}^m v_i(s) = T^{-1} \sum_{i=1}^m a_i$ for all $s \in [0, T]$. It follows from (2.9) that

$$u(\psi(s)) = v(s)\theta(\psi(s)) \quad \text{for all } s \in [0, T],$$

because $\theta(\psi(s)) = 0$ implies $u(\psi(s)) = 0$. In particular, we have

$$u(\psi(\phi(t))) = v(\phi(t))\theta(\psi(\phi(t))) \quad \text{for all } t \in [0, T].$$

Now $\psi(\phi(t)) \leq t$ for all t , and if $\psi(\phi(t)) < t$, then ϕ is constant on the interval $[\psi(\phi(t)), t]$, so that $\theta = 0$ there; hence $u = 0$ on $[\psi(\phi(t)), t]$, and $u(\psi(\phi(t))) = u(t) = 0$, $\theta(\psi(\phi(t))) = \theta(t) = 0$. Thus

$$(2.10) \quad u(t) = v(\phi(t))\theta(t) \quad \text{for all } t \in [0, T].$$

We also have

$$(2.11) \quad x(t) = y(\phi(t)) \quad \text{for all } t \in [0, T],$$

because $y(\phi(t)) = x(\psi(\phi(t)))$, and if $\psi(\phi(t)) < t$, then again ϕ is constant on $[\psi(\phi(t)), t]$ and $u = 0$ there; thus x is constant on $[\psi(\phi(t)), t]$, so that $x(\psi(\phi(t))) = x(t)$.

Since (u, x) satisfies (2.1), we have

$$x(t) = x(0) + \int_0^t A(x(\tau))u(\tau) d\tau, \quad 0 \leq t \leq T.$$

Using (2.10) and (2.11) and reading (2.6) backwards, we get

$$x(t) = y(0) + \int_0^{\phi(t)} A(y(\sigma))v(\sigma) d\sigma,$$

so that

$$y(s) = x(\psi(s)) = y(0) + \int_0^s A(y(\sigma))v(\sigma) d\sigma.$$

Thus y satisfies the equation $dy/ds = A(y)v(s)$. In the same way we get from (2.7) and (2.8) that $\int_0^T v(s) ds = a$, so that (v, y) satisfies (2.1), and that $C(v, y) = C(u, x)$. \square

LEMMA 2.3. *Let (u, x) be a time-transform of (v, y) and let $\tilde{v} = v$ a.e., $\tilde{v} \geq 0$. Then (u, x) is a time-transform of (\tilde{v}, y) .*

Proof. There is a θ such that (2.3) and (2.4) hold. We want to show that $\tilde{v}(\phi(t))\theta(t) = v(\phi(t))\theta(t)$ a.e. Assume that this is not true. Then there is a set E of positive measure such that $\theta(t) \neq 0$ and $\tilde{v}(\phi(t)) \neq v(\phi(t))$ for $t \in E$. Since $\tilde{v} = v$ a.e., this implies that $m(\phi(E)) = 0$, where m denotes the Lebesgue measure. From (2.5) we get

$$\begin{aligned} 0 = m(\phi(E)) &= \int_0^T \chi_{\phi(E)}(s) ds = \int_0^T \chi_{\phi(E)}(\phi(t))\theta(t) dt \\ &\geq \int_0^T \chi_E(t)\theta(t) dt \geq 0. \end{aligned}$$

Thus $\int_0^T \chi_E(t)\theta(t) dt = 0$, which implies that $\theta(t) = 0$ a.e. on E . This contradiction proves the lemma. \square

Let $(\hat{\mathcal{P}})$ be the problem of minimizing $C(u, x)$ over all (u, x) satisfying (2.1) and the additional constraint

$$(2.12) \quad \sum_{i=1}^m u_i(t) = T^{-1} \sum_{i=1}^m a_i, \quad 0 \leq t \leq T.$$

THEOREM 2.1. *If problem $(\hat{\mathcal{P}})$ has a solution, then it is also a solution of problem (\mathcal{P}) . In this case (u, x) is a solution of (\mathcal{P}) if and only if (u, x) is a time-transform of (v, y) , where (v, y) is a solution of $(\hat{\mathcal{P}})$. If the solution (v_0, y_0) of $(\hat{\mathcal{P}})$ is essentially unique, i.e., every other solution is of the form (v, y_0) , where $v = v_0$ a.e., then (u, x) is a solution of (\mathcal{P}) if and only if (u, x) is a time-transform of (v_0, y_0) .*

Proof. The first statement follows from Lemma 2.2, and the second follows from Lemma 2.1 and 2.2. The last statement follows from Lemma 2.3. \square

THEOREM 2.2. *Assume that (2.1) with (2.12) has solutions and that all solutions x are uniformly bounded, i.e., that there exists a constant c such that $\max_{0 \leq t \leq T} |x(t)| \leq c$ for all x such that (u, x) satisfies (2.1) and (2.12) for some u . (This is the case, for instance, if $|x^T A(x)| \leq c_1(1 + |x|^2)$.) Then problem (\mathcal{P}) has solutions.*

Proof. It follows from general existence theorems in optimal control theory that problem $(\hat{\mathcal{P}})$ has a solution, see e.g. [2, pp. 61–62], [5, pp. 259–262]. It is also easy to give a direct proof in this case: If $\{(u^k, x^k)\}_1^\infty$ is a minimizing sequence, some subsequence of $\{u^k\}$ converges weakly in L_2 ; $\{x^k\}$ is uniformly bounded and equicontinuous, so some subsequence converges uniformly. It follows from Theorem 2.1 that (\mathcal{P}) has solutions. \square

Remark 2.1. There is some redundancy in (2.1) and (2.12), because if

$$\sum_{i=1}^m u_i(t) = T^{-1} \sum_{i=1}^m a_i = \sigma \quad \text{and} \quad \int_0^T u_i(t) dt = a_i$$

for $1 \leq i \leq m-1$, then automatically

$$\int_0^T u_m(t) dt = \int_0^T \left[\sigma - \sum_{i=1}^{m-1} u_i(t) \right] dt = \sigma T - \sum_{i=1}^{m-1} a_i = a_m.$$

Therefore, when dealing with problem $(\hat{\mathcal{P}})$, we can eliminate one of the control variables, say $u_m = \sigma - \sum_{i=1}^{m-1} u_i$, and work with u_1, \dots, u_{m-1} and the constraints $u_i(t) \geq 0$, $\int_0^T u_i(t) dt = a_i$ for $1 \leq i \leq m-1$, and $\sum_{i=1}^{m-1} u_i(t) \leq \sigma$.

In the case $m = 1$, problem $(\hat{\mathcal{P}})$ is trivial, because then $u_1(t) = T^{-1}a_1$ for all t .

3. Solution of the by-pass problem when α and γ are strictly monotonic and β unimodal. Let us now consider the by-pass problem mentioned in § 1. It is described by the following equations:

$$\begin{aligned} \frac{dP}{dx} &= -\alpha(P)f(x) - \gamma(P)k(x), \\ (3.1) \quad \frac{dM}{dx} &= \alpha(P)f(x) - \beta(M)g(x) \quad (0 \leq x \leq L), \\ P(0) &= P_0 > 0, \quad M(0) = 0. \end{aligned}$$

Here α, γ and β are defined and continuously differentiable for $P \geq 0$ and $M \geq 0$, respectively, and they also satisfy

$$\begin{aligned} \alpha(P) &> 0, \quad \gamma(P) > 0 \quad \text{for } P > 0, \quad \beta(M) > 0 \quad \text{for } M > 0, \\ \alpha(0) &= \gamma(0) = \beta(0) = 0. \end{aligned}$$

The control functions f, k and g are measurable and satisfy

$$\begin{aligned} (3.2) \quad f(x) &\geq 0, \quad k(x) \geq 0, \quad g(x) \geq 0 \quad \text{for } 0 \leq x \leq L, \\ \int_0^L f(x) dx &= \int_0^L k(x) dx = \int_0^L g(x) dx = 1. \end{aligned}$$

Let f, k and g satisfy (3.2) and consider system (3.1). To be able to apply the usual existence theorem from the theory of ordinary differential equations, we would like to have β defined in a neighbourhood of 0. We therefore define $\beta(M) = -\beta(-M)$ for $M < 0$, so that $\beta \in C^1(\mathbb{R})$. Then we know that (3.1) has a unique solution in some interval $[0, \varepsilon]$ with $\varepsilon > 0$. As long as the solution exists, we find easily from (3.1) that

$$(3.3) \quad 0 < P(x) \leq P_0, \quad 0 \leq M(x) \leq \max_{0 \leq P \leq P_0} \alpha(P).$$

Therefore the solution can be continued to all of $[0, L]$, and (3.3) holds for all $x \in [0, L]$. The solution is in Carathéodory's sense, i.e., P and M are absolutely continuous and satisfy (3.1) almost everywhere. It follows from (3.3) that the solutions for various admissible f, k and g are uniformly bounded.

The optimization problem is, as in [1], [3] and [4], to minimize $M(L)$. We can now apply Theorem 2.2 and draw the conclusion that there exist optimal solutions.

As Theorem 2.1 shows, it is sufficient to solve the problem under the additional constraint

$$(3.4) \quad f(x) + g(x) + k(x) = 3/L = \sigma, \quad 0 \leq x \leq L.$$

By Remark 2.1 we shall eliminate g and work with f and k as control variables, though we will continue to write g instead of $\sigma - f - k$.

We shall apply Pontryagin's maximum principle to obtain necessary conditions for the restricted problem (that is, including (3.4)). The Hamiltonian for this problem is

$$H(P, M, \eta_P, \eta_M, f, k, \mu_1, \mu_2) \\ = \eta_P(-\alpha(P)f - \gamma(P)k) + \eta_M(\alpha(P)f - \beta(M)g) + \mu_1 f + \mu_2 k.$$

Let (f, k) denote an optimal control vector and (P, M) the corresponding solution of (3.1). From optimal control theory (see, for example, [2, pp. 185–191]) we know that there are absolutely continuous functions η_P and η_M , and constants η_0 , μ_1 and μ_2 such that $(\eta_P(x), \eta_M(x), \mu_1, \mu_2) \neq (0, 0, 0, 0)$ for all x , and

$$\eta'_P = -\frac{\partial H}{\partial P}, \quad \eta'_M = -\frac{\partial H}{\partial M},$$

$$\eta_P(L) = 0, \quad \eta_M(L) = \eta_0, \quad \eta_0 \leq 0$$

where the derivatives of H are to be taken at the point

$$(P(x), M(x), \eta_P(x), \eta_M(x), f(x), k(x), \mu_1, \mu_2),$$

and

$$(3.5) \quad H(P(x), M(x), \eta_P(x), \eta_M(x), f(x), k(x), \mu_1, \mu_2) \\ = \max_{\substack{f, k \geq 0 \\ f+k \leq \sigma}} H(P(x), M(x), \eta_P(x), \eta_M(x), f, k, \mu_1, \mu_2)$$

a.e. on $[0, L]$.

Thus

$$(3.6) \quad \eta'_P = (\eta_P - \eta_M)f\alpha'(P) + \eta_P k\gamma'(P), \quad \eta'_M = \eta_M g\beta'(M).$$

Write H as

$$(3.7) \quad H = s_1 f + s_2 k - \sigma \eta_M \beta(M)$$

where

$$(3.8) \quad s_1 = (\eta_M - \eta_P)\alpha(P) + \eta_M \beta(M) + \mu_1, \\ s_2 = \eta_M \beta(M) - \eta_P \gamma(P) + \mu_2.$$

In evaluating (3.5), it is convenient to introduce the following regions in the s -plane $[s = (s_1, s_2)]$:

$$R_1 = \{s: s_1 > 0, s_1 > s_2\}, \\ R_2 = \{s: s_2 > 0, s_2 > s_1\}, \\ R_3 = \{s: s_1 < 0, s_2 < 0\}, \\ L_1 = \{s: s_2 = 0, s_1 < 0\}, \\ L_2 = \{s: s_1 = 0, s_2 < 0\}, \\ L_3 = \{s: s_1 = s_2 > 0\}, \\ 0 = \{0 = (0, 0)\}.$$

It then follows from (3.5) and (3.7) that

$$(3.9) \quad \begin{aligned} f &= \sigma & \text{if } s \in R_1, & & f &= 0 & \text{if } s \in L_1, \\ k &= \sigma & \text{if } s \in R_2, & & k &= 0 & \text{if } s \in L_2, \\ g &= \sigma & \text{if } s \in R_3, & & g &= 0 & \text{if } s \in L_3, \end{aligned}$$

possibly after a redefinition of f , g and k on a null set (see Fig. 1).

First of all, let us show that $\eta_0 \neq 0$. Assume that $\eta_0 = 0$. Then $\eta_M(L) = 0$, and (3.6) implies first that $\eta_M(x) \equiv 0$ and then (because $\eta_P(L) = 0$) that $\eta_P(x) \equiv 0$. Then (3.8) shows that s_1 and s_2 are constants, and unless $\mu_1 = \mu_2 = 0$, (3.9) shows that one of f , g and k is identically σ or identically 0, in contradiction of (3.2). Thus $\mu_1 = \mu_2 = 0$, so that $(\eta_P(x), \eta_M(x), \mu_1, \mu_2) = (0, 0, 0, 0)$ for all x , which is impossible. Thus $\eta_M(L) = \eta_0 < 0$, and we see from (3.6) that

$$(3.10) \quad \eta_M(x) < 0 \quad \text{for all } x.$$

Of particular importance for the analysis are the derivatives of the switching functions s_1 and s_2 . From (3.8), (3.1) and (3.6) we obtain

$$\begin{aligned} s'_1 &= (\eta_M - \eta_P)\alpha'(P)P' + (\eta'_M - \eta'_P)\alpha(P) + \eta_M\beta'(M)M' + \eta'_M\beta(M) \\ &= (\eta_M - \eta_P)\alpha'(P)[- \alpha(P)f - \gamma(P)k] \\ &\quad + [\eta_M g\beta'(M) - (\eta_P - \eta_M)f\alpha'(P) - \eta_P k\gamma'(P)]\alpha(P) \\ &\quad + \eta_M\beta'(M)[\alpha(P)f - \beta(M)g] + \eta_M g\beta'(M)\beta(M), \\ s'_2 &= \eta_M\beta'(M)M' + \eta'_M\beta(M) - \eta_P\gamma'(P)P' - \eta'_P\gamma(P) \\ &= \eta_M\beta'(M)[\alpha(P)f - \beta(M)g] + \eta_M g\beta'(M)\beta(M) \\ &\quad - \eta_P\gamma'(P)[- \alpha(P)f - \gamma(P)k] - [(\eta_P - \eta_M)f\alpha'(P) + \eta_P k\gamma'(P)]\gamma(P). \end{aligned}$$

That is,

$$(3.11) \quad s'_1 = (f + g)\eta_M\alpha(P)\beta'(M) - kS, \quad s'_2 = f[\eta_M\alpha(P)\beta'(M) + S]$$

where

$$S = (\eta_M - \eta_P)\alpha'(P)\gamma(P) + \eta_P\alpha(P)\gamma'(P).$$

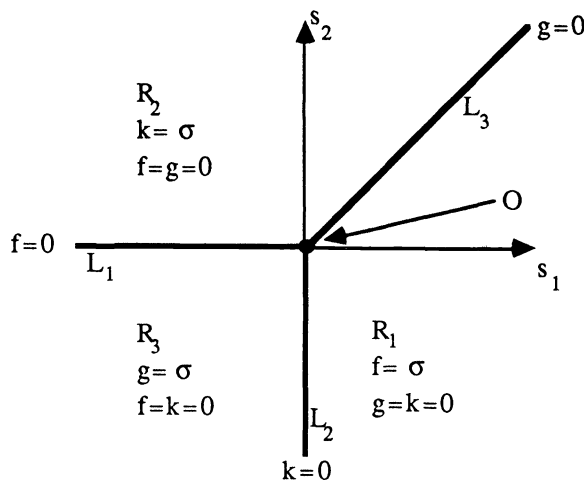


FIG. 1. Decomposition of the switching plane.

So far α, β and γ have been arbitrary (subject to our general conditions), but for the rest of this section we now assume that

$$\alpha'(P) > 0 \quad \text{and} \quad \gamma'(P) > 0 \quad \text{for all } P \geq 0,$$

and that

$$(3.12) \qquad \begin{aligned} \beta'(M) &> 0 \quad \text{for } 0 \leq M < M_0, \\ \beta'(M) &< 0 \quad \text{for } M > M_0. \end{aligned}$$

It is helpful to draw figures showing how the point $s(x)$ can move in the s -plane in different cases. If $S < 0$ and $M < M_0$ it follows from (3.9) and (3.11) that the following movements are possible (see Fig. 2. The arrows denote increasing x): In the region R_1 we have $f = \sigma$, $g = k = 0$, so that $M' > 0$, and we see from (3.11) that $s'_2 < s'_1 < 0$.

If $S < 0$ and $M > M_0$ the following movements are possible (see Fig. 3):

Note that we cannot move along L_3 in any case, because if $s_1 = s_2$ on a set E of positive measure, then $s'_1 = s'_2$ a.e. on E , but $g = 0$ implies $s'_1 - s'_2 = -\sigma S > 0$.

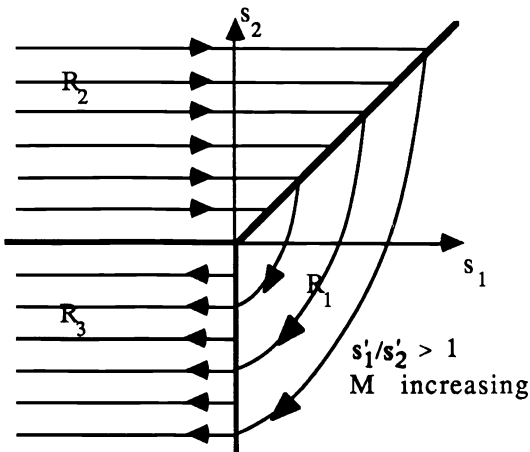


FIG. 2. Paths of (s_1, s_2) for $M < M_0$.

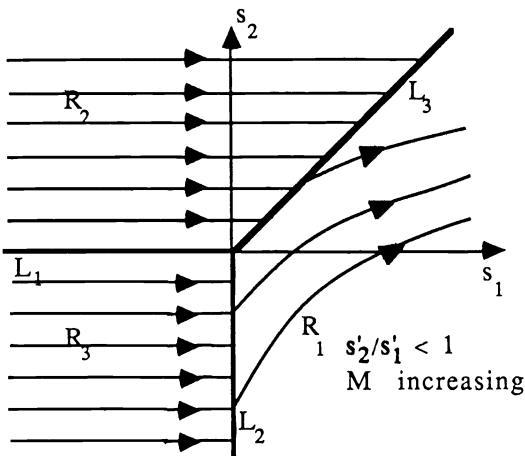


FIG. 3. Paths of (s_1, s_2) for $M > M_0$.

In what follows we will prove that $S(x) < 0$ for all x . We shall distinguish different cases depending on where $s(L)$ lies.

Case 1. $s(L) \in T = R_2 \cup R_3 \cup L_1 \cup 0$.

(i) Define $\xi_1 = \inf \{x \in [0, L]: s(t) \in T \text{ for } x \leq t \leq L\}$. Let us show that $S < 0$ on $[\xi_1, L]$. If this is not true, then, since $S(L) < 0$, there exists a $\xi'_1 \in [\xi_1, L]$ such that $S < 0$ on $(\xi'_1, L]$ and $S(\xi'_1) = 0$. In $R_2 \cup R_3 \cup L_1$ we have $f = 0$. If $s(x) \in 0$ and $f(x) \neq 0$ on a set $E_1 \subseteq (\xi'_1, L]$ of positive measure, then $s'_1 = s'_2 = 0$ a.e. on E_1 . It follows that $S = -\eta_M \alpha(P) \beta'(M)$ and $s'_1 = -(f + g + k)S = -\sigma S > 0$ a.e. on E_1 , which is a contradiction. Thus $f = 0$ a.e. on $[\xi'_1, L]$, and since $\eta_P(L) = 0$, it follows from (3.6) that $\eta_P = 0$ on $[\xi'_1, L]$; hence $S < 0$ there, which contradicts the definition of ξ'_1 . Thus $S < 0$ on $[\xi_1, L]$, and the arguments above show that $f = 0$ and $\eta_P = 0$ on $[\xi_1, L]$. From this it follows that $\xi_1 > 0$. We infer from Figs. 2 and 3 that $s(\xi_1)$ cannot belong to L_3 , because $s(x)$ cannot enter R_2 from there. It cannot belong to 0 either, because $S < 0$ in some neighbourhood to the left of ξ_1 , and, as the figures show, there is no curve leading to 0 that does not lie entirely in T (note that $s'_2 < s'_1 < 0$ in R_1 , if $M < M_0$). Thus $s(\xi_1) \in L_2$, and $s(x) \in R_3$; hence $g(x) = \sigma$, for $x \in (\xi_1, L]$. It also follows that $M(x) < M_0$ for $x \in (\xi_1, L]$, because if $M(\xi'') \geq M_0$ for some $\xi'' \in (\xi_1, L]$, then, since $M' \leq 0$, we get $M \geq M_0$ and $s'_1 \geq 0$ on (ξ_1, ξ'') which is impossible.

(ii) There is a maximal interval $[\xi_2, \xi_1]$ (ξ_2 may be equal to ξ_1) where $s_1 = 0$, $s_2 \leq 0$. If $\xi_2 < \xi_1$, then $k = 0$ and $s'_1 = \sigma \eta_M \alpha(P) \beta'(M) = 0$ on (ξ_2, ξ_1) , so that $\beta'(M) = 0$ and $M = M_0$ there. On (ξ_2, ξ_1) we have from (3.6)

$$\eta'_P = (\eta_P - \eta_M) f \alpha'(P), \quad \eta'_M = 0.$$

As $\eta_P(\xi_1) = 0$ and $\eta_M(\xi_1) < 0$, we see that $\eta_M < \eta_P < 0$ on $[\xi_2, \xi_1)$, so that $S < 0$ and $s'_2 = fS < 0$ there. Assume that $s_2(\xi_2) = 0$, and let

$$\xi'_2 = \inf \{x \geq 0: S < 0 \text{ on } [x, \xi_2]\}.$$

Figures 2 and 3 show that $s_1 \leq 0$, $s_2 = 0$ on $[\xi'_2, \xi_2]$, and, as is shown above, we must have $f = 0$ a.e. there. Then $M' \leq 0$, so that $M \geq M_0$ on $[\xi'_2, \xi_2]$. From (3.6) we obtain

$$\eta'_P = \eta_P k \gamma'(P), \quad \eta'_M = \eta_M g \beta'(M),$$

and since $\eta_P(\xi_2) < 0$, we get $\eta_P < 0$. Moreover, $\eta'_M - \eta'_P \geq 0$, and since $\eta_M(\xi_2) - \eta_P(\xi_2) < 0$, we get $\eta_M - \eta_P < 0$ on $[\xi'_2, \xi_2]$. Thus $S < 0$ on $[\xi'_2, \xi_2]$, which implies that $\xi'_2 = 0$. But it is impossible to have $M \geq M_0$ on all of $[0, \xi_2]$. Therefore $s_2(\xi_2) < 0$.

(iii) The point $s(x)$ might approach $s(\xi_2) \in L_2$ from R_3 , but with ξ'_2 defined as above, we again find that $\xi'_2 = 0$, which leads to $M \geq M_0$ on $[0, \xi_2]$. Thus $s(x)$ approaches $s(\xi_2)$ from R_1 . Let

$$\xi_3 = \inf \{x \geq 0: s(t) \in R_1 \text{ for } x < t < \xi_2\}.$$

Then $\xi_3 > 0$, since otherwise $\int_0^L k(x) dx = 0$. On $[\xi_3, \xi_2]$ we have

$$\eta'_P = \sigma(\eta_P - \eta_M) \alpha'(P), \quad \eta_P(\xi_2) \leq 0,$$

$$\eta'_M = 0, \quad \eta_M(\xi_2) < \eta_P(\xi_2).$$

Therefore $\eta_M < \eta_P < 0$ on $[\xi_3, \xi_2)$, and $S < 0$ there. Furthermore, $M < M_0$ on $[\xi_3, \xi_2)$, because $M' > 0$ there, and $M(\xi_2) \leq M_0$. By Fig. 2 $s(\xi_3) \in L_3$. Since $s(x)$ cannot move along L_3 , it must approach $s(\xi_3)$ from R_2 . Let

$$\xi'_3 = \inf \{x \geq 0: S < 0 \text{ on } [x, \xi_3]\}.$$

On $[\xi'_3, \xi_3]$ we see from Fig. 2 that $s(x) \in R_2$, and we have

$$\begin{aligned}\eta'_P &= \sigma \eta_P \gamma'(P), & \eta_P(\xi_3) &< 0, \\ \eta'_M &= 0, & \eta_M(\xi_3) &< \eta_P(\xi_3),\end{aligned}$$

so that $\eta_M < \eta_P < 0$ on $[\xi'_3, \xi_3]$. Thus $S < 0$ there, which means that $\xi'_3 = 0$. Hence $k = \sigma$ on $[0, \xi_3]$.

The whole curve in Case 1 is shown in Fig. 4.

Case 2. $s(L) \in L_3$.

The situation is the same as in step (iii) of Case 1 (except that we have $\eta_P = 0$ instead of $\eta_P < 0$, but that does not affect the argument), and we can draw the conclusion that $s(x) \in R_2$ for $x \in [0, L]$, that is $k = \sigma$ there, which is impossible.

Case 3. $s(L) \in L_2$.

If $M(L) > M_0$, $s(x)$ must approach $s(L)$ from R_3 , but as in Case 1, step (ii), this implies that $M(x) > M_0$ for all $x \in [0, L]$, which is impossible. If $M(L) \leq M_0$, the analysis is the same as in Case 1, steps (ii) and (iii). But if $M(L) < M_0$, then $g = 0$ on $[0, L]$. Thus $M(L) = M_0$.

Case 4. $s(L) \in R_1$.

Define this time

$$\xi_1 = \inf \{x \in [0, L]: s(t) \in R_1 \text{ for } x \leq t \leq L\}.$$

On $(\xi_1, L]$ we have $f = \sigma$, $g = k = 0$, so that $\xi_1 > 0$. From (3.6) we obtain on $[\xi_1, L]$

$$\eta'_P = \sigma(\eta_P - \eta_M)\alpha'(P), \quad \eta'_M = 0,$$

and since $\eta_P(L) = 0$, we get $\eta_M - \eta_P < 0$, $\eta_P < 0$; hence $S < 0$ on $[\xi_1, L]$. If $s(\xi_1) \in L_3$, we have the same situation as in step (iii) of Case 1, and we can conclude that $k = \sigma$ on $[0, \xi_1]$. But then $g = 0$ on $[0, L]$, which is impossible. Furthermore $M > M_0$ on $(\xi_1, L]$, because if $M(\xi') \leq M_0$ for some $\xi' \in (\xi_1, L]$, then, since $M' > 0$ on (ξ_1, ξ') , $M < M_0$ on $[\xi_1, \xi')$. But then $s'_1 < 0$ there, which is impossible, because $s_1(\xi_1) = 0 < s_1(\xi')$. (This is also clear from Figs. 2 and 3.) The possibility $s(\xi_1) = (0, 0)$ is ruled out by the same arguments as in Case 1, step (ii), when we showed that $s_2(\xi_2) < 0$. Thus $s(\xi_1) \in L_2$. If $M(\xi_1) > M_0$, then as in Case 1, step (ii), we find that $M(x) > M_0$ for all $x \in [0, \xi_1]$. Thus $M(\xi_1) = M_0$. From this point on the analysis is the same as in Case 1; this time $\xi_2 < \xi_1$, because it is only on $[\xi_2, \xi_1]$ that $g \neq 0$.

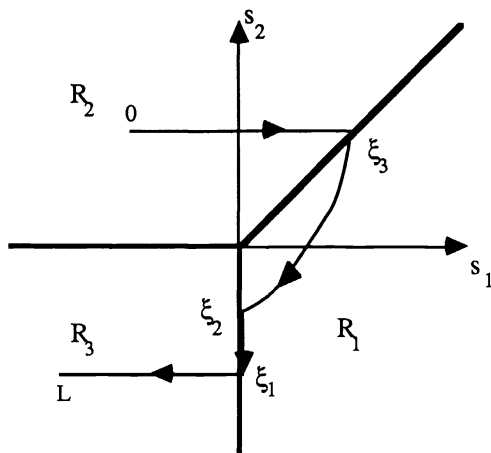


FIG. 4. Portrait of behaviour of (s_1, s_2) for $M(L) < M_0$.

The whole curve in Case 4 is shown in Fig. 5.

Now we know what possible forms the optimal solution can take. The results can be summarized as in Fig. 6.

Next we want to investigate what is the particular form of the optimal solution for different values of the parameters M_0 , $\beta_0 = \beta(M_0)$, P_0 , P_1 and P_2 , where P_1 and P_2 are defined by

$$\int_{P_1}^{P_0} \frac{ds}{\gamma(s)} = 1, \quad \int_{P_2}^{P_1} \frac{ds}{\alpha(s)} = 1.$$

In all cases $k(x) = \sigma$ for $0 \leq x < \xi_3$ and $k(x) = 0$ for $\xi_3 < x \leq L$. From (3.2) and (3.4) we obtain $\xi_3 = 1/\sigma = L/3$. On the interval $[0, L/3)$ we have $P' = -\sigma\gamma(P)$, $M' = 0$, so that

$$\int_{P(L/3)}^{P_0} \frac{ds}{\gamma(s)} = 1, \quad P\left(\frac{L}{3}\right) = P_1, \quad M\left(\frac{L}{3}\right) = 0.$$

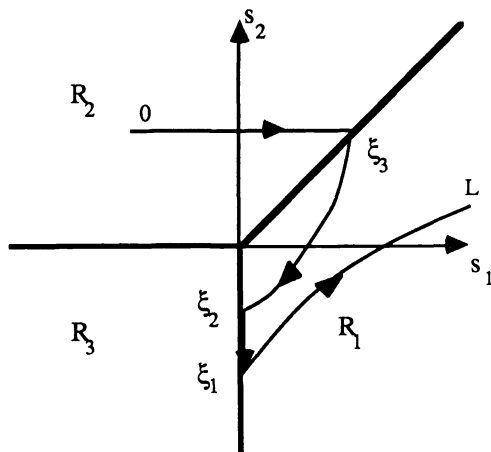


FIG. 5. Portrait of behaviour of (s_1, s_2) for $M(L) > M_0$.

I.	$k = \sigma$	$f = \sigma$	$g = \sigma$
	0	ξ_3	$\xi_2 = \xi_1$ L
II.	$k = \sigma$	$f = \sigma$	$k=0, M=M_0$ $g = \sigma$
	0	ξ_3 ξ_2	ξ_1 L
IIa.	$k = \sigma$	$f = \sigma$	$k=0, M=M_0$
	0	ξ_3 ξ_2	L
III.	$k = \sigma$	$f = \sigma$	$k=0, M=M_0$ $f = \sigma$
	0	ξ_3 ξ_2	ξ_1 L

FIG. 6. The optimal control functions.

Case I. Consider Case I in Fig. 6. From (3.2) and (3.4) we obtain $\xi_2 = \xi_1 = 2L/3$. On the interval $(L/3, 2L/3)$ we have $P' = -\sigma\alpha(P)$, $M' = \sigma\alpha(P)$, so that

$$\int_{P(2L/3)}^{P_1} \frac{ds}{\alpha(s)} = 1, \quad P\left(\frac{2L}{3}\right) = P_2, \quad M\left(\frac{2L}{3}\right) = P_1 - P_2.$$

On the interval $(2L/3, L]$ we have $P' = 0$, $M' = -\sigma\beta(M)$, so that

$$\int_{M(L)}^{P_1 - P_2} \frac{ds}{\beta(s)} = 1.$$

This can occur only if $P_1 - P_2 \leq M_0$ (since $M \leq M_0$ in Fig. 4).

Case II. Consider Case II in Fig. 6. On the interval $(L/3, \xi_2)$ we have $P' = -\sigma\alpha(P)$, $M' = \sigma\alpha(P)$, so that

$$(3.13) \quad \frac{L}{3} \int_{P(\xi_2)}^{P_1} \frac{ds}{\alpha(s)} = \xi_2 - L/3, \quad M(\xi_2) = M_0 = P_1 - P(\xi_2).$$

From this we get

$$(3.14) \quad \xi_2 = \frac{L}{3} \left(1 + \int_{P_1 - M_0}^{P_1} \frac{ds}{\alpha(s)} \right).$$

On the interval (ξ_2, ξ_1) we have $P' = -f(x)\alpha(P)$, and $0 = f(x)\alpha(P) - \beta_0 g(x)$. We obtain

$$(3.15) \quad \begin{aligned} 1 &= \int_{L/3}^{\xi_1} f(x) dx = \int_{P(\xi_1)}^{P_1} \frac{ds}{\alpha(s)}, \quad P(\xi_1) = P_2, \\ \beta_0 \int_{\xi_2}^{\xi_1} g(x) dx &= P(\xi_2) - P(\xi_1) = P_1 - P_2 - M_0. \end{aligned}$$

On the interval $(\xi_1, L]$ we have $P' = 0$, $M' = -\sigma\beta(M)$, so that

$$\begin{aligned} \frac{L}{3} \int_{M(L)}^{M_0} \frac{ds}{\beta(s)} &= L - \xi_1, \\ 1 &= \int_{\xi_2}^L g(x) dx = 3 - \sigma\xi_1 + (P_1 - P_2 - M_0)/\beta_0. \end{aligned}$$

From this we get

$$\begin{aligned} \xi_1 &= \frac{L}{3} [2 + (P_1 - P_2 - M_0)/\beta_0], \\ \int_{M(L)}^{M_0} \frac{ds}{\beta(s)} &= 1 - (P_1 - P_2 - M_0)/\beta_0. \end{aligned}$$

This case can occur only if $0 < P_1 - P_2 - M_0 < \beta_0$.

Case IIa. The limiting case where $\xi_1 = L$ can occur only if $P_1 - P_2 - M_0 = \beta_0$.

Case III. Finally, consider Case III in Fig. 6. As in Case II we obtain (3.13) and (3.14), and instead of (3.15) we get

$$\beta_0 = \beta_0 \int_{\xi_2}^{\xi_1} g(x) dx = P(\xi_2) - P(\xi_1) = P_1 - M_0 - P(\xi_1).$$

On the interval $(\xi_1, L]$ we have $P' = -\sigma\alpha(P)$, $M' = \sigma\alpha(P)$, so that

$$\begin{aligned} \int_{P(L)}^{P_1} \frac{ds}{\alpha(s)} &= 1, \quad P(L) = P_2, \\ \int_{P_2}^{P_1-M_0-\beta_0} \frac{ds}{\alpha(s)} &= 3 - \sigma\xi_1, \quad \xi_1 = L - \frac{L}{3} \int_{P_2}^{P_1-M_0-\beta_0} \frac{ds}{\alpha(s)} = \frac{L}{3} \left[2 + \int_{P_1-M_0-\beta_0}^{P_1} \frac{ds}{\alpha(s)} \right], \\ M(L) &= M_0 + P(\xi_1) - P(L) = P_1 - P_2 - \beta_0. \end{aligned}$$

This case can occur only if $P_1 - P_2 - M_0 > \beta_0$.

For any values of M_0 , β_0 , P_0 , P_1 and P_2 there is an optimal solution of (3.1.2) and (3.4), and it must be included in one of the cases above. The necessary conditions also determine (f, g, k, P, M) uniquely. This is obvious except on the interval (ξ_2, ξ_1) . But there $f + g = \sigma$, and we find that P and f are uniquely determined, P as the solution of

$$P' = -\frac{\sigma\beta_0\alpha(P)}{\beta_0 + \alpha(P)}, \quad P(\xi_2) = P_1 - M_0,$$

and $f(x)$ as

$$f(x) = \frac{\sigma\beta_0}{\beta_0 + \alpha(P(x))}.$$

When the optimal solution of (3.1.2) and (3.4) has been found, it can generate all the optimal solutions of (3.1)–(3.2) as described in Theorem 2.1.

We can summarize our results in the following theorem.

THEOREM 3.1. *Consider the problem of minimizing $M(L)$ subject to (3.1.2) and (3.4). The problem has a unique solution $(\hat{f}, \hat{g}, \hat{k}, \hat{P}, \hat{M})$. If we define P_1 and P_2 by*

$$\int_{P_1}^{P_0} \frac{ds}{\gamma(s)} = 1, \quad \int_{P_2}^{P_1} \frac{ds}{\alpha(s)} = 1,$$

the solution can be described as follows:

(I) *If $P_1 - P_2 \leq M_0$, then the optimal solution is given by Case I in Fig. 6 with $\xi_3 = L/3$, $\xi_2 = \xi_1 = 2L/3$. The optimal value $\hat{M}(L)$ is given by*

$$\int_{\hat{M}(L)}^{P_1-P_2} \frac{ds}{\beta(s)} = 1.$$

(II) *If $0 < P_1 - P_2 - M_0 < \beta_0$, then the optimal solution is given by Case II in Fig. 6 with $\xi_3 = L/3$ and*

$$\xi_2 = \frac{L}{3} \left[1 + \int_{P_1-M_0}^{P_1} \frac{ds}{\alpha(s)} \right], \quad \xi_1 = \frac{L}{3} [2 + (P_1 - P_2 - M_0)/\beta_0].$$

The optimal value $\hat{M}(L)$ is given by

$$\int_{\hat{M}(L)}^{M_0} \frac{ds}{\beta(s)} = 1 - (P_1 - P_2 - M_0)/\beta_0.$$

(IIa) *If $P_1 - P_2 - M_0 = \beta_0$, then the optimal solution is given by Case IIa in Fig. 6 with $\xi_3 = L/3$, ξ_2 as in Case II, and $\xi_1 = L$; furthermore $\hat{M}(L) = M_0$.*

(III) If $P_1 - P_2 - M_0 > \beta_0$, then the optimal solution is given by Case III in Fig. 6 with $\xi_3 = L/3$, ξ_2 as in Case II, and

$$\xi_1 = \frac{L}{3} \left[2 + \int_{P_1 - M_0 - \beta_0}^{P_1} \frac{ds}{\alpha(s)} \right].$$

The optimal value is $\hat{M}(L) = P_1 - P_2 - \beta_0$.

The optimal solutions of (3.1)–(3.2) are

$$\begin{aligned} f(x) &= \theta(x) \hat{f}(\phi(x)), & P(x) &= \hat{P}(\phi(x)), \\ g(x) &= \theta(x) \hat{g}(\phi(x)), & M(x) &= \hat{M}(\phi(x)), \\ k(x) &= \theta(x) \hat{k}(\phi(x)), \end{aligned}$$

where θ is any function in $L_1(0, L)$ such that $\theta \geq 0$,

$$\int_0^L \theta(x) dx = L \quad \text{and} \quad \phi(x) = \int_0^x \theta(t) dt.$$

The optimal value of $M(L)$ is the same as $\hat{M}(L)$.

4. General α , γ , monotonic β . We consider again the by-pass problem but this time with α and γ general and β strictly monotonic (that is, $\beta'(M) > 0$ for all $M \geq 0$). We can still use equations (3.1)–(3.11).

Outside \bar{R}_3 (the closure of R_3) $g = 0$, so there exists a smallest $\xi_1 < L$ such that $s(\xi_1)$ belongs to \bar{R}_3 . Within $R_2 \cup R_3 \cup L_1$ we have $f = 0$, so that $s'_2 = 0$, and within $R_1 \cup R_3 \cup L_2$ we have $k = 0$, so that $s'_1 = \sigma \eta_M \alpha(P) \beta'(M) < 0$. Therefore the point $s(x)$ might leave \bar{R}_3 only at 0 along L_3 . But if $s_1(x) = s_2(x) > 0$ on a set E of positive measure, then $s'_1 = s'_2$ a.e. on E , hence $S = 0$ (since $g = 0$; see (3.11)), and $s'_1 \leq 0$ there. Therefore $s(x)$ cannot leave \bar{R}_3 . When s belongs to \bar{R}_3 , we have $f = 0$ a.e., because if $s_1(x) = s_2(x) = 0$ and $f(x) \neq 0$ on a set of positive measure, then $s'_1(x) = s'_2(x) = 0$ a.e. on that set, and (3.11) gives the contradiction $\eta_M \alpha(P) \beta'(M) = 0$. This means that $\xi_1 > 0$, and that $f = 0$ on $(\xi_1, L]$.

Now, either $s(\xi_1) \in L_2$ or $s(\xi_1) \in 0$. In the first case it follows from what was said above that $s(x)$ will move into R_3 , so that $g = \sigma$ on $(\xi_1, L]$. In the second case $s(x)$ belongs to $L_1 \cup 0$ for all $x \in [\xi_1, L]$. On $(\xi_1, L]$ we have $g + k = \sigma$ a.e. and $\int_{\xi_1}^L g(x) dx = 1$, and since

$$\int_{M(L)}^{M(\xi_1)} \frac{ds}{\beta(s)} = \int_{\xi_1}^L g(x) dx,$$

any g and k satisfying these conditions on $(\xi_1, L]$ are optimal. In particular there is an optimal solution with $g = 0$ on $[\xi_1, \xi']$ and $g = \sigma$ on $(\xi', L]$ for some ξ' , and we consider that solution for the rest of this section. Thus in any case there is a ξ such that $g = 0$ on $[0, \xi]$ and $g = \sigma$ on $(\xi, L]$. It follows that $\xi = 2L/3$ and that $P = P(L)$ and $\eta_P = 0$ on $[\xi, L]$, and $\eta_M = \eta_M(\xi) < 0$ on $[0, \xi]$.

Let us return to the necessary conditions derived in § 3. Since our system is autonomous, the Hamiltonian (3.7) is equal to a constant σc on $[0, L]$. On $[0, \xi]$ we get

$$\max(\sigma s_1 - \sigma \eta_M \beta(M), \sigma s_2 - \sigma \eta_M \beta(M)) = \sigma c,$$

that is,

$$(4.1) \quad \max((\eta_M - \eta_P)\alpha(P) + \mu_1, -\eta_P\gamma(P) + \mu_2) = c,$$

and at $x = \xi = 2L/3$ in particular:

$$\max(\eta_M(\xi)\alpha(P(\xi)) + \mu_1, \mu_2) = c.$$

Case I. In the first case described (see Fig. 7) $s_2(\xi) < s_1(\xi) = 0$, and therefore

$$\mu_2 < \eta_M(\xi)\alpha(P(\xi)) + \mu_1 = c,$$

so that

$$\eta_M(\xi) = \frac{c - \mu_1}{\alpha(P(\xi))} < 0.$$

In R_1 , $s_1 > s_2$, so that

$$-\eta_P\gamma(P) + \mu_2 < \frac{c - \mu_1}{\alpha(P(\xi))}\alpha(P) - \eta_P\alpha(P) + \mu_1 = c.$$

Then

$$\eta_P = (c - \mu_1) \left[\frac{1}{\alpha(P(\xi))} - \frac{1}{\alpha(P)} \right],$$

$$\gamma(P) \left[1 - \frac{\alpha(P(\xi))}{\alpha(P)} \right] < \frac{c - \mu_2}{\mu_1 - c} \alpha(P(\xi)) = \nu,$$

and $f = \sigma$, $k = 0$.

In R_2 , $s_2 > s_1$, and

$$\frac{c - \mu_1}{\alpha(P(\xi))} \alpha(P) - \eta_P\alpha(P) + \mu_1 < -\eta_P\gamma(P) + \mu_2 = c.$$

Then

$$\eta_P = \frac{\mu_2 - c}{\gamma(P)}, \quad \gamma(P) \left[1 - \frac{\alpha(P(\xi))}{\alpha(P)} \right] > \nu,$$

and $f = 0$, $k = \sigma$.

On the line $s_1 = s_2$ we have $\gamma(P)[1 - \alpha(P(\xi))/\alpha(P)] = \nu$.

Thus we see that the function $\psi(P) = \gamma(P)[1 - \alpha(P(\xi))/\alpha(P)]$ is such that

$$(4.2) \quad \begin{aligned} \psi(P) < \nu &\Rightarrow f = \sigma, & k = 0, \\ \psi(P) > \nu &\Rightarrow f = 0, & k = \sigma. \end{aligned}$$

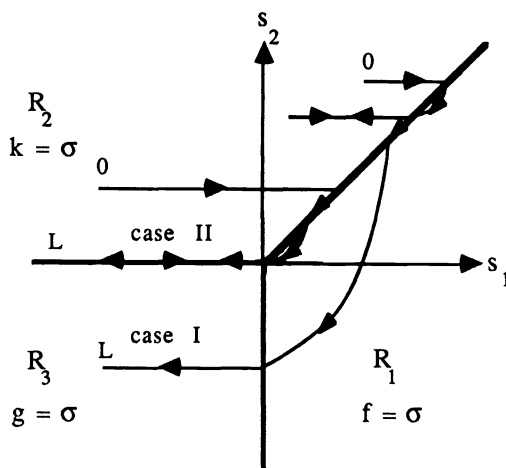


FIG. 7. Possible paths of (s_1, s_2) .

Case II. In the second case $s_1(\xi) \leq 0$, $s_2(\xi) = 0$, and

$$\eta_M(\xi)\alpha(P(\xi)) + \mu_1 \leq \mu_2 = c,$$

so that

$$\alpha(P(\xi)) \geq \frac{c - \mu_1}{\eta_M(\xi)} = \kappa.$$

In R_1 ,

$$-\eta_P\gamma(P) + c < \eta_M(\xi)\alpha(P) - \eta_P\alpha(P) + \mu_1 = c.$$

Then

$$\eta_P = \eta_M(\xi) - \frac{c - \mu_1}{\alpha(P)}, \quad \alpha(P) < \kappa,$$

and $f = \sigma$, $k = 0$.

In R_2 ,

$$\eta_M(\xi)\alpha(P) - \eta_P\alpha(P) + \mu_1 < -\eta_P\gamma(P) + c = c.$$

Then

$$\eta_P = 0, \quad \alpha(P) > \kappa,$$

and $f = 0$, $k = \sigma$.

On the line $s_1 = s_2$, $\alpha(P) = \kappa$. Consequently in this case

$$(4.3) \quad \begin{aligned} \alpha(P) < \kappa &\Rightarrow f = \sigma, & k &= 0, \\ \alpha(P) > \kappa &\Rightarrow f = 0, & k &= \sigma. \end{aligned}$$

Remark 4.1. For certain functions α and γ the condition (4.2) or (4.3) gives information on f and k almost everywhere on $[0, \xi]$, $\xi = 2L/3$. For example, if α and γ are strictly increasing in P , then so is ψ , and since $P' < 0$, (4.2) or (4.3) gives a solution of the form I in Fig. 6. Another example is the case where α and γ are (real) analytic in a neighbourhood of $[P_{\min}, P_0]$, where $P_{\min} > 0$ is the minimum of $P(L)$ over all admissible f and k . In that case there are only finitely many x in $[0, \xi]$ such that $\psi(P(x)) = \nu$ or $\alpha(P(x)) = \kappa$, unless $\alpha(P)$ is constant on $[P_{\min}, P_0]$. If this is so, then any distribution of f and k is optimal.

Remark 4.2. From the equation

$$\int_{M(L)}^{M(\xi)} \frac{ds}{\beta(s)} = 1 \quad \left(\xi = \frac{2L}{3} \right)$$

it follows that $M(\xi)$ is the minimum of

$$\int_0^\xi f(x)\alpha(P(x)) \, dx$$

subject to

$$\begin{aligned} \frac{dP}{dx} &= -\alpha(P)f(x) - \gamma(P)k(x), \\ f(x) &\geq 0, \quad k(x) \geq 0, \quad f(x) + k(x) = \sigma, \quad x \in [0, \xi], \\ \int_0^\xi f(x) \, dx &= \int_0^\xi k(x) \, dx = 1. \end{aligned}$$

5. General α , β and γ . Consider again the by-pass problem from § 3, this time in its original formulation (3.1)–(3.2). Let α , β and γ be arbitrary.

Choose an $L' > 0$ and consider the problem of minimizing

$$\int_0^{L'} f(x) \alpha(P(x)) dx,$$

when

$$P' = -\alpha(P)f(x) - \gamma(P)k(x), \quad 0 \leq x \leq L',$$

$$P(0) = P_0,$$

$$\int_0^{L'} f(x) dx = \int_0^{L'} k(x) dx = 1, \quad f(x), k(x) \geq 0.$$

According to Theorem 2.2 this problem has a solution. Let $(\bar{f}, \bar{k}, \bar{P})$ be any solution, and let λ be the minimum value, so that

$$\lambda = \int_0^{L'} \bar{f}(x) \alpha(\bar{P}(x)) dx.$$

Note that λ actually is independent of L' . If we pose the same problem with L instead of L' , then

$$\left(\frac{L'}{L} \bar{f}\left(\frac{L'x}{L}\right), \frac{L'}{L} \bar{k}\left(\frac{L'x}{L}\right), \bar{P}\left(\frac{L'x}{L}\right) \right)$$

is a solution as is easily verified.

If the restriction $f(x) + k(x) = \sigma$ is applied, Remark 4.2 shows that the solution can be characterized by (4.2) or (4.3).

Now let $L' \in (0, L)$ and extend the definition of \bar{f} , \bar{k} , and \bar{P} to all of $[0, L]$ by setting $\bar{f}(x) = \bar{k}(x) = 0$ and $\bar{P}(x) = \bar{P}(L')$ on $(L', L]$. Then \bar{P} satisfies $\bar{P}' = -\alpha(\bar{P})\bar{f}(x) - \gamma(\bar{P})\bar{k}(x)$ on $[0, L]$. Let $\mu > 0$ be given. We shall construct a particular solution $(\bar{f}, g_\mu, \bar{k}, \bar{P}, M_\mu)$ of (3.1)–(3.2) with g_μ defined as follows:

Case I. If $\lambda \leq \mu$, define $g_\mu(x) = 0$ for $0 \leq x \leq L'$, and $g_\mu(x) \geq 0$ for $L' < x \leq L$ in such a way that $\int_{L'}^L g_\mu(x) dx = 1$. Then $M_\mu(L') = \lambda$, and $M_\mu(L)$ is given by $\int_{M_\mu(L)}^\lambda ds / \beta(s) = 1$.

Case II. Assume that $\mu < \lambda < \mu + \beta(\mu)$. Consider

$$\phi(x) = \int_0^x \bar{f}(t) \alpha(\bar{P}(t)) dt, \quad 0 \leq x \leq L'.$$

Since $\phi(L') = \lambda > \mu$, there is a $y \in (0, L')$ such that $\phi(y) = \mu$. Define $g_\mu(x) = 0$ for $0 \leq x \leq y$, and

$$(5.1) \quad g_\mu(x) = \frac{1}{\beta(\mu)} \bar{f}(x) \alpha(\bar{P}(x))$$

for $y < x \leq L'$. This is permissible, because

$$\int_y^{L'} g_\mu(x) dx = [\phi(L') - \phi(y)] / \beta(\mu) = (\lambda - \mu) / \beta(\mu) < 1.$$

On $(L', L]$ we define $g_\mu(x)$ in such a way that $\int_{L'}^L g_\mu(x) dx = 1 - (\lambda - \mu)/\beta(\mu)$. By the uniqueness of the solution of the differential equation, we have $M_\mu(x) = \mu$ on $[y, L']$, and $M_\mu(L)$ is given by

$$\int_{M_\mu(L)}^\mu \frac{ds}{\beta(s)} = 1 - \frac{\lambda - \mu}{\beta(\mu)}.$$

Case III. Assume that $\lambda \geq \mu + \beta(\mu)$. Again $\phi(y) = \mu$ for some $y \in (0, L')$, and since $\phi(L') = \lambda \geq \mu + \beta(\mu)$, there is a $z \in (y, L']$ such that $\phi(z) = \mu + \beta(\mu)$. We now define $g_\mu(x)$ by expression (5.1) on (y, z) , and $g_\mu(x) = 0$ on $[0, y] \cup [z, L]$. We note that

$$\int_0^L g_\mu(x) dx = \int_y^z g_\mu(x) dx = [\phi(z) - \phi(y)]/\beta(\mu) = 1.$$

We have $M_\mu(x) = \mu$ on $[y, z]$, and

$$\begin{aligned} M_\mu(L) &= \mu + \int_z^{L'} \tilde{f}(x) \alpha(\bar{P}(x)) dx \\ &= \lambda - \int_y^z \tilde{f}(x) \alpha(\bar{P}(x)) dx \\ &= \lambda - \beta(\mu) \int_y^z g_\mu(x) dx = \lambda - \beta(\mu). \end{aligned}$$

Consider the function $F(\mu)$ defined by

$$\begin{aligned} F(\mu) &= \lambda - \beta(\mu) && \text{if } 0 \leq \mu + \beta(\mu) \leq \lambda, \\ \int_{F(\mu)}^\mu \frac{ds}{\beta(s)} &= 1 - \frac{\lambda - \mu}{\beta(\mu)} && \text{if } \mu + \beta(\mu) > \lambda \text{ and } \mu < \lambda, \\ \int_{F(\mu)}^\lambda \frac{ds}{\beta(s)} &= 1 && \text{if } \mu \geq \lambda. \end{aligned}$$

Obviously, $F(\mu) = M_\mu(L)$ for $\mu > 0$. This function is continuous and attains its minimum value for some $\mu^* > 0$. (There may be several such μ^* ; choose one of them.) Put $g^* = g_{\mu^*}$ and $M^* = M_{\mu^*}$.

THEOREM 5.1. *The solution $(\tilde{f}, g^*, \tilde{k}, \bar{P}, M^*)$ of (3.1)–(3.2) constructed above is an optimal solution.*

Proof. Let (f, g, k, P, M) be an arbitrary solution of (3.1)–(3.2). Define

$$\beta_1 = \max \{\beta(M(x)) : x \in [0, L]\},$$

and choose μ_1 and x_1 such that

$$\beta(\mu_1) = \beta_1, \quad M(x_1) = \mu_1.$$

If x_1 cannot be chosen less than L , define

$$\tilde{f}(x) = \begin{cases} 2f(2x) & \text{for } 0 \leq x \leq L/2, \\ 0 & \text{for } L/2 < x \leq L, \end{cases}$$

and $\tilde{g}(x), \tilde{k}(x)$ similarly, and

$$\begin{aligned} \tilde{P}(x) &= \begin{cases} P(2x) & \text{for } 0 \leq x \leq L/2, \\ P(L) & \text{for } L/2 < x \leq L, \end{cases} \\ \tilde{M}(x) &= \begin{cases} M(2x) & \text{for } 0 \leq x \leq L/2, \\ M(L) & \text{for } L/2 < x \leq L. \end{cases} \end{aligned}$$

Then $(\tilde{f}, \tilde{g}, \tilde{\kappa}, \tilde{P}, \tilde{M})$ is a solution of (3.1)–(3.2), and $\tilde{M}(L) = M(L)$. Therefore we may without loss of generality assume that $x_1 < L$. We then choose $L' = x_1$ and $\mu = \mu_1$ and construct $(\tilde{f}, g_{\mu_1}, \tilde{\kappa}, \tilde{P}, M_{\mu_1})$.

Case I. Assume that $\lambda \leq \mu_1$. Then

$$\int_{M_{\mu_1}(L)}^{\mu_1} \frac{ds}{\beta(s)} \geq \int_{M_{\mu_1}(L)}^{\lambda} \frac{ds}{\beta(s)} = 1 \geq \int_{L'}^L g(x) dx \geq \int_{M(L)}^{\mu_1} \frac{ds}{\beta(s)},$$

since $M' \geq -g\beta(M)$, $M(L') = M(x_1) = \mu_1$. Thus $M_{\mu_1}(L) \leq M(L)$.

Case II. Assume that $\mu_1 < \lambda < \mu_1 + \beta(\mu_1)$. By integrating $M'/\beta(M)$ over $[L', L]$ we get

$$\begin{aligned} \int_{M(L)}^{\mu_1} \frac{ds}{\beta(s)} &= \int_{M(L)}^{M(L')} \frac{ds}{\beta(s)} \\ &= \int_{L'}^L g(x) dx - \int_{L'}^L \frac{f(x)\alpha(P(x))}{\beta(M(x))} dx \\ &\leq \int_{L'}^L g(x) dx - \frac{1}{\beta_1} \int_{L'}^L f(x)\alpha(P(x)) dx \\ &= \int_{L'}^L g(x) dx - \frac{1}{\beta_1} \left[\int_0^L f(x)\alpha(P(x)) dx - \int_0^{L'} f(x)\alpha(P(x)) dx \right]. \end{aligned}$$

But $\int_0^L f(x)\alpha(P(x)) dx \geq \lambda$ according to the remark above (λ independent of L'), and

$$\int_0^{L'} f(x)\alpha(P(x)) dx = M(L') + \int_0^{L'} g(x)\beta(M(x)) dx \leq \mu_1 + \beta_1 \int_0^{L'} g(x) dx.$$

Thus

$$\begin{aligned} \int_{M(L)}^{\mu_1} \frac{ds}{\beta(s)} &\leq \int_{L'}^L g(x) dx - \frac{1}{\beta_1} \left[\lambda - \mu_1 - \beta_1 \int_0^{L'} g(x) dx \right] \\ &= 1 - \frac{\lambda - \mu_1}{\beta_1} = \int_{M_{\mu_1}(L)}^{\mu_1} \frac{ds}{\beta(s)}. \end{aligned}$$

Hence $M_{\mu_1}(L) \leq M(L)$.

Case III. Assume that $\lambda \geq \mu_1 + \beta(\mu_1)$. Then

$$\begin{aligned} M(L) &= \int_0^L f(x)\alpha(P(x)) dx - \int_0^L g(x)\beta(M(x)) dx \geq \lambda - \int_0^L g(x)\beta_1 dx \\ &= \lambda - \beta_1 = M_{\mu_1}(L). \end{aligned}$$

We have then proved that $M(L) \geq M_{\mu_1}(L) = F(\mu_1)$, and since by definition $F(\mu_1) \geq F(\mu^*) = M_{\mu^*}(L) = M^*(L)$, we see that $M^*(L)$ is the optimal value. \square

Remark 5.1. This gives a simpler proof of the result in [3].

Remark 5.2. It is easy to see that the function $F(\mu)$ defined above is continuously differentiable; indeed

$$F'(\mu) = \begin{cases} -\beta'(\mu) & \text{if } 0 \leq \mu + \beta(\mu) \leq \lambda, \\ -\frac{(\lambda - \mu)\beta'(\mu)\beta(F(\mu))}{[\beta(\mu)]^2} & \text{if } \mu + \beta(\mu) > \lambda \text{ and } \mu < \lambda, \\ 0 & \text{if } \mu \geq \lambda. \end{cases}$$

In particular $\beta'(\mu)F'(\mu) \leq 0$ for all μ , and we see that μ^* (where minimum is attained) is either λ or a point where $\beta' = 0$. In the latter case, if $\mu^* < \lambda$ and if $\{\mu: \beta'(\mu) = 0\}$ is a finite union of closed intervals (which may be single points), μ^* must be a local maximum point of β . We can also show (without any extra assumptions on β) that if $\mu^* \leq \lambda$, then

$$(5.2) \quad \beta(\mu^*) = \max \{\beta(\mu): 0 \leq \mu \leq \mu^*\}.$$

Proof of (5.2). Let $\mu_0 \in (0, \mu^*]$ be such that $\beta(\mu_0) = \max \{\beta(\mu): 0 \leq \mu \leq \mu^*\}$. Suppose that $\beta(\mu^*) < \beta(\mu_0)$, so that $\mu_0 < \mu^*$.

Case I. If $\mu^* + \beta(\mu^*) \leq \lambda$, then $F(\mu) = \lambda - \beta(\mu)$ for all $\mu \leq \mu^*$, because if $\mu' + \beta(\mu') > \lambda$ for some $\mu' < \mu^*$, then there is a $\mu'' < \mu'$ such that $\mu'' + \beta(\mu'') = \lambda$, and $F(\mu'') = \mu'' < \mu^* \leq \lambda - \beta(\mu^*) = F(\mu^*)$, which is impossible. Thus $\beta(\mu^*) < \beta(\mu_0)$ leads to the contradiction $F(\mu_0) < F(\mu^*)$.

Case II. If $\mu^* \leq \lambda < \mu^* + \beta(\mu^*)$, then

$$\int_{F(\mu^*)}^{\mu^*} \frac{ds}{\beta(s)} = 1 - \frac{\lambda - \mu^*}{\beta(\mu^*)}.$$

If $\mu_0 + \beta(\mu_0) \leq \lambda$, then $F(\mu_0) = \lambda - \beta(\mu_0)$, and

$$\frac{\mu^* - F(\mu^*)}{\beta(\mu_0)} < \int_{F(\mu^*)}^{\mu^*} \frac{ds}{\beta(s)} \leq 1 - \frac{\lambda - \mu^*}{\beta(\mu_0)};$$

hence $F(\mu_0) < F(\mu^*)$. If $\mu_0 + \beta(\mu_0) > \lambda$, then

$$\begin{aligned} \int_{F(\mu_0)}^{\mu_0} \frac{ds}{\beta(s)} &= 1 - \frac{\lambda - \mu_0}{\beta(\mu_0)} = 1 - \frac{\lambda - \mu^*}{\beta(\mu_0)} - \frac{\mu^* - \mu_0}{\beta(\mu_0)} \\ &\geq 1 - \frac{\lambda - \mu^*}{\beta(\mu^*)} - \frac{\mu^* - \mu_0}{\beta(\mu_0)} > \int_{F(\mu^*)}^{\mu^*} \frac{ds}{\beta(s)} - \int_{\mu_0}^{\mu^*} \frac{ds}{\beta(s)} \\ &= \int_{F(\mu^*)}^{\mu_0} \frac{ds}{\beta(s)}, \end{aligned}$$

hence again the contradiction $F(\mu_0) < F(\mu^*)$.

Thus we must have $\beta(\mu^*) = \beta(\mu_0)$. \square

Acknowledgments. We are grateful to L. Bass, A. J. Bracken and R. Vyborny for their encouragement and advice. We would also like to thank the referee for some helpful suggestions.

REFERENCES

- [1] L. BASS, A. J. BRACKEN AND R. VYBORNÝ, *Minimisation problems for implicit functionals defined by differential equations of liver kinetics*, J. Austral. Math. Soc. Ser. B, 25 (1984), pp. 538–562.
- [2] L. D. BERKOVITZ, *Optimal Control Theory*, Springer-Verlag, New York-Heidelberg-Berlin, 1974.
- [3] A. M. FINK, *Optimal control in liver kinetics*, J. Austral. Math. Soc. Ser. B, 27 (1986), pp. 361–369.
- [4] K. HOLMÅKER, *An optimal control problem in the study of liver kinetics*, J. Optim. Theory Appl., 48 (1986), pp. 289–302.
- [5] E. B. LEE AND L. MARKUS, *Foundations of Optimal Control Theory*, John Wiley, New York-London-Sydney, 1967.
- [6] I. P. NATANSON, *Theory of Functions of a Real Variable*, Frederick Ungar, New York, 1961.
- [7] E. C. TITCHMARSH, *The Theory of Functions*, Clarendon Press, Oxford, 1932.

SELF-TUNING TRACKERS*

P. R. KUMAR† AND L. PRALY‡

Abstract. We examine the problem of obtaining adaptive control laws which tune themselves to control laws minimizing the variance of the tracking error between the output of the linear ARMAX system and a specified reference trajectory. If the reference trajectory is sufficiently rich of order greater than or equal to the sum of the degrees of the control and noise polynomials in the ARMAX system, then an adaptive controller is exhibited for which the parameter estimates are strongly consistent. For the linear model following problem where the trajectory to be tracked is generated as the output of a linear system, it is enough for the order of sufficient richness to be greater than the degree of the noise polynomial alone. Further, if the order of sufficient richness is even smaller, as is often the case, then a lower dimensional adaptive controller which does not attempt to estimate all the coefficients of the noise polynomial is self-tuning.

Key words. adaptive control, self-tuning tracker

AMS(MOS) subject classifications. 93C40, 93E12, 93E20, 93C55

1. Introduction. The problem of stochastic adaptive control of linear ARMAX systems has received considerable attention over the past decade. The notable pioneering contributions are due to Åström and Wittenmark [1] and Ljung [2], [3]. Subsequently, Goodwin, Ramadge and Caines [4] and Goodwin and Sin [5] have proved the *self-optimality* of some adaptive control algorithms for minimum variance regulation and tracking. By *self-optimality* it is meant that the cost, the time average of the square of the tracking error, is minimal.

Recently a stochastic gradient algorithm has been proved to be *self-tuning* for the *regulation* problem (see [6]). (Recall that in the regulation problem one wants the output of the system to stay as close as possible to zero, whereas in the tracking problem one wants to track a given arbitrary trajectory.) By “self-tuning” it is meant that the adaptive control *law* converges to the optimal control law. This is clearly a property of fundamental interest since it implies that the adaptive controller can be used as a mechanism for tuning to the parameters of an optimal control law.

In this paper we examine the problem of minimum variance *tracking* where the goal is to ensure that the output of the system tracks a specified reference trajectory with minimal average squared tracking error.

From a purely technical viewpoint the analysis of the tracking problem along the lines of [6] has until now been stymied by the fact that a key geometric property of the adaptive control algorithm, which renders the regression and parameter estimate vectors orthogonal, holds only in the regulation problem and not in the case of tracking. Our first contribution here is to show how to overcome this difficulty by enlarging the dimension of the regression vector.

* Received by the editors January 27, 1986; accepted for publication July 22, 1986.

† Department of Electrical Engineering and the Coordinated Science Laboratory, University of Illinois, Urbana, Illinois 61801. The research of this author was supported in part by the National Science Foundation under grants ECS-85-06628 and ECS-84-14676 and in part by the Joint Services Electronics Program under contract N00014-84-C-0149.

‡ Centre d'Automatique et Informatique, Ecole des Mines de Paris, France. Previously visiting, Department of Electrical Engineering and the Coordinated Science Laboratory, University of Illinois, Urbana, Illinois 61801. This author is a member of the Groupe de Recherche Coordonee—Systemes Adaptatifs en Robotique, Traitement du Signal et Automatique—of the Centre National de la Recherche Scientifique.

Another well-known difficulty with the tracking problem is that when the reference trajectory to be tracked is a general nonzero trajectory (we call this the *general tracking problem*), then the control law which allows the trajectory to be tracked with minimum variance does require explicit knowledge of the coefficients of the colored noise polynomial, see [4], [6], [12]. This is another feature distinguishing the tracking problem from the regulation problem. Consequently, it is necessary to identify some additional parameters pertaining to the colored noise polynomial in order to obtain self-tuning. Such identification is established in this paper under the natural assumption that the reference trajectory is sufficiently rich of appropriate order.

The second essential contribution of this paper is the examination of how one may obtain self-tuning when the reference trajectory is not so rich as to allow one to identify all the coefficients of the colored noise polynomial. For example, in an important class of practical problems, called *set-point problems*, the output of the system is required to stay as close as possible to a certain specified level. Thus the reference trajectory is a nonzero constant, which is sufficiently rich of order one only. We examine such problems, which violate the richness assumptions of the general tracking problem, by examining the problem of following trajectories which are generated by linear models. We call these the *linear model following problems*. (The set-point problem is a special case of the linear model following problem.) Our second class of main results is to show how one may adjust the *dimension* of the regression vector to the degree of excitation present in the reference trajectory. We then provide a proof of self-tuning of the resulting reduced dimension adaptive controllers.

Our main results are therefore the following:

(i) The adaptive control laws in both the general tracking problem as well as the linear model following problem are self-optimal, i.e., the average squared tracking error is minimal (Theorem 3).

(ii) In the general tracking problem, if the reference trajectory is sufficiently rich of order at least equal to the sum of the degrees of the control and noise polynomials in the ARMAX representation of the system, then the parameter estimates are strongly consistent, i.e., they converge to the true values almost surely (Theorems 6 and 7). This result also implies that the adaptive controller is self-tuning, i.e., the adaptive control law converges to the optimal control almost surely (Theorem 7).

(iii) For the parameter estimates to be strongly consistent in the linear model following problem it is enough for the order of sufficient richness of the reference trajectory to be equal to the degree of the noise polynomial alone (Theorems 6 and 7). This again implies self-tuning (Theorem 7).

(iv) Often, the degree of sufficient richness is even smaller than the degree of the noise polynomial (e.g. the set-point problem). In such linear model following problems, a lower dimensional adaptive controller can be used. This lower dimensional adaptive controller is self-tuning (Theorem 7). The parameter estimates also converge (Theorem 6). However, since no attempt is made at estimating all the coefficients of the noise polynomial, the parameter estimates do not converge to the true values (i.e. we are using a *direct* adaptive control law).

Some comments on the nature of these results in comparison with the results in *deterministic* adaptive control are useful. In deterministic adaptive control, where there is no noise in the system, one can asymptotically obtain zero tracking error. However in stochastic adaptive control there is noise and one wants to reject as much of the noise as possible. Clearly *optimal* noise rejection will depend critically on the knowledge of the correlations inherent in the possibly colored noise. This is where the central problem of estimating the colored noise coefficients enters into the stochastic adaptive

control problem. Indeed, in the present paper, the need for richness in the reference trajectory is intimately related precisely to the need for estimating the model of the colored noise.

2. The adaptive control laws. We consider the ARMAX system

$$(1) \quad y(t) = \sum_{i=1}^p a_i y(t-i) + \sum_{i=1}^q b_i u(t-i) + \sum_{i=1}^s c_i w(t-i) + w(t)$$

where y , u and w are, respectively, the output, input and white noise. The parameters $(a_1, \dots, a_p, b_1, \dots, b_q, c_1, \dots, c_s)$ are unknown. The goal is to design an adaptive control law which ensures that the output follows a given bounded reference trajectory $\{y^*(t)\}$ with minimal average squared tracking error, and such that the adaptive control law asymptotically self-tunes to the optimal control law. It is an added bonus if the true parameters $(a_1, \dots, a_p, b_1, \dots, b_q, c_1, \dots, c_s)$ can also be asymptotically identified.

If the reference trajectory is arbitrary, we shall refer to this problem as the *general tracking problem*. In many problems however the reference trajectory is generated as the output of a *linear model*. We shall refer to such a special case as the *linear model following problem*. The special properties of a reference trajectory generated as the output of a linear model can be usefully exploited, as we will see in the sequel. We now discuss separately the general tracking problem and the linear model following problem.

2.1. The general tracking problem. In this case $\{y^*(t)\}$ is just a reference trajectory to be tracked with no special properties. We will use the following adaptive controller (with the notation $p \vee s := \max(p, s)$).

$$(2) \quad \theta(t+1) = \theta(t) + \frac{\mu \phi(t)}{r(t)} [y(t+1) - y^*(t+1)]$$

where, for the time being, $0 < \mu < 2$ is an arbitrary constant (but see the remark at the end of § 4).

$$(3) \quad r(t+1) := 1 + \sum_{k=0}^{t+1} \phi^T(k) \phi(k),$$

$$(4) \quad \phi(t) := (y(t), \dots, y(t-p \vee s+1), u(t), \dots, u(t-q+1), \\ -y^*(t+1), \dots, -y^*(t-s+1)),$$

$$(5) \quad u(t) := \frac{-1}{\beta_1(t)} \left[\sum_{i=1}^{p \vee s} \alpha_i(t) y(t-i+1) + \sum_{i=2}^q \beta_i(t) u(t-i+1) - \sum_{i=0}^s \gamma_i(t) y^*(t-i+1) \right]$$

where

$$(6) \quad (\alpha_1(t), \dots, \alpha_{p \vee s}(t), \beta_1(t), \dots, \beta_q(t), \gamma_0(t), \dots, \gamma_s(t))^T := \theta(t).$$

Note that (5) can equivalently be written as

$$(7) \quad \phi^T(t) \theta(t) = 0.$$

The motivation behind this adaptive controller is the following. Rewrite the system (1) as,

$$y(t+1) - y^*(t+1) = \left[\sum_{i=1}^p a_i y(t+1-i) + \sum_{i=1}^q b_i u(t+1-i) + \sum_{i=1}^s c_i w(t+1-i) - y^*(t+1) \right] \\ + w(t+1).$$

If one could observe the past of $w(\cdot)$ at each time t , then an optimal controller would

choose $u(t)$ so that the term in $[\cdot \cdot \cdot]$ on the right-hand side above is zero, i.e.

$$u(t) = \frac{-1}{b_1} \left[\sum_{i=1}^p a_i y(t+1-i) + \sum_{i=2}^q b_i u(t+1-i) + \sum_{i=1}^s c_i w(t+1-i) - y^*(t+1) \right],$$

for this would result in $y(t+1) = y^*(t+1) + w(t+1)$, clearly yielding the best possible tracking error. However, the sequence $w(\cdot)$ is *not* observed, and so let us replace it by $y(\cdot) - y^*(\cdot)$, which is what we hope it would be, at least asymptotically. This gives the implementable control law,

$$u(t) = \frac{-1}{b_1} \left[\sum_{i=1}^{p \vee s} (a_i + c_i) y(t+1-i) + \sum_{i=2}^q b_i u(t+1-i) - \sum_{i=1}^s c_i y^*(t+1-i) - y^*(t+1) \right].$$

It can be shown that this control law is actually optimal with respect to the long run average of the square of the tracking error; for more details, see [12]. Let us define,

$$(8) \quad \theta^o := (a_1 + c_1, \dots, a_{p \vee s} + c_{p \vee s}, b_1, \dots, b_q, 1, c_1, \dots, c_s)^T$$

(where, for convenience, we define $c_i := 0$ for $i > s$ and $a_i := 0$ for $i > p$ in (8)), and, under optimal control, the system (1) can be represented as

$$y(t+1) - y^*(t+1) = \phi^T(t) \theta^o + w(t+1),$$

while the optimal control law can be written as one which chooses $u(t)$ to satisfy,

$$\phi^T(t) \theta^o = 0.$$

Our adaptive control scheme (2)–(6) can be interpreted as trying to estimate θ^o when the system is being optimally controlled.

Remark. Note that the $(p \vee s + q + 1)$ th component of θ^o is 1, and hence is a *known* quantity. However, the estimator ignores this knowledge and estimates it anyway by $\gamma_0(t)$. We can therefore regard (2, 3) as an *unnormalized* parameter estimator. It follows that this parameter estimator is one dimension larger than that considered in Goodwin, Ramadge and Caines [4]. In this connection, it is also of interest to note that recently Wei [7] has proposed an estimator for the regulation problem which is one dimension less than [4], [6].

2.2. The linear model following problem. In many situations of interest the reference trajectory is generated, at least asymptotically, as the output of a linear model. We shall suppose that there is a sequence $\{y_m(t)\}$ such that

$$(9) \quad y_m(t) = \sum_{i=1}^l h_i y_m(t-i)$$

and the trajectory to be tracked $y^*(t)$ is asymptotically close to $y_m(t)$ in that

$$(10) \quad \sum_{t=1}^{\infty} (y^*(t) - y_m(t))^2 < +\infty.$$

Without loss of generality we can make the following two assumptions:

- (11)(i) There is no lower order difference equation satisfied by $\{y_m(t)\}$, i.e., there is *no* nontrivial polynomial $\bar{H}(z)$ of degree strictly less than l such that $\bar{H}(z)y_m(t) = 0$ for all t . (z is the *backward* shift operator.)
- (11)(ii) The roots of $H(z) := 1 - \sum_{i=1}^l h_i z^i$ are exactly on the unit circle and there are no repeated roots.

Assumption (11)(i) is without loss of generality since otherwise we could simply replace $H(z)$ in (9) by $\bar{H}(z)$. Note that this also means that the initial conditions on (9) are sufficient to excite *all* the modes of $H(z)$. Assumption (11)(ii) is also without loss of generality due to the following reasons. First, since we intend to work only with *bounded* $\{y^*(t)\}$, and since all the modes of $H(z)$ are excited, we have to assume that $H(z)$ has roots on or outside the unit circle, and also that the roots on the unit circle are not repeated. However, since we are only interested in the *asymptotic* behavior of $\{y^*(t)\}$, we can eliminate all the modes corresponding to roots of $H(z)$ which are strictly outside the unit circle, since they decay geometrically to 0. This leaves us with (11)(ii).

It is worth noting that (11)(i) and (11)(ii) together imply that

$$y_m(t) = d_0 + d_1(-1)^t + \sum d_i \sin(\omega_i t + \delta_i).$$

Depending on how large l is, we will use adaptive controllers with parameter estimators of different dimensions.

Case 1. $l \leq s$. Recall that s is the degree of the noise polynomial in (1). When $l \leq s$, we will reduce the dimension of the parameter estimator by $(s+1-l)$ components by replacing (4)–(6) by the following:

$$(12) \quad \phi(t) := (y(t), \dots, y(t-p \vee s+1), u(t), \dots, u(t-q+1), \\ -y^*(t+1), \dots, -y^*(t+2-l))^T,$$

$$(13) \quad \theta(t) := (\alpha_1(t), \dots, \alpha_{p \vee s}(t), \beta_1(t), \dots, \beta_q(t), \gamma_0(t), \dots, \gamma_{l-1}(t))^T,$$

and

$$(14) \quad u(t) = \frac{-1}{\beta_1(t)} \left[\sum_{i=1}^{p \vee s} \alpha_i(t) y(t-i+1) + \sum_{i=2}^q \beta_i(t) u(t-i+1) - \sum_{i=0}^{l-1} \gamma_i(t) y^*(t-i+1) \right],$$

or equivalently by (7).

The idea underlying the above adaptive control law is the following. If the parameters were known, the minimum variance adaptive control law would be,

$$u(t) = \frac{-1}{b_1} \left[\sum_{i=1}^{p \vee s} (a_i + c_i) y(t-i+1) + \sum_{i=2}^q b_i u(t-i+1) - y^*(t+1) - \sum_{i=1}^s c_i y^*(t-i+1) \right],$$

see [12] for details. In this control law the only terms featuring y^* are $y^*(t+1) + \sum_{i=1}^s y^*(t-i+1) = C(z)y^*(t+1)$. Thus the control law really only requires knowledge of $C(z)y^*(t)$. Let

$$(15) \quad G(z) := \sum_{i=0}^{l-1} g_i z^i$$

be a polynomial satisfying,

$$(16) \quad C(z) = F(z)H(z) + G(z)$$

for some

$$(17) \quad F(z) := \sum_{i=0}^{s-l} f_i z^i.$$

Such polynomials $G(z)$ and $F(z)$ are the remainder and quotient, respectively, when the polynomial $C(z)$ is divided by the polynomial $H(z)$. Then, asymptotically at least,

$$C(z)y^*(t) = [F(z)H(z) + G(z)]y^*(t) = F(z)H(z)y^*(t) + G(z)y^*(t) = G(z)y^*(t),$$

since by (9.10), $H(z)y^*(t) = 0$ holds asymptotically. Thus we only need knowledge of $G(z)y^*(t)$ in order to implement the true minimum variance control law. We can therefore interpret the parameter estimate (13) as trying to estimate

$$(18) \quad \theta^0 := (a_1 + c_1, \dots, a_{pvs} + c_{pvs}, b_1, \dots, b_q, g_0, g_1, \dots, g_{l-1})^T.$$

Remarks. (i) The adaptive controller need not be provided with the precise information about what the polynomial $H(z)$ is. It only needs knowledge of the *degree* of $H(z)$.

(ii) It should be noted that the parameter estimator is no more “unnormalized,” since the coefficients g_0, \dots, g_{l-1} are all unknown.

Case 2: $l \geq s + 1$. Since $(s + 1 - l) \leq 0$ when $l \geq s + 1$, no savings in dimensionality can be achieved. Hence we will use the same adaptive control law as (2)–(7). For this case also we define θ^o as in (8).

3. Sufficient richness. In the sequel we will prove that all the coefficients $(a_1, \dots, a_p, b_1, \dots, b_q, c_1, \dots, c_s)$ can be asymptotically identified when the reference trajectory $\{y^*(t)\}$ is “sufficiently rich” in an appropriate sense. We have the following definition.

DEFINITION. We shall say that a scalar sequence $\{y^*(t)\}$ is *strongly sufficiently rich of order l* if l is the largest nonnegative integer for which there exists an n and an $\varepsilon > 0$ such that

$$\sum_{k=t+1}^{t+n} (y^*(k-1), \dots, y^*(k-l))^T (y^*(k-1), \dots, y^*(k-l)) \geq \varepsilon I_l \quad \text{for all } t \text{ large enough.}$$

I_l here is the $l \times l$ identity matrix.

The following property of $\{y_m(t)\}$, and also $\{y^*(t)\}$, generated by the linear model (9), (10), (11)(i), (ii) should be noted.

LEMMA 1. Suppose $\{y^*(t)\}$ and $\{y_m(t)\}$ satisfy (9)–(11). Then both $\{y^*(t)\}$ and $\{y_m(t)\}$ are strongly sufficiently rich of order l .

Proof. We will show that there exists $\varepsilon > 0$ such that

$$\sum_{k=t+1}^{t+l} Y_l(k-1) \geq \varepsilon I_l \quad \text{for all } t \text{ large enough}$$

where

$$Y_l(k-1) := (y_m(k-1), \dots, y_m(k-l))^T (y_m(k-1), \dots, y_m(k-l)).$$

Suppose this is not true. Then there exists a sequence of vectors $\{x(t_n)\}$, with each $\|x(t_n)\| = 1$ and $x(t_n) := (x_1(t_n), \dots, x_l(t_n))^T$ such that

$$x^T(t_n) \sum_{k=t_n+1}^{t_n+l} Y_l(k-1) x(t_n) \leq \frac{1}{n}.$$

We can also assume without loss of generality that $\lim_n x(t_n) =: x$ exists with $\|x\| = 1$, $x := (x_1, \dots, x_l)^T$. Moreover, since $\{y_m(t)\}$ is bounded, $\{Y_l(k-1)\}$ is also bounded and so

$$\lim_n x^T \sum_{k=t_n+1}^{t_n+l} Y_l(k-1) x = 0.$$

Let $X(z) := \sum_{i=1}^l x_i z^i$. Interpreting z as the *backward* shift operator, we have

$$\lim_n \sum_{k=t_n+1}^{t_n+l} [X(z)y_m(k)]^2 = 0.$$

This implies that

$$\lim_n X(z)y_m(t_n + i) = 0 \quad \text{for } i = 1, 2, \dots, l.$$

Now note that $H(z)X(z)y_m(t) = X(z)H(z)y_m(t) = 0$ and so

$$X(z)y_m(t) = \sum_{k=1}^l \delta_k \lambda_k^t$$

where $\{\lambda_k\}$ is the set of roots of $H(z)$. Hence we have

$$\lim_n \sum_{k=1}^l \delta_k \lambda_k^{t_n+i} = 0 \quad \text{for } i = 1, \dots, l.$$

This can also be written as

$$\lim_n \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 \\ \lambda_1 & \lambda_2 & & & & \lambda_l \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \lambda_1^{l-1} & \lambda_2^{l-1} & & & & \lambda_l^{l-1} \end{bmatrix} \begin{bmatrix} \lambda_1^{t_n+1} & 0 & 0 & 0 & 0 & 0 \\ 0 & \lambda_2^{t_n+1} & & & & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & & & & \lambda_l^{t_n+1} \end{bmatrix} \begin{bmatrix} \delta_1 \\ \delta_2 \\ \vdots \\ \delta_l \end{bmatrix} = 0.$$

The first matrix on the left-hand side above is the Vandermonde matrix which is nonsingular since all the λ_k 's are distinct. Moreover $|\lambda_k| = 1$ for all k , and so it follows that $\delta_k = 0$ for $k = 1, \dots, l$. This however implies that $X(z)y_m(t) = 0$ for all t . However $X(z)$ is a polynomial of degree $l-1$ or less, and by (11)(i), it follows that $X(z) = 0$, i.e., $\|x\| = 0$. This is a contradiction to $\|x\| = 1$, proving that $\{y_m(t)\}$ is indeed strongly sufficiently rich of order l . By (10) it follows trivially that $\{y^*(t)\}$ is also strongly sufficiently rich of order l . (Actually it is enough that $\lim_t (y_m(t) - y^*(t)) = 0$). \square

For future reference, we also have the following result.

LEMMA 2. Let $S(t, z) := \sum_{i=0}^j s_i(t)z^i$. Suppose $\{s_i(t)\}$ is bounded for $i = 0, \dots, j$ and $\lim_t |s_i(t) - s_i(t-1)| = 0$ for $i = 0, \dots, j$. Suppose also that for some sequence $\{x(t)\}$,

$$\lim_N \frac{1}{N} \sum_{t=1}^N [S(t, z)y_m(t)]^2 = 0 \quad \text{and} \quad \lim_N \frac{1}{N} \sum_{t=1}^N x^2(t) = 0.$$

Then there exists a common subsequence $\{t_k\}$ with $\lim_k x(t_k) = 0$ and $\lim_k S(t_k, z) = K(z)H(z)$ for some polynomial $K(z)$. (By $S(t, z)y_m(t)$ we mean $\sum_{i=0}^j s_i(t)y_m(t-i)$.)

Proof. Since $\lim_t |s_i(t) - s_i(t+n)| = 0$ for every n and $\{y_m(t)\}$ is bounded, it is also true that $\lim_N \frac{1}{N} \sum_{t=1}^N [S(t+n, z)y_m(t)]^2 = 0$ for every n . Hence we can sum over n and also add $x^2(t)$ to get

$$\lim_N \frac{1}{N} \sum_{t=1}^N \left\{ x^2(t) + \sum_{n=1}^l [S(t, z)y_m(t-n)]^2 \right\} = 0.$$

Hence there is a subsequence $\{t_k\}$ such that

$$\lim_k S(t_k, z)y_m(t_k - n) = 0 \quad \text{for } n = 1, \dots, l, \quad \lim_k x(t_k) = 0.$$

Further we can also assume without loss of generality that

$$\lim_k S(t_k, z) =: S(z)$$

exists, by which we mean that $\lim_k s_i(t_k) =: s_i$ exists for $i = 0, \dots, j$ and $S(z) := \sum_{i=0}^j s_i z^i$. Further, since $\{y_m(t)\}$ is bounded, it follows that

$$\lim_k S(z)y_m(t_k - n) = 0 \quad \text{for } n = 1, \dots, l.$$

Note that $H(z)S(z)y_m(t) = 0$ for all t , and so

$$S(z)y_m(t) = \sum_{n=1}^l \delta_n \lambda_n^t$$

where $\{\lambda_n\}$ is the set of roots of $H(z)$. Proceeding just as in the proof of Lemma 1, it follows that

$$S(z)y_m(t) = 0 \quad \text{for all } t.$$

Now let $U(z)$ be the greatest common divisor of $S(z)$ and $H(z)$. Then there exist polynomials $R(z)$ and $T(z)$ such that $R(z)S(z) + T(z)H(z) = U(z)$. Hence $U(z)y_m(t) = 0$ for all t . However, since the degree of $U(z)$ is less than or equal to l , it follows from (11)(i) that $U(z) = \xi H(z)$ for some scalar ξ , and so the lemma is proved. \square

4. Assumptions. Define the polynomials

$$A(z) := 1 - \sum_{i=1}^p a_i z^i,$$

$$B(z) := \sum_{i=1}^q b_i z^{i-1},$$

$$C(z) := 1 + \sum_{i=1}^s c_i z^i.$$

Throughout this paper we employ the following assumptions only.

- (19)(i) All the roots of $B(z)$ and $C(z)$ are strictly outside the unit circle.
- (19)(ii) $\operatorname{Re} \left[C(e^{i\omega}) - \frac{\mu}{2} \right] > 0$ for $0 \leq \omega < 2\pi$.
- (19)(iii) $b_1 \neq 0$.
- (19)(iv) $z^{-1}[C(z) - A(z)]$ and $B(z)$ are polynomials of degrees respectively equal to $(p \vee s - 1)$ and $(q - 1)$, which have no common factors.
- (19)(v) $\{w(t)\}$ is a sequence of scalar random variables on a probability space $\{\Omega, F, P\}$, whose distributions are all mutually absolutely continuous with respect to Lebesgue measure.
- (19)(vi) Let $F_t := \sigma\{w(1), \dots, w(t)\}$ be the sub- σ -algebra of F generated by $\{w(1), \dots, w(t)\}$. We assume that there are $\sigma^2 > 0$ and $\delta > 0$ such that

$$E[w(t) | F_{t-1}] = 0 \quad \text{a.s.},$$

$$E[w^2(t) | F_{t-1}] = \sigma^2 \quad \text{a.s.},$$

$$\sup_t E[|w(t)|^{2+\delta} | F_{t-1}] < +\infty \quad \text{a.s.}$$

$$(19)(vii) \quad \|\theta(0)\| > 0.$$

$$(19)(viii) \quad \{y^*(t)\} \text{ is bounded.}$$

It should be noted that the condition (19)(v) guarantees that the controls are well defined a.s. through (5.14) since the event $\{\beta_1(t) = 0\}$ is a null event, see Caines and Meyn [9].

Remark. Let us consider a different constant μ_1 in place of μ in (2). It is easy to verify, see [12], that the resulting adaptive control algorithm produces parameter estimates $\theta_1(t) = (\mu_1/\mu)\theta(t)$ and *identical* inputs and outputs as the original algorithm using μ , provided $\theta_1(0)$ is chosen as $\theta_1(0) := (\mu_1/\mu)\theta(0)$. This property relies on the fact that the control input $u(t)$ is invariant with respect to scaling of $\theta(t)$ in (7). Making use of this observation, it follows that one need not restrict μ to lie in $(0, 2)$; it is enough to have $\mu \neq 0$. Further, one only needs the assumption

$$(19)(ii) \quad \operatorname{Re} C(e^{i\omega}) > 0 \quad \text{for } 0 \leq \omega < 2\pi$$

in place of (19)(ii).

5. Self-optimality. In this section we will prove the following theorem which asserts, among other things, that in all cases the adaptive controller minimizes the average squared tracking error.

THEOREM 3.

$$(20)(i) \quad \lim_{N \rightarrow \infty} \frac{1}{N} \sum_{t=1}^N [y(t) - y^*(t)]^2 = \sigma^2 \quad \text{a.s.,}$$

$$(20)(ii) \quad \lim_{N \rightarrow \infty} \frac{1}{N} \sum_{t=1}^N (E[y(t+1) - y^*(t+1) | F_t])^2 = 0 \quad \text{a.s.,}$$

$$(20)(iii) \quad \limsup_N \frac{1}{N} \sum_{t=1}^N u^2(t) < +\infty \quad \text{a.s.,}$$

$$(20)(iv) \quad \lim_t \|\theta(t) - \theta^o\|^2 \text{ exists and is finite a.s.}$$

Proof. We will abbreviate those details of the proof which are similar to those of Goodwin, Ramadge and Caines [4] or [6]. Let $\tilde{\theta}(t) := \theta(t) - \theta^o$ and define $V(t) := \|\tilde{\theta}(t)\|^2$. Using $r(t) \equiv \phi^T(t)\phi(t)$ and $\phi^T(t)\tilde{\theta}(t) = -\phi^T(t)\theta^o$, we can get

$$\begin{aligned} E[V(t+1) | F_t] &\leq V(t) - \frac{2\mu}{r(t)} \left\{ \phi^T(t)\theta^o - \frac{\mu + \delta}{2} E[y(t+1) - y^*(t+1) | F_t] \right\} \\ &\quad \cdot E[y(t+1) - y^*(t+1) | F_t] - \frac{\mu\delta}{r(t)} (E[y(t+1) - y^*(t+1) | F_t])^2 \\ &\quad + \mu^2 \frac{\phi^T(t)\phi(t)}{r^2(t)} \sigma^2 \end{aligned}$$

for all δ . Choose $\delta > 0$ so small that $[C(z) - (\mu + \delta)/2]$ is strictly positive real. Let us first consider the following case.

Case 1. *General tracking problem or the linear model following problem with $l \geq s + 1$.*

$$\begin{aligned}
 C(z)E[y(t+1) - y^*(t+1)|F_t] &= C(z)[y(t+1) - y^*(t+1) - w(t+1)] \\
 &= [y(t+1) - y^*(t+1) - w(t+1)] \\
 &\quad + [C(z) - 1][y(t+1) - y^*(t+1) - w(t+1)] \\
 &= [y(t+1) - y^*(t+1) - w(t+1)] \\
 &\quad + \sum_{i=1}^s c_i[y(t-i+1) - y^*(t-i+1) - w(t-i+1)] \\
 (21) \quad &= \left[y(t+1) - w(t+1) - \sum_{i=1}^s c_i w(t-i+1) \right] - y^*(t+1) \\
 &\quad + \sum_{i=1}^s c_i[y(t-i+1) - y^*(t-i+1)] \\
 &= \sum_{i=1}^{p \vee s} (a_i + c_i)y(t-i+1) + \sum_{i=1}^q b_i u(t-i+1) \\
 &\quad - y^*(t+1) - \sum_{i=1}^s c_i y^*(t-i+1) \\
 &= \phi^T(t)\theta^o.
 \end{aligned}$$

By the strict positive realness of $[C(z) - (\mu + \delta)/2]$ it therefore follows that

$$\begin{aligned}
 S(n) &:= 2\mu \sum_{t=1}^n \left\{ \phi^T(t)\theta^o - \frac{\mu + \delta}{2} E[y(t+1) - y^*(t+1)|F_t] \right\} E[y(t+1) - y^*(t+1)|F_t] \\
 &\geq K \quad \text{a.s. for all } n, \text{ for some } K.
 \end{aligned}$$

Defining $M(t) := V(t) + S(t-1)/r(t-1)$, and using $r(t) \geq r(t-1) > 0$, it follows that

$$E[M(t+1)|F_t] \leq M(t) - \frac{\mu\delta}{r(t)} (E[y(t+1) - y^*(t+1)|F_t])^2 + \frac{\mu^2 \phi^T(t)\phi(t)}{r^2(t)} \sigma^2.$$

The last term above is summable a.s., and so using the Positive Near Supermartingale Convergence Theorem we can get:

- (i) $\{M(t)\}$ converges a.s.,
- (ii) $\sum_{t=1}^{\infty} \frac{(E[y(t+1) - y^*(t+1)|F_t])^2}{r(t)} < +\infty$ a.s.

Now we claim that $\lim_t r(t) = +\infty$ a.s. Otherwise $r_t = 1 + \sum_{i=1}^t \phi^T(k)\phi(k)$ would lead to $\lim_t \phi(t) = 0$ on a set of positive probability. This in turn would imply $\lim_t y_t = 0$ and $\lim_t u_t = 0$, and from the system equation (1) it would then have to follow that $\lim_t C(z)w(t) = 0$ on a set of positive probability, which we will now contradict as follows. First note that $(C(z)w(t))^2$ is a linear combination of terms of the form $w^2(t-i)$ and $w(t-i)w(t-j)$. Let us first examine the first set of square terms. As a consequence of (19)(vi) and Jensen's and Minkowski's inequalities, it follows that $\sup_t E[|w^2(t) - E(w^2(t)|F_{t-1})|^{1+\delta/2}|F_{t-1}] < \infty$ a.s. Chow's Theorem [10, Thm. 3.3.1] is therefore applicable, and shows that $\lim_N 1/N \sum_{t=1}^N w^2(t) = \sigma^2$ a.s. Now we turn to the cross terms. Since $\sum_{t=1}^{\infty} w^2(t-i) = \infty$ a.s., an appeal to the Local Convergence Theorem for Martingales [11, Lemma 2.3] shows that $\sum_{t=1}^N 2w(t-i)w(t) = o(\sum_{t=1}^N w^2(t-i))$ a.s. Hence

$\lim_N 1/N \sum_{t=1}^N w(t-i)w(t) = 0$ a.s. Adding up the contributions, we get $\lim_N 1/N \sum_{t=1}^N (C(z)w(t))^2 = 1 + \sum_{i=1}^s c_i^2 > 0$ a.s. This provides the required contradiction.

Since $\lim_t r(t) = +\infty$ a.s., Kronecker's Lemma is applicable and gives

$$\lim_N \frac{1}{r(N)} \sum_{t=1}^N (E[y(t+1) - y^*(t+1) | F_t])^2 = 0 \quad \text{a.s.}$$

Utilizing the strictly minimum phase property of $B(z)$ it follows that $\{r(N)/N\}$ is bounded a.s., which proves (20)(iii) and (20)(ii). The same arguments as in Lemma 7 and Lemma 9 of [6] yield (20)(i) and (20)(iv).

Case 2. Linear model following problem with $l \leq s$. Just as in (21) we still get

$$\begin{aligned} C(z)E[y(t+1) - y^*(t+1) | F_t] &= \sum_{i=1}^{p \vee s} (a_i + c_i)y(t-i+1) + \sum_{i=1}^q b_i u(t-i+1) \\ &\quad - C(z)y^*(t+1). \end{aligned}$$

Let $\tilde{y}(t) := y_m(t) - y^*(t)$. Then from (9) and (16) we get

$$\begin{aligned} C(z)y^*(t+1) &= C(z)y_m(t+1) - C(z)\tilde{y}(t+1) \\ &= G(z)y_m(t+1) - C(z)\tilde{y}(t+1) \\ &= G(z)y^*(t+1) + [G(z) - C(z)]\tilde{y}(t+1). \end{aligned}$$

Hence

$$\begin{aligned} C(z)E[y(t+1) - y^*(t+1) | F_t] &= \sum_{i=1}^{p \vee s} (a_i + c_i)y(t-i+1) + \sum_{i=1}^q b_i u(t-i+1) \\ &\quad - G(z)y^*(t+1) + [C(z) - G(z)]\tilde{y}(t+1) \\ &= \phi^T(t)\theta^o + [C(z) - G(z)]\tilde{y}(t+1). \end{aligned}$$

By the strict positive realness property of $[C(z) - (\mu + \delta)/2]$, it follows that

$$\begin{aligned} S(n) &:= 2\mu \sum_{t=1}^n \left\{ \phi^T(t)\theta^o + [C(z) - G(z)]\tilde{y}(t+1) \right. \\ &\quad \left. - \frac{\mu + \delta}{2} E[y(t+1) - y^*(t+1) | F_t] \right\} \\ &\quad \cdot E[y(t+1) - y^*(t+1) | F_t] \\ &\geq K \quad \text{a.s. for all } n, \text{ for some } K. \end{aligned}$$

Defining $M(t) := V(t) + S(t-1)/r(t-1)$, we get

$$\begin{aligned} E[M(t+1) | F_t] &\leq M(t) - \frac{\mu\delta}{r(t)} (E[y(t+1) - y^*(t+1) | F_t])^2 + \mu^2 \frac{\phi^T(t)\phi(t)}{r^2(t)} \sigma^2 \\ &\quad + \frac{2\mu}{r(t)} E[y(t+1) - y^*(t+1) | F_t][C(z) - G(z)]\tilde{y}(t+1). \end{aligned}$$

Define $\bar{y}(t) := [C(z) - G(z)]\tilde{y}(t+1)$, and note that by (10), $\sum_{t=1}^{\infty} \bar{y}^2(t) < +\infty$. For any $\rho > 0$, we have

$$2E[y(t+1) - y^*(t+1) | F_t]\bar{y}(t) \leq \rho^2 (E[y(t+1) - y^*(t+1) | F_t])^2 + \left(\frac{\bar{y}(t)}{\rho} \right)^2.$$

Hence, choose ρ so small that $(\mu\delta - 2\mu\rho^2) > 0$, and note that

$$\begin{aligned} E[M(t+1)|F_t] \leq M(t) - \frac{(\mu\delta - 2\mu\rho^2)}{r(t)} (E[y(t+1) - y^*(t+1)|F_t])^2 \\ + \frac{\mu^2 \phi^T(t)\phi(t)}{r^2(t)} \sigma^2 + \frac{2\mu\bar{y}^2(t)}{\rho^2 r(t)}. \end{aligned}$$

Now both of the last two terms are summable, and so we can again use the Positive Near Supermartingale Convergence Theorem. The rest of the proof is similar to the previous case. \square

By (20)(i) of the above theorem, we see that usage of the adaptive controller leads to a value of σ^2 for the average of the square of the tracking error. In order to justify our claim at the beginning of this section that the adaptive controller *minimizes* the average of the square of the tracking error, we need to show that no other nonanticipative controller, including possibly controllers which utilize knowledge of the parameters (a_i, b_i, c_i) , can realize a smaller value than σ^2 for the average squared tracking error on any set of sample paths of positive measure. This is provided in the following lemma.

LEMMA 4. Consider the ARMAX system (1). Let $F_t := \sigma(w_s \text{ for } s \leq t \text{ and } y_i, u_i \text{ for } i \leq 0)$ be the σ -algebra generated by the past, and let $\{u_t\}$ be any control sequence chosen so that $u_t \in F_t$, i.e. u_t is F_t -measurable for each $t \geq 0$. Then,

$$\liminf_N \frac{1}{N} \sum_{t=1}^N (y(t) - y^*(t))^2 \geq \sigma^2 \quad \text{a.s.}$$

Proof. Define

$$g(t-1) := \left[\sum_{i=1}^p a_i y(t-i) + \sum_{i=1}^a b_i u(t-i) + \sum_{i=1}^s c_i w(t-i) \right],$$

and note that $g(t-1) \in F_{t-1}$. Rewrite the system equation (1) as $y(t) = g(t-1) + w(t)$ and get

$$y^2(t) = \frac{1}{N} \sum_{t=1}^N g^2(t-1) \left[1 + \frac{\sum_{t=1}^N 2g(t-1)w(t)}{\sum_{t=1}^N g^2(t-1)} \right] + \frac{1}{N} \sum_{t=1}^N w^2(t).$$

Appealing to the Local Convergence Theorem for Martingales [12, Lemma 2.3], we know that except on a null set,

$$\begin{aligned} \sum_{t=1}^N 2g(t-1)w(t) &= o\left(\sum_{t=1}^N g^2(t-1)\right) \quad \text{if } \sum_{t=1}^N g^2(t) = \infty, \\ &< \infty \quad \text{if } \sum_{t=1}^N g^2(t) < \infty. \end{aligned}$$

In either case, therefore, it follows that

$$\liminf_N \frac{1}{N} \sum_{t=1}^N g^2(t-1) \left[1 + \frac{\sum_{t=1}^N 2g(t-1)w(t)}{\sum_{t=1}^N g^2(t-1)} \right] \geq 0 \quad \text{a.s.}$$

Hence,

$$\begin{aligned} \liminf_N \frac{1}{N} \sum_{t=1}^N y^2(t) &\geq \lim_N \frac{1}{N} \sum_{t=1}^N w^2(t) \\ &= \sigma^2 \quad \text{a.s.} \end{aligned}$$

The last equality has been proved in the course of the proof of Theorem 3. \square

6. Self-tuning and convergence. In this section we address the self-tuning and convergence properties of the adaptive controllers.

First due to (2.7) we have the same geometrical properties as in [6]. This gives us the following lemma, see [6].

LEMMA 5.

- (22)(i) $\lim_t \|\theta(t)\|$ exists and is finite a.s.
 (22)(ii) For every n , $\lim_t \|\theta(t) - \theta(t-n)\| = 0$ a.s.
 (22)(iii) $\|\theta(t+1)\| \geq \|\theta(t)\|$.
 (22)(iv) If there is a random scalar ξ and a random subsequence $\{t_k\}$ such that
- $$\lim_k \theta(t_k) = \xi \theta^o \quad \text{a.s.}$$

then

$$\lim_t \theta(t) = \xi \theta^o \quad \text{a.s.}$$

So in order to prove that $\lim_t \theta(t) = \xi \theta^o$ it is sufficient to show that there is just one subsequence for almost every sample path along which such a limit exists.

THEOREM 6. (i) Suppose that $\{y^*(t)\}$ in the general tracking problem is strongly sufficiently rich of order $(s+q)$. Then

$$(23) \quad \lim_t \theta(t) = \xi \theta^o \quad \text{a.s.}$$

for some a.s. finite nonzero scalar random variable ξ .

(ii) The result (23) holds in the linear model following problem irrespective of the order of strong sufficient richness of $\{y^*(t)\}$ (using the appropriate definition of θ^o as in (8) or (17)).

Proof. We start with (20)(ii) which can be written as

$$(24) \quad \lim_N \frac{1}{N} \sum_{t=1}^N \{[1-A(z)]y(t+1) + zB(z)u(t+1) + [C(z)-1]w(t+1) - y^*(t+1)\}^2 = 0.$$

Define the time varying polynomials

$$P(t, z) := \sum_{i=1}^{p \vee s} \alpha_i(t) z^{i-1},$$

$$Q(t, z) := \sum_{i=1}^q \beta_i(t) z^{i-1},$$

$$R(t, z) := \begin{cases} \sum_{i=0}^{t-1} \gamma_i(t) z^i & \text{in the linear model following problem with } l \leq s, \\ \sum_{i=0}^s \gamma_i(t) z^i & \text{otherwise.} \end{cases}$$

We shall interpret z as the backward shift operator. Thus, to illustrate the notation,

$$Q(t, z)x(t) := \sum_{i=1}^q \beta_i(t)x(t-i+1); \quad Q(t, z)B(z)x(t) := \sum_{i=1}^q \beta_i(t) \sum_{j=1}^q b_j x(t-i-j+2),$$

$$B(z)Q(t, z)x(t) := \sum_{j=1}^q b_j \sum_{i=1}^q \beta_i(t-j+1)x(t-i-j+2).$$

Though $Q(t, z)B(z)x(t) \neq B(z)Q(t, z)x(t)$, it should be noted that if $\{1/N \sum_{t=1}^N x^2(t)\}$ is bounded, then it is true that

$$\lim_N \frac{1}{N} \sum_{t=1}^N [Q(t, z)B(z)x(t) - B(z)Q(t, z)x(t)]^2 = 0.$$

To verify this, one needs to use the facts that $\lim_t \|\theta(t) - \theta(t-n)\| = 0$ a.s. and $\{\theta(t)\}$ is bounded a.s.

Multiplying inside the summation in (24) by $Q(t, z)$, we have

$$\begin{aligned} \lim_N \frac{1}{N} \sum_{t=1}^N \{Q(t, z)[1 - A(z)]y(t+1) + Q(t, z)zB(z)u(t+1) \\ + Q(t, z)[C(z) - 1]w(t+1) - Q(t, z)y^*(t+1)\}^2 = 0 \quad \text{a.s.} \end{aligned}$$

Since

$$\left\{ \frac{1}{N} \sum_{t=1}^N y^2(t) \right\}, \left\{ \frac{1}{N} \sum_{t=1}^N u^2(t) \right\}, \left\{ \frac{1}{N} \sum_{t=1}^N w^2(t) \right\}, \left\{ \frac{1}{N} \sum_{t=1}^N y^{*2}(t) \right\}$$

are all bounded, we can interchange the polynomials above to get

$$\begin{aligned} (25) \quad \lim_N \frac{1}{N} \sum_{t=1}^N \{z^{-1}[1 - A(z)]Q(t, z)y(t) + B(z)Q(t, z)u(t) \\ + z^{-1}[C(z) - 1]Q(t, z)w(t) - Q(t, z)y^*(t+1)\}^2 = 0 \quad \text{a.s.} \end{aligned}$$

Now note that the control laws (5) and (14) can be written as

$$(26) \quad Q(t, z)u(t) = -P(t, z)y(t) + R(t, z)y^*(t+1).$$

Substituting (26) in (25) gives

$$\begin{aligned} \lim_N \frac{1}{N} \sum_{t=1}^N \{z^{-1}[1 - A(z)]Q(t, z) - B(z)P(t, z)\}y(t) \\ + z^{-1}[C(z) - 1]Q(t, z)w(t) \\ + \{B(z)R(t, z) - Q(t, z)\}y^*(t+1)\}^2 = 0 \quad \text{a.s.} \end{aligned}$$

Now $y(t) = w(t) + y^*(t) + E[y(t) - y^*(t) | F_{t-1}]$, and so substituting for $y(t)$ gives

$$\begin{aligned} \lim_N \frac{1}{N} \sum_{t=1}^N \{z^{-1}[C(z) - A(z)]Q(t, z) - B(z)P(t, z)\}w(t) \\ + \{B(z)R(t, z) - zB(z)P(t, z) - A(z)Q(t, z)\}y^*(t+1) \\ + \{z^{-1}[1 - A(z)]Q(t, z) - B(z)P(t, z)\}E[y(t) - y^*(t) | F_t]\}^2 = 0 \quad \text{a.s.} \end{aligned}$$

Due to (20)(ii) and the fact that $\{\theta(t)\}$ is bounded, we can drop the last term above and write

$$\begin{aligned} \lim_N \frac{1}{N} \sum_{t=1}^N \{z^{-1}[C(z) - A(z)]Q(t, z) - B(z)P(t, z)\}w(t) \\ + \{B(z)R(t, z) - zB(z)P(t, z) - A(z)Q(t, z)\}y^*(t+1)\}^2 = 0 \quad \text{a.s.} \end{aligned}$$

Since $\lim_t \|\theta(t) - \theta(t-1)\| = 0$ a.s., and since $\{y^*(t+1)\}$ is bounded, we can replace $R(t, z)$, $P(t, z)$ and $Q(t, z)$ above by $R(t-n, z)$, $P(t-n, z)$ and $Q(t-n, z)$, respectively, for any n . Thus

$$\lim_N \frac{1}{N} \sum_{t=1}^N \{ \{z^{-1}[C(z) - A(z)]Q(t-n, z) - B(z)P(t-n, z)\}w(t) + \{B(z)R(t-n, z) - zB(z)P(t-n, z) - A(z)Q(t-n, z)\}y^*(t+1) \}^2 = 0 \quad \text{a.s.}$$

Choose n larger than $(p+q+s)$, and then we can apply Lemma 11 of [6] to deduce that

$$(27) \quad \lim_N \frac{1}{N} \sum_{t=1}^N \{z^{-1}[C(z) - A(z)]Q(t-n, z) - B(z)P(t-n, z)\}^2 = 0 \quad \text{a.s.}$$

by which we mean that the average of the square of each coefficient of the polynomial in z is 0; and also

$$(28) \quad \lim_N \frac{1}{N} \sum_{t=1}^N \{ \{B(z)R(t-n, z) - zB(z)P(t-n, z) - A(z)Q(t-n, z)\}y^*(t+1) \}^2 = 0 \quad \text{a.s.}$$

Furthermore since $\{y^*(t)\}$ is bounded, (27) also implies that

$$(29) \quad \lim_N \frac{1}{N} \sum_{t=1}^N \{ \{ [C(z) - A(z)]Q(t-n, z) - zB(z)P(t-n, z) \} y^*(t+1) \}^2 = 0 \quad \text{a.s.}$$

Subtracting (28) appropriately from (29), we get

$$(30) \quad \lim_N \frac{1}{N} \sum_{t=1}^N \{ [C(z)Q(t-n, z) - B(z)R(t-n, z)] y^*(t+1) \}^2 = 0 \quad \text{a.s.}$$

Changing $t-n$ back to t in (27) and (30), we arrive at

$$(31) \quad \lim_N \frac{1}{N} \sum_{t=1}^N \{ z^{-1}[C(z) - A(z)]Q(t, z) - B(z)P(t, z) \}^2 = 0 \quad \text{a.s.,}$$

$$(32) \quad \lim_N \frac{1}{N} \sum_{t=1}^N \{ [C(z)Q(t, z) - B(z)R(t, z)] y^*(t+1) \}^2 = 0 \quad \text{a.s.}$$

Now let us treat the cases separately.

Case 1. Strong sufficient richness of order greater than or equal to $(q+s)$. This case includes the general tracking problem as well as the linear model following problem with the order of sufficient richness as shown. Since $\{y^*(t)\}$ is strongly sufficiently rich of order greater than or equal to $(q+s)$, there exist n and $\varepsilon > 0$ such that for all large t ,

$$(33) \quad \frac{1}{n} \sum_{k=t+1}^{t+n} (y^*(k+1), \dots, y^*(k-q-s+2))(y^*(k+1), \dots, y^*(k-q-s+2))^T \geq \varepsilon I_{s+q}.$$

Define

$$s_0(t) + s_1(t)z + \dots + s_{q+s-1}(t)z^{q+s-1} := S(t, z) := C(z)Q(t, z) - B(z)R(t, z).$$

Then (32) can also be written as

$$\lim_m \frac{1}{m} \sum_{j=1}^m \left\{ \frac{1}{n} \sum_{k=jn+1}^{jn+n} [S(k, z) y^*(k+1)]^2 \right\} = 0 \quad \text{a.s.}$$

Since $\lim_t \|\theta(t) - \theta(t-1)\| = 0$, we can replace $S(k, z)$ by $S(jn, z)$ to get

$$(34) \quad \lim_m \frac{1}{m} \sum_{j=1}^m \left\{ \frac{1}{n} \sum_{k=jn+1}^{jn+n} [S(jn, z) y^*(k+1)]^2 \right\} = 0 \quad \text{a.s.}$$

Define $\|S(t, z)\|^2 := \sum_{i=0}^{q+s-1} s_i^2(t)$ and (33) implies that

$$\frac{1}{n} \sum_{k=jn+1}^{jn+n} [S(jn, z) y^*(k+1)]^2 \geq \varepsilon \|S(jn, z)\|^2 \quad \text{for all large } j.$$

From (34) it follows that

$$(35) \quad \lim_m \frac{1}{m} \sum_{j=1}^m \|S(jn, z)\|^2 = 0 \quad \text{a.s.}$$

Again, since $\lim_t \|\theta(t) - \theta(t-1)\| = 0$ a.s., (35) implies that

$$(36) \quad \lim_N \frac{1}{N} \sum_{t=1}^N \|S(t, z)\|^2 = 0 \quad \text{a.s.}$$

Adding (31) and (36) gives

$$\begin{aligned} \lim_N \frac{1}{N} \sum_{t=1}^N \{z^{-1}[C(z) - A(z)]Q(t, z) - B(z)P(t, z)\}^2 \\ + \{C(z)Q(t, z) - B(z)R(t, z)\}^2 = 0 \quad \text{a.s.} \end{aligned}$$

Hence there is a common subsequence $\{t_k\}$ such that

$$(37) \quad \lim_k \{z^{-1}[C(z) - A(z)]Q(t_k, z) - B(z)P(t_k, z)\} = 0 \quad \text{a.s.}$$

and

$$(38) \quad \lim_k \{C(z)Q(t_k, z) - B(z)R(t_k, z)\} = 0 \quad \text{a.s.}$$

Since $\{\theta(t)\}$ is bounded, we can also assume without loss of generality that

$$(39) \quad \lim_k Q(t_k, z) =: Q(z); \quad \lim_k P(t_k, z) =: P(z); \quad \lim_k R(t_k, z) =: R(z) \quad \text{a.s.}$$

exist. Hence (37) and (38) imply

$$(40) \quad z^{-1}[C(z) - A(z)]Q(z) - B(z)P(z) = 0 \quad \text{a.s.,}$$

$$(41) \quad C(z)Q(z) - B(z)R(z) = 0 \quad \text{a.s.}$$

However, $Q(z)$ and $P(z)$ are polynomials of degrees less than or equal to $(q-1)$ and $(p \vee s-1)$, respectively. Hence (40) and our assumption (19)(iv) imply that

$$(42) \quad Q(z) = \xi B(z) \quad \text{and} \quad P(z) = \xi z^{-1}[C(z) - A(z)]$$

for some random scalar ξ . Then (41) also shows that $R(z) = \xi C(z)$. Moreover ξ cannot be 0, since otherwise $\lim_k \theta(t_k) = 0$, which is ruled out by (22)(iii) and (19)(vii). From (22)(iv) we obtain the desired result.

Case 2. Linear model following problem with $l < (q + s)$. Since $\lim_t (y_m(t) - y^*(t)) = 0$, we can replace $y^*(t+1)$ by $y_m(t+1)$ in (32). If $s+1 \leq l < q+s$, we shall henceforth define $G(z) := C(z)$, while if $l \leq s$, $G(z)$ is defined as previously by (15), (16). In the latter case also, from (9) and (16) we have $C(z)y_m(t+1) = G(z)y_m(t+1)$. Hence in any case,

$$(43) \quad \lim_N \frac{1}{N} \sum_{t=1}^N \{[G(z)Q(t, z) - B(z)R(t, z)]y_m(t+1)\}^2 = 0 \quad \text{a.s.}$$

Applying Lemma 2 to (43) and (31), we obtain that there is a subsequence $\{t_k\}$ such that (37) holds and also

$$\lim_k [G(z)Q(t_k, z) - B(z)R(t_k, z)] = K(z)H(z) \quad \text{a.s.}$$

Without loss of generality we can also suppose that the limits in (39) exist. Hence

$$(44) \quad G(z)Q(z) - B(z)R(z) = K(z)H(z) \quad \text{a.s.}$$

Also through (31), (40) gives (42). Substituting (42) in (44) yields

$$B(z)[\xi G(z) - R(z)] = K(z)H(z) \quad \text{a.s.}$$

Now note that by (11)(ii) all the roots of $H(z)$ are exactly on the unit circle, while all the roots of $B(z)$ are strictly outside the unit circle by (19)(i). Hence

$$\xi G(z) - R(z) = J(z)H(z) \quad \text{a.s.}$$

for some polynomial $J(z)$. However $[\xi G(z) - R(z)]$ is a polynomial of degree less than or equal to $l-1$, while $H(z)$ is a polynomial of degree exactly l . Hence

$$(45) \quad \xi G(z) - R(z) = 0 \quad \text{a.s.}$$

(42) and (45) now yield the theorem. \square

It is of interest to note that Caines and Lafortune [8] have suggested an adaptive controller which tracks $y^*(t)$ perturbed by white noise. Such a perturbed reference trajectory is strongly sufficiently rich of arbitrary large order (effectively ∞).

Having proved convergence of the parameters to $\xi\theta^0$ under the conditions of Theorem 6, we now have the following results.

THEOREM 7. (i) *In the general tracking problem suppose $\{y^*(t)\}$ is strongly sufficiently rich of order greater than or equal to $(q + s)$. Then*

$$(46) \quad \lim_t \frac{1}{\gamma_0(t)} (\alpha_1(t) - \gamma_1(t), \dots, \alpha_p(t) - \gamma_p(t), \beta_1(t), \dots, \beta_q(t), \gamma_1(t), \dots, \gamma_s(t)) \\ = (a_1, \dots, a_p, b_1, \dots, b_q, c_1, \dots, c_s) \quad \text{a.s.} \quad (\text{with } \gamma_i(t) := 0 \text{ for } i > s).$$

Thus the parameter estimates are strongly consistent. Also

$$(47) \quad \lim_t \frac{1}{\beta_1(t)} (\alpha_1(t), \dots, \alpha_{p \vee s}(t), \beta_2(t), \dots, \beta_q(t), \gamma_0(t), \dots, \gamma_s(t)) \\ = \frac{1}{b_1} (a_1 + c_1, \dots, a_{p \vee s} + c_{p \vee s}, b_2, \dots, b_q, 1, c_1, \dots, c_s) \quad \text{a.s.}$$

setting $a_i := 0$ for $i > p$ and $c_i := 0$ for $i > s$. Hence the adaptive control law (5) self-tunes to the optimal control law a.s.

(ii) In the linear model following problem with $l > s$ the results (46) and (47) continue to hold.

(iii) In the linear model following problem with $l \leq s$ we have,

$$\lim_t \frac{1}{\beta_1(t)} (\alpha_1(t), \dots, \alpha_{pvs}(t), \beta_2(t), \dots, \beta_q(t), \gamma_0(t), \dots, \gamma_{l-1}(t)) \\ = \frac{1}{b_1} (a_1 + c_1, \dots, a_{pvs} + c_{pvs}, b_2, \dots, b_q, g_0, \dots, g_{l-1})$$

setting $a_i := 0$ for $i > p$ and $c_i := 0$ for $i > s$. Here $\{g_0, \dots, g_{l-1}\}$ are defined by (15), (16). Hence the adaptive control law self-tunes to the optimal control law a.s.

7. Concluding remarks. We have proved the convergence of the parameter estimates and the self-tuning property for the *adaptive tracking* problem, justifying the name of *self-tuning trackers*.

For the *general tracking problem*, the convergence depends on whether the reference trajectory is sufficiently rich of appropriate order, as shown in Theorem 7. In the important case of reference trajectories which are not so rich, we have examined the *linear modeling problem*, and shown how one can adjust the *dimension* of the parameter estimator to the order of sufficient richness so as to obtain a self-tuning tracker. It is worth noting that the adaptive controller need not be provided with precise information such as amplitude, frequency or phases of the sinusoids in the reference trajectory. It is enough to know only the number of such components.

An important application, which is a special case of these results, is the problem of maintaining the output at a constant level, i.e., the *set-point problem*. The constant trajectory is sufficiently rich of only order 1, and only one parameter need be estimated to compensate for the colored noise and reject it optimally.

Among the outstanding problems still left unresolved are the following:

(i) Does the least squares based parameter estimation algorithm also possess the above properties? This is of vital interest because the rate of convergence of least squares based algorithms has been observed to be superior to the type of parameter estimation algorithm considered here.

(ii) What robustness properties do these types of self-tuning adaptive control laws possess?

Acknowledgment. The authors would like to thank P. Kokotovic for several stimulating discussions.

REFERENCES

- [1] K. J. ÅSTRÖM AND B. WITTENMARK, *On self-tuning regulators*, Automatica, 9 (1973), pp. 185–199.
- [2] L. LJUNG, *Analysis of recursive stochastic algorithms*, IEEE Trans. Automat. Control, AC-22 (1977), pp. 551–575.
- [3] ———, *On positive real transfer functions and the convergence of some recursive schemes*, IEEE Trans. Automat. Control, AC-22 (1977), pp. 539–550.
- [4] G. GOODWIN, P. RAMADGE AND P. CAINES, *Discrete time stochastic adaptive control*, this Journal, 19 (1981), pp. 829–853.
- [5] G. GOODWIN AND K. S. SIN, *Adaptive Filtering Prediction and Control*, Prentice-Hall, Englewood Cliffs, NJ, 1984.
- [6] A. BECKER, P. R. KUMAR AND C. Z. WEI, *Adaptive control with the stochastic approximation algorithm: Geometry and convergence*, IEEE Trans. Automat. Control, AC-30 (1985), pp. 330–338.
- [7] C. Z. WEI, *Control by stochastic approximation*, preprint, 1984.
- [8] P. CAINES AND S. LAFORTUNE, *Adaptive control with recursive identification for stochastic linear systems*, IEEE Trans. Automat. Control, AC-29 (1984), pp. 312–321.

- [9] P. CAINES AND S. P. MEYN, *The zero divisor problem of stochastic adaptive control*, preprint, August, 1984.
- [10] W. F. STOUT, *Almost Sure Convergence*, Academic Press, New York, 1974.
- [11] T. L. LAI AND C. Z. WEI, *Least squares estimates in stochastic regression with applications to identification and control of dynamic systems*, Ann. Math. Statist., 10 (1982), pp. 154–166.
- [12] P. R. KUMAR AND P. VARAIYA, *Stochastic Systems: Estimation, Identification and Adaptive Control*, Prentice-Hall, Englewood Cliffs, NJ, 1986.

SECOND-ORDER NECESSARY CONDITIONS IN CONSTRAINED SEMISMOOTH OPTIMIZATION*

ROBIN W. CHANEY†

Abstract. First- and second-order conditions are given which must necessarily be satisfied by local minimizers for certain finite-dimensional nonsmooth nonlinear programming problems. The problems considered are of standard form, having a finite number of equality and inequality constraints. The principal result does not require a constraint qualification, but does require that the functions be semismooth at the minimizer. The necessary conditions are stated in terms of the generalized gradients of nonsmooth analysis and certain second-order directional derivatives.

Key words. nonsmooth analysis, second-order necessary conditions

1. Introduction. Let $f, g_1, g_2, \dots, g_m, \dots, g_{m+p}$ be real-valued locally Lipschitzian functions on an open set W in n -dimensional real Euclidean space R^n . We consider the problem

P : Minimize $f(x)$, over all x in W such that

$$g_i(x) \leq 0 \quad \text{for } i = 1, \dots, m, \quad \text{and}$$

$$g_i(x) = 0 \quad \text{for } i = m+1, \dots, m+p.$$

Let S be the set of all points in W which are feasible for problem P . Suppose that x^* belongs to S and that the functions f, g_1, \dots, g_{m+p} are all semismooth at x^* [12, p. 961]. We present here a theorem which gives first- and second-order conditions which must necessarily be satisfied if x^* is a local solution to problem P . The theorem is stated in terms of certain Lagrangian functions; it requires no constraint qualification, because it employs a John-type Lagrangian rather than one of the Karush-Kuhn-Tucker type. It generalizes to "semismooth optimization" a theorem due to Ben-Tal [1, Thm. 3.2]; see also Ioffe [10, Thm. 6]. The proof uses a necessity theorem for unconstrained semismooth minimization derived in [5] and a strong version of the Lagrange multiplier rule (see [9, Thm. 6.1.1] and [13, Thm. 1]). In this paper, we shall also discuss a constraint qualification which involves the assumption that there is a unique "normalized" Lagrange multiplier at x^* . Finally, we give an example which illustrates an interesting distinction between the necessary conditions for unconstrained problems and those for constrained problems.

The results presented here are expressed in terms of the generalized gradients of nonsmooth analysis and certain second-order directional derivatives. For a detailed, systematic exposition of nonsmooth analysis, the reader should consult Clarke [9], to which we often refer. An interesting account of the subject (with few proofs) is given by Rockafellar [14]. Information about the second-order directional derivatives used here can be found in [4]; here, we merely give the definitions of these concepts.

We develop elsewhere two second-order sufficiency theorems for problem P which serve as companions to the results of this paper. One of these sufficiency theorems [7, Thm. 3] is a close complement to our necessity theorem, although it is more limited in scope because certain Lagrangian functions are required to be regular at x^* . The sufficiency theorems for problem P given in [7] differ substantially from those proved earlier by the author in [6] and [3].

* Received by the editors February 24, 1986; accepted for publication (in revised form) August 15, 1986.

† Department of Mathematics, Western Washington University, Bellingham, Washington 98225.

Ben-Tal and Zowe [2] have developed a general theory of necessary conditions. It is difficult to compare the work in [2] with the present work. Ben-Tal and Zowe do not use nonsmooth analysis (as set forth in [9]) and their point of view is very different from ours.

We turn now to some basic definitions drawn from earlier work. We continue to suppose that $W, f, g_1, \dots, g_{m+p}, x^*$, and S are specified as above.

DEFINITION 1. Let u be a vector in R^n . Suppose that the sequence $\{x_k\}$ in W converges to x^* , with $x_k \neq x^*$ for each k . Then we say that $\{x_k\}$ converges to x^* in direction u in case it is true that the sequence $\{|u|(x_k - x^*)/|x_k - x^*|\}$ converges to u .

DEFINITION 2. Let u be a vector in R^n . The set $\partial_u f(x^*)$ is defined to be the set of all v in R^n for each of which there exist sequences $\{x_k\}$ and $\{v_k\}$ such that $\{x_k\}$ converges to x^* in direction u , $\{v_k\}$ converges to v , and v_k belongs to $\partial f(x_k)$ for each k .

(Here, $\partial f(x)$ stands for the Clarke subdifferential [9, p. 27] of f at x . Note that $\partial_u f(x^*) \subseteq \partial f(x^*)$. The set $\partial_u f(x^*)$ can be said to consist of those generalized gradients of f at x^* which "come to x^* from" direction u .)

DEFINITION 3. The function f is semismooth at x^* if it is true that the sequence $\{v_k \cdot u\}$ converges whenever $\{x_k\}$ and $\{v_k\}$ are sequences such that $\{x_k\}$ converges to x^* in direction u and v_k belongs to $\partial f(x_k)$ for every k .

Remarks. The concept of "semismoothness" is due to Mifflin [12]. Mifflin has shown [12] that if f is semismooth at x^* , then, for each direction u , the classical directional derivative $f'(x^*; u)$ exists and is equal to the limit of every sequence $\{v_k \cdot u\}$ formed as in Definition 3 above. Mifflin has also proved [12] that convex functions, C^1 functions, and certain pointwise maxima of C^1 functions are semismooth. Furthermore, the set of semismooth functions is closed under addition, scalar multiplication, maximization over finite sets, and (in most cases) composition [12, p. 967].

Theorems 1 and 2 below are given in terms of certain second-order directional derivatives, as defined and discussed in [4]. We repeat the basic definitions here.

DEFINITION 4. Let u be a vector in R^n and suppose that v belongs to $\partial_u f(x^*)$. Then $f''_-(x^*, v, u)$ is defined to be the infimum of all numbers

$$\liminf [f(x_k) - f(x^*) - v \cdot (x_k - x^*)]/t_k^2,$$

taken over all triples of sequences $\{x_k\}$, $\{v_k\}$, and $\{t_k\}$ for which

- (a) $t_k > 0$ for each k and $\{x_k\}$ converges to x^* ,
- (b) $\{t_k\}$ converges to 0 and $\{(x_k - x^*)/t_k\}$ converges to u ,
- (c) $\{v_k\}$ converges to v with v_k in $\partial f(x_k)$ for each k .

Similarly, we define $f''_+(x^*, v, u)$ to be the supremum of all numbers

$$\limsup [f(x_k) - f(x^*) - v \cdot (x_k - x^*)]/t_k^2,$$

taken over all triples of sequences $\{x_k\}$, $\{v_k\}$, and $\{t_k\}$ for which (a), (b), and (c) all hold.

Remarks. Note that f''_- and f''_+ depend not only on x^* and u but also on the vector v in $\partial_u f(x^*)$. We term $f''_-(x^*, v, u)$ and $f''_+(x^*, v, u)$ the lower and upper (respectively) second-order directional derivatives of f at x^* and v in the direction u . More information about these concepts is found in [4].

2. The main theorem. In this section, we derive necessary conditions for optimality for problem P . We let W, f, g_i, x^* , and S be as stated in the Introduction.

We state a necessity theorem for unconstrained minimization, which is proved in [5].

Let D^* be the set of all unit vectors u^* in R^n for each of which there exists $\delta' > 0$ such that $v \cdot u^* \leq 0$ if u is a unit vector with $|u - u^*| \leq \delta'$ and if v is in $\partial_u f(x^*)$.

Let f^0 denote (as usual) the Clarke directional derivative. Given a unit vector u^* in R^n , we find that if $f^0(x^*; u^*) \leq 0$ then u^* belongs to D^* ; this is true because $f^0(x^*; \cdot)$ is the support function of $\partial f(x^*)$. Moreover, if f is semismooth at x^* and if u^* belongs to D^* , then we have $f'(x^*; u^*) \leq 0$.

THEOREM 1. *Suppose that f is semismooth at x^* in W and that x^* is an unconstrained local minimizer for $f(x)$, over x in W . If u belongs to D^* , then 0 belongs to $\partial_u f(x^*)$ and $f''_-(x^*, 0, u) \geq 0$.*

Remarks. There will be an application of Theorem 1 in the proof of Theorem 2 below. In order to apply Theorem 1 to a solution x^* to problem P , we must devise a related unconstrained problem to which x^* provides a local solution. This can be done in many ways; but the appropriate method here is to follow an approach used by Clarke [9, p. 229]. We must first make a number of definitions.

We let T be the set of all $w = (w_0, \dots, w_m, \dots, w_{m+p})$ in R^{1+m+p} such that $w_i \geq 0$ for $i = 0, \dots, m$ and $\sum_{i=0}^{m+p} w_i^2 = 1$. We let x^* belong to W (as always) and we put $g_0(x) = f(x) - f(x^*)$. Given x in W and w in T , we let

$$L(x, w) = w_0 f(x) + \sum_{i=1}^{m+p} w_i g_i(x)$$

and

$$L^*(x, w) = \sum_{i=0}^{m+p} w_i g_i(x) = L(x, w) - w_0 f(x^*).$$

For each x in W , we put

$$(1) \quad G(x) = \max \{L^*(x, w) : w \in T\}$$

and

$$T(x) = \{w \in T : G(x) = L^*(x, w)\}.$$

We denote by $\partial L(x, w)$ and $\nabla L(x, w)$, respectively, the subdifferential and gradient of the function $L(\cdot, w)$ at x ; we attach the analogous meaning to $\partial L^*(x, w)$ and $\nabla L^*(x, w)$. We also denote by $L''_+(x^*, w, v, u)$ the upper second-order directional derivative of the function $L(\cdot, w)$ at x^* and v in the direction u .

Next, let us denote by $M(x^*)$ the set of all Lagrange multipliers in T for problem P at x^* ; i.e., $M(x^*)$ is the set of all w in T such that 0 belongs to $\partial L(x^*, w)$ and $w_i g_i(x^*) = 0$ for all $i = 1, \dots, m$.

Finally, let $D^*(G)$ be the set of all unit vectors u^* in R^n for each of which there exists $\delta' > 0$ such that $v \cdot u^* \leq 0$ if u is a unit vector with $|u - u^*| \leq \delta'$ and if v belongs to $\partial_u G(x^*)$.

Thus, again, we note that if u is a unit vector for which $G^0(x^*; u) \leq 0$ then u belongs to $D^*(G)$. And, if u belongs to $D^*(G)$ and if G is semismooth at x^* then $G'(x^*; u) \leq 0$.

THEOREM 2. *Suppose that x^* provides a local solution to problem P and suppose that all the functions f and g_i are semismooth at x^* . Let u belong to $D^*(G)$. Then there exists a Lagrange multiplier w^* in $M(x^*)$ such that 0 belongs to $\partial_u L(x^*, w^*)$, $w_i^* g_i'(x^*; u) = 0$ for all $i = 1, \dots, m$, and $L''_+(x^*, w^*, 0, u) \geq 0$.*

Remarks. We shall discuss the set $D^*(G)$ later. We shall also give an example which will show that, although we can infer $f'' \geq 0$ in Theorem 1, we must settle for $L''_+ \geq 0$ in Theorem 2.

LEMMA 1. (i) If $G(x) > 0$ for x in W then the set $T(x)$ has a single element $w_x = (w_{x0}, \dots, w_{xm}, \dots, w_{x,m+p})$ given by

$$w_{xi} = [\max(0, g_i(x))]/G(x) \quad \text{for } i = 0, \dots, m$$

and

$$w_{xi} = g_i(x)/G(x) \quad \text{for } i = m+1, \dots, m+p.$$

If $G(x) \geq 0$, we have

$$(2) \quad G(x)^2 = \sum_{i=0}^m [\max(0, g_i(x))]^2 + \sum_{i=m+1}^{m+p} g_i(x)^2.$$

(ii) If x^* is a local solution to problem P , then $G(x^*) = 0$ and $G(x) \geq 0$ for all x near x^* .

(iii) x^* is a strict local solution to problem P (i.e., $f(x) > f(x^*)$ for all feasible x near x^* for which $x \neq x^*$) if and only if we have $G(x) > 0 = G(x^*)$ for all x near x^* with $x \neq x^*$.

(iv) If $G(x) > 0$ then $\partial G(x) \subseteq \partial L(x, w_x)$, with w_x as in (i).

Proof. Assertions (ii) and (iii) are easy to verify. We observe also that (iv) is an immediate consequence of (i) and [9, Thm. 2.8.2]. It remains therefore to prove (i). Fix x in W such that $G(x) > 0$ and let w belong to $T(x)$. By the standard Kuhn-Tucker theorem, we infer that multipliers t, s_0, \dots, s_m exist so that $s_i \geq 0$ and $s_i w_i = 0$ for each $i = 0, 1, \dots, m$ and so that

$$(3) \quad -g_i(x) + 2tw_i - s_i = 0 \quad \text{for } i = 0, \dots, m$$

and

$$(4) \quad -g_i(x) + 2tw_i = 0 \quad \text{for } i = m+1, \dots, m+p.$$

Since $\sum_{i=0}^{m+p} w_i^2 = 1$, we infer from (3) and (4) that

$$2t = L^*(x, w) = G(x)$$

and hence that t is positive. Therefore, $w_i = g_i(x)/G(x)$ for $i = m+1, \dots, m+p$. It follows readily that, for $i = 0, \dots, m$, we have $w_i = 0$ if $g_i(x) < 0$ and $w_i = g_i(x)/G(x)$ if $g_i(x) \geq 0$. Thus, $w = w_x$, where w_x is as defined above. Equation (2) follows at once from the formulas for w_x and the equation $G(x) = L^*(x, w_x)$.

Equation (2) also follows readily if $G(x) = 0$.

LEMMA 2. Suppose that x^* provides a local solution to problem P and that the functions f and g_i are all semismooth at x^* . Then the function G is semismooth at x^* .

Proof. Suppose that the sequence $\{x_k\}$ converges to x^* in direction u , with $|u| = 1$. Suppose also that V_k belongs to $\partial G(x_k)$ for each k . We must show that the sequence $\{V_k \cdot u\}$ converges. We know here that $x_k \neq x^*$ for all k . We let $I(x^*)$ denote the set of all $i = 0, \dots, m$ for which $g_i(x^*) = 0$. By Lemma 1, we can assume that $G(x_k) \geq 0$.

We must consider two cases. In the first case, we suppose that $g'_i(x^*; u) \leq 0$ for all i in $I(x^*)$ and $g'_i(x^*; u) = 0$ for all $i = m+1, \dots, m+p$. We proceed with the first case.

We "partition" the sequence $\{V_k \cdot u\}$ into two subsequences, one in which $G(x_k) > 0$ for all k and the other in which $G(x_k) = 0$ for all k ; we shall prove that both sequences (if they are infinite) converge to 0 and so it will follow that $\{V_k \cdot u\}$ converges to 0 in this first case.

So, we treat first the "subcase" in which we assume $G(x_k) > 0$ for all k . According to Lemma 1(i), there exists a unique $w_k = (w_{k0}, \dots, w_{km}, \dots, w_{k,m+p})$ in $T(x_k)$, with

w_{ki} satisfying the formulas given in Lemma 1(i). Furthermore, it follows from Lemma 1(iv) and [9, Prop. 2.3.3] that there exist v_{ki} in $\partial g_i(x_k)$ such that

$$(5) \quad V_k = \sum_{i=0}^{m+p} w_{ki} v_{ki}.$$

Notice that if $g_i(x^*) < 0$ then $w_{ki} = 0$ for all large k . Since each g_i is semismooth at x^* , we infer that the sequence $\{v_{ki} \cdot u\}_k$ converges to $g'_i(x^*; u)$ for each i . If $g'_i(x^*; u) = 0$, then surely $\{w_{ki}(v_{ki} \cdot u)\}_k$ converges to 0. But, if $g'_i(x^*; u) < 0$ for some i then $g_i(x_k) < 0$ for all large k and so (by Lemma 1) $w_{ki} = 0$ for all large k ; hence $\{w_{ki}(v_{ki} \cdot u)\}_k$ converges to 0. We infer from (5) that $\{V_k \cdot u\}$ must converge to 0, in this first subcase.

Next, we treat the subcase in which $G(x_k) = 0$ for all k . Again, we must show that $\{V_k \cdot u\}$ converges to 0. Let E be the set of those points in W at which at least one of f and g_i is not differentiable. Then E has measure 0. By [9, Thm. 2.8.6] and Caratheodory's theorem, we can write

$$V_k = \sum_{j=1}^{n+1} a_{kj} z_{kj}$$

where $a_{kj} \geq 0$, $\sum_{j=1}^{n+1} a_{kj} = 1$, and where, for each k and j , there are sequences $\{x_{kjq}\}_q$ in $W - E$ and $\{w_{kjq}\}_q$ in T such that $\{x_{kjq}\}_q$ converges to x_k , $\{L^*(x_k, w_{kjq})\}_q$ converges to $G(x_k) = 0$, and $\{\nabla L^*(x_{kjq}, w_{kjq})\}_q$ converges to z_{kj} . Notice that if $g_I(x_k) < 0$ for some I and k then $L^*(x_k, w_{kjq}) \leq w_{kjq, I} g_I(x_k) \leq 0$ and so the sequence $\{w_{kjq, I}\}_q$ must converge to 0. For each k and j , we can therefore select x_{kj} in $W - E$ and w_{kj} in T so that

$$|x_{kj} - x_k| \leq \frac{|x_k - x^*|^2}{k}, \quad 0 \leq G(x_k) - L^*(x_k, w_{kj}) \leq \frac{1}{k},$$

$$0 \leq w_{kj, i} < \frac{1}{k} \quad \text{if } g_i(x_k) < 0,$$

and

$$\left| V_k - \sum_{j=1}^{n+1} a_{kj} \nabla L(x_{kj}, w_{kj}) \right| < \frac{1}{k}.$$

It follows that, for each j , $\{x_{kj}\}_k$ converges to x^* in the direction u . Hence, the sequence $\{\nabla g_i(x_{kj}) \cdot u\}_k$ must converge to $g'_i(x^*; u)$, since g_i is semismooth at x^* . Now, if $g'_i(x^*; u) = 0$, it follows then that the sequence $\{a_{kj} w_{kj, i} \nabla g_i(x_{kj}) \cdot u\}_k$ converges to 0. If $g'_i(x^*; u) < 0$ or if $g_i(x^*) < 0$ then $g_i(x_k) < 0$ for all large k and for all j and so $\{w_{kj, i}\}_k$ converges to 0; therefore the sequence $\{a_{kj} w_{kj, i} \nabla g_i(x_{kj}) \cdot u\}_k$ converges to 0. It follows from all of these computations that $\{V_k \cdot u\}$ converges to 0. We have at last completed the proof for the first case.

We must now show that $\{V_k \cdot u\}$ converges if $g'_i(x^*; u) > 0$ for some i in $I(x^*)$ or if $g'_i(x^*; u) \neq 0$ for some $i > m$. It follows that $G(x_k) > 0$ for large k and so we assume $G(x_k) > 0$ for all k . Hence we again have w_k and v_{ki} so that (5) holds. From (2) and Lemma 1(ii), we infer

$$(6) \quad [G'(x^*; u)]^2 = [\max(0, f'(x^*; u))]^2 \\ + \sum_{i \in I(x^*)} [\max(0, g'_i(x^*; u))]^2 + \sum_{i=m+1}^{m+p} [g'_i(x^*; u)]^2.$$

So, we have $G'(x^*; u) > 0$; hence, Lemma 1(i) implies that $\{w_{ki}\}_k$ converges to $[\max(0, g'_i(x^*; u))]/G'(x^*; u)$ for i in $I(x^*)$, that $\{w_{ki}\}_k$ converges to

$g'_i(x^*; u)/G'(x^*; u)$ for $i > m$, and that $\{w_{ki}\}_k$ converges to 0 if $g_i(x^*) < 0$. Thus, in this second case, we find that $\{V_k \cdot u\}$ converges to $Q/G'(x^*; u)$, where

$$Q = \sum_{i \in I(x^*)} [g'_i(x^*; u) \max(0, g'_i(x^*; u))] + \sum_{i=m+1}^{m+p} [g'_i(x^*; u)]^2.$$

LEMMA 3. Suppose that x^* is a local solution to problem P. Suppose that x belongs to W , w^* belongs to T , $G(x) > 0$, and that $w_i^* = 0$ for all $i = 0, \dots, m$ for which $g_i(x) < 0$. Then, with w_x as in Lemma 1, we have

$$\frac{1}{2}|w_x - w^*|^2 = 1 - L^*(x, w^*)/L^*(x, w_x).$$

Proof. Suppose x and w are as above. We have

$$|w_x - w^*|^2 = |w_x|^2 - 2(w_x \cdot w^*) + |w^*|^2 = 2 - 2(w_x \cdot w^*).$$

Because of our assumptions about x and w^* and because of Lemma 1(i), this last equation simplifies to

$$|w_x - w^*|^2 = 2 - 2L^*(x, w^*)/L^*(x, w_x).$$

With these lemmas established, we can now turn to the proof of the theorem.

Proof of Theorem 2. It follows from Lemma 1 that x^* is a local minimizer for $G(x)$. Let u be a unit vector which belongs to the set $D^*(G)$. Because of Lemma 2, we can apply Theorem 1. We infer that 0 belongs to $\partial_u G(x^*)$ and so there exist sequences $\{x_k\}$ and $\{v_k\}$ such that $\{x_k\}$ converges to x^* in direction u , v_k belongs to $\partial G(x_k)$ for each k , and $\{v_k\}$ converges to 0. By passing to subsequences, we can assume $G(x_k) > 0$ for all k or $G(x_k) = 0$ for all k .

Case 1. Assume $G(x_k) > 0$ for all k .

In this case, it follows from Lemma 1 that v_k belongs to $\partial L^*(x_k, w_k) = \partial L(x_k, w_k)$, where w_k is the unique member of $T(x_k)$. We may assume that $\{w_k\}$ converges to w^* in T . There is a positive number M so that, given k , there exists v_k^* in $\partial L(x_k, w^*)$ so that $|v_k - v_k^*| \leq M|w_k - w^*|$. Hence $\{v_k^*\}$ converges to 0 and so 0 belongs to $\partial_u L(x^*, w^*)$. It follows from Lemma 1(i) that $w_i^* g_i(x^*) = 0$ for $i = 1, \dots, m$ and so w^* belongs to $M(x^*)$. From (6), we find that

$$|g'_i(x^*; u)| \leq G'(x^*; u) = 0 \quad \text{for } i > m$$

and

$$g'_i(x^*; u) \leq G'(x^*; u) = 0 \quad \text{for } i \text{ in } I(x^*).$$

Given $i = 1, \dots, m$ with $w_i^* > 0$, we have $w_{ki} > 0$ for large k and so $g_i(x_k) > 0$ for all large k ; therefore, $g'_i(x^*; u) = 0$. And, if $g_i(x^*) < 0$, then Lemma 1 shows that $w_{ki} = 0$ for all large k . We now infer from Lemma 3 that

$$(7) \quad \frac{1}{2}|w_k - w^*|^2 = 1 - \frac{L(x_k, w^*) - L(x^*, w^*)}{L(x_k, w_k) - L(x^*, w_k)} \quad \text{for large } k.$$

Of course, because of Lemma 1, we have

$$(8) \quad L(x_k, w_k) - L(x^*, w_k) = G(x_k) > 0.$$

Since $\{w_k\}$ converges to w^* , it follows from (7) and (8) that

$$(9) \quad L(x_k, w^*) - L(x^*, w^*) > 0 \quad \text{for all large } k.$$

Since $\{v_k^*\}$ converges to 0 and v_k^* belongs to $\partial L(x_k, w^*)$ for each k , we infer from (9) and Definition 4 that $L_+''(x^*, w^*, 0, u) \geq 0$.

Case 2. Assume $G(x_k) = 0$ for all k .

In this case, we abandon the v_k mentioned above. Instead, we observe that we may assume that each x_k is a local solution to problem P . According to the Lagrange multiplier rule (see [9, Thm. 6.1.1] or [13, Thm. 1]), there exists w_k in T so that 0 belongs to $\partial L(x_k, w_k)$ and so that $w_{ki}g_i(x_k) = 0$ for all $i = 1, \dots, m$. We may assume that $\{w_k\}$ converges to w^* in T . As in Case 1, we obtain v_k^* in $\partial L(x_k, w^*)$ such that $|v_k^*| \leq M|w_k - w^*|$. We infer that $\{v_k^*\}$ converges to 0, that 0 belongs to $\partial_u L(x^*, w^*)$, and that w^* belongs to $M(x^*)$. Moreover, since each x_k is feasible for problem P and $w_{ki}g_i(x_k) = 0$ for $i = 1, \dots, m$, it follows that $w_i^*g_i'(x^*; u) = 0$. Finally, we observe that

$$L(x_k, w^*) - L(x^*, w^*) = w_0^*(f(x_k) - f(x^*)) = 0$$

for all large k and so $L_+''(x^*, w^*, 0, u) \geq 0$ in this case also.

PROPOSITION 1. *Suppose that x^* is a local solution to problem P and that the functions f and g_i are all semismooth at x^* . Let $D(x^*)$ be the set of all unit vectors u in R^n for which $f'(x^*; u) \leq 0$, $g_i'(x^*; u) \leq 0$ for all positive i in $I(x^*)$, and $g_i'(x^*; u) = 0$ for all $i = m+1, \dots, m+p$. Then $D^*(G) \subseteq D(x^*)$. If the functions f, g_1, \dots, g_m are also regular [9, p. 39] at x^* and if the functions g_{m+1}, \dots, g_{m+p} are strictly differentiable [9, p. 30] at x^* , then $D^*(G) = D(x^*)$.*

Proof. In view of Lemma 1(ii), we know that $G'(x^*; u)$ is nonnegative for all u . Hence the relation $D^*(G) \subseteq D(x^*)$ follows from (6).

Suppose now that f, g_1, \dots, g_m are also regular at x^* and that g_{m+1}, \dots, g_{m+p} are strictly differentiable at x^* . It follows from (2), [8, Thm. 2.1], and [9, Prop. 2.3.6] that G is regular at x^* . Therefore, if u belongs to $D(x^*)$ then $G^0(x^*; u) = G'(x^*; u) = 0$ and so u belongs to $D^*(G)$.

Remarks. Proposition 1 gives a precise description of the set $D^*(G)$ provided the functions f and g_i satisfy the additional hypotheses concerning regularity and strict differentiability.

If the functions f and g_i are of class C^2 on W , then in Theorem 2 we have

$$L_+''(x^*, w^*, 0, u) = L_-''(x^*, w^*, 0, u) = \frac{1}{2}u \cdot \nabla_{xx}^2 L(x^*, w^*)u.$$

Hence, Proposition 1 shows that Theorem 2 reduces to the theorem of Ben-Tal [1, Thm. 3.2] in this case; see also Ioffe [10, Thm. 6].

COROLLARY 1. *Suppose that x^* provides a local solution to problem P and suppose that the functions f and g_i are all semismooth at x^* . Suppose also that $M(x^*)$ consists of a single vector w^* .*

If u belongs to $D^(G)$, then 0 belongs to $\partial_u L(x^*, w^*)$, $w_i^*g_i'(x^*; u) = 0$ for $i = 1, \dots, m$ and $L_+''(x^*, w^*, 0, u) \geq 0$.*

Remarks. The uniqueness of a "normalized" Lagrange multiplier w^* thus serves as a constraint qualification. To assure that the "0th component" w_0^* is positive, it appears that we need an additional assumption such as calmness [9, p. 240].

The observation just made provides an extension to semismooth optimization of a situation well understood in the case in which f and g_i are of class C^2 on W . So, let f and g_i be of class C^2 on W and suppose x^* is a local solution to P . Kyparisis has shown [11] that the set $\{w \in M(x^*): w_0 > 0\}$ consists of a single element if and only if the problem P and x^* satisfy what Kyparisis terms the *strict* Mangasarian-Fromovitz constraint qualification. Now the strict Mangasarian-Fromovitz constraint qualification implies the usual Mangasarian-Fromovitz constraint qualification [11], which in turn implies that problem P is calm at x^* [9, Cor. 5, p. 244].

To sum up, we find in the C^2 case that the combined requirement of calmness at x^* and uniqueness of the normalized Lagrange multiplier is equivalent to the requirement that the strict Mangasarian-Fromovitz constraint qualification hold for problem P at x^* .

3. An example. We conclude with a specific example which shows that in Theorem 2 we cannot replace L'_+ by L'_- ; this contrasts, of course, with the situation for unconstrained minimization described in Theorem 1. The same example illustrates an interesting point about uniqueness of the Lagrange multiplier.

Example 1. We define seven sets, whose union is R^2 . These sets do not "overlap" in the sense that their interiors are pairwise disjoint. We put

$$\begin{aligned} A_1 &= \{(x, y): x > 0 \text{ and } y \geq x^2\}, \\ A_2 &= \{(x, y): x > 0 \text{ and } 0 \leq y \leq x^2\}, \\ A_3 &= \{(x, y): x > 0 \text{ and } 0 \geq y \geq -x^2\}, \\ A_4 &= \{(x, y): x > 0 \text{ and } y \leq -x^2\}, \\ A_5 &= \{(x, y): x \leq 0 \text{ and } y > 0\}, \\ A_6 &= \{(x, y): x \leq 0 \text{ and } y < 0\}, \\ A_7 &= \{(x, y): x \leq 0 \text{ and } y = 0\}. \end{aligned}$$

The reader may find it helpful to sketch a graph of these seven sets. We define functions g_1 and f on R^2 by setting

$$g_1(x, y) = -y \quad \text{for all } (x, y) \text{ in } R^2$$

and

$$\begin{aligned} f(x, y) &= y - x^2 && \text{if } (x, y) \text{ belongs to } A_1, \\ f(x, y) &= 0 && \text{if } (x, y) \text{ belongs to } A_2 \cup A_6 \cup A_7, \\ f(x, y) &= y && \text{if } (x, y) \text{ belongs to } A_3 \cup A_5, \\ f(x, y) &= -x^2 && \text{if } (x, y) \text{ belongs to } A_4. \end{aligned}$$

It is clear that f is locally Lipschitzian on R^2 ; moreover, f is "piecewise C^2 " [5] near all points of R^2 . Use of [9, Prop. 2.2.4] and [9, Thm. 2.5.1] shows that

$$\begin{aligned} \partial f(x, y) &= \{(-2x, 1)\} && \text{if } (x, y) \text{ belongs to int } A_1, \\ \partial f(x, y) &= \{(0, 0)\} && \text{if } (x, y) \text{ belongs to } A_6 \cup \text{int } A_2, \\ \partial f(x, y) &= \{(0, 1)\} && \text{if } (x, y) \text{ belongs to } A_5 \cup \text{int } A_3, \\ \partial f(x, y) &= \{(-2x, 0)\} && \text{if } (x, y) \text{ belongs to int } A_4, \\ \partial f(x, y) &= \{(0, t): 0 \leq t \leq 1\} && \text{if } (x, y) \in A_7 \cup (A_2 \cap A_3), \\ \partial f(x, y) &= \{(-2tx, t): 0 \leq t \leq 1\} && \text{if } (x, y) \in A_1 \cap A_2, \\ \partial f(x, y) &= \{(-2tx, 1-t): 0 \leq t \leq 1\} && \text{if } (x, y) \in A_3 \cap A_4. \end{aligned}$$

From this complete information about the subdifferential of f , it follows that f is regular and semismooth at the origin.

We now form the problem P of minimizing $f(x, y)$ over all (x, y) in R^2 such that $-y = g_1(x, y) \leq 0$. It is clear that the origin $x^* = (0, 0)$ provides a local solution to problem P and so Theorem 2 is applicable here. Let $u = (1, 0)$. In view of Proposition 1, we know that u belongs to $D^*(G)$.

We now wish to identify those $w = (w_0, w_1)$ in $M(x^*)$ for which $(0, 0)$ belongs to $\partial_u L(x^*, w)$. We shall show also that for all such w we have $L''(x^*, w, 0, u) < 0$.

Thus, we suppose now that w is such that $(0, 0)$ belongs to $\partial_u L(x^*, w)$. Then there exists a sequence $\{(x_k, y_k)\}$ converging to $x^* = (0, 0)$ in direction $u = (1, 0)$ and a sequence $\{v_k\}$ converging to $(0, 0)$ such that v_k belongs to $\partial L(x_k, y_k, w)$ for each k . We may assume that $x_k > 0$ for every k . By passing to subsequences, we find that we may restrict our attention to the following *seven* cases:

Case 1. (x_k, y_k) belongs to $\text{int } A_1$ for all k .

Case 2. (x_k, y_k) belongs to $\text{int } A_2$ for all k .

Case 3. (x_k, y_k) belongs to $\text{int } A_3$ for all k .

Case 4. (x_k, y_k) belongs to $\text{int } A_4$ for all k .

Case 5. (x_k, y_k) belongs to $A_1 \cap A_2$ for all k .

Case 6. (x_k, y_k) belongs to $A_2 \cap A_3$ for all k .

Case 7. (x_k, y_k) belongs to $A_3 \cap A_4$ for all k .

We shall discuss here only the analyses of Cases 5 and 7. The interested reader can verify that no new information (in the sense to be explained below) is obtained in the other five cases.

Case 5. Here we have for each k ,

$$v_k = w_0(-2t_k x_k, t_k) + w_1(0, -1) \quad \text{with } 0 \leq t_k \leq 1.$$

Hence, the fact that $\{v_k\}$ converges to $(0, 0)$ implies $\{t_k\}$ must converge to some s in the closed interval $[0, 1]$; and, we have $w_1 = s w_0$. Since $w_0^2 + w_1^2 = 1$, we infer $w_0 \neq 0$. And, since $y_k = x_k^2$ for all k , we have

$$L(x_k, y_k, w) = w_0(y_k - x_k^2) + s w_0(-y_k) = -s w_0 x_k^2.$$

Hence, $L''(x^*, w, 0, u) = -s w_0 < 0$ if $0 < s \leq 1$.

Case 7. Here we have, for each k ,

$$v_k = w_0(-2t_k x_k, 1 - t_k) + w_1(0, -1) \quad \text{with } 0 \leq t_k \leq 1.$$

The fact that $\{v_k\}$ converges to $(0, 0)$ implies that $\{t_k\}$ converges to some c in $[0, 1]$ and $w_1 = (1 - c)w_0$. (Thus, we have obtained the *same* multipliers w as in Case 5.) Again, we have $w_0 \neq 0$, and, since $y_k = -x_k^2$ for all k , we have

$$L(x_k, y_k, w) = w_0(-x_k^2) + (1 - c)w_0(-y_k) = -c w_0 x_k^2.$$

Hence, if $c > 0$, we have $L''(x^*, w, 0, u) < 0$.

Thus far, we have found that the multipliers w in $M(x^*)$ for which 0 belongs to $\partial_u L(x^*, w)$ are of the form $(w_0, s w_0)$, with $0 \leq s \leq 1$ and $w_0^2 + (s w_0)^2 = 1$. Furthermore, for all such multipliers w , we have $L''(x^*, w, 0, u) < 0$. The interested reader will find that the analyses of the other five cases produce *no* new multipliers w for which $(0, 0)$ belongs to $\partial_u L(x^*, w)$. We can therefore conclude that in Theorem 2 we cannot, in general, replace L'_+ by L''_- .

Remarks. One other conclusion can be drawn from this example. This second conclusion is somewhat disquieting, but not surprising. Notice that the constraints here could hardly be simpler: We have a single (active) constraint given by a linear function having a nonzero gradient. If the function f were of class C^1 on R^2 , we would know that the set $M(x^*)$ is a singleton. But, in Example 1, we have seen that $M(x^*)$ contains all vectors of the form $(w_0, s w_0)$ with $0 \leq s \leq 1$ and $w_0^2 + (s w_0)^2 = 1$. So, $M(x^*)$ can be a "large" set even with the simplest of constraints. Therefore, this example suggests the possibility that the constraint qualification mentioned following Corollary 1 may be much more restrictive in the semismooth case than it is in the smooth case.

We can also examine the question of the restrictive character of our constraint qualification from another point of view. Note that if x^* is an isolated local solution to problem P and if $M(x^*)$ consists of one vector w^* , with $w_0^* > 0$, then we infer from [9, Cor. 2, p. 242] that the "value" function V is strictly differentiable (at the relevant point). Of course, it is well known that, even in the smooth case, the function V can be nondifferentiable.

REFERENCES

- [1] A. BEN-TAL, *Second-order and related extremality conditions in nonlinear programming*, J. Optim. Theory Appl., 31 (1980), pp. 143–165.
- [2] A. BEN-TAL AND J. ZOWE, *A unified theory of first and second order conditions for extremum problems in topological vector spaces*, Math. Programming Study, 19 (1982), pp. 39–76.
- [3] R. W. CHANEY, *A general sufficiency theorem for nonsmooth nonlinear programming*, Trans. Amer. Math. Soc., 276 (1983), pp. 235–245.
- [4] ———, *Second-order directional derivatives for nonsmooth functions*, J. Math. Anal. Appl., to appear.
- [5] ———, *Second-order necessary conditions in semismooth optimization*, Math. Programming, to appear.
- [6] ———, *Second-order sufficiency conditions for nondifferentiable programming problems*, this Journal, 20 (1982), pp. 20–33.
- [7] ———, *Second order sufficient conditions in nonsmooth optimization*, submitted for publication, 1985.
- [8] F. H. CLARKE, *Generalized gradients and applications*, Trans. Amer. Math. Soc., 205 (1975), pp. 247–262.
- [9] ———, *Optimization and Nonsmooth Analysis*, Wiley-Interscience, New York, 1983.
- [10] A. D. IOFFE, *Necessary and sufficient conditions for a local minimum 3: Second-order conditions and augmented duality*, this Journal, 17 (1979), pp. 266–288.
- [11] J. KYPARISIS, *On uniqueness of Kuhn-Tucker multipliers in nonlinear programming*, Math. Programming, 32 (1985), pp. 242–246.
- [12] R. MIFFLIN, *Semismooth and semiconvex functions in constrained optimization*, this Journal, 15 (1977), pp. 959–972.
- [13] R. T. ROCKAFELLAR, *Lagrange multipliers and subderivatives of optimal value functions in nonlinear programming*, Math. Programming Study, 17 (1982), pp. 28–66.
- [14] ———, *The Theory of Subgradients and its Applications: Convex and Nonconvex Functions*, Helderman Verlag, West Berlin, 1981.

SPLINE APPROXIMATION FOR RETARDED SYSTEMS AND THE RICCATI EQUATION*

F. KAPPEL† AND D. SALAMON‡

Abstract. The purpose of this paper is to introduce a new spline approximation scheme for retarded functional differential equations. The special feature of this approximation scheme is that it preserves the product space structure of retarded systems and approximates the adjoint semigroup in a strong sense. These facts guarantee the convergence of the solution operators for the differential Riccati equation in a strong sense. Numerical findings indicate a significant improvement in the convergence behaviour over both the averaging and the previous spline approximation scheme.

Key words. retarded functional differential equations, approximation, splines, Riccati equation

AMS(MOS) subject classifications. 34K35, 41A15, 93D15

1. Introduction. In this paper we introduce a new spline approximation scheme for linear time invariant retarded functional differential equations (RFDEs) and establish a number of convergence results. In particular we show that the approximate feedback law and the solution of the operator Riccati equation, associated with the linear quadratic control problem for this class of systems, converge in the uniform operator topology.

The first step of the general approach is to transform the RFDE

$$(1.1) \quad \dot{x}(t) = Lx_t + B_0u(t), \quad y(t) = C_0x(t)$$

into an abstract Cauchy problem of the form

$$(1.2) \quad \frac{d}{dt}x(t) = \mathcal{A}x(t) + \mathcal{B}u(t), \quad y(t) = \mathcal{C}x(t)$$

in the Hilbert space $\mathcal{X} = \mathbb{R}^n \times L^2[-h, 0; \mathbb{R}^n]$, $h > 0$, where \mathcal{A} is the infinitesimal generator of the strongly continuous semigroup $\mathcal{S}(t)$ which is associated with the uncontrolled delay equation. For systems of the form (1.2) there exists a general theory of the linear quadratic control problem of minimizing the cost functional

$$(1.3) \quad J(u) = \langle x(T), \mathcal{G}x(T) \rangle + \int_0^T [|y(t)|^2 + |u(t)|^2] dt$$

(see e.g. [9], [14], [19]). The optimal control can be characterized as a feedback law which is determined by an operator satisfying the differential Riccati equation (in the case $T < \infty$), respectively, the algebraic Riccati equation (in the case $T = \infty$). These operator Riccati equations involve both the original generator \mathcal{A} and its adjoint operator \mathcal{A}^* . Therefore, in order to approximate the feedback law and the Riccati operator in the strong operator topology, we have to approximate both semigroups $\mathcal{S}(t)$ and $\mathcal{S}^*(t)$ in the strong operator topology (see [20]).

* Received by the editors May 3, 1984; accepted for publication (in revised form) June 2, 1986. This work was supported in part by the Fonds zur Förderung der wissenschaftlichen Forschung, Austria, under project P4534.

† Institute for Mathematics, University of Graz, Elisabethstrasse 16, A-8010 Graz, Austria. The work of this author was supported in part by the University of Bremen, West Germany, and was begun while the author held a visiting professorship at the Forschungsschwerpunkt Dynamische Systeme, University of Bremen, in October 1982.

‡ Mathematics Research Center, University of Wisconsin, Madison, Wisconsin 53706. The work of this author was supported in part by National Science Foundation grant MCS-8210950.

For the approximation of the semigroups we use a Galerkin type scheme, i.e., we define finite dimensional subspaces \mathcal{X}^N of \mathcal{X} and operators \mathcal{A}^N on \mathcal{X}^N which generate semigroups $\mathcal{S}^N(t)$ on \mathcal{X}^N . The classical idea is to choose $\mathcal{X}^N \subset \text{dom } \mathcal{A}$ and define $\mathcal{A}^N = p^N \mathcal{A} p^N$, where p^N is the orthogonal projection of \mathcal{X} onto \mathcal{X}^N . Under appropriate consistency and stability hypotheses the convergence of $\mathcal{S}^N(t)$ to $\mathcal{S}(t)$ in the strong operator topology follows.

These ideas have been used by Banks and Kappel [7] for the development of a spline approximation scheme for RFDEs and have then been applied to problems of optimal control and parameter identification e.g. in [3], [6], [8], [27]. In particular, Kunisch [27] has established weak convergence results for the solution operators of the differential Riccati equations. Numerical findings in [8] indicate that these operators indeed do not converge strongly for the spline scheme developed in [7]. The main reason for this seems to be that the subspace \mathcal{X}^N in [7] has been chosen to be contained in the domain of \mathcal{A} which is different from the domain of \mathcal{A}^* .

In order to overcome this unequal treatment of $\mathcal{S}(t)$ and $\mathcal{S}^*(t)$, our idea was to enlarge the subspace \mathcal{X}^N such that it is neither contained in $\text{dom } \mathcal{A}$ nor in $\text{dom } \mathcal{A}^*$, but contains sufficiently many elements of both domains. Of course, in this situation the approximating operators can no longer be defined by $\mathcal{A}^N = p^N \mathcal{A} p^N$ but have to be defined directly instead (for details see § 5.1). As a result we are able to establish the desired convergence of the solution operators of the Riccati equation in the uniform operator topology for the finite time horizon problem. Despite the fact that in the case of the infinite time horizon problem our scheme always did converge numerically, we were not able to prove this convergence following the approach presented in [20]. The reason is that we do not have the uniform (with respect to N) exponential stability of the approximating semigroups for our scheme (compare the remarks at the end of § 5.3). In this respect the spline approximation scheme differs from the averaging approximation scheme in [4] for which the uniform exponential stability property has been established in [38].

In two preliminary sections we collect some basic facts from the state space and control theory for retarded systems (§ 2) and give a short survey on the theory of the linear quadratic optimal control problem for abstract systems in Hilbert space and for RFDEs (§ 3). In § 4.1 we present a general approximation scheme for abstract Cauchy problems in Banach space. In § 4.2 we consider the problem of approximating the feedback law for the finite time horizon problem following the approach given by Gibson in [20].

The main part of this paper is § 5, where we develop a special spline scheme and prove convergence results along the general ideas given in §§ 4.1 and 4.2. We also give the explicit formulae for the matrices which are necessary for the implementation of our scheme. This scheme has remarkable qualitative properties which will be published elsewhere.

Finally, in § 6 we present some of the many numerical calculations in order to demonstrate the good behaviour of our scheme and the advantage it offers over both the averaging approximation scheme [4], [20] and the spline scheme in [7], [8].

2. State space theory for linear hereditary control systems.

2.1. Linear hereditary control systems. We consider the linear hereditary control system

$$(2.1a) \quad \dot{x}(t) = Lx_t + B_0 u(t), \quad t \geq 0,$$

$$(2.1b) \quad y(t) = C_0 x(t),$$

where $x(t) \in \mathbb{R}^n$, $u(t) \in \mathbb{R}^l$, $y(t) \in \mathbb{R}^m$ and x_t is defined by $x_t(s) = x(t+s)$ for $-h \leq s \leq 0$, $h > 0$. Correspondingly B_0 and C_0 are real matrices of appropriate dimensions and L is a bounded linear functional $C(-h, 0; \mathbb{R}^n) \rightarrow \mathbb{R}^n$ given by

$$L\phi = \sum_{j=0}^p A_j \phi(-h_j) + \int_{-h}^0 A_{01}(\tau) \phi(\tau) d\tau, \quad \phi \in C(-h, 0; \mathbb{R}^n),$$

where $0 = h_0 < \dots < h_p = h$ and $A_j \in \mathbb{R}^{n \times n}$, $j = 0, \dots, p$, as well as $A_{01}(\cdot) \in L^2(-h, 0; \mathbb{R}^{n \times n})$. A solution of (2.1a) is a function $x(\cdot) \in L^2_{\text{loc}}(-h, \infty; \mathbb{R}^n)$ which is absolutely continuous with L^2 -derivative on every compact interval $[0, T]$, $T > 0$, and satisfies (2.1a) for almost all $t \geq 0$. It is well known that (2.1a) admits a unique solution $x(t) = x(t; \phi, u)$ for every input $u(\cdot) \in L^2_{\text{loc}}(0, \infty; \mathbb{R}^l)$ and every initial condition

$$(2.2) \quad x(0) = \phi^0, \quad x(\tau) = \phi^1(\tau), \quad -h \leq \tau < 0,$$

where $\phi = (\phi^0, \phi^1) \in M^2 = \mathbb{R}^n \times L^2(-h, 0; \mathbb{R}^n)$. Moreover, $x(\cdot; \phi, u)$ depends continuously on ϕ and u on compact intervals, i.e., for any $T > 0$ there exists a $K > 0$ such that

$$\sup_{0 \leq t \leq T} |x(t; \phi, u)| \leq K(\|\phi\| + \|u\|_{L^2(0, T; \mathbb{R}^l)}),$$

where $\|\phi\| = (|\phi^0|^2 + |\phi^1|_{L^2}^2)^{1/2}$ for $\phi \in M^2$ (see e.g. [12], [18]). The fundamental solution of (2.1a) will be denoted by $X(t)$ and is the $n \times n$ matrix valued solution of (2.1a) which corresponds to $u \equiv 0$ and $X(0) = I$, $X(\tau) = 0$ for $-h \leq \tau < 0$. The Laplace transform of $X(\cdot)$ is given by $\Delta^{-1}(\lambda)$, where

$$\begin{aligned} \Delta(\lambda) &= \lambda I - L(e^{\lambda \cdot} I) \\ &= \lambda I - \sum_{j=0}^p A_j e^{-\lambda h_j} - \int_{-h}^0 A_{01}(\tau) e^{\lambda \tau} d\tau, \quad \lambda \in \mathbb{C} \end{aligned}$$

is the characteristic matrix of (2.1a). Again it is well known that the forced motion of (2.1a) (in case $\phi = 0$) can be written as

$$(2.3) \quad x(t; 0, u) = \int_0^t X(t-\tau) B_0 u(\tau) d\tau, \quad t \geq 0.$$

2.2. Semigroups and state space description. Existence, uniqueness and continuous dependence results for solutions of RFDEs motivate the definition of the state of system (2.1) to be the pair

$$(2.4) \quad w(t) = (x(t), x_t) \in M^2,$$

which completely describes the past history of the solution at time $t \geq 0$. The evolution of this state is governed by the variation-of-constants formula

$$(2.5) \quad w(t) = S(t)\phi + \int_0^t S(t-s)Bu(s) ds, \quad t \geq 0,$$

which is the infinite dimensional version of (2.3). The input operator $B: \mathbb{R}^l \rightarrow M^2$ is given by

$$Bu = (B_0 u, 0) \in M^2, \quad u \in \mathbb{R}^l,$$

and the semigroup $S(\cdot)$ corresponds to the free motion of the system, i.e., $S(t): M^2 \rightarrow M^2$, $t \geq 0$ is defined by

$$S(t)\phi = (x(t; \phi, 0), x_t(\phi, 0)), \quad t \geq 0, \quad \phi \in M^2.$$

The infinitesimal generator of $S(\cdot)$ is given by

$$(2.6) \quad \begin{aligned} \text{dom } A &= \{\phi \in M^2 \mid \phi^1 \in W^{1,2}, \phi^0 = \phi^1(0)\}, \\ A\phi &:= (L\phi^1, \dot{\phi}^1), \end{aligned}$$

where $W^{1,2}$ denotes the Sobolev space $W^{1,2}(-h, 0; \mathbb{R}^n)$. The function $w(t)$ as defined in (2.5) is a mild solution of the abstract system

$$(2.7) \quad \begin{aligned} \dot{w}(t) &= Aw(t) + Bu(t), \\ y(t) &= Cw(t), \quad w(0) = \phi. \end{aligned}$$

The output operator $C: M^2 \rightarrow \mathbb{R}^m$ is defined by $C\phi = C_0\phi^0$, $\phi \in M^2$.

The operator A^* dual to A is explicitly given by (see e.g. [17]):

$$(2.7) \quad \begin{aligned} \text{dom } A^* &= \left\{ f \in M^2 \mid f^1 + \sum_{j=1}^{p-1} A_j^T f^0 \chi_{[-h, -h_j]} \in W^{1,2}, f^1(-h) = A_p^T f^0 \right\}, \\ [A^*f]^0 &= f^1(0) + A_0^T f^0, \\ [A^*f]^1(\tau) &= A_{01}^T(\tau) f^0 - \frac{d}{d\tau} \left[f^1(\tau) + \sum_{j=1}^{p-1} A_j^T f^0 \chi_{[-h, -h_j]}(\tau) \right]. \end{aligned}$$

The characteristic function of an interval I is denoted by χ_I .

2.3. Stability, stabilizability and controllability. System (2.1) is said to be *stable* if every solution $x(t)$ of the free system (i.e. $u(t) \equiv 0$) tends to zero as t goes to infinity. Equivalently, the semigroup $S(\cdot)$ is exponentially stable, i.e.,

$$\omega_0 = \lim_{t \rightarrow \infty} \frac{1}{t} \ln \|S(t)\| = \sup \{ \text{Re } \lambda \mid \lambda \in \sigma(A) \} < 0$$

(see for instance [22]). The spectrum of A is given by $\sigma(A) = \{\lambda \in \mathbb{C} \mid \det \Delta(\lambda) = 0\}$. Note that $\sigma(A^*) = \sigma(A)$.

The control system (2.1) is said to be *stabilizable* if there exists a control law

$$(2.8) \quad \begin{aligned} u(t) &= K(x(t), x_t) \\ &= K_0 x(t) + \int_{-h}^0 K_1(\tau) x(t+\tau) d\tau, \end{aligned}$$

where $K_0 \in \mathbb{R}^{l \times n}$, $K_1(\cdot) \in L^2(-h, 0; \mathbb{R}^{l \times n})$, such that the closed loop system (2.1), (2.8) is stable. We have the following important characterization (see [33], [36]).

THEOREM 2.1. *The following statements are equivalent:*

- (i) System (2.1) is stabilizable.
- (ii) There exists a $K \in \mathcal{L}(M^2, \mathbb{R}^l)$ such that the operator $A + BK$ generates an exponentially stable C_0 -semigroup.

- (iii) $\text{rank} [\Delta(\lambda), B_0] = n$ for all $\lambda \in \mathbb{C}$ with $\text{Re } \lambda \geq 0$.

The dual result is the following (see e.g. [10] or [36], [37]):

THEOREM 2.2. *The following statements are equivalent:*

- (i) There exists a $H \in \mathcal{L}(\mathbb{R}^m, M^2)$ such that the operator $A + HC$ generates an exponentially stable semigroup.

- (ii) $\text{rank} \begin{pmatrix} \Delta(\lambda) \\ C_0 \end{pmatrix} = n$ for all $\lambda \in \mathbb{C}$ with $\text{Re } \lambda \geq 0$.

System (2.1) is called *detectable* if the statements of the previous theorem are satisfied. A detailed discussion of the duality relations between feedback stabilization and dynamic observation in the product space framework can be found in [36].

3. The linear quadratic control problem.

3.1. Control systems in Hilbert spaces. Let us first deal with general linear control systems in Hilbert spaces \mathcal{X} , \mathcal{U} and \mathcal{Y} described by

$$(3.1) \quad \begin{aligned} \dot{x}(t) &= \mathcal{A}x(t) + \mathcal{B}u(t), & x(0) &= x_0, \\ y(t) &= \mathcal{C}x(t). \end{aligned}$$

We assume that $\mathcal{B} \in \mathcal{L}(\mathcal{U}, \mathcal{X})$, $\mathcal{C} \in \mathcal{L}(\mathcal{X}, \mathcal{Y})$ and that \mathcal{A} is the infinitesimal generator of a C_0 -semigroup $\mathcal{S}(t)$ on \mathcal{X} . System (3.1) will be understood in the sense of mild solutions, i.e., the trajectories of the system are given by

$$(3.2) \quad x(t) = \mathcal{S}(t)x_0 + \int_0^t \mathcal{S}(t-s)\mathcal{B}u(s) ds, \quad t \geq 0$$

for any $x_0 \in \mathcal{X}$ and any input $u(\cdot) \in L^2_{\text{loc}}(0, \infty; \mathcal{U})$.

Let $\mathcal{R} : \mathcal{U} \rightarrow \mathcal{U}$ and $\mathcal{G} : \mathcal{X} \rightarrow \mathcal{X}$ be selfadjoint linear operators satisfying

$$\langle x, \mathcal{G}x \rangle \geq 0 \quad \text{for all } x \in \mathcal{X}$$

and

$$\langle u, \mathcal{R}u \rangle \geq \varepsilon \|u\|^2 \quad \text{for all } u \in \mathcal{U}$$

with some $\varepsilon > 0$. In this section we look at the control problem of minimizing the cost functional

$$(3.3) \quad J(u) = \langle x(T), \mathcal{G}x(T) \rangle + \int_0^T [\|\mathcal{C}x(t)\|^2 + \langle u(t), \mathcal{R}u(t) \rangle] dt,$$

where $x(t)$ is given by (3.2) and $T > 0$ is a fixed final time. For the proof of the following result see [14] and [19].

THEOREM 3.1. *For any $x_0 \in \mathcal{X}$ there exists a unique control function $\bar{u}(\cdot) \in L^2(0, T; \mathcal{U})$ which minimizes the cost functional (3.3) under the constraint (3.2). The optimal control is of feedback form and is given by*

$$(3.4) \quad \bar{u}(t) = -\mathcal{R}^{-1}\mathcal{B}^*\mathcal{P}(t)\bar{x}(t), \quad t \geq 0,$$

where $\bar{x}(t)$ is the mild solution of the Cauchy problem $\dot{x} = (\mathcal{A} - \mathcal{B}\mathcal{R}^{-1}\mathcal{B}^*\mathcal{P}(t))x$, $x(0) = x_0$, and $t \rightarrow \mathcal{P}(t) \in \mathcal{L}(\mathcal{X})$ is the unique operator valued function on $[0, T]$ with the following properties:

- (i) $\mathcal{P}(t)$ is positive semidefinite for every $t \in [0, T]$.
- (ii) The function $t \rightarrow \mathcal{P}(t)x$ is continuous on $[0, T]$ for every $x \in \mathcal{X}$, and satisfies the Riccati integral equation

$$(3.5) \quad \begin{aligned} \mathcal{P}(t)x &= \mathcal{S}^*(T-t)\mathcal{G}\mathcal{S}(T-t)x \\ &+ \int_t^T \mathcal{S}^*(\tau-t)[\mathcal{C}^*\mathcal{C} - \mathcal{P}(\tau)\mathcal{B}\mathcal{R}^{-1}\mathcal{B}^*\mathcal{P}(\tau)]\mathcal{S}(\tau-t)x d\tau, \\ &0 \leq t \leq T, \quad x \in \mathcal{X}. \end{aligned}$$

Moreover, the optimal cost is given by

$$J(\bar{u}) = \langle x_0, \mathcal{P}(0)x_0 \rangle.$$

In [19] it is also shown that $\mathcal{P}(t)$ satisfies

$$(3.6) \quad \mathcal{P}(t)x = \mathcal{S}^*(T-t)\mathcal{G}\Phi(T, t)x + \int_t^T \mathcal{S}^*(\tau-t)\mathcal{C}^*\mathcal{C}\Phi(\tau, t)x d\tau, \quad 0 \leq t \leq T, \quad x \in \mathcal{X}$$

where $\Phi(\tau, t)$ is the evolution operator given by

$$(3.7) \quad \Phi(\tau, t)x = \mathcal{S}(\tau - t)x - \int_t^\tau \mathcal{S}(\tau - \sigma)\mathcal{B}\mathcal{R}^{-1}\mathcal{B}^*\mathcal{P}(\sigma)\Phi(\sigma, t)x \, d\sigma, \\ 0 \leq t \leq \tau \leq T, \quad x \in \mathcal{X}.$$

Let us now consider the problem of minimizing the cost functional

$$(3.8) \quad J(u) = \int_0^\infty [\|\mathcal{C}x(t)\|^2 + \langle u(t), \mathcal{R}u(t) \rangle] \, dt,$$

where again $x(t)$ is given by (3.2). For this situation the following result has been proved (see [14], [15], [41]; further references can be found in the survey paper [9]):

THEOREM 3.2. (a) *The following statements are equivalent:*

(i) *For any $x_0 \in \mathcal{X}$ there exists an input $u(\cdot) \in L^2(0, \infty; \mathcal{U})$ such that the corresponding cost $J(u)$ given by (3.8) and (3.2) is finite.*

(ii) *There exists a positive semidefinite operator $\mathcal{P} \in \mathcal{L}(\mathcal{X})$ satisfying the algebraic Riccati operator equation*

$$(3.9) \quad \langle \mathcal{A}y, \mathcal{P}x \rangle + \langle \mathcal{P}y, \mathcal{A}x \rangle + \langle \mathcal{C}y, \mathcal{C}x \rangle - \langle \mathcal{P}y, \mathcal{B}\mathcal{R}^{-1}\mathcal{B}^*\mathcal{P}x \rangle = 0$$

for all $x, y \in \text{dom } \mathcal{A}$.

(b) *If the statements under (a) are valid, then there exists a unique optimal control $\bar{u}(t)$ which is given by the feedback law*

$$(3.10) \quad \bar{u}(t) = -\mathcal{R}^{-1}\mathcal{B}^*\mathcal{P}\bar{x}(t), \quad t \geq 0,$$

where $\bar{x}(t)$ is the mild solution of the Cauchy problem $\dot{x} = (\mathcal{A} - \mathcal{B}\mathcal{R}^{-1}\mathcal{B}^*\mathcal{P})x$, $x(0) = x_0$, and \mathcal{P} is the minimal solution of (3.9). Moreover, the optimal cost is given by

$$J(\bar{u}) = \langle x_0, \mathcal{P}x_0 \rangle.$$

(c) *Suppose that the statements under (a) are satisfied and let \mathcal{P} be the minimal positive semidefinite solution of (3.9). Moreover, let $\mathcal{P}_T(t)$, $0 \leq t \leq T$, be the unique positive semidefinite solution of (3.5) with $\mathcal{G} = 0$. Then \mathcal{P} is the strong limit of $\mathcal{P}_T(0)$ as T goes to infinity.*

(d) *Suppose that there exists some $\mathcal{K} \in L(\mathcal{Y}, \mathcal{X})$ such that the operator $\mathcal{A} + \mathcal{K}\mathcal{C}$ generates an exponentially stable semigroup. Then there exists at most one positive semidefinite solution of (3.9). Moreover, if such a solution exists, then the closed loop semigroup generated by $\mathcal{A} - \mathcal{B}\mathcal{R}^{-1}\mathcal{B}^*\mathcal{P}$ is exponentially stable.*

3.2. Applications to hereditary systems. Let us first apply Theorem 3.1 to system (Σ) which is associated to system (2.1) in terms of the state concept introduced in § 2. The cost functional for system (2.1) is assumed to be

$$(3.11) \quad J(u) = x(T; \phi, u)^T G_0 x(T; \phi, u) \\ + \int_0^T [\|C_0 x(t; \phi, u)\|^2 + u(t)^T R u(t)] \, dt,$$

where $R \in \mathbb{R}^{l \times l}$ is positive definite and $G_0 \in \mathbb{R}^{n \times n}$ is positive semidefinite. If the operator $G: M^2 \rightarrow M^2$ is defined by $G\phi = (G_0\phi^0, 0)$, $\phi \in M^2$, then the cost functional for system (Σ) is given by (3.3) (with $\mathcal{G} = G$, $\mathcal{C} = C$ and $\mathcal{R} = R$, of course). According

to Theorem 3.1 there exists a unique, positive semidefinite, strongly continuous family $\Pi(\cdot)$ of operators in $\mathcal{L}(M^2)$ which satisfies the Riccati integral equation

$$(3.12) \quad \begin{aligned} \Pi(t)\phi &= S^*(T-t)GS(T-t)\phi \\ &+ \int_t^T S^*(\tau-t)[C^*C - \Pi(\tau)BR^{-1}B^*\Pi(\tau)]S(\tau-t)\phi \, d\tau, \\ \phi &\in M^2, \quad 0 \leq t \leq T. \end{aligned}$$

Let us now look at the structure of the operator $\Pi(t)$. Due to the product space structure of the state space M^2 we can write

$$\Pi(t) = \begin{pmatrix} \Pi_{00}(t) & \Pi_{01}(t) \\ \Pi_{10}(t) & \Pi_{11}(t) \end{pmatrix},$$

where $\Pi_{00}(t)$ is a selfadjoint operator $\mathbb{R}^n \rightarrow \mathbb{R}^n$ which can be represented by a symmetric matrix and $\Pi_{11}(t)$ is a selfadjoint operator $L^2 \rightarrow L^2$. The operator $\Pi_{10}(t)$ can be represented by a matrix-valued function $\Pi_{10}(t, \cdot) \in L^2(-h, 0; \mathbb{R}^{n \times n})$. The adjoint operator $\Pi_{01}(t) = \Pi_{10}^*(t)$ from $L^2 \rightarrow \mathbb{R}^n$ is given by

$$\Pi_{01}(t)\phi = \int_{-h}^0 \Pi_{10}^T(t, \tau)\phi(\tau) \, d\tau, \quad \phi \in L^2.$$

We are mainly interested in the matrices $\Pi_{00}(t)$ and $\Pi_{10}(t, \tau)$, which determine the optimal feedback law

$$(3.13) \quad \bar{u}(t) = -R^{-1}B_0^T \left[\Pi_{00}(t)x(t) + \int_{-h}^0 \Pi_{10}^T(t, \tau)x(t+\tau) \, d\tau \right]$$

for system (2.1). Recall that B^* maps $\phi \in M^2$ to $B_0^T \phi^0 \in \mathbb{R}^l$.

For the rest of this section we assume that system (2.1) is stabilizable and detectable, so that system (Σ) satisfies the assumptions of Theorem 3.2. Hence there exists a positive semi-definite operator $\Pi \in \mathcal{L}(M^2)$ satisfying the algebraic Riccati equation

$$(3.14) \quad A^*\Pi\phi + \Pi A\phi - \Pi BR^{-1}B^*\Pi\phi + C^*C\phi = 0,$$

$\phi \in \text{dom } A$. The equation can be written in this form since every solution \mathcal{P} of (3.9) maps $\text{dom } \mathcal{A}$ into $\text{dom } \mathcal{A}^*$, i.e.,

$$(3.15) \quad \text{range } \Pi \subset \text{dom } A^*.$$

Again the operator Π can be written in block form

$$\Pi = \begin{pmatrix} \Pi_{00} & \Pi_{01} \\ \Pi_{10} & \Pi_{11} \end{pmatrix},$$

where $\Pi_{01} = \Pi_{10}^*$ maps L^2 into \mathbb{R}^n . Hence the optimal feedback law is of the form

$$(3.16) \quad \begin{aligned} u(t) &= -R^{-1}B^*\Pi(x(t), x_t) \\ &= -R^{-1}B_0^T \left[\Pi_{00}x(t) + \int_{-h}^0 \Pi_{10}^T(\tau)x(t+\tau) \, d\tau \right]. \end{aligned}$$

Finally note that the closed loop system (2.1), (3.16) is stable (Theorem 3.2).

4. A general approximation scheme.

4.1. Approximation of the state. In this section we present a general approximation scheme for linear abstract Cauchy problems restricting ourselves to a situation which is of sufficient generality for our purposes.

Let \mathcal{X} be a real Hilbert space with inner product $\langle \cdot, \cdot \rangle$ and corresponding norm $\|\cdot\|$. Furthermore let \mathcal{A} be the infinitesimal generator of the C_0 -semigroup $\mathcal{S}(t)$, $t \geq 0$, on \mathcal{X} . It is a fundamental result that there exists constants $M \geq 1$ and $\omega \in \mathbb{R}$ such that

$$\|\mathcal{S}(t)\| \leq Me^{\omega t}, \quad t \geq 0.$$

In order to approximate the trajectories $\mathcal{S}(t)x_0$, $x_0 \in \mathcal{X}$, it is a standard idea to choose a sequence $\{\mathcal{X}^N\}$ of finite dimensional subspaces of \mathcal{X} with corresponding orthogonal projections

$$p^N: \mathcal{X} \rightarrow \mathcal{X}^N, \quad N = 1, 2, \dots,$$

and to define (in an appropriate way) a sequence $\{\mathcal{A}^N\}$ of linear operators

$$\mathcal{A}^N: \mathcal{X}^N \rightarrow \mathcal{X}^N, \quad N = 1, 2, \dots.$$

With \mathcal{A}^N and $x_0 \in \mathcal{X}$ we associate the Cauchy problem

$$(4.1) \quad \begin{aligned} \dot{x}^N(t) &= \mathcal{A}^N x^N(t), & t \geq 0, \\ x^N(0) &= p^N x_0 \end{aligned}$$

on \mathcal{X}^N . We extend the definition of \mathcal{A}^N to all of \mathcal{X} by $\mathcal{A}^N x = \mathcal{A}^N p^N x$ and define the C_0 -semigroup $\mathcal{S}^N(t)$, $t \geq 0$, on \mathcal{X} by

$$\mathcal{S}^N(t)x_0 = e^{\mathcal{A}^N t} p^N x_0 = e^{\mathcal{A}^N t} p^N x_0 + x_0 - p^N x_0, \quad t \geq 0, \quad x_0 \in \mathcal{X}.$$

The following hypotheses will be used in order to guarantee the desired convergence $\mathcal{S}^N(t)p^N x_0 \rightarrow \mathcal{S}(t)x_0$:

(H1) $\lim_{N \rightarrow \infty} p^N x = x$ for all $x \in \mathcal{X}$.

(H2) There exists constants $\tilde{M} \geq 1$ and $\tilde{\omega} \in \mathbb{R}$ such that

$$\|\mathcal{S}^N(t)x\| \leq \tilde{M}e^{\tilde{\omega}t}\|x\|$$

for all $t \geq 0$, $x \in \mathcal{X}^N$ and $N = 1, 2, \dots$.

(H3) There exists a dense subset $D \subset \text{dom } \mathcal{A}$ which is invariant with respect to $\mathcal{S}(t)$, $t \geq 0$, such that

(i) $\lim_{N \rightarrow \infty} \mathcal{A}^N p^N x = \mathcal{A}x$ for all $x \in D$, and

(ii) for any $x \in D$ there exists a function $m(\cdot, x) \in L^1_{\text{loc}}(0, \infty; \mathbb{R})$ such that

$$\|\mathcal{A}^N p^N \mathcal{S}(t)x\| \leq m(t; x) \quad \text{a.e. on } [0, \infty)$$

for all N .

Hypothesis (H2) is equivalent to

(H2*) For any N there exists a norm $\|\cdot\|_N$ on \mathcal{X}^N such that

(i) For some constant $\tilde{M} \geq 1$

$$\|x\| \leq \|x\|_N \leq \tilde{M}\|x\|, \quad x \in \mathcal{X}^N, \quad N = 1, 2, \dots, \quad \text{and}$$

(ii) for some constant $\tilde{\omega} \in \mathbb{R}$ all operators $\mathcal{A}^N - \tilde{\omega}I$ are dissipative on $(\mathcal{X}^N, \|\cdot\|_N)$ (i.e. $\|(\mathcal{A}^N - \mu I)x\|_N \geq (\mu - \tilde{\omega})\|x\|_N$ for $x \in \mathcal{X}^N$, $\mu > \tilde{\omega}$ or, in case $(\mathcal{X}^N, \|\cdot\|_N)$ is a Hilbert space with inner product $\langle \cdot, \cdot \rangle_N$, $\langle \mathcal{A}^N x, x \rangle_N \leq \tilde{\omega}\|x\|_N^2$ for $x \in \mathcal{X}^N$; cf. [32, Thm. 4.2], [30, p. 244]).

The equivalence of (H2) and (H2*) follows from the well-known relation between exponential estimates for a semigroup and dissipativity properties of its generator (cf., for instance, [32, Thm. 4.3]).

THEOREM 4.1. *Let (H1)–(H3) be satisfied for the sequences \mathcal{X}^N , p^N , \mathcal{A}^N , $N = 1, 2, \dots$. Then for all $x_0 \in \mathcal{X}$*

$$(4.2) \quad \lim_{N \rightarrow \infty} e^{\mathcal{A}^N t} p^N x_0 = \mathcal{S}(t)x_0$$

uniformly for t in bounded intervals.

Proof. If $x_0 \in \text{dom } \mathcal{A}$ then $d/dt \mathcal{S}(t)x_0 = \mathcal{A}\mathcal{S}(t)x_0$. This together with (4.1) implies

$$\mathcal{S}(t)x_0 - \mathcal{S}^N(t)p^N x_0 = \mathcal{S}^N(t)[x_0 - p^N x_0] + \int_0^t \mathcal{S}^N(t-\tau)(\mathcal{A} - \mathcal{A}^N)\mathcal{S}(\tau)x_0 d\tau,$$

$x_0 \in \text{dom } \mathcal{A}$, $0 \leq t \leq T$. Let $x_0 \in D$. Using (H2), we immediately get

$$\|\mathcal{S}(t)x_0 - \mathcal{S}^N(t)p^N x_0\| \leq \tilde{M} e^{\tilde{\omega}T} \left\{ \|x_0 - p^N x_0\| + \int_0^T \|(\mathcal{A} - \mathcal{A}^N)\mathcal{S}(\tau)x_0\| d\tau \right\}$$

for $0 \leq t \leq T$. Then (4.2) follows from (H1), (H3) and Lebesgue's dominated convergence theorem. Note that $\mathcal{S}(\tau)x_0 \in D$ for $\tau \in [0, T]$. A density argument using (H2) completes the proof. \square

The methods in the proof of the previous theorem are well known in connection with numerical approximation of partial differential equations (see for instance the proof of the Lax-Richtmyer equivalence theorem in [21]). For delay equations this approach appears for the first time in [3], [6] and has later on been used in [24]. We equally well could have used the Trotter-Kato theorem [32].

Next we consider the nonhomogeneous problem

$$(4.3a) \quad \dot{x}(t) = \mathcal{A}x(t) + \mathcal{B}u(t), \quad t \geq s,$$

$$(4.3b) \quad x(s) = x_0 \in \mathcal{X},$$

where $u \in L^2_{\text{loc}}(s, \infty; \mathbb{R}^l)$ and \mathcal{B} is a linear operator $\mathbb{R}^l \rightarrow \mathcal{X}$. The unique mild solution $x(t) = x(t; s, x_0, u)$ of (4.3) is given by

$$(4.4) \quad x(t) = \mathcal{S}(t-s)x_0 + \int_s^t \mathcal{S}(t-\tau)\mathcal{B}u(\tau) d\tau, \quad t \geq s.$$

In addition to the approximating sequence \mathcal{X}^N , p^N , \mathcal{A}^N , $N = 1, 2, \dots$, introduced above let us assume that \mathcal{B}^N , $N = 1, 2, \dots$, is a sequence of corresponding input operators $\mathbb{R}^l \rightarrow \mathcal{X}^N$. Then we consider the approximating systems

$$(4.5a) \quad \dot{x}^N(t) = \mathcal{A}^N x^N(t) + \mathcal{B}^N u(t), \quad t \geq s,$$

$$(4.5b) \quad x^N(s) = p^N x_0, \quad x_0 \in \mathcal{X},$$

on \mathcal{X}^N with the unique solution $x^N(t) = x^N(t; s, p^N x_0, u)$ given by

$$(4.6) \quad x^N(t) = \mathcal{S}^N(t-s)p^N x_0 + \int_s^t \mathcal{S}^N(t-\tau)\mathcal{B}^N u(\tau) d\tau, \quad t \geq s,$$

where $\mathcal{S}^N(t): \mathcal{X} \rightarrow \mathcal{X}$ is defined as above.

THEOREM 4.2. *Assume that $\mathcal{S}^N(\cdot)$, $N = 1, 2, \dots$, and $\mathcal{S}(\cdot)$ are C_0 -semigroups on \mathcal{X} such that for constants $M \geq 1$, $\omega \in \mathbb{R}$*

$$\|\mathcal{S}^N(t)\| \leq M e^{\omega t}, \quad t \geq 0, \quad N = 1, 2, \dots,$$

and for all $x_0 \in \mathcal{X}$

$$\lim_{N \rightarrow \infty} \mathcal{S}^N(t)p^N x_0 = \mathcal{S}(t)x_0$$

uniformly on bounded t -intervals. Furthermore assume that

$$\lim_{N \rightarrow \infty} \mathcal{B}^N \xi = \mathcal{B} \xi \quad \text{for all } \xi \in \mathbb{R}^l.$$

Then for all $x_0 \in \mathcal{X}$, $T > 0$ and $\gamma > 0$

$$\lim_{N \rightarrow \infty} x^N(t; s, p^N x_0, u) = x(t; s, x_0, u)$$

uniformly for $0 \leq s \leq t \leq T$ and for $u \in L^2(s, T; \mathbb{R}^l)$ with $\|u\|_{L^2(s, T; \mathbb{R}^l)} \leq \gamma$.

We omit the proof of this result since it is only a slight modification of that given in [4] for a very similar result.

It is clear that in this section without any changes \mathcal{X} could have been a real Banach space.

4.2. Approximation of the feedback law in the optimal problem. We restrict ourselves to the finite time control problem of minimizing the cost functional

$$(4.7) \quad J_s(u) = \langle x(T), \mathcal{G}x(T) \rangle + \int_s^T [\|\mathcal{C}x(t)\|^2 + u(t)^T \mathcal{R}u(t)] dt$$

associated with the Cauchy problem (4.3). We assume that the operators $\mathcal{G}: \mathcal{X} \rightarrow \mathcal{X}$, $\mathcal{R}: \mathbb{R}^l \rightarrow \mathbb{R}^l$, $\mathcal{C}: \mathcal{X} \rightarrow \mathbb{R}^m$ are defined as in § 3.1. As we have seen in that section (with obvious modifications for the case when the initial time s is not necessarily zero), the unique solution of this problem is given by the feedback law

$$(4.8) \quad \bar{u}_s(t) = -\mathcal{R}^{-1} \mathcal{B}^* \mathcal{P}(t) \Phi(t, s) x_0, \quad s \leq t \leq T,$$

where $\mathcal{P}(t): \mathcal{X} \rightarrow \mathcal{X}$ is the unique positive semidefinite solution of the Riccati differential equation (3.5) and $\Phi(t, s)$ is given by (3.7).

Correspondingly, we consider the sequence of control problems of minimizing

$$(4.9) \quad J_s^N(u) = \langle x^N(T), \mathcal{G}x^N(T) \rangle + \int_s^T [\|\mathcal{C}x^N(t)\|^2 + u(t)^T \mathcal{R}u(t)] dt,$$

where $x^N(t) = x^N(t; s, p^N x_0, u)$ is the unique solution of (4.5). The optimal control is given by the feedback law

$$(4.10) \quad \begin{aligned} \bar{u}_s^N(t) &= -\mathcal{R}^{-1} (\mathcal{B}^N)^* \mathcal{P}^N(t) \Phi^N(t, s) p^N x_0 \\ &= -\mathcal{R}^{-1} (\mathcal{B}^N)^* \mathcal{P}^N(t) \Phi^N(t, s) x_0, \quad s \leq t \leq T, \end{aligned}$$

where the strongly continuous, positive semidefinite operator $\mathcal{P}^N(t): \mathcal{X} \rightarrow \mathcal{X}$ and the strongly continuous evolution operator $\Phi^N(t, s): \mathcal{X} \rightarrow \mathcal{X}$ are defined by the equations

$$(4.11) \quad \begin{aligned} \mathcal{P}^N(t)x &= \mathcal{P}^N(T-t)^* p^N \mathcal{G} p^N \Phi^N(T, t)x \\ &\quad + \int_t^T \mathcal{P}^N(\tau-t)^* p^N \mathcal{C}^* \mathcal{C} p^N \Phi^N(\tau, t)x d\tau, \quad t \leq T, \end{aligned}$$

and

$$(4.12) \quad \Phi^N(t, s)x = \mathcal{P}^N(t-s)x - \int_s^t \mathcal{P}^N(t-\tau) \mathcal{B}^N \mathcal{R}^{-1} (\mathcal{B}^N)^* \mathcal{P}^N(\tau) \Phi^N(\tau, s)x d\tau, \quad t \geq s$$

for $x \in \mathcal{X}$. It follows immediately from (4.11) and the fact that $\mathcal{P}^N(t)$ is selfadjoint that

$$(4.13) \quad \mathcal{P}^N(t) = p^N \mathcal{P}^N(t) p^N, \quad t \leq T.$$

This in turn implies, by (4.12), that

$$(4.14) \quad p^N \Phi^N(t, s) = \Phi^N(t, s) p^N, \quad s \leq t \leq T.$$

Note that these two facts justify the second equation in (4.12). Moreover, the optimal cost of (4.9), (4.5) is given by

$$(4.15) \quad J_s^N(\bar{u}_s^N) = \langle x_0, \mathcal{P}^N(s) x_0 \rangle.$$

We remark that $\mathcal{P}^N(t)$, regarded as an operator on \mathcal{X}^N , satisfies the following finite dimensional Riccati differential equation

$$(4.16) \quad \begin{aligned} & \frac{d}{dt} \mathcal{P}^N(t) + (\mathcal{A}^N)^* \mathcal{P}^N(t) + \mathcal{P}^N(t) \mathcal{A}^N \\ & - \mathcal{P}^N(t) \mathcal{B}^N \mathcal{R}^{-1}(\mathcal{B}^N)^* \mathcal{P}^N(t) + p^N \mathcal{C}^* \mathcal{C} p^N = 0, \quad t \leq T, \\ & \mathcal{P}^N(t) = p^N \mathcal{G} p^N. \end{aligned}$$

Obviously, the most interesting question is how the original system (4.3) behaves when the optimal feedback control (4.8) is replaced by the approximate control law

$$(4.17) \quad \hat{u}_s^N(t) = -\mathcal{R}^{-1}(\mathcal{B}^N)^* \mathcal{P}^N(t) \hat{\Phi}^N(t, s) x_0,$$

where $\hat{\Phi}^N(t, s)$ denotes the corresponding closed loop evolution operator on \mathcal{X} which is defined by

$$(4.18) \quad \hat{\Phi}^N(t, s) x = \mathcal{S}(t-s) x - \int_s^t \mathcal{S}(t-\tau) \mathcal{B} \mathcal{R}^{-1}(\mathcal{B}^N)^* \mathcal{P}^N(\tau) \hat{\Phi}^N(\tau, s) x \, d\tau$$

for $x \in \mathcal{X}$ and $s \leq t \leq T$.

All the desired convergence results are contained in the next theorem which is a straight forward consequence of Theorems 6.1–6.3 in [20]. For the convenience of the reader we present the main ideas of the proof.

THEOREM 4.3. *Let us assume that*

- (i) *There exist constants $M \geq 1$, $\omega \in \mathbb{R}$ such that*

$$\|\mathcal{S}^N(t)\| \leq M e^{\omega t}, \quad t \geq 0, \quad N = 1, 2, \dots;$$

- (ii) *For every $x \in \mathcal{X}$*

$$\lim_{N \rightarrow \infty} \mathcal{S}^N(t) p^N x = \mathcal{S}(t) x, \quad \lim_{N \rightarrow \infty} \mathcal{S}^N(t)^* p^N x = \mathcal{S}(t)^* x$$

uniformly on $[0, T]$;

- (iii) *$\lim_{N \rightarrow \infty} \mathcal{B}^N \xi = \mathcal{B} \xi$ for every $\xi \in \mathbb{R}^l$.*

Then, for every $x_0 \in \mathcal{X}$,

- (a) $\lim_{N \rightarrow \infty} J_s^N(\bar{u}_s^N) = \lim_{N \rightarrow \infty} J_s(\hat{u}_s^N) = J_s(\bar{u}_s),$
 (b) $\lim_{N \rightarrow \infty} \bar{u}_s^N(t) = \lim_{N \rightarrow \infty} \hat{u}_s^N(t) = \bar{u}_s(t),$
 (c) $\lim_{N \rightarrow \infty} \Phi^N(t, s) x_0 = \lim_{N \rightarrow \infty} \hat{\Phi}^N(t, s) x_0 = \Phi(t, s) x_0,$
 (d) $\lim_{N \rightarrow \infty} \mathcal{P}^N(s) x_0 = \mathcal{P}(s) x_0$

and the limits are uniform on the domain $0 \leq s \leq t \leq T$. If range \mathcal{G} is finite dimensional, then $\mathcal{P}^N(s)$ converges to $\mathcal{P}(s)$ in the uniform operator topology, uniformly on the interval $[0, T]$.

Proof. Let us introduce the operators $\mathcal{F}_s(t): L^2(s, T; \mathbb{R}^l) \rightarrow \mathcal{X}$, $\mathcal{G}_s: \mathcal{X} \rightarrow L^2(s, T; \mathbb{R}^l)$, $\mathcal{R}_s: L^2(s, T; \mathbb{R}^l) \rightarrow L^2(s, T; \mathbb{R}^l)$ by defining

$$\begin{aligned} \mathcal{F}_s(t)u &= \int_s^t \mathcal{P}(t-\tau)\mathcal{B}u(\tau) d\tau, \\ (4.19) \quad \mathcal{G}_s x &= \mathcal{F}_s(T)^* \mathcal{G}\mathcal{P}(T-s)x + \int_s^T \mathcal{F}_s(\tau)^* \mathcal{C}^* \mathcal{C}\mathcal{P}(\tau-s)x d\tau, \\ \mathcal{R}_s u &= \mathcal{F}_s(T)^* \mathcal{G}\mathcal{F}_s(T)u + \int_s^T \mathcal{F}_s(\tau)^* \mathcal{C}^* \mathcal{C}\mathcal{F}_s(\tau)u d\tau + \mathcal{R}u \end{aligned}$$

for $u \in L^2(s, T; \mathbb{R}^l)$ and $x \in \mathcal{X}$. Of course, $\mathcal{R}u$ is defined by $(\mathcal{R}u)(t) = \mathcal{R}u(t)$, $s \leq t \leq T$. Then it is easy to see that the Fréchet derivative of J_s with respect to u is given by $J'_s(u) = 2\mathcal{R}_s u + 2\mathcal{G}_s x_0$. Since the optimal control \bar{u}_s satisfies $J'_s(\bar{u}_s) = 0$, this implies

$$(4.20) \quad \bar{u}_s = -\mathcal{R}_s^{-1} \mathcal{G}_s x_0.$$

Analogously, we get

$$(4.21) \quad \bar{u}_s^N = -(\mathcal{R}_s^N)^{-1} \mathcal{G}_s^N p^N x_0 = -(\mathcal{R}_s^N)^{-1} \mathcal{G}_s^N x_0$$

where \mathcal{R}_s^N , \mathcal{G}_s^N , \mathcal{F}_s^N are defined as above with $\mathcal{P}(t)$, \mathcal{B} , \mathcal{C} , \mathcal{G} replaced by $\mathcal{P}^N(t)$, \mathcal{B}^N , \mathcal{C}^N , \mathcal{G}^N , respectively. Combining these formulae with (3.7), (4.8) and (4.12), (4.10), we get

$$(4.22) \quad \Phi(t, s)x_0 = \mathcal{P}(t-s)x_0 - \mathcal{F}_s(t)\mathcal{R}_s^{-1}\mathcal{G}_s x_0,$$

$$(4.23) \quad \Phi^N(t, s)x_0 = \mathcal{P}^N(t-s)x_0 - \mathcal{F}_s^N(t)(\mathcal{R}_s^N)^{-1}\mathcal{G}_s^N x_0$$

for every $s \in [0, T]$ and every $t \in [s, T]$.

We have shown in Theorem 4.2 that $\mathcal{F}_s^N(t)$ converges to $\mathcal{F}_s(t)$ in the uniform operator topology, uniformly for $0 \leq s \leq t \leq T$. This implies that for every $x \in \mathcal{X}$

$$(4.24) \quad \lim_{N \rightarrow \infty} \mathcal{G}_s^N x = \mathcal{G}_s x$$

uniformly on $[0, T]$ and moreover $\|\mathcal{R}_s^N - \mathcal{R}_s\| \rightarrow 0$, also uniformly on $[0, T]$. Choosing $\varepsilon > 0$ such that $\xi^T \mathcal{R} \xi \geq \varepsilon \|\xi\|^2$ for $\xi \in \mathbb{R}^l$, we obtain

$$\|\mathcal{R}_s^N u\| \geq \varepsilon \|u\|, \quad u \in L^2(s, T; \mathbb{R}^l), \quad N = 1, 2, \dots$$

and hence

$$(4.25) \quad \lim_{N \rightarrow \infty} \|(\mathcal{R}_s^N)^{-1} - \mathcal{R}_s^{-1}\| = 0$$

uniformly on $[0, T]$.

It follows immediately from (4.22)–(4.25) that $\Phi^N(t, s)$ converges strongly to $\Phi(t, s)$. By (4.11) and (3.6), this implies the strong convergence of the Riccati operators $\mathcal{P}^N(s)$ to $\mathcal{P}(s)$. Now the convergence result on $\hat{\Phi}^N(t, s)$ follows from the inequality

$$\begin{aligned} &\|\Phi(t, s)x - \hat{\Phi}^N(t, s)x\| \\ &\leq \int_s^t \|\mathcal{P}(t-\tau)\mathcal{B}\mathcal{R}^{-1}\| \|[(\mathcal{B}^N)^* \mathcal{P}^N(\tau) - \mathcal{B}^* \mathcal{P}(\tau)]\Phi(\tau, s)x\| d\tau \\ &\quad + \int_s^t \|\mathcal{P}(t-\tau)\mathcal{B}\mathcal{R}^{-1}(\mathcal{B}^N)^* \mathcal{P}^N(\tau)\| \|\hat{\Phi}^N(\tau, s)x - \Phi(\tau, s)x\| d\tau \end{aligned}$$

and Gronwall's lemma. Thus we have established the statements (c) and (d). Statement (b) follows from (c) and (d), since the control functions \bar{u}_s , \bar{u}_s^N , \hat{u}_s^N are given by (4.10), (4.12), (4.19) respectively. Statement (a) is an immediate consequence of (b) and (d), since $J_s^N(\bar{u}_s^N) = \langle x_0, \mathcal{P}^N(s)x_0 \rangle$ and $J_s(\bar{u}_s) = \langle x_0, \mathcal{P}(s)x_0 \rangle$. If range \mathcal{G} is finite dimensional, then the convergence of $\mathcal{P}^N(s)$ in the uniform operator topology can be established by analogous considerations as those in the proof of Theorem 4.2 (see [4]), again by the use of (4.13) and (3.6). \square

5. A special approximation scheme.

5.1. General ideas. In this section we present a general idea how to construct special approximation schemes satisfying the assumptions of § 4.

Let the sequence X^N , $N = 1, 2, \dots$, of subspaces of M^2 be defined by

$$X^N = \left\{ \phi \in M^2 \left| \phi = \hat{e}_0^N \alpha_0 + \sum_{i=1}^p \sum_{j=1}^{k_N} \hat{e}_{ij}^N \alpha_{ij}, \alpha_0, \alpha_{ij} \in \mathbb{R}^n \right. \right\},$$

where the "basis elements" \hat{e}_0^N , \hat{e}_{ij}^N are given by

$$\hat{e}_0^N = (I, 0), \quad \hat{e}_{ij}^N = (0, e_{ij}^N I),$$

i.e., $X^N = \mathbb{R}^n \times Y^N$ with $Y^N = \text{span}(e_{11}^N I, \dots, e_{pk_N}^N I) \subset L^2(-h, 0; \mathbb{R}^n)$. It is clear that $\dim X^N = (pk_N + 1)n$. We assume that

$$(5.1) \quad \begin{aligned} e_{ij}^N | [-h_i, -h_{i-1}] &\in W^{1,2}(-h_i, -h_{i-1}; \mathbb{R}), \\ e_{ij}^N(\tau) &= 0 \quad \text{for } \tau \notin [-h_i, -h_{i-1}]. \end{aligned}$$

Without restriction we may further assume that e_{ij}^N is right-hand continuous on $[-h, 0)$.

Because of the product space structure of the subspaces X^N the orthogonal projections $p^N: M^2 \rightarrow X^N$ are given by

$$p^N \phi = (\phi^0, \pi^N \phi^1) \quad \text{for } \phi = (\phi^0, \phi^1) \in M^2,$$

where π^N is the orthogonal projection $L^2(-h, 0; \mathbb{R}^n) \rightarrow Y^N$. We introduce

$$\hat{E}^N = (\hat{e}_0^N, \hat{e}_{11}^N, \dots, \hat{e}_{pk_N}^N)$$

and denote by $\alpha^N(\phi) = \text{col}(\alpha_0^N, \alpha_{11}^N, \dots, \alpha_{pk_N}^N) \in \mathbb{R}^{n(pk_N+1)}$ the coordinate vector of an element $\phi \in X^N$, i.e.,

$$\phi = \hat{E}^N \alpha^N(\phi), \quad \phi \in X^N.$$

An easy calculation shows

$$(5.2) \quad \alpha^N(p^N \phi) = (Q^N)^{-1} d^N(\phi), \quad \phi \in M^2,$$

where

$$\begin{aligned} d^N(\phi) &= \langle \hat{E}^N, \phi \rangle_{M^2} = \text{col}(\phi^0, \langle e_{11}^N, \phi^1 \rangle_{L^2}, \dots, \langle e_{pk_N}^N, \phi^1 \rangle_{L^2}), \\ Q^N &= \langle \hat{E}^N, \hat{E}^N \rangle_{M^2} = \text{diag}(I, q_1^N \otimes I, \dots, q_p^N \otimes I), \end{aligned}$$

where $q_i^N = (\langle e_{ij}^N, e_{ik}^N \rangle_{L^2})_{j,k=1, \dots, k_N}$ and I is the $n \times n$ identity matrix. For elements in X^N the inner product has the representation

$$(5.3) \quad \langle \phi, \psi \rangle_{M^2} = \alpha^N(\phi)^T Q^N \alpha^N(\psi), \quad \phi, \psi \in X^N.$$

Since for elements $\phi = (\phi^0, \phi^1) \in X^N$ in general $\phi^0 \neq \lim_{\tau \uparrow 0} \phi^1(\tau)$ and ϕ^1 may have jumps of arbitrary size at the delay points $-h_i$, it is clear that X^N is not contained in $\text{dom } A$ nor in $\text{dom } A^*$. However, the operators A and A^* can formally be extended to all of X^N in the following way:

$$(5.4) \quad [A\phi]^0 = A_0\phi^0 + \sum_{i=1}^p A_i\phi^1(-h_i) + \int_{-h}^0 A_{01}(\tau)\phi^1(\tau) d\tau,$$

$$(5.5) \quad [A\phi]^1(\tau) = \frac{d^+}{d\theta} \phi^1(\tau) + \delta_0(\tau)(\phi^0 - \lim_{\tau \uparrow 0} \phi^1(\tau)) \\ + \sum_{i=1}^{p-1} \delta_i(\tau)(\phi^1(-h_i) - \lim_{\tau \uparrow -h_i} \phi^1(\tau)),$$

$$(5.6) \quad [A^*\psi]^0 = \lim_{\tau \uparrow 0} \psi^1(\tau) + A_0^T\psi^0,$$

$$(5.7) \quad [A^*\psi]^1(\tau) = A_{01}^T(\tau)\psi^0 - \frac{d^+}{d\theta} \psi^1(\tau) \\ + \sum_{i=1}^{p-1} \delta_i(\tau)(A_i^T\psi^0 - \psi^1(-h_i) + \lim_{\tau \uparrow -h_i} \psi^1(\tau)) \\ + \delta_p(\tau)(A_p^T\psi^0 - \psi^1(-h))$$

for $\phi, \psi \in X^N$, where δ_i denotes the Dirac delta impulse at $-h_i$, $i = 0, \dots, p$. Below we shall introduce the operators A^N and $(A^N)^*$ by projecting these formal extensions formally back into the subspace X^N . Since jumps of the function components of elements in X^N may occur at $\tau = -h_1$, we have two possible interpretations of δ_i as a functional on X^N , namely the evaluation of either the right-hand or the left-hand limit at $-h_i$. Correspondingly we introduce the following two types of approximate delta impulses which can be obtained. We define

$$\delta_{i,+}^N = \hat{E}^N \gamma_{i,+}^N, \quad i = 1, \dots, p, \\ \delta_{i,-}^N = \hat{E}^N \gamma_{i,-}^N, \quad i = 0, \dots, p-1,$$

where

$$Q^N \gamma_{i,+}^N = \text{col}(0, e_{11}^N(-h_i), \dots, e_{pk_N}^N(-h_i)), \\ Q^N \gamma_{i,-}^N = \text{col}(0, \lim_{\tau \uparrow -h_i} e_{11}^N(\tau), \dots, \lim_{\tau \uparrow -h_i} e_{pk_N}^N(\tau)).$$

The following lemma describes the action of the approximating delta-impulses.

LEMMA 5.1. For any $x \in \mathbb{R}^n$ and $\phi \in M^2$

$$\langle \delta_{i,+}^N x, \phi \rangle_{M^2} = x^T (\pi^N \phi^1)(-h_i), \quad i = 1, \dots, p, \\ \langle \delta_{i,-}^N x, \phi \rangle_{M^2} = x^T \lim_{\tau \uparrow -h_i} (\pi^N \phi^1)(\tau), \quad i = 0, \dots, p-1.$$

Proof. Using (5.2), (5.3) and the definition of $\delta_{i,+}^N$, we get

$$\langle \delta_{i,+}^N x, \phi \rangle_{M^2} = \langle \hat{E}^N \gamma_{i,+}^N x, p^N \phi \rangle_{M^2} = (\gamma_{i,+}^N x)^T Q^N \alpha(p^N \phi) \\ = x^T (0, e_{11}^N(-h_i), \dots, e_{pk_N}^N(-h_i)) \alpha^N(p^N \phi) \\ = x^T (\pi^N \phi^1)(-h_i).$$

The proof for $\delta_{i,-}^N$ is analogous. \square

The following definition of the operators A^N is obtained by formally projecting $A\phi$ as given by (5.4), (5.5) into X^N and putting $p^N\delta_i = \delta_{i,-}^N$, $i = 0, \dots, p-1$.

DEFINITION 5.2. For any $\phi = (\phi^0, \phi^1) \in X^N$ we define

$$\begin{aligned} A^N\phi = & \left(A_0\phi^0 + \sum_{i=1}^p A_i\phi^1(-h_i) + \int_{-h}^0 A_{01}(\tau)\phi^1(\tau) d\tau, \pi^N\left(\frac{d^+}{d\theta}\phi^1\right) \right) \\ & + \delta_{0,-}^N(\phi^0 - \lim_{\tau \uparrow 0} \phi^1(\tau)) + \sum_{i=1}^{p-1} \delta_{i,-}^N(\phi^1(-h_i) - \lim_{\tau \uparrow -h_i} \phi^1(\tau)). \end{aligned}$$

The adjoint operators $(A^N)^*$ are given in

LEMMA 5.3. For any $\psi = (\psi^0, \psi^1) \in X^N$ the operator $(A^N)^*$ is given by

$$\begin{aligned} (A^N)^*\psi = & \left(\lim_{\tau \uparrow 0} \psi^1(\tau) + A_0^T\psi^0, \pi^N\left(A_{01}^T\psi^0 - \frac{d^+}{d\theta}\psi^1\right) \right) \\ & + \sum_{i=1}^{p-1} \delta_{i,+}^N(A_i^T\psi^0 + \lim_{\tau \uparrow -h_i} \psi^1(\tau) - \psi^1(-h_i)) \\ & + \delta_{p,+}^N(A_p^T\psi^0 - \psi^1(-h)). \end{aligned}$$

Proof. By definition of the adjoint operator we get

$$\begin{aligned} \langle (A^N)^*\psi, \phi \rangle_{M^2} &= \langle \psi, A^N\phi \rangle_{M^2} \\ &= (\psi^0)^T \left[A_0\phi^0 + \sum_{i=1}^p A_i\phi^1(-h_i) + \int_{-h}^0 A_{01}(\tau)\phi^1(\tau) d\tau \right] \\ &\quad + \left\langle \psi^1, \pi^N\left(\frac{d^+}{d\theta}\phi^1\right) \right\rangle_{L^2} + \langle \psi, \delta_{0,-}^N(\phi^0 - \lim_{\tau \uparrow 0} \phi^1(\tau)) \rangle_{M^2} \\ &\quad + \sum_{i=1}^{p-1} \langle \psi, \delta_{i,-}^N(\phi^1(-h_i) - \lim_{\tau \uparrow -h_i} \phi^1(\tau)) \rangle_{M^2} \end{aligned}$$

for any $\phi = (\phi^0, \phi^1)$, $\psi = (\psi^0, \psi^1)$ in X^N . By Lemma 5.1 we see

$$\begin{aligned} \langle \psi, \delta_{0,-}^N(\phi^0 - \lim_{\tau \uparrow 0} \phi^1(\tau)) \rangle_{M^2} &= \lim_{\tau \uparrow 0} \psi^1(\tau)^T (\phi^0 - \lim_{\tau \uparrow 0} \phi^1(\tau)), \\ \langle \psi, \delta_{i,-}^N(\phi^1(-h_i) - \lim_{\tau \uparrow -h_i} \phi^1(\tau)) \rangle_{M^2} \\ &= \lim_{\tau \uparrow -h_i} \psi^1(\tau)^T (\phi^1(-h_i) - \lim_{\tau \uparrow -h_i} \phi^1(\tau)), \quad i = 1, \dots, p-1. \end{aligned}$$

Furthermore,

$$\begin{aligned} \left\langle \psi^1, \pi^N\left(\frac{d^+}{d\theta}\phi^1\right) \right\rangle_{L^2} &= \left\langle \psi^1, \frac{d^+}{d\theta}\phi^1 \right\rangle_{L^2} = \sum_{i=1}^p \int_{-h_i}^{-h_{i-1}} \psi^1(\tau)^T \left(\frac{d^+}{d\theta}\phi^1\right)(\tau) d\tau \\ &= \sum_{i=1}^p \left[\lim_{\tau \uparrow -h_{i-1}} \psi^1(\tau)^T \phi^1(\tau) - \psi^1(-h_i)^T \phi^1(-h_i) \right] \\ &\quad - \left\langle \pi^N\left(\frac{d^+}{d\theta}\psi^1\right), \phi^1 \right\rangle_{L^2} \end{aligned}$$

and

$$(\psi^0)^T \int_{-h}^0 A_{01}(\tau)\phi^1(\tau) d\tau = \langle A_{01}^T\psi^0, \phi^1 \rangle_{L^2} = \langle \pi^N(A_{01}^T\psi^0), \phi^1 \rangle_{L^2}.$$

Altogether we have

$$\begin{aligned} \langle (A^N)^* \psi, \phi \rangle_{M^2} = & [\lim_{\tau \uparrow 0} \psi^1(\tau) + A_0^T \psi^0]^T \phi^0 \\ & + \sum_{i=1}^{p-1} [A_i^T \psi^0 + \lim_{\tau \uparrow -h_i} \psi^1(\tau) - \psi^1(-h_i)]^T \phi^1(-h_i) \\ & + [A_p^T \psi^0 - \psi^1(-h)]^T \phi^1(-h) + \left\langle \pi^N \left(A_{01}^T \psi^0 - \frac{d^+}{d\theta} \psi^1 \right), \phi^1 \right\rangle_{L^2}. \end{aligned}$$

The result now follows using Lemma 5.1. \square

Note, that $(A^N)^* \psi$ can formally be obtained by projecting $A^* \psi$ as given by (5.6), (5.7) into X^N but now putting $p^N \delta_i = \delta_{i+}^N$, $i = 1, \dots, p$. Without any additional assumption we have the following.

LEMMA 5.4. *Hypothesis (H2) is valid for the sequence A^N , $N = 1, 2, \dots$ and therefore also for the sequence $(A^N)^*$, $N = 1, 2, \dots$.*

Proof. We introduce an equivalent inner product on M^2 by

$$\langle \phi, \psi \rangle_g = (\phi^0)^T \psi^0 + \int_{-h}^0 \phi^1(\tau)^T \psi^1(\tau) g(\tau) d\tau, \quad \phi, \psi \in M^2,$$

where the weighting function g is right-hand continuous on $[-h, 0)$ and

$$g(\tau) = p - i + 1 \quad \text{for } \tau \in [-h_i, -h_{i-1}), \quad i = 1, \dots, p.$$

It is clear that the corresponding norm $\|\cdot\|_g$ on M^2 is equivalent to the original norm,

$$\|\phi\| \leq \|\phi\|_g \leq \sqrt{p} \|\phi\|, \quad \phi \in M^2.$$

Since $(\phi^0, \phi^1 g) \in X^N$ for any $\phi \in X^N$, we obtain from Lemma 5.1

$$\langle \delta_{i-}^N x, \phi \rangle_g = (p - i) x^T \lim_{\tau \uparrow -h_i} \phi^1(\tau), \quad i = 0, \dots, p - 1,$$

for $x \in \mathbb{R}^n$ and $\phi \in X^N$. Using this equation and Definition 5.2, we get for $\phi \in X^N$

$$\begin{aligned} \langle A^N \phi, \phi \rangle_g = & \left[A_0 \phi^0 + \sum_{i=1}^p A_i \phi^i(-h_i) + \int_{-h}^0 A_{01}(\tau) \phi^1(\tau) d\tau \right]^T \phi^0 \\ & + \left\langle \pi^N \left(\frac{d^+}{d\theta} \phi^1 \right), \phi^1 g \right\rangle_{L^2} + p [\phi^0 - \lim_{\tau \uparrow 0} \phi^1(\tau)]^T \lim_{\tau \uparrow 0} \phi^1(\tau) \\ & + \sum_{i=1}^{p-1} (p - i) [\phi^1(-h_i) - \lim_{\tau \uparrow -h_i} \phi^1(\tau)]^T \lim_{\tau \uparrow -h_i} \phi^1(\tau). \end{aligned}$$

Obviously $\pi^N(\phi^1 g) = \phi^1 g$ and hence

$$\begin{aligned} \left\langle \pi^N \left(\frac{d^+}{d\theta} \phi^1 \right), \phi^1 g \right\rangle_{L^2} &= \left\langle \frac{d^+}{d\theta} \phi^1, \phi^1 g \right\rangle_{L^2} \\ &= \sum_{i=1}^p (p - i + 1) \int_{-h_i}^{-h_{i-1}} \phi^1(\tau)^T \phi^1(\tau) d\tau \\ &= \frac{1}{2} \sum_{i=1}^p (p - i + 1) \left[\lim_{\tau \uparrow -h_{i-1}} |\phi^1(\tau)|^2 - |\phi^1(-h_i)|^2 \right]. \end{aligned}$$

Using this and several times the inequality $\alpha\beta \leq \frac{1}{2}\alpha^2 + \frac{1}{2}\beta^2$, we get for $\phi \in X^N$

$$\begin{aligned} \langle A^N \phi, \phi \rangle_g &\leq \left(|A_0| + \frac{1}{2} \sum_{i=1}^p |A_i|^2 + \|A_{01}\|_{L^2} \right) \|\phi\|_g^2 \\ &\quad + \frac{1}{2} \sum_{i=1}^p |\phi^1(-h_i)|^2 + \frac{p}{2} |\phi^0|^2 - \frac{p}{2} \lim_{\tau \uparrow 0} |\phi^1(\tau)|^2 \\ &\quad + \frac{1}{2} \sum_{i=1}^p (p-i+1) \lim_{\tau \uparrow -h_{i-1}} |\phi^1(\tau)|^2 - |\phi^1(-h_i)|^2 \\ &\quad + \frac{1}{2} \sum_{i=2}^p (p-i+1) [|\phi^1(-h_{i-1})|^2 - \lim_{\tau \uparrow -h_{i-1}} |\phi^1(\tau)|^2] \\ &\leq \omega \|\phi\|_g^2 \end{aligned}$$

with $\omega = p/2 + |A_0| + \frac{1}{2} \sum_{i=1}^p |A_i|^2 + \|A_{01}\|_{L^2}$. This proves (H2*) with $\|\cdot\|_N = \|\cdot\|_g$ for all N . Since $\|S^N(t)\| = \|S^N(t)^*\|$ the proof is finished. \square

In order to verify hypothesis (H3), we need additional assumptions concerning the convergence properties of $\pi^N \phi^1$, $N = 1, 2, \dots$, for ϕ in a suitably restricted subset of M^2 . Observe, that we get from (5.1)

$$\pi^N \phi^1 = \sum_{j=1}^p \pi_j^N(\phi^1 | [-h_j, -h_{j-1})),$$

where π_j^N is the orthogonal projection $L^2(-h_j, -h_{j-1}; \mathbb{R}^n) \rightarrow \text{span}(e_{j1}^N I, \dots, e_{jk_N}^N I)$. We define the sets \tilde{D} and \tilde{D}^* by

$$\begin{aligned} \tilde{D} &= \{\phi \in M^2 \mid \phi^0 = \phi^1(0), \phi^1 \in W^{2,2}(-h, 0; \mathbb{R}^n)\}, \\ \tilde{D}^* &= \left\{ \psi \in M^2 \mid \psi^1(-h) = A_p^T \psi^0, \psi^1 + \sum_{j=1}^{p-1} A_i^T \psi^0 \chi_{[-h_i, -h_{i-1})} \in W^{1,2}(-h, 0; \mathbb{R}^n), \right. \\ &\quad \left. \text{and } A_{01}^T \psi^0 - \frac{d^+}{d\theta} \psi^1 \in W^{1,2}(-h_i, -h_{i-1}; \mathbb{R}^n), i = 1, \dots, p \right\}. \end{aligned}$$

Obviously, $\text{dom } A^2 \subset \tilde{D} \subset \text{dom } A$ and $\text{dom } (A^*)^2 \subset \tilde{D}^* \subset \text{dom } A^*$. Furthermore we put

$$\kappa(N) = \max(\|\delta_{0,-}^N\|, \dots, \|\delta_{p-1,-}^N\|, \|\delta_{1,+}^N\|, \dots, \|\delta_{p,+}^N\|),$$

where $\|\delta_{i,\pm}^N\|$ is the norm of $\delta_{i,\pm}^N$ considered as an operator $\mathbb{R}^n \rightarrow M^2$. We impose the following hypothesis:

(A) There exists a sequence $\rho(N)$ with $\lim_{N \rightarrow \infty} \rho(N) = 0$ such that for $i = 1, \dots, p$

$$\begin{aligned} \text{(i)} \quad (1 + \kappa(N)) \|f - \pi_i^N f\|_{L^\infty} + \|f - \pi_i^N f\|_{L^2} \\ + \left\| \frac{d}{d\theta} (f - \pi_i^N f) \right\|_{L^2} \leq \rho(N) \|f\|_{W^{2,2}} \end{aligned}$$

for all $f \in W^{2,2}(-h_i, -h_{i-1}; \mathbb{R}^n)$, and

$$\text{(ii)} \quad \|f - \pi_i^N f\|_{L^2} \leq \rho(N) \|f\|_{W^{1,2}}$$

for all $f \in W^{1,2}(-h_i, -h_{i-1}; \mathbb{R}^n)$.

LEMMA 5.5. Assume that (A) is satisfied. Then the following is true:

(a) Assumption (H1) is satisfied.

(b) For any $\phi \in \tilde{D}$

$$\|A^N p^N \phi - A\phi\|_{M^2} \leq \gamma_0 \rho(N) \|\phi^1\|_{W^{2,2}}, \quad N = 1, 2, \dots,$$

where γ_0 is a constant independent of ϕ and N .

(c) If in addition $A_{01} \in W^{1,2}(-h_i, -h_{i-1}; \mathbb{R}^{n \times n})$, $i = 1, \dots, p$, then for any $\psi \in \tilde{D}^*$

$$\begin{aligned} & \| (A^N)^* p^N \psi - A^* \psi \|_{M^2} \\ & \leq \gamma_0 \rho(N) \sum_{i=1}^p (\| \psi^1 \|_{W^{2,2}(-h_i, -h_{i-1}; \mathbb{R}^n)} + \| A_{01}^T \psi^0 \|_{W^{1,2}(-h_i, -h_{i-1}; \mathbb{R}^n)}), \end{aligned}$$

$N = 1, 2, \dots$, where again γ_0 is a constant independent of ϕ and N .

Proof. (a) is obvious from (A). In order to prove (b), we put $\phi^N = \pi^N \phi^1$ and get, using (2.6) and Definition 5.2, for $\phi \in \tilde{D}$

$$\begin{aligned} & \| A^N p^N \phi - A \phi \|_{M^2} \\ & \leq \sum_{i=1}^p |A_i| \| \phi^N - \phi^1 \|_{L^\infty} + \| A_{01} \|_{L^2} \| \phi^N - \phi^1 \|_{L^2} \\ & \quad + \left\| \pi^N \left(\frac{d^+}{d\theta} \phi^N - \dot{\phi}^1 \right) \right\|_{L^2} + \| \pi^N(\dot{\phi}^1) - \dot{\phi}^1 \|_{L^2} \\ & \quad + \| \delta_{0,-}^N \| | \phi^0 - \lim_{\tau \uparrow 0} \phi^N(\tau) | + \sum_{i=1}^{p-1} \| \delta_{i,+}^N \| | \phi^N(-h_i) - \lim_{\tau \uparrow -h_i} \phi^N(\tau) | \\ & \leq \sum_{i=1}^p |A_i| \| \phi^N - \phi^1 \|_{L^\infty} + \| A_{01} \|_{L^2} \| \phi^N - \phi^1 \|_{L^2} \\ & \quad + \left\| \frac{d^+}{d\theta} (\phi^N - \phi^1) \right\|_{L^2} + \| \pi^N(\dot{\phi}^1) - \dot{\phi}^1 \|_{L^2} \\ & \quad + \kappa(N) \left\{ | \phi^1(0) - \lim_{\tau \uparrow 0} \phi^N(\tau) | + \sum_{i=1}^{p-1} | \phi^N(-h_i) - \phi^1(-h_i) | + \sum_{i=1}^{p-1} | \phi^1(-h_i) - \lim_{\tau \uparrow -h_i} \phi^N(\tau) | \right\} \\ & \leq \left[\sum_{i=1}^p |A_i| + (2p-1)\kappa(N) \right] \| \phi^N - \phi^1 \|_{L^\infty} \\ & \quad + \| A_{01} \|_{L^2} \| \phi^N - \phi^1 \|_{L^2} + \left\| \frac{d^+}{d\theta} (\phi^N - \phi^1) \right\|_{L^2} + \| \pi^N(\dot{\phi}^1) - \dot{\phi}^1 \|_{L^2} \\ & \leq \gamma_0 \rho(N) \| \phi^1 \|_{W^{2,2}}, \end{aligned}$$

where in the last step we have used assumption (A). γ_0 is an appropriately chosen constant.

(c) Using (2.7), Lemma 5.3 and $\psi \in D(A^*)$, we get

$$\begin{aligned} & \| (A^N)^* p^N \psi - A^* \psi \|_{M^2} \leq | \lim_{\tau \uparrow 0} \psi^N(\tau) - \psi^1(0) | \\ & \quad + \left\| \pi^N \left(A_{01}^T \psi^0 - \frac{d^+}{d\theta} \psi^N \right) - A_{01}^T \psi^0 + \frac{d^+}{d\theta} \psi^1 \right\|_{L^2} \\ & \quad + \sum_{i=1}^{p-1} \| \delta_{i,+}^N \| | A_i^T \psi^0 + \lim_{\tau \uparrow -h_i} \psi^N(\tau) - \psi^N(-h_i) | \\ & \quad + \| \delta_{p,+}^N \| | A_p^T \psi^0 - \psi^N(-h) | \\ & \leq (1 + (2p-1)\kappa(N)) \| \psi^1 - \psi^N \|_{L^\infty} \\ & \quad + \left\| \pi^N \left(A_{01}^T \psi^0 - \frac{d^+}{d\theta} \psi^1 \right) - \left(A_{01}^T \psi^0 - \frac{d^+}{d\theta} \psi^1 \right) \right\|_{L^2} \\ & \quad + \left\| \frac{d^+}{d\theta} (\psi^1 - \psi^N) \right\|_{L^2}. \end{aligned}$$

Under the given assumptions we have $\psi^1 \in W^{2,2}(-h_i, -h_{i-1}; \mathbb{R}^n)$, $i = 1, \dots, p$. Therefore we get from (A), (ii)

$$\begin{aligned} & \left\| \pi^N \left(A_{01}^T \psi^0 - \frac{d^+}{d\theta} \psi^1 \right) - \left(A_{01}^T \psi^0 - \frac{d^+}{d\theta} \psi^1 \right) \right\|_{L^2} \\ & \leq \rho(N) \sum_{i=1}^p \left\| A_{01}^T \psi^0 - \frac{d^+}{d\theta} \psi^1 \right\|_{W^{1,2}(-h_i, -h_{i-1}; \mathbb{R}^n)} \\ & \leq \rho(N) \sum_{i=1}^p (\|A_{01}^T \psi^0\|_{W^{1,2}(-h_i, -h_{i-1}; \mathbb{R}^n)} + \|\psi^1\|_{W^{2,2}(-h_i, -h_{i-1}; \mathbb{R}^n)}) \end{aligned}$$

and

$$\begin{aligned} \|(A^N)^* p^N \psi - A^* \psi\|_{M^2} & \leq \gamma_0 \rho(N) \sum_{i=1}^p (\|A_{01}^T \psi^0\|_{W^{1,2}(-h_i, -h_{i-1}; \mathbb{R}^n)} \\ & \quad + \|\psi^1\|_{W^{2,2}(-h_i, -h_{i-1}; \mathbb{R}^n)}). \quad \square \end{aligned}$$

It is clear that under the conditions of Lemma 5.5, hypothesis (H3), (i) is satisfied for X^N , p^N , A^N , $N = 1, 2, \dots$, if we take $D = \text{dom } A^2$, and for X^N , p^N , $(A^N)^*$, $N = 1, 2, \dots$, if we take $D = \text{dom } (A^*)^2$. The next lemma establishes (H3), (ii).

LEMMA 5.6. Assume that (A) is satisfied.

(a) There exist constants $\tilde{M} \geq 1$ and $\omega \in \mathbb{R}$ such that for all $\phi \in \text{dom } A^2$

$$\|A^N p^N S(t) \phi\|_{M^2} \leq \tilde{M} e^{\omega t} |\phi|_2, \quad t \geq 0, \quad N = 1, 2, \dots,$$

where $|\phi|_2 = \|\phi\| + \|A\phi\| + \|A^2\phi\|$.

(b) If in addition $A_{01} \in W^{1,2}(-h_i, -h_{i-1}; \mathbb{R}^{n \times n})$, $i = 1, \dots, p$, then there exist constants $M^* \geq 1$ and $\omega \in \mathbb{R}$ such that for all $\psi \in \text{dom } (A^*)^2$

$$\|(A^N)^* p^N S^*(t) \psi\|_{M^2} \leq M^* e^{\omega t} |\psi|_2^*, \quad t \geq 0, \quad N = 1, 2, \dots,$$

where $|\psi|_2^* = \|\psi\| + \|A^* \psi\| + \|(A^*)^2 \psi\|$.

Proof. (a) Since $S(\cdot)$ restricted to $\text{dom } A^2$ is a C_0 -semigroup on $\text{dom } A^2$ equipped with the graph norm $|\cdot|_2$, we have

$$|S(t) \phi|_2 \leq M e^{\omega t} |\phi|_2, \quad t \geq 0, \quad \phi \in \text{dom } A^2,$$

with some constants $M \geq 1$, $\omega \in \mathbb{R}$. From

$$A^k \phi = \left(L \left(\frac{d^{k-1}}{d\theta^{k-1}} \phi^1 \right), \frac{d^k}{d\theta^k} \phi^1 \right),$$

$k = 1, 2, \dots$, $\phi \in \text{dom } A^k$, we see that

$$\|\phi^1\|_{W^{2,2}} \leq |\phi|_2, \quad \phi \in \text{dom } A^2.$$

Therefore, for $\phi \in \text{dom } A^2$

$$\|(S(t) \phi)^1\|_{W^{2,2}} \leq M e^{\omega t} |\phi|_2, \quad t \geq 0$$

and by Lemma 5.5, (b)

$$\begin{aligned} \|A^N p^N S(\phi)\| & \leq \|AS(t) \phi\| + \|A^N p^N S(t) \phi - AS(t) \phi\| \\ & \leq M e^{\omega t} \|A\phi\| + \gamma_0 \rho(N) M e^{\omega t} |\phi|_2 \\ & \leq M(1 + \gamma_0 \rho(N)) e^{\omega t} |\phi|_2, \quad t \geq 0, \quad N = 1, 2, \dots \end{aligned}$$

(b) As in part (a) we have

$$|S^*(t)\psi|_2^* \leq M e^{\omega t} |\psi|_2^*, \quad t \geq 0, \quad \psi \in \text{dom}(A^*)^2.$$

Using (2.7), $(A^*)^2\psi = (\dots, A_{01}^T(\psi^1(0) + A_0^T\psi^0) - d^+/d\theta(A_{01}^T\psi^0 - d^+/d\theta\psi^1))$ and $d^+/d\theta(A_{01}^T\psi^0 - d^+/d\theta\psi^1) = A_{01}^T\psi^0 - \dot{\psi}^1$ on the intervals $[-h_i, -h_{i-1}]$, $i = 1, \dots, p$, it is not difficult to see that for a constant $\kappa = \kappa(\|A_{01}\|_{L^2}, \|d^+/d\theta A_{01}\|_{L^2})$ the following estimate is valid:

$$\sum_{i=1}^p \|\psi^1\|_{W^{2,2}(-h_i, -h_{i-1}; \mathbb{R}^n)} \leq \kappa |\psi|_2^*, \quad \psi \in \text{dom}(A^*)^2.$$

The rest of the proof is analogous to that for part (a) but now using Lemma 5.5, (c). \square

From Lemmas 5.4–5.6 it immediately follows that under the assumption specified in these lemmas Theorem 4.1 applies to the sequences $X^N, p^N, A^N, (A^N)^*$, $N = 1, 2, 3, \dots$, defined in this section. The approximating control systems on X^N are given by (compare (4.5) and (4.9))

$$\begin{aligned} \dot{w}^N(t) &= A^N w^N(t) + B^N u(t), \\ (\Sigma^N) \quad y^N(t) &= C^N w^N(t), \quad t \geq 0, \\ w^N(0) &= p^N \phi, \quad \phi \in M^2, \end{aligned}$$

with the cost functional

$$(5.8) \quad J^N(u) = \langle w^N(T), G w^N(T) \rangle + \int_0^T [|y^N(t)|^2 + u(t)^T R u(t)] dt,$$

where the input and output operators are given by

$$\begin{aligned} B^N u &= p^N B u = B u, \quad u \in \mathbb{R}^l, \\ C^N \phi &= C p^N \phi = C \phi, \quad \phi \in X^N. \end{aligned}$$

As in § 3.2 we assume that $R \in \mathbb{R}^{l \times l}$ is positive definite, $G_0 \in \mathbb{R}^{n \times n}$ is positive semidefinite and $G: M^2 \rightarrow M^2$ is defined by $G\phi = (G_0\phi^0, 0)$. The Riccati operators corresponding to (Σ^N) and (5.8) satisfy

$$\Pi^N(t) = p^N \Pi(t) p^N, \quad 0 \leq t \leq T,$$

and (restricted to X^N)

$$\begin{aligned} &\frac{d}{dt} \Pi^N(t) + (A^N)^* \Pi^N(t) + \Pi^N(t) A^N \\ (5.9) \quad &- \Pi^N(t) B^N R^{-1} (B^N)^* \Pi^N(t) + (C^N)^* C^N = 0, \quad 0 \leq t \leq T, \\ &\Pi^N(T) = G^N. \end{aligned}$$

Here G^N is the restriction of G to X^N .

It follows from the results of this section that Theorem 4.3 applies to system (Σ^N) with (5.8). More precisely, we have the following theorem which may be considered as the main result of this paper.

THEOREM 5.7. *Let hypothesis (A) be satisfied and assume that $A_{01} \in W^{1,2}(-h_i, -h_{i-1}; \mathbb{R}^{n \times n})$ for $i = 1, \dots, p$. Then*

$$\lim_{N \rightarrow \infty} \|\Pi(t) - \Pi^N(t)\| = 0$$

uniformly for $0 \leq t \leq T$.

Of course, also assertions (a)–(c) of Theorem 4.3 are true for system (Σ^N) , (5.8) and (Σ) , (3.1) (or equivalently (2.1), (3.1)).

5.2. Matrix representations. For implementation of a scheme obtained along the lines described in the previous section we have to compute matrix representations $[A^N]$, $[(A^N)^*]$, $[B^N]$, $[C^N]$ and $[G^N]$ of the operators A^N , $(A^N)^*$, B^N , C^N and G^N , respectively, with respect to the basis \hat{E}^N . How to compute the coordinate vector of $p^N \phi$ for $\phi \in M^2$ has already been shown (see (5.2)).

Define the $k_N \times k_N$ -matrices h_i^N , $i = 1, \dots, p$, and g_i^N , $i = 1, \dots, p-1$, by

$$h_i^N = \left(\left\langle e_{il}^N, \frac{d^+}{d\theta} e_{im}^N \right\rangle_{L^2} - \lim_{\tau \uparrow -h_{i-1}} e_{il}^N(\tau) e_{im}^N(\tau) \right)_{l,m=1, \dots, k_N},$$

$$g_i^N = (\lim_{\tau \uparrow -h_i} e_{i+1,l}^N(\tau) e_{i,m}^N(-h_i))_{l,m=1, \dots, k_N}$$

and put

$$A_{ij}^N = \langle A_{01}, e_{ij}^N I \rangle_{L^2}, \quad i = 1, \dots, p, \quad j = 1, \dots, k_N,$$

$$\alpha_i^N = (A_i e_{i1}^N(-h_i) + A_{i1}^N, \dots, A_i e_{ik_N}^N(-h_i) + A_{ik_N}^N) \in \mathbb{R}^{n \times nk_N},$$

$i = 1, \dots, p$, and

$$\beta^N = \text{col} \left(\lim_{\tau \uparrow 0} e_{11}^N(\tau) I, \dots, \lim_{\tau \uparrow 0} e_{1k_N}^N(\tau) I \right) \in \mathbb{R}^{nk_N \times n}.$$

Furthermore, define the $(pk_N + 1)n \times (pk_N + 1)n$ -matrix H^N by

$$H^N = \begin{pmatrix} A_0 & \alpha_1^N & \cdots & \alpha_p^N \\ \beta^N & h_1^N \otimes I & \begin{array}{c} 0 \text{---} 0 \\ \diagdown \quad \diagup \\ 0 \end{array} & 0 \\ 0 & g_1^N \otimes I & \begin{array}{c} \diagdown \quad \diagup \\ 0 \end{array} & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \begin{array}{c} 0 \text{---} 0 \\ \diagdown \quad \diagup \\ 0 \end{array} & 0 \end{pmatrix}.$$

PROPOSITION 5.8. (a) $[A^N] = (Q^N)^{-1} H^N$.

(b) $[(A^N)^*] = (Q^N)^{-1} (H^N)^T$.

(c) $[B^N] = \text{col} (B_0, 0, \dots, 0) \in \mathbb{R}^{n(pk_N+1) \times l}$.

(d) $[C^N] = (C_0, 0, \dots, 0) \in \mathbb{R}^{m \times n(pk_N+1)}$.

(e) $[G^N] = \begin{pmatrix} G_0 & 0 & \cdots & 0 \\ 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 0 \end{pmatrix} \in \mathbb{R}^{n(pk_N+1) \times n(pk_N+1)}$.

Proof. $[A^N]$ is characterized by

$$\alpha^N(A^N \phi) = [A^N] \alpha^N(\phi), \quad \phi \in X^N.$$

On the other hand we get from (5.2) and $\phi = \hat{E}^N \alpha^N(\phi)$

$$\alpha^N(A^N \phi) = (Q^N)^{-1} d^N(A^N \phi) = (Q^N)^{-1} \langle \hat{E}^N, A^N \hat{E}^N \rangle \alpha^N(\phi),$$

i.e.,

$$[A^N] = (Q^N)^{-1} \langle \hat{E}^N, A^N \hat{E}^N \rangle.$$

From Definition 5.2 we get

$$\begin{aligned} A^N \hat{e}_0^N &= (A_0, 0) + \delta_{0,-}^N, \\ A^N \hat{e}_{ij}^N &= \left(A_i e_{ij}^N(-h_i) + A_{ij}^N, \pi^N \left(\frac{d^+}{d\theta} e_{ij}^N \right) \right) \\ &\quad - \delta_{i-1,-}^N \lim_{\tau \uparrow -h_{i-1}} e_{ij}^N(\tau) + \delta_{i,-}^N e_{ij}^N(-h_i), \end{aligned}$$

$i = 1, \dots, p-1, j = 1, \dots, k_N$, and

$$\begin{aligned} A^N \hat{e}_{pj}^N &= \left(A_p e_{pj}^N(-h) + A_{pj}^N, \pi^N \left(\frac{d^+}{d\theta} e_{pj}^N \right) \right) \\ &\quad - \delta_{p-1,-}^N \lim_{\tau \uparrow -h_{p-1}} e_{pj}^N(\tau), \quad j = 1, \dots, k_N. \end{aligned}$$

Observing

$$\left\langle e_{il}^N, \pi^N \left(\frac{d^+}{d\theta} e_{km}^N \right) \right\rangle_{L^2} = \left\langle e_{il}^N, \frac{d^+}{d\theta} e_{km}^N \right\rangle_{L^2},$$

Lemma 5.1 and (5.2) we obtain by straightforward calculation

$$\langle \hat{E}^N, A^N \hat{E}^N \rangle = H^N.$$

In order to prove the representation for $[(A^N)^*]$, we use (5.3) and get

$$\begin{aligned} \alpha^N(\phi)^T Q^N [(A^N)^*] \alpha^N(\psi) &= \alpha^N(\phi)^T Q^N \alpha^N((A^N)^* \psi) \\ &= \langle \phi, (A^N)^* \psi \rangle = \langle A^N \phi, \psi \rangle = \alpha^N(A^N \phi)^T Q^N \alpha^N(\psi) \\ &= \alpha^N(\phi)^T [A^N]^T Q^N \alpha^N(\psi) = \alpha^N(\phi)^T (H^N)^T \alpha^N(\psi) \end{aligned}$$

for all $\phi, \psi \in X^N$, i.e.,

$$[(A^N)^*] = (Q^N)^{-1} (H^N)^T.$$

For $u \in \mathbb{R}^l$ we have

$$\begin{aligned} \alpha^N(B^N u) &= \alpha^N((B_0 u, 0)) = (Q^N)^{-1} d^N((B_0 u, 0)) \\ &= (Q^N)^{-1} \text{col}(B_0 u, 0, \dots, 0) = \text{col}(B_0, 0, \dots, 0) u, \end{aligned}$$

which proves the given form of $[B^N]$. The proofs for $[C^N]$ and $[G^N]$ are analogous. \square

It is obvious from Proposition 5.8 that $[(A^N)^*] = (Q^N)^{-1} [A^N] Q^N$. Therefore we do not get the standard Riccati matrix differential equation if we take everywhere in (5.9) the matrix representations of the operators involved. In order to overcome this difficulty we define

$$(5.10) \quad \Gamma^N(t) = Q^N [\Pi^N(T-t)], \quad 0 \leq t \leq T,$$

and get from (5.9) the standard Riccati equation for $\Gamma^N(t)$

$$\begin{aligned} (5.11) \quad \frac{d}{dt} \Gamma^N &= [A^N]^T \Gamma^N + \Gamma^N [A^N] \\ &\quad - \Gamma^N [B^N] R^{-1} [B^N]^T \Gamma^N + [C^N]^T [C^N], \quad 0 \leq t \leq T, \\ \Gamma^N(0) &= [G^N]. \end{aligned}$$

Note, that $[(B^N)^*] = [B^N]^T$, $[(C^N)^*] = [C^N]^T$ and that $\Pi^N(t)^* = \Pi^N(t)$ implies $\Gamma^N(t)^T = \Gamma^N(t)$.

Equation (5.11) can advantageously be solved using a method due to Casti and Kailath (see for instance [35, pp. 304 ff.]) which we indicate here for $p = 1$ and $A_{01} \equiv 0$. We define

$$W_0 = A_0^T G_0 + G_0 A - G_0 B_0 R^{-1} B_0^T G_0 + C_0^T C_0$$

and the $2n \times (k_N + 1)n$ -matrices

$$F_1^N = \begin{pmatrix} W_0 & 0 & \cdots & 0 & G_0 A_1 \\ A_1^T G_0 & 0 & \cdots & 0 & 0 \end{pmatrix}, \quad F_2^N = \begin{pmatrix} I & 0 & \cdots & 0 \\ 0 & \cdots & 0 & I \end{pmatrix}.$$

Then

$$\Gamma^N(t) = [G^N] + \int_0^t L_1^N(\tau)^T L_2^N(\tau) d\tau,$$

where

$$\begin{aligned} \frac{d}{dt} L_i^N(t) &= L_i^N(t) ([A^N] - [B^N] R^{-1} [B^N]^T \Gamma^N(t)), \\ L_i^N(0) &= F_i^N, \quad i = 1, 2. \end{aligned}$$

Note that this is a system of $4n^2(k_N + 1)$ differential equations compared to the $n^2(k_N + 1)^2$ differential equations of system (5.11) (in case $p = 1$, $A_{01} \equiv 0$).

If in the optimal feedback law (3.13) for the delay system (2.1) we use $\Pi^N(t)$ instead of $\Pi(t)$ we get the suboptimal controls (compare (4.17))

$$(5.12) \quad \hat{u}^N(t) = -R^{-1} B^* \Pi^N(t) p^N(\hat{x}^N(t), \hat{x}_t^N), \quad 0 \leq t \leq T,$$

where $\hat{x}^N(t)$ is the solution of (2.1) with $u(t) = \hat{u}^N(t)$. We introduce the $n \times n$ -matrices $\Pi_0^N(t)$, $\Pi_{11}^N(t)$, \dots , $\Pi_{pk_N}^N(t)$ by

$$(5.13) \quad \Pi^N(t) = \begin{pmatrix} \Pi_0^N(t) & * & \cdots & * \\ \Pi_{11}^N(t) & & & \\ \vdots & & & \\ \Pi_{pk_N}^N(t) & * & \cdots & * \end{pmatrix}$$

and define

$$(5.14) \quad \Pi_1^N(t, \tau) = \sum_{i=1}^p \sum_{j=1}^{k_N} \Pi_{ij}^N(t)^T e_{ij}^N(\tau)$$

for $-h \leq \tau \leq 0$ and $0 \leq t \leq T$. Then the control law (5.12) takes the following form:

$$\begin{aligned} \hat{u}^N(t) &= -R^{-1} [B^N]^T [\Pi^N(t)] (Q^N)^{-1} d^N((\hat{x}^N(t), \hat{x}_t^N)) \\ &= -R^{-1} (B_0^T, 0, \dots, 0) [\Pi^N(t)]^T d^N((\hat{x}^N(t), \hat{x}_t^N)) \\ (5.15) \quad &= -R^{-1} B_0^T \left\{ \Pi_0^N(t)^T \hat{x}^N(t) \right. \\ &\quad \left. + \sum_{i=1}^p \sum_{j=1}^{k_N} \Pi_{ij}^N(t)^T \int_{-h}^0 e_{ij}^N(\tau) \hat{x}^N(t+\tau) d\tau \right\} \\ &= -R^{-1} B_0^T \left\{ \Pi_0^N(t) \hat{x}^N(t) + \int_{-h}^0 \Pi_1^N(t, \tau) \hat{x}^N(t+\tau) d\tau \right\}. \end{aligned}$$

The additional condition used in [20] in order to prove convergence results analogous to those contained in Theorem 4.3 for the infinite time horizon problem in general cannot be satisfied for concrete realizations of the approximation scheme presented in this section. Especially this condition is not satisfied for the concrete realization of the scheme using spline functions as described in the next section. Despite this fact we did also numerical computations using the spline scheme for the infinite time horizon problem (see § 6.2). Therefore we conclude this section with a short description of the equations governing the approximation of this problem. Consider system (Σ^N) with the cost functional

$$(5.16) \quad J^N(u) = \int_0^\infty [|y^N(t)|^2 + u(t)^T R u(t)] dt.$$

The feedback law which minimizes $J^N(u)$ subject to (Σ^N) is governed by the Riccati operator Π^N which satisfies $\Pi^N = p^N \Pi^N p^N$ and (restricted to X^N) the algebraic Riccati equation

$$(5.17) \quad (A^N)^* \Pi^N + \Pi^N A^N - \Pi^N B R^{-1} B^* \Pi^N + C^* C = 0.$$

Analogously to (5.10) we define $\Gamma^N = Q^N [\Pi^N]$ and obtain the standard algebraic Riccati matrix equation

$$(5.18) \quad [A^N]^T \Gamma^N + \Gamma^N [A^N] - \Gamma^N [B^N] R^{-1} [B^N]^T \Gamma^N + [C^N]^T [C^N] = 0.$$

Note, that as for the finite time horizon problem, $\Gamma^N = (\Gamma^N)^T$ follows from $\Pi^N = (\Pi^N)^*$. Using Π^N instead of Π in the feedback law (3.16) we get by analogous computations as in (5.15) the suboptimal control law

$$(5.19) \quad \hat{u}^N = -R^{-1} B_0^T \left\{ \Pi_0^N \hat{x}^N(t) + \int_{-h}^0 \Pi_1^N(\tau) \hat{x}^N(t+\tau) d\tau \right\}, \quad t \geq 0,$$

where $\Pi_1^N(\tau) = \sum_{i=1}^p \sum_{j=1}^{k_N} (\Pi_{ij}^N)^T e_{ij}(\tau)$ and the $n \times n$ -matrices Π_0^N , Π_{ij}^N , $i=1, \dots, p$, $j=1, \dots, k_N$, are defined by

$$[\Pi^N] = \begin{pmatrix} \Pi_0^N & * & \cdots & * \\ \Pi_{11}^N & \vdots & \ddots & \vdots \\ \vdots & \vdots & \ddots & \vdots \\ \Pi_{pk_N}^N & * & \cdots & * \end{pmatrix}$$

(compare (5.13) and (5.14)).

5.3. The spline scheme. In this section we give a realization of the scheme developed in § 5.1 by using first order spline functions.

For $N=1, 2, \dots$ we choose the meshpoints

$$t_{ij}^N = -h_{i-1} - j \frac{r_i}{N}, \quad i=1, \dots, p, \quad j=0, \dots, N,$$

where $r_i = h_i - h_{i-1}$ and define the basis splines

$$e_{i0}^N(\tau) = \begin{cases} \frac{N}{r_i}(\tau - t_{i1}^N) & \text{for } t_{i1}^N \leq \tau < t_{i0}^N, \\ 0 & \text{elsewhere,} \end{cases}$$

$$e_{ij}^N(\tau) = \begin{cases} -\frac{N}{r_i}(\tau - t_{i,j-1}^N) & \text{for } t_{ij}^N \leq \tau < t_{i,j-1}^N, \\ \frac{N}{r_i}(\tau - t_{i,j+1}^N) & \text{for } t_{i,j+1}^N \leq \tau < t_{i,j}^N, \\ 0 & \text{elsewhere,} \end{cases}$$

$j = 1, \dots, N-1$, and

$$e_{iN}^N(\tau) = \begin{cases} -\frac{N}{r_i}(\tau - t_{i,N-1}^N) & \text{for } t_{iN}^N \leq \tau < t_{i,N-1}^N, \\ 0 & \text{elsewhere,} \end{cases}$$

$i = 1, \dots, p$. Note, that compared to § 5.1 we slightly have changed the enumeration of the e_{ij}^N (j running from 0 to N now). We have $k_N = N+1$ and therefore $\dim X^N = (p(N+1)+1)n$.

An easy computation shows that the matrices q_i^N which determine Q^N (recall $Q^N = \text{diag}(I, q_1^N \otimes I, \dots, q_p^N \otimes I)$) are given by $q_i^N = (r_i/N)q^N$, where

$$q^N = \begin{pmatrix} \frac{1}{3} & \frac{1}{6} & 0 & 0 \\ \frac{1}{6} & \frac{2}{3} & 0 & 0 \\ 0 & 0 & \frac{2}{3} & \frac{1}{6} \\ 0 & 0 & \frac{1}{6} & \frac{1}{3} \end{pmatrix} \in \mathbb{R}^{(N+1) \times (N+1)}.$$

For the approximating delta impulses we have

LEMMA 5.9. For all $N = 1, 2, \dots$,

$$\begin{aligned} \|\delta_{i+}^N\| &\leq \left(\frac{6N}{r_i}\right)^{1/2}, & i = 1, \dots, p, \\ \|\delta_{i-}^N\| &\leq \left(\frac{6N}{r_{i+1}}\right)^{1/2}, & i = 0, \dots, p-1. \end{aligned}$$

Proof. Using (5.3) and the definition of δ_{i+}^N and Q^N we get for any $x \in \mathbb{R}^n$

$$\begin{aligned} \|\delta_{i+}^N x\|^2 &= (\gamma_{i+}^N x)^T Q^N (Q^N)^{-1} Q^N \gamma_{i+}^N x \\ &= \frac{N}{r_i} x^T (0, \dots, 0, I) (q^N \otimes I)^{-1} \text{col}(0, \dots, 0, I) x \\ &\leq \frac{6N}{r_i} |x|^2, \end{aligned}$$

where we have used $\lambda_{\min}(q^N) \geq 1/6$. The estimate for δ_{i-}^N is analogous. \square

As a consequence of Lemma 5.9 we have

$$\kappa(N) \leq \left(\frac{6N}{\rho}\right)^{1/2} \quad \text{with } \rho = \min_{i=1, \dots, p} r_i.$$

PROPOSITION 5.10. Hypothesis (A) is satisfied for the spline scheme with $\rho(N) = \text{const}/N$. As a consequence Theorem 5.7 is valid for the spline scheme.

Proof. The following estimates are standard estimates for spline functions:

$$\|f - \pi_i^N f\|_{L^2} \leq \frac{\text{const}}{N^2} \|\ddot{f}\|_{L^2},$$

$$\left\| \frac{d}{d\theta} (f - \pi_i^N f) \right\|_{L^2} \leq \frac{\text{const}}{N} \|\dot{f}\|_{L^2}$$

for $f \in W^{2,2}(-h_i, -h_{i-1}; \mathbb{R}^n)$ (see [39, Thm. 6.5]) and

$$\|f - \pi_i^N f\|_{L^2} \leq \frac{\text{const}}{N} \|\dot{f}\|_{L^2}$$

for $f \in W^{1,2}(-h_i, -h_{i-1}; \mathbb{R}^n)$ (see [39, Exercise 6.1]). In order to get the estimate for the L^∞ -norm, we observe that $\pi_i^N f = \chi^N$, where χ^N is the cubic type *I* interpolating spline for $\Phi(\tau) = \int_{-h_i}^\tau \int_{-h_i}^\theta f(\sigma) d\sigma d\theta$, $\tau \in [-h_i, -h_{i-1}]$ (see, for instance, [39, Proof of Thm. 6.6]). Note, that interpolating cubic splines in [39] are always type *I* interpolating splines in the terminology of [1]. Then we get from [23, Thm. 5.7.1] (with $L = d^2/d\theta^2$ and $m = 2$) or [11, p. 235] (with $m = r = 2$, $q = \infty$)

$$\|f - \pi_i^N f\|_{L^\infty} \leq \text{const} \left(\frac{1}{N} \right)^{3/2} \|f\|_{W^{2,2}}$$

for $f \in W^{2,2}(-h_i, -h_{i-1}; \mathbb{R}^n)$. These estimates together with Lemma 5.9 imply (A). \square

Using $e_{i,j}^N(-h_i) = 1$ for $j = N$ and $= 0$ for $j = 0, \dots, N-1$, $\lim_{\tau \uparrow -h_i} e_{i+1,j}^N(\tau) = 1$ for $j = 0$ and $= 0$ for $j = 1, \dots, N$, we immediately get

$$\alpha_i^N = (A_{i0}^N, \dots, A_{i,N-1}^N, A_{iN}^N + A_i), \quad i = 1, \dots, p,$$

$$\beta^N = \text{col}(I, 0, \dots, 0),$$

$$h_i^N = \begin{pmatrix} -\frac{1}{2} & -\frac{1}{2} & 0 & 0 \\ & \frac{1}{2} & 0 & 0 \\ 0 & 0 & 0 & -\frac{1}{2} \\ 0 & 0 & \frac{1}{2} & -\frac{1}{2} \end{pmatrix} \in \mathbb{R}^{(N+1) \times (N+1)}, \quad i = 1, \dots, p,$$

and

$$g_i^N = \begin{pmatrix} 0 & 0 & 1 \\ & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} \in \mathbb{R}^{(N+1) \times (N+1)}, \quad i = 1, \dots, p-1.$$

Recall Proposition 5.8 for the matrix representations of A^N and $(A^N)^*$.

We conclude this section with some remarks:

(1) As we already mentioned in the Introduction, the spline scheme developed in this paper has interesting qualitative properties. For instance the relations between the state concept as defined in (2.4) and the dual state concept which are governed by the so called structural operator F are preserved under approximation (see [10], [13], [16], [17], [28], [31]). These results will be published elsewhere (see also [26]).

(2) Another very important property of the spline scheme is that the approximating systems (Σ^N) are stable, stabilizable or detectable for all N sufficiently large whenever the delay system has the property (see [26]).

(3) In order to use the results of [20] for the infinite time horizon problem, the approximation scheme should have the following property: If the delay system (2.1) is stable, i.e. $\|S(t)\| \leq M e^{-\varepsilon t}$, $t \geq 0$, with $M \geq 1$, $\varepsilon > 0$, then there exist constants $\tilde{M} \geq 1$ and $\tilde{\varepsilon} > 0$ such that for all N sufficiently large

$$\|S^N(t)\| \leq \tilde{M} e^{-\tilde{\varepsilon} t}, \quad t \geq 0.$$

It can be shown that this property is not satisfied for our spline scheme (see [26] for details).

(4) The assumption $e_{ij}^N |[-h_i, -h_{i-1}] \in W^{1,2}(-h_i, -h_{i-1}; \mathbb{R})$ can easily be weakened. One can allow jumps of e_{ij}^N in the interval $[-h_i, -h_{i-1}]$. Then the definition of A^N has to be modified by adding additional terms containing the approximating delta impulses corresponding to these additional jumps. This idea has been used in [34], where a realization of the scheme using piecewise linear functions is investigated. If one uses stepfunctions then one obtains the well-known averaging scheme.

(5) An important feature of the scheme developed in § 5.1 is that due to the product space structure of X^N orthogonality of the functions e_{ij}^N , $i = 1, \dots, p$, $j = 1, \dots, k_N$, implies orthogonality of the "basis elements"

$$\hat{e}_0^N, \hat{e}_{ij}^N, \quad i = 1, \dots, p, \quad j = 1, \dots, k_N.$$

This property has been exploited in [25] where a very efficient realization of the scheme by using Legendre polynomials is discussed.

6. Numerical results for the optimal control problem. The spline algorithm presented in § 5.3 was applied to a large number of examples. In this section we present the numerical findings for some of those examples. The numerical results confirm the theoretical results in case of the finite time horizon problem. The scheme performs also very well in case of the infinite time horizon problem. This is shown by two examples which already have been considered in the literature [8].

6.1. An example with finite final time. This is the problem of minimizing

$$J(u) = \frac{3}{2}x(3)^2 + \frac{1}{2} \int_0^3 u(t)^2 dt$$

subject to

$$(6.1) \quad \begin{aligned} \dot{x}(t) &= x(t-1) + u(t), & 0 \leq t \leq T=3, \\ \phi^0 &= \phi^1(0), & \phi^1(t) \equiv 1. \end{aligned}$$

For this example we have $n=1$, $p=1$, $A_{01} \equiv 0$, $A_0 = C_0 = 0$, $A_1 = B_0 = 1$, $G_0 = \frac{3}{2}$ and $R = \frac{1}{2}$. The optimal controls, trajectories and costs were calculated in [5] using the maximum principle and are given by

$$\bar{u}(t) = \begin{cases} -\frac{\delta}{2}[(t-2)^2 + 3], & 0 \leq t \leq 1, \\ \delta(t-3), & 1 \leq t \leq 2, \\ -\delta, & 2 \leq t \leq 3, \end{cases}$$

$$\bar{x}(t) = \begin{cases} 1+t-\delta\left[\frac{3t}{2}+\frac{1}{6}(t-2)^3+\frac{4}{3}\right], & 0 \leq t \leq 1, \\ \frac{3}{2}+\frac{t^2}{2}-\delta\left[4+\frac{3}{4}(t-1)^2+\frac{1}{24}(t-3)^4+\frac{4}{3}(t-1)-\frac{1}{2}(t-3)^2\right], & 1 \leq t \leq 2, \\ -\delta\left[\frac{547}{120}+5(t-2)+\frac{3}{2}(t-2)^2+\frac{1}{4}(t-2)^3-\frac{1}{6}(t-4)^3+\frac{1}{120}(t-4)^5\right], & 2 \leq t \leq 3, \end{cases}$$

$$J(\bar{u}) = \frac{329}{60} \delta^2 \quad \text{where } \delta = \frac{185}{329}.$$

The matrix valued function $[\Pi^N(t)]$, $0 \leq t \leq 3$, was computed as indicated in § 5.2. The suboptimal trajectories $\hat{x}^N(t)$ were obtained solving (6.1) with $u(t) = \hat{u}^N(t)$, $\hat{u}^N(t)$ given by (5.15), by a modified Runge-Kutta procedure. Then $\hat{u}^N(t)$ was computed from (5.15).

The numerical results we obtained are presented in Tables 6.1 and 6.2. We observe that the error $|\hat{u}^N(t) - \bar{u}(t)|$ is larger around $t = 1$ and $t = 2$ compared to other points in $[0, 3]$ because there $\bar{u}(t)$ has jumps in the derivative whereas $\hat{u}^N(t)$ is continuously differentiable on $[0, 3]$. In Table 6.2 we didn't include the values for $t = 0$ because always $\hat{x}^N(0) = \bar{x}(0)$ for our algorithm.

6.2. Two examples for the infinite time horizon problem. For the examples equation (5.18) was solved using the Newton-Kleinman algorithm as presented in [35], for instance. The Lyapunov matrix equation which has to be solved in each step of this algorithm was solved using the quadratically convergent procedure given by R. A. Smith [40] (see also [35, p. 297]). The suboptimal trajectories $\hat{x}^N(t)$ and controls $\hat{u}^N(t)$ were calculated as for the example in § 6.1 using (5.19). The two examples were already considered in [8] where the approximation was done based on the spline algorithm developed in [7].

TABLE 6.1

\hat{u} t	$\hat{u}^4(t)$	$\hat{u}^8(t)$	$\hat{u}^{16}(t)$	$\bar{u}(t)$
0	-1.9694	-1.9676	-1.9679	-1.9681
0.25	-1.7049	-1.7049	-1.7043	-1.7045
0.5	-1.4740	-1.4758	-1.4760	-1.4761
0.75	-1.2817	-1.2824	-1.2828	-1.2828
1.0	-1.1267	-1.1252	-1.1250	-1.1246
1.25	-0.9882	-0.9832	-0.9846	-0.9840
1.5	-0.8410	-0.8448	-0.8445	-0.8435
1.75	-0.6922	-0.7002	-0.7050	-0.7029
2.0	-0.5885	-0.5769	-0.5704	-0.5623
2.25	-0.5572	-0.5611	-0.5623	
2.5	-0.5620	-0.5623	-0.5623	
2.75	-0.5623	-0.5623	-0.5623	
3.0	-0.5621	-0.5622	-0.5623	
$J(\hat{u})$	1.7338	1.7338	1.7338	1.7338

TABLE 6.2

$t \backslash \hat{x}$	$\hat{x}^4(t)$	$\hat{x}^8(t)$	$\hat{x}^{16}(t)$	$\bar{x}(t)$
0.25	0.7914	0.7916	0.7916	0.7917
0.5	0.6448	0.6448	0.6448	0.6448
0.75	0.5511	0.5506	0.5507	0.5507
1.0	0.5007	0.5005	0.5005	0.5005
1.25	0.4589	0.4593	0.4595	0.4595
1.5	0.4083	0.4096	0.4094	0.4094
1.75	0.3655	0.3642	0.3646	0.3646
2.0	0.3375	0.3372	0.3371	0.3370
2.25	0.3159	0.3167	0.3168	0.3168
2.5	0.2845	0.2848	0.2848	0.2849
2.75	0.2403	0.2407	0.2408	0.2408
3.0	0.1874	0.1874	0.1874	0.1874

The first example is Example 4.1 in [8] and considers the minimization of

$$J(u) = \int_0^\infty [x(t)^2 + u(t)^2] dt$$

subject to

$$\begin{aligned} \dot{x}(t) &= x(t) + x(t-1) + u(t), \quad t \geq 0, \\ \phi^0 &= 0, \quad \phi^1(t) = \sin \pi t, \quad -1 \leq t \leq 0. \end{aligned}$$

In this case we have $n = p = 1$, $A_0 = A_1 = B_0 = C_0 = R = 1$. In Table 6.3 we give the values for $J(\hat{u}^N)$ and the optimal costs $J^N = \langle \Pi^N p^N(\phi^0, \phi^1), p^N(\phi^0, \phi^1) \rangle$ for the approximating systems (Σ^N) with cost functional (5.16) and the corresponding values obtained in [8].

In Table 6.4 we show the values of Π_0^N , Π_{1N}^N and for Π_0^N as obtained in [8]. Since $\text{range } \Pi \subset \text{dom } A^*$ (cf. (3.15)), we have $A_1^T \Pi_{00} \phi^0 = (\Pi_{10} \phi^0)(-1)$ for all $\phi^0 \in \mathbb{R}^n$. Therefore in case of this example we should have

$$\Pi_0^N - \Pi_{1N}^N \rightarrow 0 \quad \text{as } N \rightarrow \infty.$$

In Table 6.5 we give the values for $\Pi_1^N(\tau)$, which governs the distributed feedback in

TABLE 6.3

N	$J(\hat{u}^N)$	J^N	$J(\hat{u}^N), [8]$	$J^N, [8]$
4	0.321439	0.321430	0.3272	0.2484
8	0.321439	0.321432	0.3271	0.3027
16	0.321439	0.321430	0.3272	0.3163

TABLE 6.4

N	Π_0^N	Π_{1N}^N	$\Pi_0^N, [8]$
4	2.80886	2.77538	2.7940
8	2.809328	2.80096	2.8054
16	2.809390	2.80729	2.8084
32	2.809396	2.80887	2.8091

TABLE 6.5

J	$\Pi_1^4\left(-\frac{j}{32}\right)$	$\Pi_1^8\left(-\frac{j}{32}\right)$	$\Pi_1^{16}\left(-\frac{j}{32}\right)$	$\Pi_1^{32}\left(-\frac{j}{32}\right)$
0	0.63598	0.63683	0.63694	0.63696
1	—	—	—	0.66132
2	—	—	0.68698	0.68764
3	—	—	—	0.71558
4	—	0.74240	0.74512	0.74553
5	—	—	—	0.77722
6	—	—	0.81040	0.81114
7	—	—	—	0.84695
8	0.87064	0.88269	0.88474	0.88517
9	—	—	—	0.92547
10	—	—	0.96750	0.96839
11	—	—	—	1.01361
12	—	1.05757	1.06113	1.06165
13	—	—	—	1.11225
14	—	—	1.16491	1.16591
15	—	—	—	1.22238
16	1.26664	1.27879	1.28151	1.28218
17	—	—	—	1.34508
18	—	—	1.41048	1.41162
19	—	—	—	1.48157
20	—	1.54972	1.55463	1.55547
21	—	—	—	1.63314
22	—	—	1.71381	1.71512
23	—	—	—	1.80125
24	1.86588	1.88693	1.89104	1.89209
25	—	—	—	1.98748
26	—	—	2.08649	2.08802
27	—	—	—	2.19358
28	—	2.29692	2.30348	2.30477
29	—	—	—	2.42147
30	—	—	2.54253	2.54432
31	—	—	—	2.67323
32	2.77538	2.80096	2.80729	2.80887

(6.3), at the knots $-j/N$, $j=0, \dots, N$, for $N=4, 8, 16$ and 32 . We clearly see that $\Pi_1^N(\tau)$ converges uniformly on $-1 \leq \tau \leq 0$ as $N \rightarrow \infty$. Note, that $\Pi_1^N(\tau)$ is a continuous piecewise linear function on $[-1, 0]$ with knots at $-j/N$, $j=0, \dots, N$.

In Tables 6.6 and 6.7 we present the values for $\hat{u}^N(t)$ on $0 \leq t \leq 4$ and for $\hat{x}^N(t)$ on $0 \leq t \leq 3$, respectively.

The results of this example show a significant improvement in the qualitative behavior of our spline scheme compared to the scheme presented in [8]. In both schemes $\Pi_1^N(\tau)$ is a continuous piecewise linear function on $[-1, 0]$. But in [8] this function is increasingly oscillatory with increasing N (compare Figs. 4.1–4.4 in [8]), whereas in our scheme $\Pi_1^N(\tau)$ is strictly monotone and obviously converging in the supremum norm. This property of our scheme becomes very important if one wants to implement the approximating feedback law in a real system. Our scheme seems also to be more accurate as far as approximation of Π_{00} by Π_0^N and of $J(\bar{u})$ by $J^N(\hat{u}^N)$ or J^N is concerned.

The next example is Example 4.2 in [8] and considers a simplified model for the Mach number control loop for the National Transonic Facility at NASA Langley

TABLE 6.6

$t \backslash \hat{u}^N$	$\hat{u}^4(t)$	$\hat{u}^8(t)$	$\hat{u}^{16}(t)$
0	0.86836	0.86817	0.86816
0.25	0.64894	0.64891	0.64891
0.5	0.49650	0.49657	0.49658
0.75	0.35400	0.36400	0.36401
1.0	0.24627	0.24618	0.24618
1.25	0.16154	0.16146	0.16146
1.5	0.10999	0.10993	0.10993
1.75	0.08024	0.08021	0.08021
2.0	0.06015	0.06015	0.06015
2.25	0.04348	0.04347	0.04347
2.5	0.02983	0.02982	0.02982
2.75	0.01996	0.01995	0.01995
3.0	0.01373	0.01372	0.01372
3.25	0.00991	0.00991	0.00991
3.5	0.00729	0.00729	0.00729
3.75	0.00524	0.00523	0.00523
4.0	0.00362	0.00362	0.00362

TABLE 6.7

$t \backslash \hat{x}^N$	$\hat{x}^4(t)$	$\hat{x}^8(t)$	$\hat{x}^{16}(t)$
0.25	0.11259	0.11258	0.11258
0.5	0.05332	0.05331	0.05332
0.75	-0.06628	-0.06626	-0.06626
1.0	-0.10850	-0.10846	-0.10846
1.25	-0.06160	-0.06158	-0.06158
1.5	-0.01397	-0.01397	-0.01397
1.75	0.00753	0.00752	0.00752
2.0	0.00178	0.00178	0.00178
2.25	-0.00784	-0.00784	-0.00784
2.5	-0.01030	-0.01029	-0.01029
2.75	-0.00646	-0.00646	-0.00646
3.0	-0.00178	-0.00178	-0.00178

Research Center. For details see [2] or [8]. The problem is to

minimize

$$J(u) = \int_0^\infty [x^T(t) C_0^T C_0 x(t) + u^2(t)] dt$$

subject to

$$(6.2) \quad \dot{x}(t) = \begin{pmatrix} -a & 0 & 0 \\ 0 & 0 & 1 \\ 0 & -\omega^2 & -2\xi\omega \end{pmatrix} x(t) + \begin{pmatrix} 0 & ka & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} x(t-0.33) + \begin{pmatrix} 0 \\ 0 \\ \omega^2 \end{pmatrix} u(t), \quad t \geq 0,$$

$$\phi^0 = \text{col}(-0.1, 8.547, 0) = \phi^1(t), \quad -0.33 \leq t \leq 0.$$

We have $C_0 = (100, 0, 0)$, $n = 3$, $p = 1$, $k = -0.0117$, $\xi = 0.8$, $\omega = 0.6$ and $1/a = 1.964$. Because of the simple structure of this problem it is possible to calculate the true solution following an idea contained in [29]. If we put for $t \geq 0$, $h = 0.33$

$$y_1(t) = x_1(t+h), \quad y_2(t) = x_2(t), \quad y_3(t) = x_3(t),$$

we obtain by a simple calculation

$$(6.3) \quad \frac{d}{dt} \begin{pmatrix} y_1(t) \\ y_2(t) \\ y_3(t) \end{pmatrix} = \begin{pmatrix} -a & ka & 0 \\ 0 & 0 & 1 \\ 0 & -\omega^2 & -2\xi\omega \end{pmatrix} \begin{pmatrix} y_1(t) \\ y_2(t) \\ y_3(t) \end{pmatrix} + \begin{pmatrix} 0 \\ 0 \\ \omega^2 \end{pmatrix} u(t).$$

The cost functional takes the form

$$(6.4) \quad J(u) = 10^4 \int_0^h x_1(t)^2 dt + \int_0^\infty [10^4 y_1(t)^2 + u(t)^2] dt,$$

where

$$(6.5) \quad x_1(t) = e^{-at} \phi_1^0 + ak \int_0^t e^{-a(t-\tau)} \phi_2^1(\tau-h) d\tau, \quad 0 \leq t \leq h$$

is not dependent on $u(t)$ on the interval $[0, h]$. Therefore minimizing $J(u)$ subject to (6.2) is equivalent to minimizing

$$\tilde{J}(u) = \int_0^\infty [10^4 y_1(t)^2 + u(t)^2] dt$$

subject to (6.3) with initial data

$$(6.6) \quad y_1(0) = x_1(h), \quad y_2(0) = x_2(0), \quad y_3(0) = x_3(0).$$

The solution of the latter problem is given by the feedback law

$$\bar{u}(t) = -(0, 0, \omega^2) \tilde{\Pi}_0 \bar{y}(t),$$

where $\bar{y}(t)$ is the solution of (6.3) with $u(t) = \bar{u}(t)$ and initial data (6.6). $\tilde{\Pi}_0$ is the solution of the algebraic Riccati equation

$$(6.7) \quad \tilde{A}^T \tilde{\Pi}_0 + \tilde{\Pi}_0 \tilde{A} - \tilde{\Pi}_0 B_0 B_0^T \tilde{\Pi}_0 + C_0^T C_0 = 0,$$

where \tilde{A} is the system matrix in (6.5) and B_0 , C_0 are the same as for (6.2).

Equation (6.7) was solved numerically to give

$$\tilde{\Pi}_0 = \begin{pmatrix} 8220.51099 & -11.61086 & -1.12107 \\ -11.61086 & 0.01851 & 0.00186 \\ -1.12107 & 0.00186 & 0.00019 \end{pmatrix}.$$

The optimal costs for the original problem are given by (see (6.4))

$$(6.8) \quad \begin{aligned} J(\bar{u}) &= \tilde{J}(\bar{u}) + 10^4 \int_0^h \bar{x}_1(t)^2 dt \\ &= \bar{y}^T(0) \tilde{\Pi}_0 \bar{y}(0) + 10^4 \int_0^h \bar{x}_1(t)^2 dt. \end{aligned}$$

Using (6.5) it is easy to calculate $J(\bar{u})$. In Table 6.8 we give the values for $J(\bar{u})$, $J(\hat{u}^N)$ and $J^N = \langle \Pi^N p^N(\phi^0, \phi^1), p^N(\phi^0, \phi^1) \rangle$ and the values available in [8].

TABLE 6.8

N	$J(\hat{u}^N)$	J^N	$J(\hat{u}^N), [8]$	$J^N, [8]$
4	136.39587	136.40499	136.7354	138.7345
8	136.40094	136.40509	136.7354	138.7624
16	136.40250	136.40521	—	—
$J(\bar{u})$		136.40490		

Computing $J(\bar{u})$ for general initial data (ϕ^0, ϕ^1) by using (6.8), (6.6) and (6.5) and comparing the result with

$$J(\bar{u}) = (\phi^0)^T \Pi_{00} \phi^0 + 2(\phi^0)^T \Pi_{01} \phi^1 + \langle \phi^1, \Pi_{11} \phi^1 \rangle_{L^2},$$

we immediately obtain an explicit representation of Π :

$$\begin{aligned} \Pi_{00} &= \begin{pmatrix} e^{-ah} & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \tilde{\Pi}_0 \begin{pmatrix} e^{-ah} & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} + 10^4 \frac{1 - e^{-2ah}}{2a} \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}, \\ \Pi_{01} \phi^1 &= ak \begin{pmatrix} e^{-ah} & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \tilde{\Pi}_0 \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} \int_{-h}^0 e^{a\tau} \phi^1(\tau) d\tau \\ &\quad + 10^4 k e^{-ah} \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} \int_{-h}^0 \frac{e^{-a\tau} - e^{a\tau}}{2} \phi^1(\tau) d\tau \end{aligned}$$

or, equivalently, $(\Pi_{01}^* \phi^0)(\theta) = \Pi_1(\theta) \phi^0$ with

$$\Pi_1(\theta) = ak \begin{pmatrix} 0 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} \tilde{\Pi}_0 \begin{pmatrix} e^{-ah} & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} e^{a\theta} + 10^4 k e^{-ah} \begin{pmatrix} 0 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} \frac{e^{-a\theta} - e^{a\theta}}{2},$$

$-h \leq \theta \leq 0$, and

$$\begin{aligned} (\Pi_{11} \phi^1)(\theta) &= a^2 k^2 e^{a\theta} \begin{pmatrix} 0 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} \tilde{\Pi}_0 \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} \int_{-h}^0 e^{a\tau} \phi^1(\tau) d\tau \\ &\quad + ak^2 \begin{pmatrix} 0 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{pmatrix} \int_{-h}^0 \frac{e^{-a|\tau-\theta|} - e^{a(\tau+\theta)}}{2} \phi^1(\tau) d\tau, \quad -h \leq \theta \leq 0. \end{aligned}$$

In Table 6.9 we present the values for Π_0^N , Π_0^N as computed in [8] and Π_{00} , whereas in Table 6.10 we give the values for the second row of $\Pi_1^N(-jh/4)$ and $\Pi_1(-jh/4)$ for $j=0, \dots, 4$. The other rows of these matrices are always zero. Again, our scheme is more accurate compared to the scheme in [8] and $\Pi_1^N(\theta)$ converges uniformly on $[-h, 0]$ to $\Pi_1(\theta)^T$.

TABLE 6.9

N	Π_0^N	$\Pi_0^N, [8]$
4	$\begin{pmatrix} 8677.02417 & -9.81502 & -0.94768 \\ -9.81502 & 0.01851 & 0.00186 \\ -0.94768 & 0.00186 & 0.00019 \end{pmatrix}$	$\begin{pmatrix} 8676.9237 & -9.8164 & -0.9477 \\ -9.8164 & 0.0185 & 0.0019 \\ -0.9477 & 0.0019 & 0.0002 \end{pmatrix}$
8	$\begin{pmatrix} 8677.02698 & -9.81505 & -0.94768 \\ -9.81505 & 0.01851 & 0.00186 \\ -0.94768 & 0.00186 & 0.00019 \end{pmatrix}$	$\begin{pmatrix} 8676.9829 & -9.8154 & -0.9477 \\ -9.8154 & 0.0185 & 0.0019 \\ -0.9477 & 0.0019 & 0.0002 \end{pmatrix}$
16	$\begin{pmatrix} 8677.03516 & -9.81506 & -0.94768 \\ -9.81506 & 0.01851 & 0.00186 \\ -0.94768 & 0.00186 & 0.00019 \end{pmatrix}$	—
Π_{00}	$\begin{pmatrix} 8677.02405 & -9.81505 & -0.94768 \\ -9.81505 & 0.01851 & 0.00186 \\ -0.94768 & 0.00186 & 0.00019 \end{pmatrix}$	

TABLE 6.10

j	$\Pi_1^4\left(-\frac{j\hbar}{4}\right)$		
0	-41.39697	0.06916	0.00668
1	-43.83789	0.06652	0.00640
2	-46.37943	0.06334	0.00613
3	-48.97898	0.06118	0.00590
4	-51.69006	0.05828	0.00563

j	$\Pi_1^8\left(-\frac{j\hbar}{4}\right)$		
0	-41.39721	0.06917	0.00668
1	-43.84998	0.06626	0.00640
2	-46.38019	0.06355	0.00614
3	-48.99226	0.06093	0.00588
4	-51.69080	0.05843	0.00564

j	$\Pi_1^{16}\left(-\frac{j\hbar}{4}\right)$		
0	-41.39727	0.06917	0.00668
1	-43.85012	0.06631	0.00640
2	-46.38036	0.06358	0.00614
3	-48.99246	0.06097	0.00589
4	-51.69102	0.05846	0.00564

j	$\Pi_1\left(-\frac{j\hbar}{4}\right)$		
0	-41.39721	0.06917	0.00668
1	-43.85008	0.06632	0.00641
2	-46.38034	0.06360	0.00614
3	-48.99246	0.06098	0.00589
4	-51.69103	0.05847	0.00565

Acknowledgment. F. Kappel acknowledges the hospitality provided by the members of the Forschungsschwerpunkt Dynamische Systeme, University of Bremen.

REFERENCES

- [1] J. H. AHLBERG, E. N. NILSEN AND J. C. WALSH, *The Theory of Splines and Their Applications*, Academic Press, New York, 1967.
- [2] E. S. ARMSTRONG AND J. S. TRIPP, *An application of multivariable design techniques to the control of the National Transonic Facility*, NASA Technical Paper 1887, NASA Langley Research Center, Hampton, VA, 1981.
- [3] H. T. BANKS, *Identification of nonlinear delay systems using spline methods*, Proc. Internat. Conf. on Nonlinear Phenomena in Math. Sci., Arlington, TX, June 1980.
- [4] H. T. BANKS AND J. A. BURNS, *Hereditary control problems: Numerical methods based on averaging approximation*, this Journal, 16 (1978), pp. 169–208.
- [5] H. T. BANKS, J. A. BURNS, E. M. CLIFF AND P. R. THRIFT, *Numerical solutions of hereditary control problems via an approximation technique*, Brown University, LCDS Technical Report 15–6, Providence, RI, 1975.
- [6] H. T. BANKS AND P. L. DANIEL, *Parameter estimation of nonlinear nonautonomous distributed systems*, Proc. 20th IEEE Conf. Decision and Control, San Diego, CA, December 1981, pp. 228–232.
- [7] H. T. BANKS AND F. KAPPEL, *Spline approximation for functional differential equations*, J. Differential Equations, 34 (1979), pp. 496–522.
- [8] H. T. BANKS, G. I. ROSEN AND K. ITO, *A spline based technique for computing Riccati operators and feedback controls in regulator problems for delay equations*, SIAM J. Sci. Statist. Comput., 5 (1984), pp. 830–855.
- [9] A. BENSOUSSAN, M. C. DELFOUR AND S. K. MITTER, *The linear quadratic optimal control problem for infinite dimensional systems over an infinite horizon; survey and examples*, Electronic Systems Laboratory Reports, Massachusetts Inst. Technology, Cambridge, MA, 1977.
- [10] C. BERNIER AND A. MANITIUS, *On semigroups in $\mathbb{R}^n \times L^p$ corresponding to differential equations with delays*, Canad. J. Math., 30 (1978), pp. 897–914.
- [11] K. BOEHMER, *Spline-Funktionen, Theorie und Anwendungen*, Teubner, Stuttgart, 1974.
- [12] J. BORISOVIĆ AND A. S. TURBABIN, *On the Cauchy problem for linear nonhomogeneous differential equations with retarded argument*, Soviet Math. Dokl., 10 (1969), pp. 401–405.
- [13] J. A. BURNS AND T. L. HERDMAN, *Adjoint semigroup theory for a class of functional differential equations*, SIAM J. Math. Anal., 5 (1976), pp. 729–745.
- [14] R. F. CURTAIN AND A. J. PRITCHARD, *Infinite Dimensional Linear Systems Theory*, Springer, Berlin, New York, 1978.
- [15] R. DATKO, *Unconstrained control problems with quadratic cost*, this Journal, 11 (1973), pp. 32–52.
- [16] M. C. DELFOUR, E. B. LEE AND A. MANITIUS, *F-reduction of the operator Riccati equation*, Automatica, 14 (1978), pp. 385–395.
- [17] M. C. DELFOUR AND A. MANITIUS, *The structural operator F and its role in the theory of retarded systems*, Part I: J. Math. Anal. Appl., 73 (1980), pp. 466–490; Part II: J. Math. Anal. Appl., 74 (1980), pp. 359–381.
- [18] M. C. DELFOUR AND S. K. MITTER, *Hereditary differential systems with constant delays I: General case*, J. Differential Equations, 12 (1972), pp. 213–235.
- [19] J. S. GIBSON, *The Riccati integral equations for optimal control problems in Hilbert spaces*, this Journal, 17 (1979), pp. 537–565.
- [20] ———, *Linear quadratic optimal control of hereditary differential systems: infinite dimensional Riccati equations and Numerical approximations*, this Journal, 21 (1983), pp. 95–139.
- [21] D. GOTTlieb AND S. A. ORSZAG, *Numerical Analysis of Spectral Methods: Theory and Applications*, CBMS-NSF Regional Conference Series in Applied Mathematics, Society for Industrial and Applied Mathematics, Philadelphia, PA, 1977.
- [22] J. K. HALE, *Theory of Functional Differential Equations*, Springer, Berlin, New York, 1977.
- [23] A. S. B. HOLLAND AND B. N. SAHNEY, *The General Problem of Approximation and Spline Functions*, Krieger, Huntington, New York, 1979.
- [24] F. KAPPEL, *Finite dimensional approximation of systems with infinite dimensional state space*, in EQUADIFF 82, H. W. Knobloch and K. Schmitt, eds., Lecture Notes in Mathematics 1017, Springer, Berlin, 1983, pp. 287–299.
- [25] F. KAPPEL AND G. PROBST, *Approximation of feedback controls for delay systems using Legendre polynomials*, Confer. Sem. Mat. Univ. Bari, 201 (1984), pp. 1–36.

- [26] F. KAPPEL AND D. SALAMON, *Spline approximation for retarded systems and the Riccati equation*, Institute for Mathematics, University of Graz, Preprint No. 42-1984, and Mathematics Research Center, University of Wisconsin, Technical Summary Report No. 2680, April 1984.
- [27] K. KUNISCH, *Approximation schemes for the linear quadratic optimal control problem associated with delay equations*, this Journal, 20 (1982), pp. 506-540.
- [28] A. MANITIUS, *Completeness and F-completeness of eigenfunctions associated with retarded functional differential equations*, J. Differential Equations, 35 (1980), pp. 1-29.
- [29] A. MANITIUS AND H. TRAN, *Numerical simulation of a nonlinear feedback controller for a wind tunnel model involving a time delay*, Optimal Control Appl. Meth., 7 (1986), pp. 19-39.
- [30] R. H. MARTIN, JR., *Nonlinear Operators and Differential Equations in Banach Spaces*, John Wiley, New York, 1976.
- [31] R. K. MILLER, *Linear Volterra integro-differential equations as semigroups*, Funkcial. Ekvac., 17 (1974), pp. 39-55.
- [32] A. PAZY, *Semigroups of Linear Operators and Applications to Partial Differential Equations*, Springer, Berlin, New York, 1983.
- [33] L. PANDOLFI, *Feedback stabilization of functional differential equations*, Boll. Un. Mat. Ital., 12 (1975), pp. 626-635.
- [34] G. PROBST, *Piecewise linear approximation for hereditary control problems*, Institute for Mathematics, University of Graz, Technical Report No. 60-1985, Graz, 1985.
- [35] D. L. RUSSELL, *Mathematics of Finite Dimensional Control Systems, Theory and Design*, Marcel Dekker, New York, 1979.
- [36] D. SALAMON, *On dynamic observation and state feedback for time delay systems*, in *Evolution Equations and Their Applications*, F. Kappel and W. Schappacher, eds., Research Notes in Math., Pitman, London, 1982, pp. 202-219.
- [37] ———, *Control and Observation of Neutral Systems*, Research Notes in Math., 91, Pitman, London, 1984.
- [38] ———, *Structure and stability of finite dimensional approximations for functional differential equations*, University of Wisconsin-Madison, TSR# 2586, 1983.
- [39] M. H. SCHULTZ, *Spline Analysis*, Prentice-Hall, Englewood Cliffs, NJ, 1973.
- [40] R. A. SMITH, *Matrix equation $XA + BX = C$* , SIAM J. Appl. Math., 16 (1968), pp. 198-201.
- [41] J. ZABCZYK, *Remarks on the algebraic Riccati equation in Hilbert space*, Appl. Math. Optim., 2 (1976), pp. 251-258.

STATE CONSTRAINED CONTROL PROBLEMS GOVERNED BY VARIATIONAL INEQUALITIES*

ZHENG-XU HE†

Abstract. We will consider various types of problems of minimizing some given cost functional over all pairs of states and controls subject to some parabolic or elliptic variational inequality and some state constraints. Our aim is to derive the first order optimality conditions for optimal pairs of these problems.

Key words. optimal control, state constraints, parabolic variational inequality, elliptic variational inequality, optimal pair, optimality conditions

AMS(MOS) subject classifications. Primary 49B22; secondary 49A29

1. Introduction. We will study some classes of control problems governed by the nonlinear parabolic system

$$y'(t) + Ay(t) + Fy(t) \ni Bu(t), \quad t \in (0, 1)$$

with the state constraints

$$y(t) \in K, \quad t \in [0, 1], \quad y(1) \in K_1,$$

or governed by the elliptic system

$$Ay + Fy \ni Bu$$

with

$$y \in K,$$

where A is a linear self-adjoint positive definite operator in the state space H , F is the subdifferential of some lower-semicontinuous (l.s.c.) convex function on H , B is a linear continuous operator from the space of controls to H , and both K and K_1 are closed subsets of H . In the absence of the state constraints, these problems have been treated by V. Barbu [1], [2] and [3], where first order necessary conditions were derived under suitable conditions. For those results, we refer to [4] where a unified approach to the theory of optimality conditions was presented. Using somewhat different methods, the above problems were also studied in [6], [7], [8], [14] and [15].

In this paper, we will extend the results of V. Barbu to the above problems with state constraints. For this purpose, we will employ the notion of "normal cone" defined in [9] and [10]. Our results may also be compared with those of [5, Chap. 4] (and of [12] in the finite dimensional case), but there are important differences between them due to the nonconvexity of the data and the nonlinearity of state equation in the present problems.

To avoid a lengthy exposition, we will consider mainly the following problem:

(P) Minimize

$$(1.1) \quad G_0(y, u) = \int_0^1 (g(t, y(t)) + h(u(t))) dt + \phi_0(y(1))$$

over all $(y, u) \in W^{1,2}([0, 1]; H) \times L^2(0, 1; U)$ subject to the equation

* Received by the editors July 24, 1985; accepted for publication November 26, 1985.

† Faculty of Mathematics, University of Iași, R-6600, Iași, Romania. Present address, Department of Mathematics, University of California at San Diego, La Jolla, California 92093.

$$(1.2) \quad \begin{aligned} y'(t) + Ay(t) + \beta(y(t) - \Psi) &\ni Bu(t) + f(t) \quad \text{a.e. } t \in (0, 1), \\ y(0) &= y_0, \end{aligned}$$

and the state constraints

$$(1.3) \quad y(t) \in K \quad \forall t \in [0, 1],$$

$$(1.4) \quad y(1) \in K_1.$$

Here $H = L^2(\Omega)$, Ω denotes a fixed bounded domain in R^N with smooth boundary Γ ; U is another Hilbert space; $g: [0, 1] \times H \rightarrow R$, $h: U \rightarrow R$ and $\phi_0: H \rightarrow R$ are given functions; K and K_1 are closed subsets of H ; $A: D(A_H) \subseteq H \rightarrow H$ is a symmetric linear elliptic operator; β is a maximal monotone graph in $R \times R$; B is a linear continuous operator from U to H ; $f \in L^2(0, 1; H)$, $\Psi \in W^{1,2}([0, 1]; H)$ and $y_0 \in H$.

Some other types of state constrained control problems will be briefly discussed. Since the proofs in these cases are very similar to that for (P), they will be omitted. Practically, all control problems in [4] may be generalized by adding the state constraints, and the optimality conditions could be derived by our method.

The content of the paper is the following. In § 2, we will discuss problem (P) and give the main results. Sections 3 and 4 will be devoted to the proof of these results. In § 5, we will give some results for the state constrained boundary control problems of parabolic type, while in § 6 we will discuss the state constrained control problems governed by elliptic variational inequalities. Finally, some examples will be given in § 7, the emphasis will be placed in verifying hypothesis (H6) given in § 2.

We will keep the notation of [4, Chaps. 5, 6 and 3]. So, the norms of H , V and U will be denoted by $\|\cdot\|_2$, $\|\cdot\|$ and $\|\cdot\|_U$, respectively (V will be described in § 2); (\cdot, \cdot) will denote the pairing between V and V^* as well as the scalar product in $H \times H$; $\langle \cdot, \cdot \rangle$ will denote the scalar product in $U \times U$.

We conclude this section by recalling some facts concerning the normal cones and the subgradients defined in [9] and [10]. Let E be a Banach space with the norm $\|\cdot\|$, and let $\varphi: E \rightarrow [-\infty, +\infty]$ be any function. If φ is finite at a point $z \in E$, then the *upper subderivative* of φ at z with respect to $y \in E$ is defined to be

$$(1.5) \quad \varphi^\uparrow(z; y) = \sup_{\rho > 0} \inf_{\delta > 0} \sup \left\{ \inf_{\|y - y'\| \leq \rho} \frac{\varphi(z' + ty') - \alpha'}{t}; \alpha' \geq \varphi(z'), \right. \\ \left. t > 0, \|z' - z\| + |\alpha' - \varphi(z)| + t < \delta \right\}.$$

The function $y \mapsto \varphi^\uparrow(z; y)$ is sublinear and l.s.c. [9, Thm. 2], and the *subdifferential* (or the set of *subgradients*) of φ at z is

$$(1.6) \quad \partial\varphi(z) = \{z^* \in E^*; z^*(y) \leq \varphi^\uparrow(z; y) \quad \forall y \in E\}.$$

The set $\partial\varphi(z)$ is weak-star closed and convex in E^* , $\partial\varphi(z)$ is not empty if and only if $\varphi^\uparrow(z; 0)$ is not $-\infty$ (see [9, Thm. 4]).

Let C be any subset of E , and denote by ψ_C the indicator function of C , i.e.,

$$\psi_C(y) = \begin{cases} 0 & \text{if } y \in C, \\ +\infty & \text{if } y \notin C. \end{cases}$$

Then by [9, Eq. (7.9)], for any $z \in C$, the *normal cone* $N_C(z)$ to C at z is characterized by

$$(1.7) \quad N_C(z) = \partial\psi_C(z).$$

But, if we let $d_C: E \rightarrow R$ be the function

$$(1.8) \quad d_C(z) = \inf \{ \|y' - y\|; y' \in C \},$$

then it is not difficult to verify that for any $y \in E$ and $z \in C$,

$$d_C^\uparrow(z; y) > 0 \text{ implies } \psi_C^\uparrow(z; y) = +\infty,$$

and so

$$d_C^\uparrow(z; y) \leq \psi_C^\uparrow(z; y) \quad \forall y \in E.$$

It follows that

$$(1.9) \quad \partial d_C(z) \subseteq \partial \psi_C(z) = N_C(z) \quad \forall z \in C.$$

Since $N_C(z)$ is a cone with vertex in the origin, by (1.9) we see that the cone spanned by $\partial d_C(z)$ is included in $N_C(z)$.

2. Hypotheses and the main results. As in [4, § 5.1], we assume that there is a Hilbert space V , which is dense in H , such that

$$V \subseteq H = H^* \subseteq V^*$$

algebraically and topologically, and the following conditions hold:

(H1) The injection of V in H is compact.

(H2) A is a linear and symmetric operator from V to V^* satisfying $(Ay, y) \geq \omega \|y\|^2 - \alpha |y|_2^2$ for all $y \in V$, for some $\omega > 0$ and $\alpha \in R$.

(H3) β is a maximal monotone graph in $R \times R$ such that $0 \in \beta(0)$, and there exists $C \geq 0$ such that

$$(Ay, \xi(y - \Psi(\cdot, t))) \geq -C(1 + |\xi(y - \Psi(\cdot, t))|_2)(1 + |y - \Psi(\cdot, t)|_2) \quad \forall t \in [0, 1],$$

for any $y \in D(A_H) = \{z \in V; Az \in H\}$ and every increasing function $\xi \in C^1(R)$ with $\xi(0) = 0$, $\dot{\xi} \leq 1$.

(H4) $h: U \rightarrow (-\infty, +\infty]$ is l.s.c. and convex.

(H5) $g: [0, 1] \times H \rightarrow R$ is measurable in t , $g(\cdot, 0) \in L^\infty(0, 1)$, and for every $r > 0$, there exists $L_r > 0$ independent of t , such that

$$|g(t, y) - g(t, z)| + |\phi_0(y) - \phi_0(z)| \leq L_r |y - z|_2 \quad \forall t \in [0, 1], |y|_2 + |z|_2 \leq r.$$

We will assume throughout that $y_0 \in V$ and

$$\int_\Omega j(y_0(x) - \Psi(x, 0)) \, dx < +\infty,$$

where $j: R \rightarrow R$ is the convex function such that $\partial j = \beta$. The equation (1.2) has for every $u \in L^2(0, 1; U)$ a unique solution $y = y(u)$ belonging to $W^{1,2}([0, 1]; H) \cap L^2(0, 1; D(A_H)) \cap C([0, 1]; V)$.

If there exists an admissible pair (y, u) , i.e. a pair satisfying (1.2), (1.3) and (1.4), such that

$$(2.1) \quad H_0(u) = \int_0^1 h(u(t)) \, dt < +\infty,$$

and if $G_0(y, u) \rightarrow +\infty$ for (y, u) admissible and $\|u\|_{L^2(0,1;U)} \rightarrow +\infty$, then the control problem (P) admits at least one optimal pair. The proof is the same with that of [4, Prop. 5.1]. However, in order to derive the optimality conditions, we need an additional hypothesis.

Let (y^*, u^*) be a fixed optimal pair for (P). Let $\varepsilon \in (0, +\infty]$. We define $\phi^\varepsilon: L^\infty(0, 1; H) \times H \rightarrow [-\infty, +\infty]$ by

$$(2.2) \quad \phi^\varepsilon(z, z_1) = \inf \{ G_0(y, u); y(t) \in z(t) + K \text{ a.e. } t \in (0, 1), y(1) \in z_1 + K_1, (y, u) \text{ satisfies (1.2), } \|y - y^*\|_\infty \leq \varepsilon, \|u - u^*\|_2 \leq \varepsilon \},$$

where $\|\cdot\|_\infty$ and $\|\cdot\|_2$ denote the norms of $L^\infty(0, 1; H)$ and $L^2(0, 1; U)$, respectively, and the infimum of the empty set is $+\infty$. Note that $\phi^\varepsilon(0, 0) = G_0(y^*, u^*)$ is finite and independent of ε , and $\phi^\varepsilon \leq \phi^{\varepsilon'}$ if $\varepsilon' \leq \varepsilon$.

The last hypothesis is:

(H6) There exists $\varepsilon \in (0, +\infty]$ such that

$$\liminf_{\substack{\|z\|_\infty \rightarrow 0 \\ |z_1|_2 \rightarrow 0}} (\phi^\varepsilon(z, z_1) - \phi^\varepsilon(0, 0)) / (\|z\|_\infty + |z_1|_2) > -\infty.$$

Let $s > N/2$ and let $Y = H^s(\Omega) \cap V$. Denote $Q = \Omega \times (0, 1)$, and

$$(2.3) \quad \mathcal{K} = \{y \in L^\infty(0, 1; H); y(t) \in K \text{ a.e. } t \in (0, 1)\}.$$

\mathcal{K} is evidently a closed subset of $L^\infty(0, 1; H)$. We have (compare with [4, Thms. 5.1 and 5.2]):

THEOREM 2.1. *Let (y^*, u^*) be any optimal pair of the control problem (P) where β is locally Lipschitz. Suppose that (H1)–(H6) are satisfied. Then there exist $p \in BV([0, 1]; Y^*) \cap L^\infty(0, 1; H) \cap L^2(0, 1; V)$, $\mu \in (L^\infty(Q))^*$, $\lambda \in (L^\infty(0, 1; H))^*$ and $q_1 \in H$ such that*

$$(2.4) \quad p' - Ap - \mu - \lambda \in \partial g(t, y^*) \quad \text{a.e. in } Q,$$

$$(2.5) \quad \mu_a(x, t) \in p(x, t) \partial \beta(y^*(x, t) - \Psi) \quad \text{a.e. in } Q,$$

$$(2.6) \quad p(1) + q_1 + \partial \phi_0(y^*(1)) \ni 0,$$

$$(2.7) \quad \lambda \in N_{\mathcal{K}}(y^*), \quad q_1 \in N_{K_1}(y^*(1)),$$

$$(2.8) \quad B^*p(t) \in \partial h(u^*(t)) \quad \text{a.e. } t \in (0, 1),$$

where by μ_a we denote the absolutely continuous part of μ .

Furthermore, if in addition β satisfies

$$(2.9) \quad \dot{\beta}(r) \leq C(|\beta(r)| + |r| + 1) \quad \text{a.e. } r \in \mathbb{R},$$

for some $C \geq 0$, then $\mu = \mu_a \in L^1(Q)$.

THEOREM 2.2. *Let (y^*, u^*) be any optimal pair for the control problem (P) with*

$$(2.10) \quad \beta(r) = \begin{cases} 0 & \text{if } r > 0, \\ (-\infty, 0] & \text{if } r = 0, \\ \phi & \text{if } r < 0. \end{cases}$$

Suppose (H1)–(H6) hold. Then there exist $p \in BV([0, 1]; Y^) \cap L^\infty(0, 1; H) \cap L^2(0, 1; V)$, $\lambda \in (L^\infty(0, 1; H))^*$ and $q_1 \in H$ such that*

$$(2.11) \quad (p' - Ap - \lambda)_a \in \partial g(t, y^*) \quad \text{a.e. in } \{y^* > \Psi\},$$

$$(2.12) \quad p(f + Bu^* - Ay^* - y^{*'}) = 0 \quad \text{a.e. in } Q,$$

$$(2.13) \quad p(1) + q_1 + \partial \phi_0(y^*(1)) \ni 0,$$

$$(2.14) \quad \lambda \in N_{\mathcal{K}}(y^*), \quad q_1 \in N_{K_1}(y^*(1)),$$

$$(2.15) \quad B^*p(t) \in \partial h(u^*(t)) \quad \text{a.e. } t \in (0, 1).$$

Remark 2.1. It may be seen from the proofs given in §§ 3 and 4 that Theorems 2.1 and 2.2 also hold for local solutions (y^*, u^*) for the problem (P), i.e. for the pair (y^*, u^*) which minimizes $G_0(y, u)$ over all (y, u) satisfying (1.2), (1.3), (1.4) and $\|y - y^*\|_\infty < \varepsilon_0$, $\|u - u^*\|_2 \leq \varepsilon_0$, for some prescribed $\varepsilon_0 > 0$.

Remark 2.2. Our results also apply (with some minor modifications) to the problem (P) where G_0 has the following form

$$(2.16) \quad G_0(y, u) = \int_0^1 (g(t, y(t)) + h(u(t))) dt + \psi_0(y, y(1)) + H'_0(u),$$

where $H'_0: L^2(0, 1; U) \rightarrow (-\infty, +\infty]$ is convex and l.s.c., and $\psi_0: L^\infty(0, 1; H) \times H \rightarrow R$ is locally Lipschitz. We may also replace the state constraint (1.3) by the following more general one:

$$(2.17) \quad y \in \mathcal{K},$$

where \mathcal{K} is an arbitrary closed subset of $L^2(0, 1; H)$.

Remark 2.3. Hypothesis (H6) is essential to deduce the optimality conditions in our problem where state constraints are present. As a matter of fact, if $\beta = 0$ and if $g(t, \cdot)$, ϕ_0 , K and K_1 are convex, then (H6) is implied by the optimality conditions (of Theorem 2.1).

Here is the argument: Suppose that (2.4)–(2.8) hold. Let $(z, z_1) \in L^\infty(0, 1; H) \times H$, and let (y, u) be a pair satisfying (1.2) and

$$y(t) \in z(t) + K, \quad y(1) \in z_1 + K_1.$$

Multiplying (2.4) by $y - y^*$, and then integrating by parts over $[0, 1]$, we get by (1.2) that

$$(2.18) \quad \begin{aligned} (p(1), y(1) - y^*(1)) - \int_0^1 (p, B(u - u^*)) dt - \lambda(y - y^*) \\ \leq \int_0^1 g(t, y) dt - \int_0^1 g(t, y^*) dt. \end{aligned}$$

As $y - z \in \mathcal{K}$ and $y(1) - z_1 \in K_1$, we have by (2.7) that

$$\lambda(y^* - y) = \lambda(y^* - (y - z)) - \lambda(z) \geq -\lambda(z),$$

and

$$(q_1, y^*(1) - y(1)) = (q_1, y^*(1) - (y(1) - z_1)) - (q, z_1) \geq -(q_1, z_1).$$

It follows by (2.18), (2.6) and (2.8) (note that $\mu = \mu_a = 0$) that

$$\phi^\infty(z, z_1) - \phi^\infty(0, 0) \geq -\lambda(z) - (q_1, z),$$

which implies (H6) with $\varepsilon = \infty$.

3. Some lemmas. For the proof of Theorems 2.1 and 2.2, we will use a similar argument as that in [4, §§ 5.1, 5.2, 5.3 and 5.4]. But before that, some lemmas will be necessary.

LEMMA 3.1. Let $\gamma \in L^\infty(Q)$ be any positive function, let $\zeta \in (L^\infty(0, 1; H))^*$ and $p_1 \in H$. Then there exists a unique solution $p \in BV([0, 1]; Y^*) \cap L^\infty(0, 1; H) \cap L^2(0, 1; V)$ to the following equation:

$$(3.1) \quad \begin{aligned} p'(t) - Ap(t) - \gamma p(t) &= \zeta \quad \text{in } (0, 1), \\ p(1) &= p_1. \end{aligned}$$

Furthermore, there exists a constant $C \geq 0$ independent of γ such that

$$(3.2) \quad \|p\|_{L^\infty(0,1;H)} + \|p\|_{L^2(0,1;V)} + \|\gamma p\|_{L^1(Q)} \leq C(\|\zeta\|_{(L^\infty(0,1;H))^*} + |p_1|_2).$$

Remark 3.1. Equation (3.1) should be interpreted as follows: for any $z \in W^{1,2}([0, 1]; Y)$ (or equivalently, for any $z \in W^{1,2}([0, 1]; H) \cap L^2(0, 1; D(A_H))$) with $z(0) = 0$,

$$(3.3) \quad (p_1, z(1)) - \int_0^1 (p, z' + Az + \gamma z) dt = \zeta(z).$$

Proof of Lemma 3.1. The uniqueness follows by the fact that for any $y \in L^2(0, 1; H)$ there exists some $z \in W^{1,2}([0, 1]; H) \cap L^2(0, 1; D(A_H))$ such that $z(0) = 0$ and

$$z' + Az + \gamma z = y.$$

To establish the existence and relation (3.2), we first assume that

$$\zeta \in L^2(0, 1; H) \quad \text{and} \quad p_1 \in V.$$

In this case, the solution p to (3.1) exists and belongs to $W^{1,2}([0, 1]; H) \cap L^2(0, 1; D(A_H))$. Multiplying (3.1) by p , and then integrating over $[t, 1]$, we obtain

$$|p(t)|_2^2 + \int_t^1 (Ap(\tau), p(\tau)) d\tau + \int_t^1 \int_\Omega \gamma p(\tau)^2 dx d\tau = |p(1)|_2^2 + \int_t^1 (\zeta(\tau), p(\tau)) d\tau.$$

Using (H2) and $\gamma \geq 0$, we get

$$|p(t)|_2^2 + \omega \int_t^1 \|p(\tau)\|^2 d\tau \leq |p_1|_2^2 + \alpha \int_t^1 |p(\tau)|_2^2 d\tau + \int_t^1 (\zeta(\tau), p(\tau)) d\tau,$$

by which we deduce

$$\|p\|_{L^\infty(0,1;H)} + \|p\|_{L^2(0,1;V)} \leq C_1(\|\zeta\|_{L^1(0,1;H)} + |p_1|_2)$$

for some $C_1 \geq 0$ depending on ω and α . An argument similar to that of [4, Lemma 5.3] also shows

$$\int_Q |\gamma p| dx dt \leq C_2(\|\zeta\|_{L^1(0,1;H)} + |p_1|_2),$$

with $C_2 \geq 0$ independent of γ . Combining the last two relations, we conclude (3.2) with $C = C_1 + C_2$.

Now let $\zeta \in (L^\infty(0, 1; H))^*$ and $p_1 \in H$. Then there exist some sequences $\zeta^n \in L^2(0, 1; H)$ and $p_1^n \in V$ such that

$$(3.4) \quad \begin{aligned} \zeta^n &\rightarrow \zeta \quad \text{weak star in } (L^\infty(0, 1; H))^*, \\ p_1^n &\rightarrow p_1 \quad \text{weakly in } H. \end{aligned}$$

Let p^n be the solution of:

$$(3.5) \quad \begin{aligned} (p^n)' - Ap^n - \gamma p^n &= \zeta^n \quad \text{in } (0, 1), \\ p^n(1) &= p_1^n. \end{aligned}$$

Then as we have shown

$$(3.6) \quad \|p^n\|_{L^\infty(0,1;H)} + \|p^n\|_{L^2(0,1;V)} + \|\gamma p^n\|_{L^1(Q)} \leq C(\|\zeta^n\|_{(L^\infty(0,1;H))^*} + |p_1^n|_2).$$

So there is a subsequence of p^n (still denoted by p^n) such that

$$(3.7) \quad \begin{aligned} p^n &\rightarrow p \quad \text{weak star in } L^\infty(0, 1; H), \\ &\quad \text{weakly in } L^2(0, 1; V), \quad \text{strongly in } L^2(0, 1; H), \end{aligned}$$

$$(3.8) \quad (p^n)' = Ap^n - \gamma p^n + \zeta^n \rightarrow p' \quad \text{weak star in } (L^\infty(0, 1; Y))^*,$$

for some $p \in L^\infty(0, 1; H) \cap L^2(0, 1; V)$. Since $p^n(t)$ is bounded in H for each t , and the inclusion of $H = H^*$ in Y^* is compact, we may apply the theorem of Helly to conclude that

$$(3.9) \quad p^n(t) \rightarrow p(t) \quad \text{in } Y^* \text{ for any } t \in [0, 1],$$

and $p \in BV([0, 1]; Y^*)$. Tending to the limit in (3.5) and (3.6), and using (3.4), (3.7) and (3.8), we may deduce (3.1) and (3.2). Q.E.D.

Remark 3.2. From the proof of Lemma 3.1, we see that if $\zeta^n \in (L^\infty(0, 1; H))^*$ and $p_1^n \in H$ satisfy (3.4), then we have (3.7), (3.8) and (3.9), where p^n and p are solutions to (3.5) and (3.1) respectively.

LEMMA 3.2. Let $g(t, \cdot)$, ϕ_0 and β be Fréchet differentiable with $\dot{\beta} \in L^\infty(R) \cap C(R)$ and let $h = h(t, u): [0, 1] \times U \rightarrow R$ be a function which is measurable in t , Fréchet differentiable and convex in u and satisfies

$$(3.10) \quad \nabla h(\cdot, u(\cdot)) \in L^2(0, 1; U) \quad \text{whenever } u \in L^2(0, 1; U).$$

Assume that (\bar{y}, \bar{u}) is a local solution (in the sense of Remark 2.1) for the problem:

Minimize

$$\int_0^1 (g(t, y(t)) + h(t, u(t))) dt + \phi_0(y(1)) + \psi_0(y, y(1))$$

over all $(y, u) \in W^{1,2}([0, 1]; H) \times L^2(0, 1; U)$ subject to (1.2),

where $\psi_0: L^\infty(0, 1; H) \times H \rightarrow R$ is Lipschitz in some neighborhood of $(\bar{y}, \bar{y}(1))$. Then there exist $p \in BV([0, 1]; Y^*) \cap L^\infty(0, 1; H) \cap L^2(0, 1; V)$, $\lambda \in (L^\infty(0, 1; H))^*$ and $q_1 \in H$ such that

$$(3.11) \quad p' - Ap - \dot{\beta}(\bar{y} - \Psi)p - \lambda = \nabla g(t, \bar{y}(t)) \quad \text{in } (0, 1),$$

$$(3.12) \quad p(1) + q_1 + \nabla \phi_0(\bar{y}(1)) = 0,$$

$$(3.13) \quad (\lambda, q_1) \in \partial \psi_0(\bar{y}, \bar{y}(1)),$$

$$(3.14) \quad B^*p(t) = \nabla h(t, \bar{u}(t)) \quad \text{a.e. } t \in (0, 1).$$

Proof. For any $v \in L^2(0, 1; U)$ and $\rho > 0$, we let y_ρ be the solution to equation (1.2) with $u = u_\rho = \bar{u} + \rho v$. For ρ small enough, we have:

$$(3.15) \quad \begin{aligned} &\int_0^1 (g(t, y_\rho(t)) + h(t, \bar{u}(t) + \rho v(t))) dt + \phi_0(y_\rho(1)) + \psi_0(y_\rho, y_\rho(1)) \\ &\quad \geq \int_0^1 (g(t, \bar{y}(t)) + h(t, \bar{u}(t))) dt + \phi_0(\bar{y}, \bar{y}(1)) + \psi_0(\bar{y}(1)). \end{aligned}$$

It is not difficult to see that

$$(3.16) \quad \lim_{\rho \searrow 0} \frac{1}{\rho} (y_\rho - \bar{y}) = z(v) \quad \text{in } C([0, 1]; H)$$

where $z(v) \in W^{1,2}([0, 1]; H) \cap L^2(0, 1; D(A_H))$ denotes the solution to

$$(3.17) \quad \begin{aligned} z(v)' + Az(v) + \beta(\bar{y} - \Psi)z(v) &= Bv, \\ z(v)(0) &= 0. \end{aligned}$$

So

$$(3.18) \quad \begin{aligned} \lim_{\rho \searrow 0} \frac{1}{\rho} \int_0^1 (g(t, y_\rho(t)) - g(t, \bar{y}(t))) dt &= \int_0^1 (\nabla g(t, \bar{y}(t)), z(v)) dt, \\ \lim_{\rho \searrow 0} \frac{1}{\rho} (\phi_0(y_\rho(1)) - \phi_0(\bar{y}(1))) &= (\nabla \phi_0(\bar{y}(1)), z(v)(1)), \end{aligned}$$

and since ψ_0 is Lipschitz near $(y, y(1))$,

$$(3.19) \quad \limsup_{\rho \searrow 0} \frac{1}{\rho} (\psi_0(y_\rho, y_\rho(1)) - \psi_0(\bar{y}, \bar{y}(1))) \leq \psi_0^\dagger(\bar{y}, \bar{y}(1); z(v), z(v)(1)).$$

On the other hand, we have for $\rho \in (0, 1]$,

$$\begin{aligned} \int_0^1 \langle \nabla h(t, \bar{u}(t)), v(t) \rangle dt &\leq \int_0^1 \frac{1}{\rho} (h(t, \bar{u}(t) + \rho v(t)) - h(t, \bar{u}(t))) dt \\ &\leq - \int_0^1 \langle \nabla h(t, \bar{u}(t) + v(t)), v(t) \rangle dt, \end{aligned}$$

and

$$\lim_{\rho \searrow 0} \frac{1}{\rho} (h(t, \bar{u}(t) + \rho v(t)) - h(t, \bar{u}(t))) = \langle \nabla h(t, \bar{u}(t)), v(t) \rangle \quad \text{a.e. } t \in (0, 1).$$

Then it follows by (3.10) that

$$(3.20) \quad \lim_{\rho \searrow 0} \frac{1}{\rho} \int_0^1 (h(t, \bar{u}(t) + \rho v(t)) - h(t, \bar{u}(t))) dt = \int_0^1 \langle \nabla h(t, \bar{u}(t)), v(t) \rangle dt.$$

Now let Z be the following linear space:

$$Z = \{z \in W^{1,2}([0, 1]; H); z = z(v) \text{ for some } v \in L^2(0, 1; U)\},$$

and define $T: Z \rightarrow \mathbb{R}$ by

$$(3.21) \quad \begin{aligned} T(z(v)) &= - \int_0^1 ((\nabla g(t, \bar{y}(t)), z(v)) + \langle \nabla h(t, \bar{u}(t)), v \rangle) dt \\ &\quad - (\nabla \phi_0(\bar{y}(1)), z(v)(1)). \end{aligned}$$

Then we obtain by (3.15), (3.18), (3.19), (3.20) and the Lipschitz continuity of ψ_0 near $(\bar{y}, \bar{y}(1))$ that

$$(3.22) \quad T(z) \leq \psi_0^\dagger(\bar{y}, \bar{y}(1); z, z(1)) \leq C(\|z\|_\infty + |z(1)|_2),$$

for any $z \in Z$, where C is the Lipschitz constant of ψ_0 in a small neighborhood of $(\bar{y}, \bar{y}(1))$. The last inequality also shows that T is well defined, i.e., for any $v, w \in L^2(0, 1; U)$ we have

$$T(z(v)) = T(z(w)) \quad \text{whenever } z(v) = z(w).$$

T is obviously linear, then by the Hahn-Banach theorem, there exists some $(\lambda, q_1) \in (L^\infty(0, 1; H))^* \times H$ such that

$$(3.23) \quad \lambda(z) + (q_1, z(1)) = T(z) \quad \forall z \in Z,$$

and

$$\lambda(y) + (q_1, y_1) \leq \psi_0^*(\bar{y}, \bar{y}(1); y, y_1) \quad \forall (y, y_1) \in L^\infty(0, 1; H) \times H.$$

The last inequality shows (3.13). By Lemma 3.1, there exists a unique solution p to (3.11) and (3.12). From (3.21), (3.23) and (3.17), we may deduce

$$\int_0^1 \langle \nabla h(t, u(t)) - B^*p(t), v(t) \rangle dt = 0 \quad \forall v \in L^2(0, 1; U),$$

which yields (3.14). Q.E.D.

LEMMA 3.3. Let $g(t, \cdot)$, ϕ_0 and β be Fréchet differentiable with $\dot{\beta} \in L^\infty(R) \cap C(R)$. Let (\bar{y}, \bar{u}) be a local solution for the problem:

Minimize

$$(3.24) \quad G(y, u) = G_0(y, u) + \psi_0(y, y(1)) + \frac{1}{2} \int_0^1 |u - u^*|_U^2 dt$$

over all $(y, u) \in W^{1,2}([0, 1]; H) \times L^2(0, 1; U)$ subject to (1.2),

where G_0 is defined by (1.1) and $\psi_0: L^\infty(0, 1; H) \times H \rightarrow R$ is Lipschitz in some neighborhood of $(\bar{y}, \bar{y}(1))$. Then there exist $p \in BV([0, 1]; Y^*) \cap L^\infty(0, 1; H) \cap L^2(0, 1; V)$, $\lambda \in (L^\infty(0, 1; H))^*$ and $q_1 \in H$ such that

$$(3.25) \quad p' - Ap - \dot{\beta}(\bar{y} - \Psi)p - \lambda = \nabla g(t, \bar{y}(t)) \quad \text{in } (0, 1),$$

$$(3.26) \quad p(1) + q_1 + \nabla \phi_0(\bar{y}(1)) = 0,$$

$$(3.27) \quad (\lambda, q_1) \in \partial \psi_0(\bar{y}, \bar{y}(1)),$$

$$(3.28) \quad B^*p(t) \in \partial h(\bar{u}(t)) + \bar{u} - u^* \quad \text{a.e. } t \in (0, 1).$$

Proof. Let $h_\rho: U \rightarrow R$ be the function

$$h_\rho(u) = \inf \left\{ \frac{1}{2\rho} |u - v|_U^2 + h(v); v \in U \right\},$$

where $\rho > 0$. Consider the approximating problem:

Minimize

$$\int_0^1 (g(t, y(t)) + h_\rho(u(t)) + \frac{1}{2} |u(t) - u^*(t)|_U^2 + \frac{1}{2} |u(t) - \bar{u}(t)|_U^2) dt \\ + \phi_0(y(1)) + \psi_0(y, y(1))$$

over all $(y, u) \in W^{1,2}([0, 1]; H) \times L^2(0, 1; U)$ subject to (1, 2) and

$$\|y - \bar{y}\|_\infty \leq \varepsilon_0, \quad \|u - \bar{u}\|_2 \leq \varepsilon_0,$$

with $\varepsilon_0 > 0$ small enough. Let $(\bar{y}_\rho, \bar{u}_\rho)$ be an optional pair to the above problem which obviously exists. Then we may show that

$$(3.29) \quad \begin{aligned} \bar{u}_\rho &\rightarrow \bar{u} \quad \text{in } L^2(0, 1; U) \text{ for } \rho \searrow 0, \\ \bar{y}_\rho &\rightarrow \bar{y} \quad \text{in } W^{1,2}([0, 1]; H) \cap L^2(0, 1; D(A_H)) \text{ for } \rho \searrow 0. \end{aligned}$$

Applying Lemma 3.2, we deduce that there exist for any ρ small enough some p_ρ , λ_ρ and $q_{1\rho}$ such that

$$\begin{aligned}
 (3.30) \quad & p'_\rho - Ap_\rho - \dot{\beta}(\bar{y}_\rho - \Psi)p_\rho - \lambda_\rho = \nabla g(t, \bar{y}_\rho) \quad \text{in } (0, 1), \\
 & p_\rho(1) + q_{1\rho} + \nabla \phi_0(\bar{y}_\rho(1)) = 0, \\
 & (\lambda_\rho, q_{1\rho}) \in \partial \psi_0(\bar{y}_\rho, \bar{y}_\rho(1)), \\
 & B^*p_\rho(t) = \nabla h_\rho(\bar{u}_\rho(t)) + \bar{u}_\rho - u^* + \bar{u}_\rho - \bar{u} \quad \text{a.e. } t \in (0, 1).
 \end{aligned}$$

In virtue of (3.29) and the Lipschitz continuity of ψ_0 near $(\bar{y}, \bar{y}(1))$, we have

$$(3.31) \quad \|\lambda_\rho\|_{(L^\infty(0,1;H))^*} + |q_{1\rho}|_2 \leq C_1,$$

here and in the following C_i , $i = 1, 2, \dots$, are some positive constants independent of $\rho > 0$ (ρ small enough). By (H5), we also get

$$(3.32) \quad \|\nabla g(t, \bar{y}_\rho)\|_\infty + |\nabla \phi_0(\bar{y}_\rho(1))|_2 \leq C_2.$$

Since $\dot{\beta} \geq 0$, we deduce from (3.30), (3.31), (3.32) and Lemma 3.1 that p_ρ is bounded in $BV([0, 1]; Y^*) \cap L^\infty(0, 1; H) \cap L^2(0, 1; V)$ for $\rho \searrow 0$. An argument similar to the last part of the proof of Lemma 3.1 shows for some sequence $\rho_m \rightarrow 0$, and some $p \in BV([0, 1]; Y^*) \cap L^\infty(0, 1; H) \cap L^2(0, 1; V)$,

$$\begin{aligned}
 (3.33) \quad & p_{\rho_m} \rightarrow p \quad \text{weak star in } L^\infty(0, 1; H), \\
 & \text{weakly in } L^2(0, 1; V), \quad \text{strongly in } L^2(0, 1; H),
 \end{aligned}$$

$$(3.34) \quad (p_{\rho_m})' \rightarrow p' \quad \text{weak star in } (L^\infty(0, 1; Y))^*,$$

$$(3.35) \quad p_{\rho_m}(t) \rightarrow p(t) \quad \text{in } Y^* \quad \forall t \in [0, 1].$$

Since $\dot{\beta}$ is continuous and bounded, we may get by (3.29) that

$$\dot{\beta}(\bar{y}_{\rho_m} - \Psi) \rightarrow \dot{\beta}(\bar{y} - \Psi) \quad \text{weak star in } L^\infty(Q).$$

Combining this with (3.33), we get

$$(3.36) \quad \dot{\beta}(\bar{y}_{\rho_m} - \Psi)p_{\rho_m} \rightarrow \dot{\beta}(\bar{y} - \Psi)p \quad \text{weakly in } L^2(0, 1; H).$$

By (3.31), (3.32) and (3.29), we may assume that

$$\begin{aligned}
 (3.37) \quad & \lambda_{\rho_m} \rightarrow \lambda \quad \text{weak star in } (L^\infty(0, 1; H))^*, \\
 & q_{1\rho_m} \rightarrow q_1 \quad \text{weakly in } H. \\
 (3.38) \quad & \nabla g(t, \bar{y}_{\rho_m}) \rightarrow \nabla g(t, \bar{y}) \quad \text{weak star in } L^\infty(0, 1; H), \\
 & \nabla \phi_0(\bar{y}_{\rho_m}(1)) \rightarrow \nabla \phi_0(\bar{y}(1)) \quad \text{weakly in } H.
 \end{aligned}$$

Now tending to the limit in (3.30) for $\rho = \rho_m \searrow 0$, using (3.29), (3.33), (3.34), (3.35), (3.36), (3.37) and (3.38), we may conclude that p , λ and q_1 satisfy (3.25)–(3.28) (for the proof of (3.27), see also [4, Prop. 1.11(ii)]). Q.E.D.

4. Proof of Theorems 2.1 and 2.2. By (H6), there exist $\varepsilon > 0$ and $r_0 > 0$ such that

$$(4.1) \quad M = - \inf_{\|z\|_\infty + |z|_2 \leq r_0} (\phi^\varepsilon(z, z_1) - \phi^\varepsilon(0, 0)) / (\|z\|_\infty + |z|_2)$$

is not $+\infty$. Define the functions $d: L^\infty(0, 1; H) \rightarrow [0, +\infty)$ and $d_1: H \rightarrow [0, \infty)$ by

$$(4.2) \quad d(y) = \inf \{\|y - z\|_\infty, z \in \mathcal{K}\},$$

where \mathcal{K} is the set (2.3), and

$$(4.3) \quad d_1(y_1) = \inf \{|y_1 - z_1|_2; a_1 \in K_1\},$$

respectively. Let $\tilde{\phi}: [0, +\infty) \rightarrow [0, +\infty]$ be the function:

$$(4.4) \quad \tilde{\phi}(r) = -\inf \{\phi^\varepsilon(z, z_1) - \phi^\varepsilon(0, 0); \|z\|_\infty + |z_1|_2 \leq r\}.$$

Then $\tilde{\phi}$ is nonnegative and increasing in r . By (4.1), we have

$$\tilde{\phi}(r) \leq Mr \quad \text{if } 0 \leq r \leq r_0.$$

Then we may construct a continuous function $\tilde{\psi}: [0, +\infty) \rightarrow [0, +\infty]$ satisfying the properties:

$$(4.5) \quad \tilde{\psi}(r) = Mr \quad \forall r \in [0, r_0/2],$$

$$(4.6) \quad \tilde{\psi}(r) \geq \tilde{\phi}(r) \quad \forall r \in [0, +\infty).$$

Let $\psi_0: L^\infty(0, 1; H) \times H \rightarrow [0, +\infty]$ be the function defined by:

$$(4.7) \quad \psi_0(y, y_1) = \tilde{\psi}(d(y) + d_1(y_1)).$$

We have:

LEMMA 4.1. *Let (y^*, u^*) be an arbitrary optimal pair for the control problem (P) and suppose (H6) holds. Then (y^*, u^*) solves locally the following problem:*

Minimize

$$(4.8) \quad G(y, u) = G_0(y, u) + \psi_0(y, y(1))$$

over all $(y, u) \in W^{1,2}([0, 1]; H) \times L^2(0, 1; U)$ subject to (1.2).

Proof. Assume contrarily that there exists a pair (y, u) satisfying (1.2) such that

$$G(y, u) < G(y^*, u^*),$$

and

$$(4.9) \quad \|y - y^*\|_\infty \leq \varepsilon_0 = \varepsilon, \quad \|u - u^*\|_2 \leq \varepsilon_0 = \varepsilon.$$

Then by (4.7) and (4.8),

$$G_0(y, u) + \tilde{\psi}(d(y) + d_1(y(1))) < G_0(y^*, u^*).$$

As $\tilde{\psi}$ is continuous, there exists some $\delta > 0$ such that

$$G_0(y, u) + \tilde{\psi}(d(y) + d_1(y(1)) + \delta) < G_0(y^*, u^*).$$

By (4.6), we get

$$\tilde{\phi}(d(y) + d_1(y(1)) + \delta) < -(G_0(y, u) - G_0(y^*, u^*)),$$

which in virtue of (2.2), (4.9), (4.2) and (4.3) contradicts (4.4). Lemma 4.1 is proved.

It is easy to see from (4.5), (4.7) and (1.9) that

$$\partial\psi_0(y^*, y^*(1)) = M(\partial d(y^*) \times \partial d_1(y^*(1))) \subseteq N_{\mathcal{K}}(y^*) \times N_{K_1}(y^*(1)).$$

So in virtue of Lemma 4.1, Theorems 2.1 and 2.2 follow from Propositions 4.1 and 4.2 below.

PROPOSITION 4.1. *Let (y^*, u^*) be any local optimal pair for the following problem:*

(P') *Minimize*

$$G(y, u) = G_0(y, u) + \psi_0(y, y(1))$$

over all $(y, u) \in W^{1,2}([0, 1]; H) \times L^2(0, 1; U)$ subject to (1.2).

Suppose that (H1)–(H5) hold and $\psi_0: L^\infty(0, 1; H) \times H \rightarrow (-\infty, +\infty]$ is Lipschitz in some neighborhood of $(y^*, y^*(1))$. If β is locally Lipschitz, then there exist $p \in BV([0, 1]; Y^*) \cap L^\infty(0, 1; H) \cap L^2(0, 1; V)$, $\mu \in (L^\infty(Q))^*$, $\lambda \in (L^\infty(0, 1; H))^*$ and $q_1 \in H$ such that

$$(4.10) \quad p' - Ap - \mu - \lambda \in \partial g(t, y^*) \quad \text{a.e. in } Q,$$

$$(4.11) \quad \mu_a(x, t) \in p(y, t) \partial \beta(y^*(x, t) - \Psi) \quad \text{a.e. in } Q,$$

$$(4.12) \quad p(1) + q_1 + \partial \phi_0(y^*(1)) \ni 0,$$

$$(4.13) \quad (\lambda, q_1) \in \partial \psi_0(y^*, y^*(1)),$$

$$(4.14) \quad B^*p(t) \in \partial h(u^*(t)) \quad \text{a.e. } t \in (0, 1).$$

Furthermore, if in addition β satisfies (2.9), then $\mu = \mu_a \in L^q(Q)$.

PROPOSITION 4.2. Let β be defined by (2.10). Under the assumptions of Proposition 4.1, there exist $p \in BV([0, 1]; Y^*) \cap L^\infty(0, 1; H) \cap L^2(0, 1; V)$, $\lambda \in (L^\infty(0, 1; H))^*$ and $q_1 \in H$ such that

$$(4.15) \quad (p' - Ap - \lambda)_a \in \partial g(t, y^*) \quad \text{a.e. in } \{y^* > \Psi\},$$

$$(4.16) \quad p(f + Bu^* - Ay^* - y^{*'}) = 0 \quad \text{a.e. in } Q,$$

$$(4.17) \quad p(1) + q_1 + \partial \phi_0(y^*(1)) \ni 0,$$

$$(4.18) \quad (\lambda, q_1) \in \partial \psi_0(y^*, y^*(1)),$$

$$(4.19) \quad B^*p(t) \in \partial h(u^*(t)) \quad \text{a.e. } t \in (0, 1).$$

Proof of Propositions 4.1 and 4.2. We will proceed as in the proof of [4, Thms. 5.1 and 5.2]. The notations of [4, § 5.2] will be implicitly used.

Let $\rho > 0$. Then there exists at least one solution (y_ρ, u_ρ) to the following approximating problem:

$$(P'_\rho) \quad \begin{aligned} &\text{Minimize} \\ &G^\rho(y, u) = \int_0^1 (g^\rho(t, y(t)) + h(t, u(t)) + \tfrac{1}{2}\|u(t) - u^*(t)\|_U^2) dt \\ &\quad + \phi_0^\rho(y(1)) + \psi_0(y, y(1)) \\ &\text{over all } (y, u) \in W^{1,2}([0, 1]; H) \times L^2(0, 1; U) \text{ subject to} \\ &y' + Ay + \beta^\rho(y - \Psi) = Bu + f \quad \text{a.e. } t \in (0, 1), \\ &y(0) = y_0, \\ &\text{and } \|y - y^*\|_\infty \leq \varepsilon_0, \|u - u^*\|_2 \leq \varepsilon_0. \end{aligned}$$

As in the proof of [4, Lemma 5.2], we may show that for $\rho \searrow 0$,

$$(4.21) \quad u_\rho \rightarrow u^* \quad \text{in } L^2(0, 1; U),$$

$$(4.22) \quad y_\rho \rightarrow y^* \quad \text{in } L^2(0, 1; V) \cap C([0, 1]; H),$$

$$\text{weakly in } W^{1,2}([0, 1]; H) \cap L^2(0, 1; D(A_H)),$$

$$(4.23) \quad \beta^\rho(y_\rho - \Psi) \rightarrow \eta \quad \text{weakly in } L^2(0, 1; H)$$

where $\eta = f + Bu^* - Ay^* - y^{*'} \in \beta(y^* - \Psi)$.

By Lemma 3.3, there exist for any $\rho > 0$ small enough, some $p_\rho \in BV([0, 1]; Y^*) \cap L^\infty(0, 1; H) \cap L^2(0, 1; V)$, $\lambda_\rho \in (L^\infty(0, 1; H))^*$ and $q_{1\rho} \in H$ such that

$$(4.24) \quad p'_\rho - Ap_\rho - \dot{\beta}^\rho(y_\rho - \Psi)p_\rho - \lambda_\rho = \nabla g^\rho(t, y_\rho(t)) \quad \text{in } (0, 1),$$

$$(4.25) \quad p_\rho(1) + q_{1\rho} + \nabla \phi_0^\rho(y_\rho(1)) = 0,$$

$$(4.26) \quad (\lambda_\rho, q_{1\rho}) \in \partial \psi_0(y_\rho, y_\rho(1)),$$

$$(4.27) \quad B^*p_\rho(t) \in \partial h(u_\rho(t)) + u_\rho - u^*(t) \quad \text{a.e. in } (0, 1).$$

In virtue of the Lipschitz continuity of ψ_0 near $(y^*, y^*(1))$ and (4.22), we have for ρ small,

$$(4.28) \quad \|\lambda_\rho\|_{(L^\infty(0,1;H))^*} + |q_{1\rho}|_2 \leq C_1.$$

(As before, C_i , $i = 1, 2, \dots$, denote constants independent of ρ .) By hypothesis (H5),

$$(4.29) \quad \|\nabla g^\rho(\cdot, y_\rho(\cdot))\|_\infty + |\nabla \phi_0^\rho(y_\rho(1))|_2 \leq C_2.$$

Using Lemma 3.1, we deduce from (4.24), (4.25), (4.28) and (4.29) that

$$(4.30) \quad \|p_\rho\|_{L^\infty(0,1;H)} + \|p_\rho\|_{L^2(0,1;V)} + \int_Q |\dot{\beta}^\rho(y_\rho - \Psi)p_\rho| \, dx \, dt \leq C_3.$$

By (4.22), (4.28), (4.29) and (4.30), we may select a sequence $\rho_m \searrow 0$ such that

$$(4.31) \quad \lambda_{\rho_m} \rightarrow \lambda \quad \text{weak star in } (L^\infty(0, 1; H))^*,$$

$$(4.32) \quad q_{1\rho_m} \rightarrow q_1 \quad \text{weakly in } H,$$

$$(4.33) \quad \nabla g^{\rho_m}(t, y_{\rho_m}(t)) \rightarrow \zeta \quad \text{weak star in } L^\infty(0, 1; H),$$

$$(4.34) \quad \nabla \phi_0^{\rho_m}(y_{\rho_m}(1)) \rightarrow q_2 \quad \text{weakly in } H,$$

$$(4.35) \quad p_{\rho_m} \rightarrow p \quad \text{weak star in } L^\infty(0, 1; H),$$

$$\text{weakly in } L^2(0, 1; V),$$

$$(4.36) \quad \dot{\beta}^{\rho_m}(y_{\rho_m} - \Psi)p_{\rho_m} \rightarrow \mu \quad \text{weak star in } (L^\infty(Q))^*,$$

for some $\lambda \in (L^\infty(0, 1; H))^*$, $q_1 \in H$, $\zeta \in L^\infty(0, 1; H)$, $q_2 \in H$, $p \in L^\infty(0, 1; H) \cap L^2(0, 1; V)$ and $\mu \in (L^\infty(Q))^*$. As in the last part of the proof of Lemma 3.1, we may infer that $p \in BV([0, 1]; Y^*)$ and

$$p_{\rho_m}(t) \rightarrow p(t) \quad \text{in } Y^* \quad \forall t \in [0, 1].$$

As in [5, Lemma 5.4], we get

$$(4.37) \quad \zeta(t) \in \partial g(t, y^*(t)) \quad \text{a.e. } t \in (0, 1),$$

and

$$(4.38) \quad q_2 \in \partial \phi_0(y^*(1)).$$

Tending to the limit in (4.24)–(4.27), it follows that p , λ and q_1 satisfy the following equations

$$(4.39) \quad p' - Ap - \mu - \lambda \in \partial g(t, y^*),$$

$$(4.40) \quad p(1) + q_1 + \partial \phi_0(y^*(1)) \ni 0,$$

$$(4.41) \quad (\lambda, q_1) \in \partial \psi_0(y^*, y^*(1)),$$

$$(4.42) \quad B^*p(t) \in \partial h(u^*(t)) \quad \text{a.e. } t \in (0, 1),$$

where μ is the limit in (4.36).

Now the argument of [4, § 5.3] and [4, § 5.4] may be repeated here word for word, and it completes the proof of Propositions 4.1 and 4.2. Q.E.D.

5. State constrained boundary control problems. We will consider here the state constrained control systems with nonlinear boundary value conditions. The other types of state constrained boundary control problems may be studied in the same way. Since the proof of the results is similar to that in the previous case (using [4, § 6.1]), it will be omitted. Our results constitute extensions of [4, Thms. 6.1 and 6.2].

Let Ω be a bounded domain of R^N with smooth boundary $\Gamma = \partial\Omega$. Assume Γ_1 and Γ_2 are two disjoint and smooth parts of Γ such that $\Gamma = \Gamma_1 \cup \Gamma_2$. Let $a_{jk} \in C^1(\bar{\Omega})$, $a_0 \in L^\infty(\Omega)$, with $a_{jk} = a_{kj}$ for $j, k = 1, 2, \dots, N$. Assume that for some $\omega > 0$, we have

$$(5.1) \quad \sum_{j,k=1}^N a_{jk}(x) \xi_j \xi_k \geq \omega \|\xi\|^2 \quad \forall x \in \Omega, \quad \xi \in R^N.$$

Define A_0 to be the following second order differential operator

$$(5.2) \quad A_0 y = - \sum_{j,k=1}^N (a_{jk}(x) y_{x_j})_{x_k} + a_0(x) y(x).$$

We will consider the problem:

(P_b) Minimize

$$(5.3) \quad G_1(y, u_1, u_2) = \int_0^1 (g(t, y(t)) + h_1(u_1(t)) + h_2(u_2(t))) dt + \phi_0(y(1))$$

over all $y \in W([0, 1]; H^1(\Omega)) = L^2(0, 1; H^1(\Omega)) \cap W^{1,2}([0, 1]; H^1(\Omega)^*)$
and $(u_1, u_2) \in L^2(0, 1; U_1) \times L^2(0, 1; U_2)$ subject to

$$(5.4) \quad \begin{cases} y_t + A_0 y = f_0 & \text{in } Q = \Omega \times (0, 1), \\ y(x, 0) = y_0(x), & x \in \Omega, \\ \frac{\partial y}{\partial \nu} + \beta_i(y) \ni B_i u_i + f_i & \text{in } \Sigma_i = \Gamma_i \times (0, 1), \quad i = 1, 2, \dots, \end{cases}$$

and

$$(5.5) \quad y(t) \in K, \quad t \in [0, 1],$$

$$(5.6) \quad y(1) \in K_1.$$

Here $f_0 \in L^2(Q)$, $f_i \in L^2(\Gamma_i)$, $i = 1, 2$; U_i are Hilbert spaces with the scalar products $\langle \cdot, \cdot \rangle_i$; B_i are linear continuous operators from U_i to $L^2(\Gamma_i)$; $\partial y / \partial \nu$ denotes the conormal derivative on Γ ; y_t is the derivative of y with respect to t ; β_i are maximal graphs in $R \times R$ such that $0 \in \beta_i(0)$; $g: [0, 1] \times H \rightarrow R$ and $\phi_0: H \rightarrow R$ satisfy (H5) of § 2; $h_i: U_i \rightarrow (-\infty, +\infty]$ satisfy (H4) of § 2; and $y_0 \in H^1(\Omega)$ is such that

$$j_i(y_0) \in L^1(\Gamma_i),$$

where

$$\partial j_i = \beta_i, \quad i = 1, 2.$$

Let (y^*, u^*) be any optimal pair to the control problem (P_b). We define for each $\varepsilon \in (0, +\infty]$ the function $\phi_b^\varepsilon: L^\infty(0, 1; H) \times H \rightarrow [-\infty, +\infty]$ by

$$(5.7) \quad \phi_b^\varepsilon(z, z_1) = \inf \{ G_1(y, u); y(t) \in z(t) + K, y(1) \in z_1 + K_1, \\ (y, u) \text{ satisfies (5.4), } \|y - y^*\|_\infty \leq \varepsilon, \|u - u^*\|_2 \leq \varepsilon \}.$$

Then we have:

THEOREM 5.1. *Let (y^*, u_1^*, u_2^*) be any optimal solution for the control problem (P_b) where β_i are locally Lipschitz functions satisfying (2.9). Assume that for some $\varepsilon > 0$,*

$$(5.8) \quad \liminf_{\|z\|_\infty + |z_1|_2 \rightarrow 0} (\phi_b^\varepsilon(z, z_1) - \phi_b^\varepsilon(0, 0)) / (\|z\|_\infty + |z_1|_2) \neq -\infty.$$

Then there exist $p \in BV([0, 1]; (H^1(\Omega))^) \cap L^2(0, 1; L^2(\Omega)) \cap L^2(0, 1; H^1(\Omega))$, $\lambda \in (L^\infty(0, 1; H))^*$ and $q_1 \in L^2(\Omega)$ such that $\partial p / \partial \nu \in L^1(\Sigma)$ and*

$$(5.9) \quad p_t - A_0 p - \lambda \in \partial g(t, y^*) \quad \text{in } Q,$$

$$(5.10) \quad \partial p / \partial \nu + p \partial \beta_i(y^*) \ni 0 \quad \text{on } \Gamma_i, \quad i = 1, 2,$$

$$(5.11) \quad p(1) + q_1 + \partial \phi_0(y^*(1)) \ni 0,$$

$$(5.12) \quad \lambda \in N_{\mathcal{H}}(y^*), \quad q_1 \in N_{K_1}(y^*(1)),$$

$$(5.13) \quad B_i^* p(t) \in \partial h_i(u_i^*(t)) \quad \text{a.e. } t \in (0, 1), \quad i = 1, 2,$$

where \mathcal{H} is the set (2.3).

THEOREM 5.2. *Suppose that $\Gamma_1 = \Gamma$, $\Gamma_2 = \phi$, $h_1 = h$ and $\beta_1 = \beta$ is given by (2.10). Let (y^*, u^*) be an optimal pair for the control problem (P_b) and assume that (5.8) is satisfied for some $\varepsilon > 0$. Then there exist $p \in BV([0, 1]; (H^1(\Omega))^*) \cap L^\infty(0, 1; L^2(\Omega)) \cap L^2(0, 1; H^1(\Omega))$, $\lambda \in (L^\infty(0, 1; H))^*$ and $q_1 \in L^2(\Omega)$ such that $\partial p / \partial \nu \in (L^\infty(\Sigma))^*$ and*

$$(5.14) \quad p_t - A_0 p - \lambda \in \partial g(t, y^*) \quad \text{in } Q,$$

$$(5.15) \quad \left(\frac{\partial p}{\partial \nu} \right)_a = 0 \quad \text{in } \{(\sigma, t) \in \Sigma; y^*(\sigma, t) > 0\},$$

$$(5.16) \quad p = 0 \quad \text{a.e. in } \left\{ (\sigma, t) \in \Sigma; y^*(\sigma, t) = 0, Bu^* - \frac{\partial y^*}{\partial \nu} - f_1 > 0 \right\},$$

$$(5.17) \quad p(1) + q_1 + \partial \phi_0(y^*(1)) \ni 0,$$

$$(5.18) \quad \lambda \in N_{\mathcal{H}}(y^*), \quad q_1 \in N_{K_1}(y^*(1)),$$

$$(5.19) \quad B^* p(t) \in \partial h(u^*(t)) \quad \text{a.e. } t \in (0, 1).$$

Remark 5.1. A function $p \in BV([0, 1]; (H^1(\Omega))^*) \cap L^2(0, 1; H^1(\Omega))$ is said to be the solution of the equation

$$(5.20) \quad \begin{aligned} p_t - A_0 p &= \zeta \quad \text{in } Q, \\ \frac{\partial p}{\partial \nu} &= \eta \quad \text{on } \Sigma, \\ p(1) &= p_1, \end{aligned}$$

with $\zeta \in (L^\infty(0, 1; H))^*$, $\eta \in (L^\infty(\Sigma))^*$ and $p_1 \in L^2(\Omega)$, if for any $y \in W^{1,2}([0, 1]; H^1(\Omega))$ with $y / \Sigma \in L^\infty(\Sigma)$ and $y(0) = 0$,

$$(5.21) \quad \sum_{j,k=1}^N \int_0^1 \int_\Omega a_{jk} \frac{\partial y}{\partial x_j} \frac{\partial y}{\partial x_k} dx dt + \int_0^1 (p, y_t) dt + (p_1, y(1)) = \zeta(y) + \eta(y / \Sigma).$$

6. State constrained control of elliptic variational inequalities. The state constrained control problems governed by elliptic variational inequalities are in parallel with those governed by parabolic variational inequalities. Here we will consider as a model the

distributed control problem. The proof for the results (i.e. for Theorems 6.1 and 6.2 below) may be simply a combination of a lemma like Lemma 4.1 and [4, Thms. 3.2 and 3.3].

The state constrained distributed control problem is of the following type.

(P_e) Minimize

$$(6.1) \quad g(y) + h(u)$$

over all $y \in H_0^1(\Omega) \cap H^2(\Omega)$, $u \in U$ subject to

$$(6.2) \quad A_0 y + \beta(y - \Psi) \ni f + Bu \quad \text{a.e. in } \Omega,$$

and

$$(6.3) \quad y \in K,$$

where $g: L^2(\Omega) \rightarrow \mathbb{R}$ is a locally Lipschitz function; U is a Hilbert space; $h: U \rightarrow (-\infty, +\infty]$ is convex and l.s.c.; A_0 has the same meaning as in § 5; β is a maximal graph in $\mathbb{R} \times \mathbb{R}$ with $0 \in \beta(0)$; $\Psi \in H^2(\Omega)$, $\Psi \leq 0$ on Γ ; $f \in L^2(\Omega)$; $B: U \rightarrow L^2(\Omega)$ is linear and continuous; K is a closed subset of $H = L^2(\Omega)$.

Let (y^*, u^*) be any optimal pair for the control problem (P_e). Define for any $\varepsilon \in (0, +\infty]$ the function $\phi_\varepsilon^*: H \rightarrow [-\infty, +\infty]$ by

$$(6.4) \quad \phi_\varepsilon^*(z) = \inf \{g(y) + h(u); y \in z + K, (y, u) \text{ satisfies (6.2)}, \\ |y - y^*|_2 \leq \varepsilon, |u - u^*|_U \leq \varepsilon\}.$$

We have (compare with [4, Thms. 3.2 and 3.3]):

THEOREM 6.1. *Let $(y^*, u^*) \in H_0^1(\Omega) \times U$ be any optimal pair for the control problem (P_e) where $\beta: \mathbb{R} \rightarrow \mathbb{R}$ is locally Lipschitz and monotonically increasing function. Assume that for some $\varepsilon \in (0, +\infty]$,*

$$(6.5) \quad \liminf_{|z|_2 \rightarrow 0} (\phi_\varepsilon^*(z) - \phi_\varepsilon^*(0)) / |z|_2 \neq -\infty.$$

Then there exist $p \in H_0^1(\Omega)$, $\lambda, \xi \in L^2(\Omega)$ such that $A_0 p \in (L^\infty(\Omega))^$ and*

$$(6.6) \quad -(A_0 p)_a - p \partial \beta(y^* - \Psi) \ni \lambda + \xi, \quad \xi \in \partial g(y^*) \quad \text{in } \Omega,$$

$$(6.7) \quad \lambda \in N_K(y^*),$$

$$(6.8) \quad B^* p \in \partial h(u^*).$$

Moreover, if either $1 \leq N \leq 3$ or β satisfies (2.9), then $A_0 p \in L^1(\Omega)$ and (6.6) becomes

$$(6.9) \quad -A_0 p - p \partial \beta(y^* - \Psi) \ni \lambda + \xi, \quad \xi \in \partial g(y^*) \quad \text{in } \Omega.$$

THEOREM 6.2. *Let (y^*, u^*) be any optimal pair for the control problem (P_e) where β is defined by (2.10). Assume that (6.5) holds for some $\varepsilon \in (0, +\infty]$. Then there exist $p \in H_0^1(\Omega)$, $\lambda, \xi \in L^2(\Omega)$ with $A_0 p \in (L^\infty(\Omega))^*$, $\xi \in \partial g(y^*)$ and*

$$(6.10) \quad (A_0 p)_a + \xi + \lambda = 0 \quad \text{a.e. in } \{y^* > \Psi\},$$

$$(6.11) \quad p(A_0 y^* - f - Bu^*) = 0 \quad \text{a.e. in } \Omega,$$

$$(6.12) \quad (A_0 p + \xi + \lambda, \chi(y^* - \Psi)) = 0 \quad \forall \chi \in C^1(\bar{\Omega}),$$

$$(6.13) \quad \lambda \in N_K(y^*),$$

$$(6.14) \quad B^* p \in \partial h(u^*),$$

$$(6.15) \quad (A_0 p + \xi + \lambda, p) \leq 0.$$

If $1 \leq N \leq 3$, then (6.10) reduces to

$$(6.16) \quad (y^* - \Psi)(A_0 p + \xi + \lambda) = 0.$$

7. Examples. We will discuss some specific cases of the control problems studied in the previous sections.

Example 1. (The state constrained convex control problem.) Consider the following problem

(Q₁) Minimize

$$(7.1) \quad G_0(y, u) = \int_0^1 (g(t, y) + h(u)) dt + \phi_0(y(1))$$

over $(y, u) \in W^{1,2}([0, 1]; H) \times L^2(0, 1; U)$ subject to

$$(7.2) \quad y' + Ay = Bu(t) + f(t),$$

and

$$(7.3) \quad y(t) \in K, \quad \forall t \in [0, 1],$$

$$(7.4) \quad y(1) \in K_1.$$

Here $H, U, g, h, \phi_0, A, B, f, y_0, K$ and K_1 have the same meaning as in § 2. In addition, we assume that

$g(t, \cdot)$ and ϕ_0 are convex functions,

K and K_1 are convex sets.

Problem (Q₁) is referred to as state constrained convex control problem.

We note that in this case ϕ^ε defined by (2.2) is l.s.c. (if $\varepsilon \in (0, +\infty)$), proper and convex. So we have

PROPOSITION 7.1. *Hypothesis (H6) is satisfied for $\varepsilon \in (0, +\infty)$ if and only if*

$$(7.5) \quad \partial\phi^\varepsilon(0) \text{ is not empty.}$$

In particular, if $0 \in \text{int } D(\phi^\varepsilon)$, then (H6) holds.

Proof. The “if” part is obvious. Let us prove the “only if” part. Since ϕ^ε is convex, it is subdifferentially regular at $(0, 0)$ (see [9, Prop. 3]), i.e.

$$(\phi^\varepsilon)^\dagger(0, 0; z, z_1) = \liminf_{\substack{(z', z'_1) \rightarrow (z, z_1) \\ \rho \searrow 0}} \frac{\phi^\varepsilon(\rho z', \rho z'_1) - \phi^\varepsilon(0, 0)}{\rho}.$$

So hypothesis (H6) says that

$$(\phi^\varepsilon)^\dagger(0, 0; 0, 0) \neq -\infty,$$

which implies (7.5) by [9, Thm. 4]. Q.E.D.

Let K_h be the subset

$$(7.6) \quad K_h = \{y(1) \in H; (y, u) \text{ satisfies (7.2), (7.3) and } H_0(u) < +\infty\},$$

where $H_0: L^2(0, 1; U) \rightarrow (-\infty, +\infty]$ is defined by

$$(7.7) \quad H_0(u) = \int_0^1 h(u(t)) dt.$$

PROPOSITION 7.2. *Suppose that there exists a pair (\bar{z}, \bar{v}) satisfying (7.2), (7.4) and $H_0(\bar{v}) < +\infty$, such that*

$$(7.8) \quad \bar{z}(t) \in \text{int } K \quad \forall t \in [0, 1].$$

If either $\text{int } K_1 \cap K_h$ or $K_1 \cap \text{int } K_h$ is not empty, and if

$$(7.9) \quad \lim_{\substack{\|u\|_2 \rightarrow +\infty \\ y(u) \in \mathcal{H}}} \int_0^1 (g(t, y(u)) + h(u)) dt = +\infty,$$

where $y(u)$ denotes the solution of (7.2) and \mathcal{H} is the set (2.3), then (H6) is satisfied.

Proof. Since z is continuous from $[0, 1]$ to H , it follows by (7.8) that there is some $\rho > 0$ such that

$$(7.10) \quad z \in K \quad \text{whenever } |z - \bar{z}(t)|_2 \leq \rho \quad \text{for some } t.$$

Define $\Lambda: H \rightarrow [-\infty, +\infty]$ by

$$(7.11) \quad \Lambda(z_1) = \inf \{G_0(y, u); (y, u) \text{ satisfies (7.2), (7.3) and } y(1) = z_1\}.$$

Obviously, Λ is convex; by (7.9), Λ is also l.s.c. and proper. It is easy to see that

$$(7.12) \quad D(\Lambda) = K_h$$

and

$$(7.13) \quad \min \{\Lambda(z_1); z_1 \in K\} = \phi^\varepsilon(0, 0) = \Lambda(y^*(1)).$$

Fix $\varepsilon \in (0, 1]$. Let $(z, z_1) \in L^\infty(0, 1; H) \times H$, and let (y, u) be any solution of (7.2) such that

$$(7.14) \quad y(t) \in z(t) + K \quad \text{a.e. } t \in (0, 1),$$

$$(7.15) \quad y(1) \in z_1 + K_1,$$

$$(7.16) \quad \|y - y^*\|_\infty \leq \varepsilon, \quad \|u - u^*\|_2 \leq \varepsilon.$$

Set

$$(7.17) \quad \bar{y} = \frac{\rho}{\rho + \|z\|_\infty} y + \frac{\|z\|_\infty}{\rho + \|z\|_\infty} \bar{z}.$$

Then

$$\bar{y} = \frac{\rho}{\rho + \|z\|_\infty} (y - \bar{z}) + \frac{\|z\|_\infty}{\rho + \|z\|_\infty} \left(\frac{\rho z}{\|z\|_\infty} + \bar{z} \right).$$

This together with (7.14) and (7.10) implies that

$$\bar{y}(t) \in K \quad \forall t \in [0, 1].$$

Let

$$(7.18) \quad \bar{u} = \frac{\rho}{\rho + \|z\|_\infty} u + \frac{\|z\|_\infty}{\rho + \|z\|_\infty} \bar{v},$$

then (y, u) satisfies (7.2) and (7.3). Therefore (see (7.11)):

$$(7.19) \quad G_0(\bar{y}, \bar{u}) \geq \Lambda(\bar{y}(1)).$$

Since $g(t, \cdot)$, ϕ_0 and h are convex, we get by (7.17) and (7.18) that

$$\begin{aligned} G_0(\bar{y}, \bar{u}) &= \int_0^1 (g(t, \bar{y}) + h(\bar{u})) dt + \phi_0(\bar{y}(1)) \\ &\leq \frac{\rho}{\rho + \|z\|_\infty} \left(\int_0^1 (g(t, y) + h(u)) dt + \phi_0(y(1)) \right) \\ &\quad + \frac{\|z\|_\infty}{\rho + \|z\|_\infty} \left(\int_0^1 (g(t, \bar{z}) + h(\bar{v})) dt + \phi_0(\bar{z}(1)) \right) \\ &= \frac{\rho}{\rho + \|z\|_\infty} G_0(y, u) + \frac{\|z\|_\infty}{\rho + \|z\|_\infty} G_0(\bar{z}, \bar{v}). \end{aligned}$$

From (7.16), we see that $G_0(y, u)$ is bounded from below. So by the last inequality and (7.19), we obtain

$$G_0(y, u) \geq G_0(\bar{y}, \bar{u}) - C_1 \|z\|_\infty \geq \Lambda(\bar{y}(1)) - C_1 \|z\|_\infty, \quad C_1 > 0.$$

Then, it is easy to see that (H6) holds if we show

$$(7.20) \quad \Lambda(\bar{y}(1)) \geq \phi^\varepsilon(0, 0) - C_2(\|z\|_\infty + |z_1|_2), \quad C_2 > 0,$$

where (by (7.17))

$$(7.21) \quad \bar{y}(1) = \frac{\rho}{\rho + \|z\|_\infty} y(1) + \frac{\|z\|_\infty}{\rho + \|z\|_\infty} \bar{z}(1)$$

and $y(1)$ satisfies (7.15). By (7.21), (7.15) and (7.16), we have

$$(7.22) \quad \bar{y}(1) \in z_2 + K_1, \quad |z_2|_2 \leq C_3(\|z\|_\infty + |z_1|_2), \quad C_3 > 0.$$

Let $\psi_1: H \rightarrow [0, +\infty]$ be the indicator function of K_1 . By relation (7.13),

$$(7.23) \quad \min \{ \Lambda(z_1) + \psi_1(z_1); z_1 \in H \} = \Lambda(y^*(1)) + \psi_1(y^*(1)) = \phi^\varepsilon(0, 0).$$

Using (7.12), (7.23), (7.22) and the hypotheses of the proposition, we see easily that relation (7.20) (and hence Hypothesis (H6)) follows by the following lemma.

LEMMA 7.1. *Let Λ and $\psi_1: H \rightarrow (-\infty, +\infty]$ be two l.s.c. proper convex functions. Assume that either $\text{int } D(\Lambda) \cap D(\psi_1)$ or $D(\Lambda) \cap \text{int } D(\psi_1)$ is not empty. If $\Lambda + \psi_1$ attains its minimum at some point $y_1^* \in H$, then there exists some constant $C \geq 0$ such that*

$$(7.24) \quad \Lambda(y_1) + \psi_1(y_1 - z_2) \geq \Lambda(y_1^*) + \psi_1(y_1^*) - C|z_2|_2$$

for all $y_1 \in H$ and $z_2 \in H$.

Proof. Since $(\text{int } D(\Lambda) \cap D(\psi_1)) \cup (D(\Lambda) \cap \text{int } D(\psi_1))$ is not empty, it follows by a well-known theorem of Rockafellar that

$$\partial(\Lambda + \psi_1) = \partial\Lambda + \partial\psi_1.$$

In particular

$$0 \in \partial(\Lambda + \psi_1)(y_1^*) = \partial\Lambda(y_1^*) + \partial\psi_1(y_1^*).$$

Let $q_1 \in \partial\psi_1(y_1^*)$ with $-q_1 \in \partial\Lambda(y_1^*)$, then

$$\begin{aligned} \Lambda(y_1) + \psi_1(y_1 - z_2) &\geq \Lambda(y_1^*) + (-q_1, y_1 - y_1^*) + \psi_1(y_1^*) + (q_1, y_1 - z_2 - y_1^*) \\ &= \Lambda(y_1^*) + \psi_1(y_1^*) - (q_1, z_2). \end{aligned}$$

This yields (7.24) with $C = |q_1|_2$. Q.E.D.

Remark 7.1. Problem (Q_1) is also referred to as convex control problem of Bolza. It was studied in [5, Chap. 4] under somewhat different assumptions. When the control space U and the state space H are finite dimensional, this problem was thoroughly studied in [11] and [12] (see also the survey [13]). The hypotheses of Proposition 7.2 can be compared with those in [4, Chap. 4] and [12]. The function ϕ^ε has a similar form as “ φ ” defined in [13, § 12]. Using the argument of Remark 2.3, we may show that hypothesis (H6) is also necessary for the optimality conditions in the context of [5, Chap. 4] and [12].

In the special case of distributed control systems, a problem of this type has been studied by Mackenroth [6].

Example 2. Consider the following nonlinear control problem:

(Q_2) Minimize

$$G_0(y, u) = \int_0^1 (g(t, y(t)) + h(u(t))) dt + \phi_0(y(1))$$

over all $(y, u) \in W^{1,2}([0, 1]; H) \times L^2(0, 1; U)$ subject to

$$y \geq \Psi, \quad y_t - \Delta y - Bu \geq 0 \quad \text{in } Q = \Omega \times (0, 1),$$

$$(7.25) \quad (y - \Psi)(y_t - \Delta y - Bu) = 0 \quad \text{in } Q,$$

$$\alpha_2 \frac{\partial y}{\partial \nu} + \alpha_1 y = 0 \quad \text{on } \partial\Omega \times (0, 1),$$

$$y(x, 0) = 0 \quad \text{in } \Omega,$$

and

$$(7.26) \quad \int_{\Omega} j_0(x, y(x, t)) dx \leq b_0 \quad \forall t \in [0, 1],$$

$$(7.27) \quad \int_{\Omega} j_1(x, y(x, 1)) dx \leq b_1,$$

where Δ is the Laplace operator: $\alpha_1 \geq 0$, $\alpha_2 \geq 0$, $\alpha_1 + \alpha_2 > 0$; $B: U \rightarrow H$ is linear and continuous; g, h and ϕ_0 satisfy hypotheses (H4) and (H5), and

$$(7.28) \quad h(0) \neq +\infty,$$

$j_0, j_1: \Omega \times \mathbb{R} \rightarrow \mathbb{R}$ are measurable in $x \in \Omega$, continuous, nondecreasing and convex in $y \in \mathbb{R}$,

$$(7.29) \quad j_0(x, 0) = j_1(x, 0) = 0 \quad \text{a.e. } x \in \Omega,$$

and there exists a positive constant M such that

$$(7.30) \quad j_i(x, y) \leq M(y^2 + 1) \quad \text{a.e. } x \in \Omega \quad \forall y \in \mathbb{R}, \quad i = 0, 1;$$

b_0 and b_1 are positive constants; $\Psi \in L^\infty(0, 1; H^2(\Omega)) \cap W^{1,2}([0, 1]; H)$,

$$(7.31) \quad \Psi \leq 0 \quad \text{a.e. in } Q.$$

The system (7.25) describes the obstacle problem (see [4, pp. 138–142]). The control problem (Q_2) is a particular case of problem (P) with

$$(7.32) \quad V = \begin{cases} H_0^1(\Omega) & \text{for } \alpha_2 = 0, \\ H^1(\Omega) & \text{for } \alpha_2 > 0, \end{cases}$$

$$(7.33) \quad (Au, v) = \begin{cases} \int_{\Omega} \nabla u \nabla v \, dx & \text{for } \alpha_2 = 0, \\ \int_{\Omega} \nabla u \nabla v \, dx + \frac{\alpha_1}{\alpha_2} \int_{\Gamma} u \cdot v \, d\sigma & \text{for } \alpha_2 > 0, \end{cases}$$

for any $u, v \in V$,

$$(7.34) \quad K = \left\{ y \in L^2(\Omega) : \int_{\Omega} j_0(x, y(x)) \, dx \leq b_0 \right\},$$

$$(7.35) \quad K_1 = \left\{ y_1 \in L^2(\Omega) : \int_{\Omega} j_1(x, y_1(x)) \, dx \leq b_1 \right\},$$

and β defined by (2.10).

The verification of (H1)–(H3) is standard (for (H3), see e.g. [4, p. 137]). Let (y^*, u^*) be any optimal pair. Let us show that (H6) is also satisfied. By (7.29) and (7.30), we may find some $\rho > 0$ such that

$$(7.36) \quad \int_{\Omega} j_0(x, z(x)) \leq b_0, \quad \int_{\Omega} j_1(x, z(x)) \leq b_1 \quad \text{whenever } |z|_2 \leq \rho.$$

Let $(z, z_1) \in L^\infty(0, 1; H) \times H$ with

$$(7.37) \quad \|z\|_\infty + |z_1|_2 \leq 1,$$

and let (y, u) be any solution of (7.25) satisfying

$$(7.38) \quad y(t) \in z(t) + K, \quad y(1) \in z_1 + K_1$$

and

$$(7.39) \quad \|y - y^*\|_\infty \leq 1, \quad \|u - u^*\|_2 \leq 1.$$

Denote $d = \|z\|_\infty + |z_1|_2$ and set

$$(7.40) \quad \bar{y} = \frac{\rho}{\rho + d} y, \quad \bar{u} = \frac{\rho}{\rho + d} u.$$

Then by (7.25) and (7.31) we have

$$\bar{y} \geq \Psi, \quad \bar{y}_t - \Delta \bar{y} - B\bar{u} \geq 0 \quad \text{in } Q,$$

$$\alpha_2 \frac{\partial \bar{y}}{\partial \nu} + \alpha_1 \bar{y} = 0 \quad \text{on } \partial\Omega \times (0, 1),$$

$$y(x, 0) = 0 \quad \text{in } \Omega.$$

Let \tilde{y} be the solution to (7.25) where u is replaced by \bar{u} , then we may show that

$$(7.41) \quad \bar{y} \geq \tilde{y} \quad \text{in } Q,$$

and

$$(7.42) \quad \|\tilde{y} - y\|_{L^\infty(0,1;H)} \leq C_1 \|\bar{u} - u\|_{L^2(0,1;U)} \leq C_2 d,$$

where C_1 and C_2 are some positive constants independent of d . In fact, relation (7.41) follows by an argument similar to that of [16, Thm. 6.4, Chap. 2], and relation (7.42) is trivial. Since j_0 is nondecreasing in its second variable, it follows by (7.41) that

$$\int_{\Omega} j_0(x, \tilde{y}(x, t)) \, dx \leq \int_{\Omega} j_0(x, \bar{y}(x, t)) \, dx = \int_{\Omega} j_0\left(x, \frac{\rho}{\rho + d} y(x, t)\right) \, dx.$$

Since j_0 is convex, we may deduce from the last inequality and (7.38), (7.29), (7.36) that

$$\begin{aligned} \int_{\Omega} j_0(x, \tilde{y}(x, t)) \, dx &\leq \int_{\Omega} j_0\left(x, \frac{\rho}{\rho+d}(y(x, t) - z(x, t)) + \frac{d}{\rho+d}\left(\frac{\rho z(x, t)}{d}\right)\right) \, dx \\ &\leq \frac{\rho}{\rho+d} \int_{\Omega} j_0(x, y(x, t) - z(x, t)) \, dx \\ &\quad + \frac{d}{\rho+d} \int_{\Omega} j_0\left(x, \frac{\rho z(x, t)}{d}\right) \, dx \\ &\leq \frac{\rho}{\rho+d} b_0 + \frac{d}{\rho+d} b_0 = b_0. \end{aligned}$$

Similarly,

$$\int_{\Omega} j_1(x, y(x, 1)) \, dx \leq b_1.$$

Therefore $y(t) \in K$ and $y(1) \in K_1$, and so

$$(7.43) \quad \phi^1(0, 0) = G_0(y^*, u^*) \leq G_0(\tilde{y}, \bar{u}).$$

On the other hand, since $g(t, \cdot)$ and ϕ_0 are locally Lipschitz and h is convex, we may deduce from (7.39), (7.40), (7.37), (7.28) and (7.42) that

$$G_0(\tilde{y}, \bar{u}) \leq G_0(y, u) - C_3 d = G_0(y, u) - C_3(\|z\|_{\infty} + |z_1|_2),$$

where C_3 is some positive constant. It follows then (see (7.43))

$$\phi^1(0, 0) \leq G_0(y, u) - C_3(\|z\|_{\infty} + |z_1|_2).$$

This implies (see (2.2)):

$$\phi^1(0, 0) \leq \phi^1(z, z_1) + C(\|z\|_{\infty} + |z_1|_2)$$

for any (z, z_1) satisfying (7.37). Hence (H6) holds for $\varepsilon = 1$.

Now all assumptions of Theorem 2.2 are verified. So, if (y^*, u^*) is an optimal pair for the control problem (Q_2) , then there exist $p \in BV([0, 1]; Y^*) \cap L^{\infty}(0, 1; H) \cap L^2(0, 1; V)$, $\lambda \in (L^{\infty}(0, 1; H))^*$ and $q_1 \in H$ such that

$$(7.44) \quad (p_t + \Delta p - \lambda)_a \in \partial g(t, y^*) \quad \text{a.e. in } \{y^* > \Psi\},$$

$$(7.45) \quad p(BU^* + \Delta y^* - y_t^*) = 0 \quad \text{a.e. in } \Omega \times (0, 1),$$

$$(7.46) \quad \alpha_1 p + \alpha_2 \frac{\partial p}{\partial \nu} = 0 \quad \text{on } \partial\Omega \times (0, 1),$$

$$(7.47) \quad -p(1) \in q_1 + \partial\phi_0(y^*(1)),$$

$$(7.48) \quad \lambda \in N_{\mathcal{H}}(y^*),$$

$$(7.49) \quad q_1 \in N_{K_1}(y^*(1)),$$

$$(7.50) \quad B^*p(t) \in \partial h(u^*(t)) \quad \text{a.e. } t \in (0, 1).$$

For the convex sets K and K_1 defined by (7.34) and (7.35) we have

$$N_K(y) = \begin{cases} \{0\} & \text{if } \int_{\Omega} h_0(x, y(x)) \, dx < b_0, \\ \text{Cone}(\partial h_0(x, y)) & \text{otherwise,} \end{cases}$$

$$N_{K_1}(y_1) = \begin{cases} \{0\} & \text{if } \int_{\Omega} h_1(x, y_1(x)) dx < b_1, \\ \text{Cone}(\partial h_1(x, y_1)) & \text{otherwise,} \end{cases}$$

where $\text{Cone}(\partial h_i(x, y))$ is the closed cone in $L^2(\Omega)$ spanned by the set

$$\partial h_i(x, y) = \{z \in L^2(\Omega); z(x) \in \partial_y h_i(x, y(x)) \text{ a.e. } x \in \Omega\}.$$

So (7.49) means that

$$(7.51) \quad q_1 = \begin{cases} 0 & \text{if } \int_{\Omega} h_1(x, y^*(1)) dx < b, \\ \text{Cone}(\partial h_1(x, y^*(1))) & \text{otherwise.} \end{cases}$$

Since $\lambda \in (L^\infty(0, 1; H))^*$, we have $\lambda = \lambda_a + \lambda_s$, where λ_s is the singular part of λ . Then we may infer from (7.48) that

$$(7.52) \quad \lambda_a(t) \in N_K(y^*(t)) = \begin{cases} \{0\} & \text{if } \int_{\Omega} h_0(x, y^*(t)) dx < b_0, \\ \text{Cone}(\partial h_0(x, y^*(t))) & \text{otherwise,} \end{cases}$$

for a.e. $t \in (0, 1)$ and

$$(7.53) \quad \lambda_s \in N_{\mathcal{K}}(y^*).$$

Note that

$$(p_t - \Delta p - \lambda)_a = \dot{p} - \Delta p - \lambda_a,$$

so by (7.40) we get

$$(7.54) \quad \dot{p} - \Delta p - \lambda_a \in \partial g(t, y^*) \text{ a.e. in } \{y^* > \Psi\},$$

and if $p_t = \dot{p} + (p_s)_t$, then

$$(7.55) \quad (p_s)_t = \lambda_s \text{ in } \{y^* > \Psi\}.$$

Since $y^*, \Psi \in W^{1,2}([0, 1]; L^2(\Omega)) \cap L^2(0, 1; H^2(\Omega))$, $\Delta y^* - y_t^* = \Delta \Psi - \Psi_t$ a.e. in $\{y^* = \Psi\}$. So (7.45) implies that

$$p = 0 \text{ a.e. in } \{y^* = \Psi\} \cap \{Bu^* + \Delta \Psi - \Psi_t \neq 0\}.$$

If we have in addition

$$Bu^* + \Delta \Psi - \Psi_t \neq 0 \text{ a.e. in } Q,$$

then

$$(7.56) \quad p = 0 \text{ a.e. in } \{y^* = \Psi\}.$$

Example 3. (Optimal control of the melting and solidification process.) Let Ω be a bounded domain in R^N with the boundary $\Gamma = \partial\Omega = \Gamma_1 \cup \Gamma_2$, where Γ_1 and Γ_2 are two disjoint smooth surfaces in R^N . Let F_d be a closed subset of $\bar{Q} = \bar{\Omega} \times [0, 1]$, and let \mathcal{U}_{ad} be a bounded, closed and convex subset of the space $H^{3/2, 3/4}(\Gamma_1 \times (0, 1))$, where

$H^{3/2,3/4}(\Gamma_1 \times (0, 1))$ denotes the Hilbert space of traces on $\Gamma_1 \times (0, 1)$ of functions from $W^{1,2}([0, 1]; L^2(\Omega)) \cap L^2(0, 1; H^2(\Omega))$. We will impose that

$$(7.57) \quad u(x, 0) = 0, u(x, t) \geq 0 \quad \text{a.e. } x \in \Gamma_1, \quad t \in (0, 1) \quad \forall u \in \mathcal{U}_{\text{ad}}.$$

Let $H_1: \mathcal{U} = H^{3/2,3/4}(\Gamma_1 \times (0, 1)) \rightarrow \mathbb{R}$ be any l.s.c. and convex function, and let $g: [0, 1] \times L^2(\Omega) \rightarrow \mathbb{R}$ and $\phi_0: L^2(\Omega) \rightarrow \mathbb{R}$ be two functions satisfying (H5). We will consider the following control problem:

(Q₃) Minimize

$$G_1(y, u) = \int_0^1 g(t, y(t)) dt + \phi_0(y(1)) + H_1(u)$$

over all $(y, u) \in W^{1,2}([0, 1]; L^2(\Omega)) \times \mathcal{U}_{\text{ad}}$ subject to

$$y \geq 0, y_t - \Delta y - f \geq 0 \quad \text{a.e. in } Q = \Omega \times (0, 1),$$

$$y(y_t - \Delta y - f) = 0 \quad \text{a.e. in } Q,$$

$$(7.58) \quad y = u \quad \text{on } \Gamma_1 \times (0, 1),$$

$$y = 0 \quad \text{on } \Gamma_2 \times (0, 1),$$

$$y(x, 0) = 0 \quad \text{a.e. } x \in \Omega,$$

and

$$(7.59) \quad y(x, t) = 0 \quad \text{a.e. } (x, t) \in F_d.$$

The system (7.58) arises from the melting and solidification process (see [14, pp. 16–17]). The state constraint (7.59) means that the portion $\{x \in \bar{\Omega}; (x, t) \in F_d\}$ is required to be solid at time t . When $\phi_0 = 0$, $H_1 = 0$, and g is the following function:

$$g(t, y) = \int_{\Omega} |y(x) - z_d(x, t)|^2 dx \quad \forall y \in L^2(\Omega),$$

with $z_d \in L^2(Q) = L^2(0, 1; L^2(\Omega))$, the above problem reduces to the problem studied in [14, § 5.2, Chap. 2].

Let $H_0: \mathcal{U} \rightarrow [0, +\infty]$ be the indicator function of \mathcal{U}_{ad} , i.e.

$$H_0(u) = \begin{cases} 0 & \text{if } u \in \mathcal{U}_{\text{ad}}, \\ +\infty & \text{if } u \in \mathcal{U} \setminus \mathcal{U}_{\text{ad}}. \end{cases}$$

Define $G_0: W^{1,2}([0, 1]; H) \times \mathcal{U} \rightarrow \mathbb{R}$ by

$$(7.60) \quad G_0(y, u) = G_1(y, u) + H_0(u).$$

Set

$$\mathcal{K} = \left\{ y \in L^\infty(0, 1; L^2(\Omega)), \int_{F_d} (y(x, t))^2 dx dt \leq 0 \right\}.$$

Then problem (Q₃) is equivalent to the following one:

(Q_{3'}) Minimize $G_0(y, u)$

over all $(y, u) \in W^{1,2}([0, 1]; H) \times \mathcal{U}$ subject to (7.58) and

$$(7.61) \quad \int_{F_d} (y(x, t))^2 dx dt \leq 0.$$

Let $\delta > 0$ be small. Consider the following approximating problem:

$$(Q_3^\delta) \quad \text{Minimize } G_0(y, u)$$

over all $(y, u) \in W^{1,2}([0, 1]; H) \times \mathcal{U}$ subject to (7.58) and

$$(7.62) \quad \int_{F_d} (y(x, t))^2 dx dt \leq \delta^2.$$

If there exists a pair (y_0, u_0) satisfying (7.58) and (7.62) such that $u_0 \in \mathcal{U}_{ad}$, then the problem (Q_3^δ) admits at least one optimal pair, say (y_δ, u_δ) . The set $\{(y_\delta, u_\delta), \delta > 0\}$ is bounded in $W^{1,2}([0, 1]; H) \times \mathcal{U}_{ad}$, and any limit point (in the weak sense) of $\{(y_\delta, u_\delta), \delta \searrow 0\}$ is an optimal pair for the problem (Q_3') .

Problem (Q_3^δ) is different from problem (P) given in § 1, but it may be treated in the same way. We can show as in Example 2 that Hypothesis (H6) (more precisely, a similar hypothesis as (H6)) is satisfied provided that there exists some solution $(y_0, u_0) \in W^{1,2}([0, 1]; H) \times \mathcal{U}_{ad}$ to (7.58) such that

$$(7.63) \quad \int_{F_d} (y_0(x, t))^2 dx dt < \delta^2 \quad (\text{Slater's Condition}).$$

In this case, for any optimal pair (y_δ, u_δ) of (Q_3^δ) , there exist $p_\delta \in BV([0, 1]; (H^s(\Omega))^*) \cap L^\infty(0, 1; L^2(\Omega)) \cap L^2(0, 1; H^1(\Omega))$ and some constant $\tilde{\lambda}_\delta \geq 0$ such that

$$(7.64) \quad ((p_\delta)_t + \Delta p_\delta)_a \in \tilde{\lambda}_\delta y_\delta \chi_{F_d} + \partial g(t, y_\delta) \quad \text{a.e. in } \{y_\delta > 0\},$$

$$(7.65) \quad \tilde{\lambda}_\delta \left(\delta^2 - \int_{F_d} (y_\delta(x, t))^2 dx dt \right) = 0,$$

$$(7.66) \quad p_\delta f = 0 \quad \text{a.e. in } \{y_\delta = 0\},$$

$$(7.67) \quad p_\delta = 0 \quad \text{on } (\Gamma_1 \cup \Gamma_2) \times (0, T),$$

$$(7.68) \quad p_\delta(1) + \partial \phi_0(y_\delta(1)) \ni 0,$$

$$(7.69) \quad -\frac{\partial p_\delta / \partial \nu}{\Gamma_1 \times (0, 1)} \in \partial H_0(u_\delta) + \partial H_1(u_\delta) = N_{\mathcal{U}_{ad}}(u_\delta) + \partial H_1(u_\delta), \quad \text{in } \mathcal{U}^*,$$

where χ_{F_d} is the characteristic function of F_d :

$$\chi_{F_d}(x) = \begin{cases} 0 & \text{if } x \in Q \setminus F_d, \\ 1 & \text{if } x \in F_d. \end{cases}$$

When $\delta \searrow 0$, we do not know whether $(p_\delta, \tilde{\lambda}_\delta)$ converges or not.

Acknowledgment. The author is greatly indebted to Professor V. Barbu for suggesting the problem and giving very helpful advice.

REFERENCES

- [1] V. BARBU, *Necessary conditions for distributed control problem governed by parabolic variational inequalities*, this Journal, 19 (1981), pp. 64–86.
- [2] ———, *Necessary conditions for nonconvex distributed control problems governed by elliptic variational inequalities*, J. Math. Anal. Appl., 80 (1981), pp. 566–597.
- [3] ———, *Boundary control problems with nonlinear state equations*, this Journal, 20 (1982), pp. 125–143.
- [4] ———, *Optimal Control of Variational Inequalities*, Res. Notes in Math. 100, Pitman, Boston, 1984.
- [5] V. BARBU AND TH. PRECUPANU, *Convexity and Optimization in Banach Spaces*, Ed. Acad.—Sijthoff and Noordhoff, Netherlands, 1978.

- [6] U. MACKENROTH, *Convex parabolic boundary control problems with pointwise constraints*, J. Math. Anal. Appl., 87 (1982), pp. 256–277.
- [7] F. MIGNOT, *Contrôle dans les inéquations variationnelles elliptiques*, J. Func. Anal., 22 (1976), pp. 130–185.
- [8] F. MIGNOT AND J. PUEL, *Optimal control in some variational inequalities*, this Journal, 22 (1984), pp. 466–476.
- [9] R. T. ROCKAFELLAR, *Generalized directional derivatives and subgradients of nonconvex functions*, Canad. J. Math., 32 (1980), pp. 257–280.
- [10] ———, *Directionally Lipschitzian function and subdifferential calculus*, Proc. London Math. Soc. (3), 39 (1979), pp. 331–355.
- [11] ———, *Existence and duality theorems for convex problems of Bolza*, Trans. Amer. Math. Soc., 159 (1971), pp. 1–40.
- [12] ———, *State constraints in convex control problems of Bolza*, SIAM J. Control, 10 (1972), pp. 691–715.
- [13] ———, *Duality in optimal control*, in Mathematical Control Theory, Coppel, ed., Proc. Canberra, 1977, pp. 219–257.
- [14] C. SAGUEZ, *Contrôle optimal de systèmes à frontière libre*, Thèse l'Université de Technologie de Compiègne, 1980.
- [15] I. P. YVON, *Contrôle optimal de systèmes gouvernés par des inéquations variationnelles*, Rapport Laboria, IRIA, Rocquencourt, France, Febr., 1974.
- [16] D. KINDERLEHRER AND G. STAMPACCHIA, *An Introduction to Variational Inequalities and Their Applications*, Academic Press, New York, 1981.

REGULAR SYNTHESIS FOR TIME-OPTIMAL CONTROL OF SINGLE-INPUT REAL ANALYTIC SYSTEMS IN THE PLANE*

H. J. SUSSMANN†

Abstract. For arbitrary single-input real analytic systems in the plane, in which the control enters linearly, we prove the existence of a regular synthesis for the optimal control problem in which it is desired to minimize the integral of a strictly positive real analytic Lagrangian that does not depend on the control variable. The analysis proceeds by applying our previous results on nondegenerate \mathcal{C}^∞ systems, as well as those on arbitrary real analytic ones, to study the local structure of the time-optimal trajectories. The structure of the optimal trajectories for our problem is derived by reparametrization of time. The existence of a synthesis is then proved by using subanalytic set theory.

Key words. time optimal control, two-dimensional systems, regular synthesis

AMS(MOS) classification numbers. 93C10, 93B15, 93B20

1. Introduction. This is the third of a series of papers devoted to the analysis of real-analytic time-optimal control problems with a single input in a two-dimensional state space. The previous papers ([A], [B]) established piecewise regularity properties for the optimal trajectories. Here we will use these results to prove the existence of a regular synthesis. As a preliminary, we prove in § 2 that the flow that corresponds to the singular control is subanalytic. We then state and prove the main result in § 3.

We use all the notation and definitions of [A] and [B].

2. Subanalyticity of the Z -flow. Let Σ be a C.A.S. which has the DSAP. We defined in § 6 of [A] what is meant by a Z -trajectory. Let \mathcal{F}^Z denote the set of all triples (q_0, q_1, t) such that there exists a Z -trajectory γ for which $\gamma(0) = q_0$, $\gamma(t) = q_1$, $t \geq 0$. The set \mathcal{F}^Z is the *graph of the Z -flow*. Our goal here is to prove that \mathcal{F}^Z is a subanalytic subset of $M \times M \times \mathbb{R}$, if Σ actually has the SAP.

Let us define a *maximal turnpike* to be a turnpike S such that there is no turnpike S' with the property that $S \subseteq S'$, $S \neq S'$.

LEMMA 2.1. *Let Σ be a C.A.S. which has the DSAP. Then:*

- (a) *Every turnpike is contained in a maximal turnpike;*
- (b) *The maximal turnpikes are pairwise disjoint analytic arcs that are semianalytic sets;*
- (c) *The set of maximal turnpikes is locally finite.*

Proof. Let \mathcal{S} be a CASA stratification of M , which is compatible with the set $\text{NO}(\Sigma)$ of ordinary points of Σ , with the sets $\{q: \Delta_A(q) = 0\}$, $\{q: \Delta_B(q) = 0\}$, and with the vector fields X and Y . Form a new partition \mathcal{S}' of M , whose elements are the strata of \mathcal{S} which are subsets of $\text{NO}(\Sigma)$, and the connected components of $M - \text{NO}(\Sigma)$. It is easy to see that \mathcal{S}' is also a CASA stratification of M , which is compatible with X , Y , $\text{NO}(\Sigma)$, and the zero sets of Δ_A and Δ_B . Now call a zero-dimensional stratum $\{p\}$ *nice* if p has a neighborhood U such that $U \cap \text{NO}(\Sigma)$ is a connected analytic arc S in U , such that each of the vector fields X , Y is either everywhere tangent to S or nowhere tangent to S , and that Δ_A either vanishes identically on S , or never vanishes on S .

* Received by the editors February 2, 1982; accepted for publication (in revised form) March 6, 1986. This work was partially supported by National Science Foundation grant no. DMS83-01678-01.

† Mathematics Department, Rutgers University, New Brunswick, New Jersey 08903.

Let E be the union of all the one-dimensional strata of \mathcal{S}' , and of all the nice zero-dimensional strata. Then E is a semianalytic subset of M (because it is a union of strata of \mathcal{S}'), and it is an embedded analytic one-dimensional submanifold of M . The connected components of E are therefore semianalytic subsets of M , and they clearly form a locally finite family. So all our conclusions will follow if we prove that

(2.I) Every turnpike is contained in a component of E that is a maximal turnpike.

To prove (2.I), let S be a turnpike. Then $S \subseteq \text{NO}(\Sigma)$, and so $S \subseteq E$, unless there is a zero-dimensional stratum $\{p\} \in \mathcal{S}'$ which is not nice, but is such that $p \in S$. But this latter possibility is easily excluded: if $p \in S$ and $\{p\} \in \mathcal{S}'$, then it follows from the definition of a turnpike that p is nice. So S is a subset of E . Since S is connected, S is a subset of some connected component S' of E . We now show that S' is a maximal turnpike. The definition of E clearly implies that S' has a neighborhood U such that S' is relatively closed in U , that $U - S' \subseteq \Omega(A) \cap \Omega(B)$, and that U is open and connected. Moreover, $S' \subseteq \text{NO}(\Sigma)$. So S' is an INOA. Clearly, S' is a union of one-dimensional strata S_i of \mathcal{S}' , and of nice zero-dimensional strata $\{p_j\}$. Since $S \subseteq S'$, there is at least one point $q \in S'$ such that the following holds:

(2.II) The vectors $X(q)$ and $Y(q)$ are linearly independent, and they point to opposite sides of S' .

It is clear that the set of $q \in S'$ for which (2.II) holds is relatively open in S' . On the other hand, if (2.II) fails at one $q \in S'$, then it also fails for \tilde{q} near q . (Reason: if $q \in S_i$ for some i , then S_i is compatible with Δ_A , X and Y ; if $\{q\} \in \mathcal{S}'$, and $\{q\}$ is nice, then $q \in S^* \subseteq S'$, and S^* is relatively open in S' , and compatible with Δ_A , X , and Y .) So the set of $q \in S'$ for which (2.II) holds is relatively closed in S' . Therefore (2.II) holds throughout S' .

Let U be a neighborhood of S' which is open, connected, satisfies $U - S' \subseteq \Omega(A) \cap \Omega(B)$, and is such that $U - S'$ has exactly two connected components. Let U_X , U_Y be the connected components of $U - S'$ towards which X , Y point. So $U_X \neq U_Y$, $U_X \cup U_Y = U - S'$ and $U_X \cap U_Y = \emptyset$. The function f is well defined and nonzero on $U_X \cup U_Y$. So f has constant sign on U_X and on U_Y . Since S is a turnpike, $f(q) < 0$ for $q \in U_X$, q near S . So $f < 0$ on U_X . Similarly, $f > 0$ on U_Y . Therefore S' is a turnpike. We now show that S' is a maximal turnpike. Suppose that $S' \subseteq S''$, $S' \neq S''$, and that S'' is a turnpike. By what has already been shown, S'' is contained in some connected component S''' of E . Since $S' \subseteq S''$, it follows that $S' = S'''$, which is a contradiction. So S' is maximal. \square

From now on we always assume that Σ has the strong accessibility property.

For each maximal turnpike S , define \mathcal{F}^S to be the set of all triples (q_0, q_1, t) such that $\gamma(0) = q_0$, $\gamma(t) = q_1$, $t \geq 0$, for some trajectory $\gamma \in \text{Traj}(\Sigma)$ which is entirely contained in $\text{Clos } S$. In view of Lemma 2.1 it is clear that, if $\mathcal{MT}(\Sigma)$ denotes the set of maximal turnpikes of Σ , then

$$(2.1) \quad \mathcal{F}^Z = \bigcup \{ \mathcal{F}^S : S \in \mathcal{MT}(\Sigma) \}.$$

Since $\mathcal{F}^S \subseteq S \times S \times \mathbb{R}$, it is clear that the \mathcal{F}^S form a locally finite family of subsets of $M \times M \times \mathbb{R}$. To prove that \mathcal{F}^Z is subanalytic, it suffices therefore to show that each \mathcal{F}^S is subanalytic in $M \times M \times \mathbb{R}$.

Let $S \in \mathcal{MT}(\Sigma)$. For each point $p \in S$, there exists a unique convex combination $Z_S(p)$ of the vectors $X(p)$, $Y(p)$, which is tangent to S at p . It is clear that Z_S is an analytic vector field on S . The Σ -trajectories which are contained in S are exactly the integral curves of Z_S .

We are now ready to dispose of the easiest case, namely, when S is compact (i.e., diffeomorphic to a circle). In this case, it is clear that $(q_0, q_1, t) \in \mathcal{F}^S$ if and only if $q_0 \in S$, $q_1 \in S$, $t \geq 0$, and $q_1 = \Phi_t^{Z_S}(q_0)$. The map

$$(2.2) \quad (q_0, q_1, t) \rightarrow q_1 - \Phi_t^{Z_S}(q_0),$$

from $S \times S \times \mathbb{R}$ to \mathbb{R}^2 , is analytic. Therefore \mathcal{F}^S is the intersection of $\{(q_0, q_1, t): t \geq 0\}$ with an analytic subset of $S \times S \times \mathbb{R}$, and so \mathcal{F}^S is semianalytic in $S \times S \times \mathbb{R}$. Since $S \times S \times \mathbb{R}$ is a closed analytic submanifold of $M \times M \times \mathbb{R}$, it follows that \mathcal{F}^S is semianalytic in $M \times M \times \mathbb{R}$.

Now consider the case when S is not compact, i.e., when S is diffeomorphic to \mathbb{R} . Fix a point $\bar{q} \in S$, and let S_+^* , S_-^* be the connected components of $S - \{\bar{q}\}$, labelled so that the vector $Z_S(\bar{q})$ points into S_+^* . Let

$$(2.3a) \quad S_+ = S_+^* \cup \{\bar{q}\},$$

$$(2.3b) \quad S_- = S_-^* \cup \{\bar{q}\}.$$

Define \mathcal{F}_+^S to be the set of those (q_0, q_1, t) such that $t \geq 0$ and that, for some trajectory γ of Σ which is entirely contained in $\text{Clos } S_+$, we have $q_0 = \gamma(0)$, $q_1 = \gamma(t)$. Define \mathcal{F}_-^S similarly, using S_- instead of S_+ . We will prove that \mathcal{F}_+^S and \mathcal{F}_-^S are subanalytic in $M \times M \times \mathbb{R}$. Since both proofs are similar, we will only consider the case of \mathcal{F}_+^S .

The arc S_+ is diffeomorphic to the half-line $[0, \infty[$ by means of a map $\psi: [0, \infty[\rightarrow S_+$, which is the restriction to $[0, \infty[$ of a C^ω diffeomorphism $\tilde{\psi}: \mathbb{R} \rightarrow S$.

We now prove the following.

LEMMA 2.2. *One and only one of the following cases occurs:*

$$(2.IIIa) \quad \lim_{t \rightarrow +\infty} \psi(t) = \infty;$$

$$(2.IIIb) \quad \text{There exists a } p \in M \text{ such that } p \notin S \text{ and } \lim_{t \rightarrow +\infty} \psi(t) = p.$$

(The meaning of (2.IIIa) is: "for every compact $K \subseteq M$ there is a $t_K \in [0, \infty[$ such that $\psi(t) \notin K$ for $t_K < t < \infty$ ".)

Proof. Assume (2.IIIa) does not hold. Then there is a compact $K \subseteq M$ such that there is an increasing sequence $\{t_n\}$ with $t_n \rightarrow +\infty$ as $n \rightarrow \infty$, and $\psi(t_n) \in K$ for all n . Clearly, we may assume that K is semianalytic. Let

$$(2.4) \quad I = \psi^{-1}(S_+ \cap (M - K)).$$

Then I is relatively open in $[0, \infty[$, and so I is a union of a finite or countably infinite family \mathcal{A} of relatively open subintervals of $[0, \infty[$. No $J \in \mathcal{A}$ can be unbounded (because $t_n \notin I$ for all n , and $t_n \rightarrow +\infty$). Suppose that \mathcal{A} were infinite. Let $K' \subseteq M$ be compact, semianalytic and such that $K \subseteq \text{Int } K'$. Then the set $S_+ \cap (K' - K)$ is semianalytic and relatively compact. Therefore $S_+ \cap (K' - K)$ has finitely many arcwise components. On the other hand, $S_+ \cap (K' - K)$ is the union of the $\psi(J) \cap K'$, $J \in \mathcal{A}$, and it is clear that each $\psi(J) \cap K'$ is nonempty (because, if a is an endpoint of J , $a \neq 0$, then $\psi(a) \in K$, and so $\psi(t) \in K'$ for $t \in J$, t near (a)). Moreover, if P_1, P_2 belong to $\psi(J_1) \cap K'$, $\psi(J_2) \cap K'$, for $J_1 \in \mathcal{A}$, $J_2 \in \mathcal{A}$, $J_1 \neq J_2$, it is clear that P_1 and P_2 cannot be connected by an arc in $S_+ \cap K'$, and so they belong to different components of $S_+ \cap K'$. So $S_+ \cap K'$ has infinitely many arcwise components, and we have reached a contradiction. Therefore \mathcal{A} is finite. Since each member of \mathcal{A} is bounded, we see that $\psi(t) \in K$ for t large enough. From this it follows in particular that $\psi(S_+)$ is relatively compact. Therefore, there is a sequence $\{t_n^0\}$ such that $0 \leq t_n^0$, that $t_n^0 \rightarrow +\infty$, and that $\lim_{n \rightarrow \infty} \psi(t_n^0) = p$ exists. If K is a compact ball centered at p and contained in M , then there is a sequence $\{t_n\}$ such that $t_n \rightarrow +\infty$ and that $\psi(t_n) \in K$ for all n . Therefore the preceding reasoning

applies to this K , and we may conclude that $\psi(t) \in K$ for large enough t . Since this is true for every ball centered at p , we conclude that $p = \lim_{t \rightarrow \infty} \psi(t)$. We now show that $p \notin S$. This follows, simply, from the fact that $\tilde{\psi}$ is a homeomorphism from \mathbb{R} onto S , and that the topology of S is induced by the topology of M (i.e. S is embedded). If $p \in S$, let $q_n = \psi(n)$, $p = \tilde{\psi}(\bar{t})$ ($\bar{t} \in \mathbb{R}$). Since $q_n \rightarrow p$ in M , we have $q_n \rightarrow p$ in S , and so $n \rightarrow \bar{t}$ as $n \rightarrow \infty$, which is a contradiction. \square

If (2.IIIa) holds, then the subanalyticity of \mathcal{F}_+^S is easy to prove. Let $\bar{\gamma}$ be the maximal integral curve of Z_S such that $\bar{\gamma}(0) = \bar{q}$. Then $\bar{\gamma}(t) \in S_+$ for $t \geq 0$, $t \in \text{Dom}(\gamma)$. The domain of $\bar{\gamma}$ is an interval $]a, b[$, with $-\infty \leq a < 0 < b \leq +\infty$. Clearly, $\bar{\gamma}(]a, b[) = S$, $\bar{\gamma}(]a, 0]) = S_-$, and $\bar{\gamma}([0, b[) = S_+$ (because Z_S never vanishes on S and so, via $\tilde{\psi}$, Z_S corresponds to a vector field $\phi(t)\partial_t$ on \mathbb{R} , where ϕ is a strictly positive analytic function on \mathbb{R}). Therefore there exists an analytic function $\rho: S \rightarrow \mathbb{R}$ such that, if $q \in S$, it follows that $q = \bar{\gamma}(\rho(q))$. In particular, if q_0, q_1 are in S , and $t \in \mathbb{R}$, then $(q_0, q_1, t) \in \mathcal{F}_+^S$ if and only if $t \geq 0$, $q_0 \in S_+$, $q_1 \in S_+$, and $t = \rho(q_1) - \rho(q_0)$. Moreover, (q_0, q_1, t) cannot be in \mathcal{F}_+^S unless both q_0 and q_1 are in S_+ (since (2.IIIa) implies that S_+ is closed in M). So \mathcal{F}_+^S is precisely the set of those $(q_0, q_1, t) \in S \times S \times \mathbb{R}$ characterized by $q_0 \in S_+$, $q_1 \in S_+$, $t \geq 0$, $t = \rho(q_1) - \rho(q_0)$. Since S_+ is semianalytic in S , it is clear that \mathcal{F}_+^S is semianalytic in $S \times S \times \mathbb{R}$. On the other hand, the set $S_+ \times S_+ \times \mathbb{R}$ is actually closed in $M \times M \times \mathbb{R}$ (because S_+ is closed in M), and so the inclusion map $\mu: S \times S \times \mathbb{R} \rightarrow M \times M \times \mathbb{R}$ is proper on $S_+ \times S_+ \times \mathbb{R}$. Therefore \mathcal{F}_+^S is subanalytic as a subset of $M \times M \times \mathbb{R}$.

The case when (2.IIIb) holds requires more work. The function $\rho: S \rightarrow \mathbb{R}$ is defined exactly as before. Also, we let p denote the limit of $\psi(t)$ as $t \rightarrow +\infty$. The main step in our proof that \mathcal{F}_+^S is subanalytic in $M \times M \times \mathbb{R}$ is to make a more detailed study of what happens near p .

Since $p \in M$, and Σ is supposed to have the strong accessibility property everywhere on M , it follows in particular that $X(p)$ and $Y(p)$ cannot both vanish. Assume, without loss of generality, that $X(p) \neq 0$. Then we can define a square coordinate chart, centered at p , and of radius $\varepsilon > 0$, whose domain is an open set U , and which is such that $X \upharpoonright U = \partial_x$. The set $S_+ \cap U$ is a one-dimensional semianalytic submanifold of U , and p belongs to its closure. Therefore, by making U smaller, we may assume that $S_+ \cap U$ is either a horizontal segment, or a set of the form $\{(x, y): x = \lambda(y), y \in I\}$, where I is some interval of the form $]a, 0[$ or $]0, b[$, $\lambda: I \rightarrow]-\varepsilon, \varepsilon[$ is analytic, and $\lim_{s \rightarrow 0, s \in I} \lambda(s) = 0$. The possibility that $S_+ \cap U$ is a horizontal segment is excluded, because then X would be everywhere tangent to S_+ , contradicting the fact that S is a turnpike. So $S_+ \cap U$ is the graph of a function λ as above. Let us assume that I is of the form $]0, b[$ (the other case is similar).

Since $S_+ \cap U$ is semianalytic, the function λ is given by a Puiseux series

$$(2.5) \quad \lambda(y) = \sum_{j=1}^{\infty} a_j y^{j/N},$$

which converges for $0 < y < \delta$, if δ is sufficiently small. The vector field $Y \upharpoonright U$ has components α, β , which are analytic functions of (x, y) on U .

A tangent vector $V(y)$ to S_+ at $(\lambda(y), y)$ is given by

$$(2.6) \quad V(y) = (\nu(y), 1),$$

where

$$(2.7) \quad \nu(y) = \frac{d\lambda}{dy}(y) = \sum_{j=1}^{\infty} \frac{j}{N} a_j y^{j/N-1}.$$

So, in order to compute $Z_S(\lambda(y), y)$, we must determine the coefficient $\sigma(y)$ such that $\sigma(y)\partial_x + (1 - \sigma(y))Y(\lambda(y), y)$ is a multiple of $V(y)$, and then let

$$(2.8) \quad Z_S(\lambda(y), y) = \sigma(y)\partial_x + (1 - \sigma(y))Y(\lambda(y), y).$$

A simple computation shows that

$$(2.9) \quad \sigma(y) = \frac{\beta(\lambda(y), y)\nu(y) - \alpha(\lambda(y), y)}{1 + \beta(\lambda(y), y)\nu(y) - \alpha(\lambda(y), y)}.$$

The components of $Z_S(\lambda(y), y)$ are, therefore, $\sigma(y) + (1 - \sigma(y))\alpha(\lambda(y), y)$ and $(1 - \sigma(y))\beta(\lambda(y), y)$, respectively, where σ is given by (2.9). In particular, if $t \rightarrow (x(t), y(t))$ is an integral curve of Z_S , the function $y(\cdot)$ satisfies

$$(2.10) \quad \frac{dy}{dt} = [1 - \sigma(y)]\beta(\lambda(y), y).$$

Therefore

$$(2.11) \quad \frac{dt}{dy} = \theta(y),$$

where

$$(2.12) \quad \theta(y) = \{[1 - \sigma(y)]\beta(\lambda(y), y)\}^{-1}.$$

It is clear that λ is an analytic function of $y^{1/N}$, and so $y \rightarrow \alpha(\lambda(y), y)$ and $y \rightarrow \beta(\lambda(y), y)$ are analytic functions of $y^{1/N}$. On the other hand, ν is a meromorphic function of $y^{1/N}$, and so σ is meromorphic in $y^{1/N}$ as well. So θ is a meromorphic function of $y^{1/N}$. That is, θ satisfies

$$(2.13) \quad \theta(y) = \sum_{j=j_0}^{\infty} \theta_j y^{j/N},$$

where the series in (2.13) converges for $0 < y < \delta_0$, if δ_0 is small enough, and where j_0 is an integer, and $\theta_{j_0} \neq 0$.

If we let

$$(2.14) \quad \tau^*(y) = \sum_{\substack{j_0 \leq j < \infty \\ j \neq -N}} \left(\frac{j}{N} + 1\right)^{-1} \theta_j y^{j/N+1} + \bar{\theta}$$

(where $\bar{\theta}$ is an arbitrary constant) and

$$(2.15) \quad \tau(y) = \tau^*(y) + \theta_{-N} \log y$$

(where “log” denotes natural logarithm), then $\tau:]0, \delta_0[\rightarrow \mathbb{R}$ is analytic, and its y -derivative is θ , so that τ is, up to a constant, the time along the Z_S -trajectory, which is obtained by suitably reparametrizing the set $\{(\lambda(y), y): 0 < y < \delta_0\}$ (that is: $\tau(y) = \rho(\lambda(y), y) + \text{constant}$).

Now fix a \hat{y} such that $0 < \hat{y} < \delta_0$, and let \hat{S}_+ denote the compact set

$$(2.16) \quad \{p\} \cup \{(\lambda(y), y): 0 < y \leq \hat{y}\}.$$

Let $\hat{\mathcal{F}}_+^S$ denote the set of those (q_0, q_1, t) such that $q_0 \in \hat{S}_+$, $q_1 \in \hat{S}_+$, $t \geq 0$, and $q_0 = \gamma(0)$, $q_1 = \gamma(t)$ for some $\gamma \in \text{Traj}(\Sigma)$ which is contained in \hat{S} . We will prove that $\hat{\mathcal{F}}_+^S$ is subanalytic in $U \times U \times \mathbb{R}$. It is clear that the function τ is strictly decreasing on $]0, \delta_0[$ (because, as y decreases, the point $(\lambda(y), y)$ moves towards p , i.e., in the direction of the Z_S -trajectory). So the limit $\tau(0+)$ exists, and is either finite or equal to $+\infty$. If

$\tau(0+) = +\infty$, then no trajectory of Σ which is contained in \hat{S}_+ can contain both p and a point in $\hat{S}_+ - \{p\}$. So, in this case, if $q_i = (x_i, y_i)$, $i = 0, 1$, we have that $(q_0, q_1, t) \in \hat{\mathcal{F}}_+^S$ if and only if either

$$(2.IV_i) \quad q_0 \in \hat{S}_+, q_1 \in \hat{S}_+, q_0 \neq p \neq q_1, t \geq 0, \text{ and } t = \tau(y_1) - \tau(y_0), \text{ or}$$

$$(2.IV_{ii}) \quad q_0 = q_1 = p \text{ and } t = 0 \text{ or}$$

$$(2.IV_{iii}) \quad q_0 = q_1 = p, t \geq 0, \text{ and } 0 \text{ is a convex combination of } X(p) \text{ and } Y(p).$$

Let us define P to be the set $\{p\} \times \{p\} \times [0, \infty)$ if 0 is a convex combination of $X(p)$ and $Y(p)$, and let P be the empty set otherwise. Then we have shown that, if $\tau(0+) = +\infty$, the set $\hat{\mathcal{F}}_+^S$ is the union of P and of the set $^*\hat{\mathcal{F}}_+^S$ of those (q_0, q_1, t) for which (2.IV_i) holds.

If $\tau(0+) < +\infty$, then conditions (2.IV_i), (2.IV_{ii}) or (2.IV_{iii}) imply that $(q_0, q_1, t) \in \hat{\mathcal{F}}_+^S$, but there are other possibilities as well, namely

$$(2.IV_{iv}) \quad q_0 \in \hat{S}_+, q_0 \neq p, q_1 = p, t \geq 0, \text{ and } t = \tau(0+) - \tau(y_0), \text{ and}$$

$$(2.IV_v) \quad q_0 \in \hat{S}_+, q_0 \neq p, q_1 = p, t \geq 0, t \geq \tau(0+) - \tau(y_0) \text{ and } 0 \text{ is a convex combination of } X(p) \text{ and } Y(p).$$

Precisely, if $\tau(0+) < +\infty$, the point (q_0, q_1, t) is in $\hat{\mathcal{F}}_+^S$ if and only if (2.IV_i), (2.IV_{ii}), (2.IV_{iii}), (2.IV_{iv}) or (2.IV_v) holds. The set of points where (2.IV_{ii}) or (2.IV_{iii}) holds is P . The set Q of points where (2.IV_{iv}) holds is the intersection of $\text{Clos } ^*\hat{\mathcal{F}}_+^S$ and of $(\hat{S}_+ - \{p\}) \times \{p\} \times \mathbb{R}$, so Q is subanalytic in $U \times U \times \mathbb{R}$ if $^*\hat{\mathcal{F}}_+^S$ is. Finally, the set Q' of points where (2.IV_v) holds is empty if 0 is not a convex combination of $X(p)$ and $Y(p)$, and is equal to

$$(2.17) \quad \{(q_0, q_1, t): (\exists t')(0 \leq t' \leq t \wedge (q_0, q_1, t') \in Q)\}$$

if 0 is a combination of $X(p)$ and $Y(p)$. In either case, Q' is subanalytic in $U \times U \times \mathbb{R}$ if Q is.

Summarizing the preceding observations, we have shown that, whether $\tau(0+)$ is finite or infinite, the set $\hat{\mathcal{F}}_+^S$ is the union of $^*\hat{\mathcal{F}}_+^S$ and of finitely many sets that are subanalytic in $U \times U \times \mathbb{R}$ if $^*\hat{\mathcal{F}}_+^S$ is. So the subanalyticity of $\hat{\mathcal{F}}_+^S$ in $U \times U \times \mathbb{R}$ will follow if we prove that $^*\hat{\mathcal{F}}_+^S$ is subanalytic in $U \times U \times \mathbb{R}$.

To prove that $^*\hat{\mathcal{F}}_+^S$ is subanalytic in $U \times U \times \mathbb{R}$, we first write

$$(2.18) \quad \tau^*(y) = y^{-k/N} \tau^{**}(y),$$

where $\tau^{**}(y)$ is given by a power series in $y^{1/N}$ which converges for $0 < y < \delta_0$, and k is a nonnegative integer. The equation

$$(2.19) \quad t = \tau(y_1) - \tau(y_0)$$

is therefore equivalent to

$$(2.20) \quad \zeta_0^k \zeta_1^k t = \zeta_0^k \tau^{**}(\zeta_1^N) - \zeta_1^k \tau^{**}(\zeta_0^N) + N \theta_{-N} \zeta_0^k \zeta_1^k (\log \zeta_1 - \log \zeta_0)$$

if $t \neq 0$, $0 < y_0 < \delta_0$, $0 < y_1 < \delta_0$, $\zeta_0 = y_0^{1/N}$ and $\zeta_1 = y_1^{1/N}$.

If $\theta_{-N} = 0$, the subanalyticity of $^*\hat{\mathcal{F}}_+^S$ in $U \times U \times \mathbb{R}$ follows easily. Indeed, if $t \neq 0$, $0 < y_0 < \delta_0$, $0 < y_1 < \delta_0$, then the point $(q_0, q_1, t) \in U \times U \times \mathbb{R}$ is in $^*\hat{\mathcal{F}}_+^S$ if and only if the following hold:

$$(2.V_i) \quad q_0 \in \hat{S}_+, q_1 \in \hat{S}_+, q_0 \neq p \neq q_1 \text{ and } t > 0;$$

$$(2.V_{ii}) \quad \text{There exist } \zeta_0, \zeta_1 \text{ such that } 0 \leq \zeta_0 \leq \hat{y}^{1/N}, 0 \leq \zeta_1 \leq \hat{y}^{1/N}, \zeta_0^N = y_0, \zeta_1^N = y_1, \text{ and}$$

$$(2.21) \quad \zeta_0^k \zeta_1^k t = \zeta_0^k \tau^{**}(\zeta_1^N) - \zeta_1^k \tau^{**}(\zeta_0^N).$$

Equation (2.21) is of the form $\eta(\zeta_0, \zeta_1, t) = 0$, where η is an analytic function on W , where

$$(2.22) \quad W =]-\delta_0^{1/N}, \delta_0^{1/N}[\times]-\delta_0^{1/N}, \delta_0^{1/N}[\times \mathbb{R}.$$

So, if we let D denote the set of all $(\zeta_0, \zeta_1, t, y_0, y_1)$ in $W \times]-\delta_0, \delta_0[\times]-\delta_0, \delta_0[$ that satisfy (2.21) and $0 < y_0 \leq \hat{y}$, $0 < y_1 \leq \hat{y}$, $\zeta_0^N = y_0$, $\zeta_1^N = y_1$, then the projection $\pi: (\zeta_0, \zeta_1, t, y_0, y_1) \rightarrow (t, y_0, y_1)$, from $W \times]-\delta_0, \delta_0[\times]-\delta_0, \delta_0[$ to $\mathbb{R} \times]-\delta_0, \delta_0[\times]-\delta_0, \delta_0[$, is proper on $\text{Clos } D$, and therefore maps D to a subanalytic subset of $\mathbb{R} \times]-\delta_0, \delta_0[\times]-\delta_0, \delta_0[$. This clearly implies that ${}^*\hat{\mathcal{F}}_+^S$ is subanalytic in $U \times U \times \mathbb{R}$ (since ${}^*\mathcal{F}_+^S$ is the set of those $(q_0, q_1, t) = (x_0, y_0, x_1, y_1, t)$ such that either $q_0 \in \hat{S}_+$, $q_1 \in \hat{S}_+$, $(y_0, y_1, t) \in \pi(D)$, and $t > 0$, or $q_0 \in \hat{S}$, $q_0 = q_1$, $t = 0$).

If $\theta_{-N} \neq 0$, then a slightly more complicated reasoning is needed. Clearly, a point $(q_0, q_1, t) \in U \times U \times \mathbb{R}$ for which $t > 0$ is in ${}^*\hat{\mathcal{F}}_+^S$ if and only if (2.Vi) holds and, in addition:

$$(2.Viii) \quad \begin{aligned} &\text{There exist } \zeta_0, \zeta_1, \omega \text{ such that } 0 \leq \zeta_0 \leq \hat{y}^{1/N}, 0 \leq \zeta_1 \leq \hat{y}^{1/N}, \zeta_0^N = y_0, \\ &\zeta_1^N = y_1, \zeta_0 e^\omega = \zeta_1, \text{ and} \\ &\zeta_0^k \zeta_1^k t = \zeta_0^k \tau^{**}(\zeta_1^N) - \zeta_1^k \tau^{**}(\zeta_0^N) + N\theta_{-N} \zeta_0^k \zeta_1^k \omega. \end{aligned}$$

(Naturally, the equation $\zeta_0 e^\omega = \zeta_1$ is included to make sure that ω is equal to $\log \zeta_1 - \log \zeta_0$.)

The proof that ${}^*\hat{\mathcal{F}}_+^S$ is subanalytic in $U \times U \times \mathbb{R}$ proceeds as before, except that W is replaced by $\mathbb{R} \times W$, because of the new coordinate ω . We now let D denote the set of all $(\omega, \zeta_0, \zeta_1, t, y_0, y_1)$ that satisfy $t > 0$ and all the equalities and inequalities of (2.Viii). The subanalyticity of ${}^*\hat{\mathcal{F}}_+^S$ then follows if we prove that the projection π onto the last three coordinates, regarded as a map from $\mathbb{R} \times W \times]-\delta_0, \delta_0[\times]-\delta_0, \delta_0[$ to $\mathbb{R} \times]-\delta_0, \delta_0[\times]-\delta_0, \delta_0[$, is proper on $\text{Clos } D$. This requires that we prove that, as long as $(\omega, \zeta_0, \zeta_1, t, y_0, y_1) \in D$, and t remains bounded, it follows that ω is bounded, and that $|\zeta_0|, |\zeta_1|$ are bounded by a constant c which is less than $\delta_0^{1/N}$. The bound for $|\zeta_0|, |\zeta_1|$ is trivial (just let $c = \hat{y}^{1/N}$). Therefore the only problem is to prove that, as long as $(\omega, \zeta_0, \zeta_1, t, y_0, y_1) \in D$, and t is bounded, it follows that ω is bounded. Equivalently (since $\omega = \log \zeta_1 - \log \zeta_0 = 1/N \log(y_1/y_0)$) we need an a priori bound of the form

$$(2.23) \quad \left| \log \left(\frac{y_1}{y_0} \right) \right| \leq C(T)$$

for $(q_0, q_1, t) \in {}^*\hat{\mathcal{F}}_+^S$, $0 < t \leq T$, where the constant $C(T)$ is only allowed to depend on T .

To get the bound (2.23), consider first the case when $j_0 = -N$. Then $\tau^*(y)$ is bounded for $y \in [0, \hat{y}]$ and so, if $(q_0, q_1, t) \in {}^*\hat{\mathcal{F}}_+^S$, we have

$$(2.24) \quad t = \tau^*(y_1) - \tau^*(y_0) + \theta_{-N} \left(\log \frac{y_1}{y_0} \right),$$

so that

$$\left| \log \left(\frac{y_1}{y_0} \right) \right| \leq \frac{T + 2C}{|\theta_{-N}|}$$

as long as $0 < t \leq T$, if C is an upper bound for $|\tau^*|$ on $[0, \hat{y}]$. So the desired bound holds.

Next consider the case when $j_0 \neq -N$. Since we are assuming that $\theta_{-N} \neq 0$, it necessarily follows that $j_0 < -N$. Therefore $\tau^*(y)$ is asymptotic to y^{-r} as $y \rightarrow 0$, for some $r > 0$. Choose the constant $\bar{\theta}$ (which so far was arbitrary) so that $\tau(\hat{y}) = 1$. Then

$\tau(y) \geq 1$ for $0 < y \leq \hat{y}$, and $\tau(y)$ is asymptotic to y^{-r} as $y \rightarrow 0+$. So there are constants C_0, C_1 such that $0 < C_0 < C_1$, and that

$$(2.25) \quad C_0 y^{-r} \leq \tau(y) \leq C_1 y^{-r}$$

for $0 < y \leq \hat{y}$. So, if $0 < y_1 \leq y_0 \leq \hat{y}$, and $t = \tau(y_1) - \tau(y_0)$, we have

$$t \geq C_0 y_1^{-r} - C_1 y_0^{-r},$$

so that

$$(2.26) \quad C_0 \left(\frac{y_0}{y_1} \right)^r \leq C_1 + t y_0^r.$$

If $0 < t \leq T$, (2.26) gives

$$(2.27) \quad \log \left(\frac{y_0}{y_1} \right) \leq \frac{1}{r} \log \left(\frac{C_1 + T \hat{y}^r}{C_0} \right).$$

Since $y_1 \leq y_0$, we also have $\log(y_0/y_1) \geq 0$, and so the bound (2.23) follows. As explained before, this completes the proof that ${}^*\hat{\mathcal{F}}_+^S$ is subanalytic in $U \times U \times \mathbb{R}$. Therefore, we have proved that $\hat{\mathcal{F}}_+^S$ is subanalytic in $U \times U \times \mathbb{R}$. Since $U \times U \times \mathbb{R}$ is open in $M \times M \times \mathbb{R}$, and $\hat{\mathcal{F}}_+^S$ is a subset of $\hat{S}_+ \times \hat{S}_+ \times \mathbb{R}$, which is contained in $U \times U \times \mathbb{R}$ and is closed in $M \times M \times \mathbb{R}$, it follows that $\hat{\mathcal{F}}_+^S$ is subanalytic in $M \times M \times \mathbb{R}$.

Now we are ready to complete the proof that \mathcal{F}_+^S is subanalytic in $M \times M \times \mathbb{R}$. Let S^* be the closed subarc of S which goes from \bar{q} to \hat{q} . Let \mathcal{B} be the set of those (q_0, q_1, t) such that $q_0 \in S^*$, $q_1 \in S^*$ and $(q_0, q_1, t) \in \mathcal{F}_+^S$. Let S^{**} be a slightly larger open arc, containing S^* . Then the function ρ is analytic on S^{**} , and \mathcal{B} is the subset of $S^{**} \times S^{**} \times \mathbb{R}$ characterized by the conditions $q_0 \in S^*$, $q_1 \in S^*$, $t \geq 0$ and $t = \rho(q_1) - \rho(q_0)$. So \mathcal{B} is subanalytic in $S^{**} \times S^{**} \times \mathbb{R}$. On the other hand, the inclusion from $S^{**} \times S^{**} \times \mathbb{R}$ to $M \times M \times \mathbb{R}$ is proper on $S^* \times S^* \times \mathbb{R}$, and $\mathcal{B} \subseteq S^* \times S^* \times \mathbb{R}$. So \mathcal{B} is subanalytic in $M \times M \times \mathbb{R}$.

Now, if $(q_0, q_1, t) \in M \times M \times \mathbb{R}$, then it is clear that (q_0, q_1, t) is in \mathcal{F}_+^S if and only if one of the following three conditions holds:

- (2.VIi) $(q_0, q_1, t) \in \mathcal{B}$;
- (2.VIii) $(q_0, q_1, t) \in \hat{\mathcal{F}}_+^S$;
- (2.VIiii) There exist t', t'' such that $0 \leq t' \leq t$, $0 \leq t'' \leq t$, $t' + t'' = t$ and that $(q_0, \hat{q}, t') \in \mathcal{B}$, $(\hat{q}, q_1, t'') \in \hat{\mathcal{F}}_+^S$.

So \mathcal{F}_+^S is the union of three subanalytic subsets of $M \times M \times \mathbb{R}$, and is therefore subanalytic in $M \times M \times \mathbb{R}$.

We have now proved that, if S is a noncompact maximal turnpike, and if \bar{q} , S_+ , S_- , \mathcal{F}_+^S , \mathcal{F}_-^S are defined as above, then \mathcal{F}_+^S is subanalytic in $M \times M \times \mathbb{R}$. A similar proof shows that \mathcal{F}_-^S is subanalytic in $M \times M \times \mathbb{R}$ as well. (Or, equivalently, the proof for \mathcal{F}_-^S can be reduced to that for \mathcal{F}_+^S by changing the signs of X and Y .) We are now ready to prove that \mathcal{F}^S is subanalytic. With the diffeomorphism $\tilde{\psi}: \mathbb{R} \rightarrow S$ defined as before, let $p_+ = \lim_{t \rightarrow +\infty} \tilde{\psi}(t)$, $p_- = \lim_{t \rightarrow -\infty} \tilde{\psi}(t)$, so that each of p_+ , p_- exists, and is either infinite or a point of M . (Recall that the meaning of " $p_+ = \infty$ " was explained earlier. A similar interpretation is understood for " $p_- = \infty$ ".) If both p_- , p_+ are infinite, or if exactly one of them is finite, or if both are finite but $p_- \neq p_+$, then it is easy to see that, if ξ is any trajectory of Σ which is contained in $\text{Clos } S$, then either ξ is contained in $\text{Clos } S_-$, or it is contained in $\text{Clos } S_+$, or is a concatenation of two

trajectories ξ_- , ξ_+ , contained in $\text{Clos } S_-$, $\text{Clos } S_+$, respectively. So, if $(q_0, q_1, t) \in M \times M \times \mathbb{R}$, the point (q_0, q_1, t) belongs to \mathcal{F}^S if and only if one of the following three conditions holds:

- (2.VIIi) $(q_0, q_1, t) \in \mathcal{F}_-^S$;
- (2.VIIii) $(q_0, q_1, t) \in \mathcal{F}_+^S$;
- (2.VIIiii) There exist t' , t'' such that $0 \leq t' \leq t$, $0 \leq t'' \leq t$, $t' + t'' = t$ and that $(q_0, \bar{q}, t') \in \mathcal{F}_-^S$ and $(\bar{q}, q_1, t'') \in \mathcal{F}_+^S$.

This clearly shows that \mathcal{F}^S is subanalytic in $M \times M \times \mathbb{R}$.

There remains the exceptional case when p_- , p_+ are both finite and equal (e.g., if S is a circle minus one point). In this case, if the time parameter along the integral curve of Z_S approaches infinity as one approaches p_+ via either S_+ or S_- , then it is still true that $(q_0, q_1, t) \in \mathcal{F}^S$ if and only if one of (2.VIIi, ii, iii) holds, and this implies, exactly as before, that \mathcal{F}^S is subanalytic in $M \times M \times \mathbb{R}$. However, if the time parameter remains finite as we approach p_+ via S_- and via S_+ , then $\text{Clos } S$ carries a periodic singular trajectory. If we let $T > 0$ denote the period, then, if $(q_0, q_1, t) \in M \times M \times \mathbb{R}$, and if $0 \leq t \leq T$, we have that $(q_0, q_1, t) \in \mathcal{F}^S$ if and only if (2.VIIi), (2.VIIii) or (2.VIIiii) holds, or

- (2.VIIiv) $(q_1, q_0, T - t)$ satisfies one of (2.VIIi, ii, iii).

This shows that $\mathcal{F}^S \cap (M \times M \times [0, T])$ is subanalytic in $M \times M \times \mathbb{R}$. Since $\mathcal{F}^S \cap (M \times M \times [nT, (n+1)T])$ is obtained from $\mathcal{F}^S \cap (M \times M \times [0, T])$ by the translation $(q_0, q_1, t) \rightarrow (q_0, q_1, t + nT)$, this set is subanalytic as well. So \mathcal{F}^S is subanalytic in $M \times M \times \mathbb{R}$.

We have proved the following.

THEOREM 2.3. *Let Σ be a C.A.S. that has the SAP. If S is a maximal turnpike of Σ , then \mathcal{F}^S is a subanalytic subset of $M \times M \times \mathbb{R}$. \square*

From this we also get the following.

COROLLARY 2.4. *Let Σ be a C.A.S. that has the SAP. Then \mathcal{F}^Z is a subanalytic subset of $M \times M \times \mathbb{R}$.*

3. Existence of regular synthesis. We are now ready to piece together all our preceding results and prove the existence of a regular synthesis. We will rely heavily on the general abstract theorems of [Su 6] (cf. also [Su 3] for an outline, without proof, of some of these results). Here we will limit ourselves to a precise statement of all the definitions involved, so as to show that the general theorems of [Su 6] apply.

First, consider a control system

$$(3.1) \quad \dot{x} = F(x, u), \quad x \in M, \quad u \in U,$$

where M is a manifold of class C^k ($k = \infty$ or $k = \omega$), U is a compact subset of \mathbb{R}^m for some m , and F is of class C^k jointly in x, u . A *feedback control flow* for (3.1) on a subset S of M is a choice $\Gamma = \{\Gamma_x : x \in S\}$ of a family of admissible pairs $\Gamma_x = (u_x(\cdot), \gamma_x)$, such that each γ_x is a trajectory of (3.1) starting at x and entirely contained in S , and the following *compatibility condition* holds:

$$(3.I) \quad \text{If } t \in \text{Dom}(\gamma_x) \text{ and } \gamma_x(t) = y, \text{ then } u_y(\cdot) = u_x(\cdot) \upharpoonright (\text{Dom}(\gamma_x) \cap [t, \infty[) \\ \text{and } \gamma_y = \gamma_x \upharpoonright (\text{Dom}(\gamma_x) \cap [t, \infty[).$$

(Clearly, Condition (3.I) says that Γ is “memoryless.”)

A feedback control flow Γ on S is said to *steer* S to a point $p \in S$ if, for every $x \in S$, the trajectory γ_x ends at p (i.e., $\gamma_x(\max \text{Dom}(\gamma_x)) = p$). Clearly, if Γ steers S to

p , then it is always possible to make a time translation, so as to assume that, for every x , the domain $\text{Dom}(\gamma_x)$ is an interval of the form $[-\tau_x, 0]$, with $\tau_x \geq 0$. Whenever we talk about a Γ steering S to p , we will always assume that this modification has been carried out.

If $p \in M$, let $\text{Contr}(p)$ denote the set of all points in M that can be steered to p , in finite time, by a trajectory of (3.1). A *feedback control flow with target p* is a feedback control flow on $\text{Contr}(p)$, which steers $\text{Contr}(p)$ to p .

Now suppose that a Lagrangian function $L: M \times U \rightarrow \mathbb{R}$ is given, and that L is of class C^k and nonnegative. For each x, y , we may consider the optimal control problem $P_L(x, y)$ of minimizing the integral

$$\int_{t \in \text{Dom}(\gamma)} L(\gamma(t), u(t)) dt$$

among all admissible pairs $(u(\cdot), \gamma)$ such that γ starts at x and ends at y . An *L -optimal feedback control flow with target p* is a feedback control flow Γ with target p , which has the property that, for each $x \in \text{Contr}(p)$, the admissible pair Γ_x is a solution of $P_L(x, p)$. An *L -extremal feedback control flow with target p* is a feedback control flow Γ with target p such that, for each $x \in \text{Contr}(p)$, the pair Γ_x satisfies the Maximum Principle for the problem $P_L(x, p)$.

It is clear that if Γ is L -optimal then Γ is L -extremal. For individual trajectories, it is well known that extremality does not imply optimality. The main point of the theory of the *regular synthesis* is that, for *feedback control flows* Γ with target p , L -optimality is actually equivalent to L -extremality, provided that Γ satisfies some extra regularity conditions. A *regular synthesis* is, roughly, an L -extremal feedback control flow Γ with target p , which satisfies the regularity conditions (so that, in particular, it follows that Γ is optimal). In order to give a precise definition of regular synthesis we will first define what it means for a Γ (L -extremal or not) to be “regular,” and we will then define a *regular synthesis* to be a Γ which is both regular and L -extremal. Our definition of regularity will be such that, when Γ is regular, L -extremality implies L -optimality. Therefore, the class of regular feedback control flows has the nice property that, for Γ in this class, the Maximum Principle is both a necessary and a sufficient condition for optimality. If, in addition, one can prove general existence theorems within this class, then it will be clear that regular feedback control flows are a natural class of objects at which to look.

Our definition of regularity is taken from [Su 6], and is a modification of definitions given by other authors (e.g., Boltyanskii [Bo], Baytman [Ba], Brunovsky [Br1]). First define a *piecewise C^k vector field* on a set S to be a 6-tuple $(\mathcal{P}, \mathcal{P}_1, \mathcal{P}_2, \mathbf{X}, \mathbf{E}, \epsilon)$ where:

- (3.IIi) \mathcal{P} is a finite or countably infinite partition of S into connected, embedded submanifolds of class C^k ;
- (3.IIii) $\mathcal{P}_1, \mathcal{P}_2$ are subsets of \mathcal{P} such that $\mathcal{P}_1 \cup \mathcal{P}_2 = \mathcal{P}$, $\mathcal{P}_1 \cap \mathcal{P}_2 = \emptyset$;
- (3.IIiii) \mathbf{X} is a family $\{X_P: P \in \mathcal{P}_1\}$ such that, for each $P \in \mathcal{P}_1$, X_P is a vector field of class C^k on P ;
- (3.IIiv) ϵ is a family $\{\epsilon_P: P \in \mathcal{P}_2\}$ of continuous functions $\epsilon_P: P \rightarrow]0, \infty[$ and \mathbf{E} is a family $\{E_P: P \in \mathcal{P}_2\}$, such that each E_P is a continuous map $E_P: \{(q, t): q \in P, 0 \leq t < \epsilon_P(q)\} \rightarrow S$, with the property that there exists a $P' \in \mathcal{P}_1$ such that E_P maps $\{(q, t): q \in P, 0 < t < \epsilon_P(q)\}$ into P' in a C^k fashion, and that each curve $t \rightarrow E_P(q, t)$, $0 < t < \epsilon_P(q)$ is an integral curve of $X_{P'}$;

- (3.IIv) If $P \in \mathcal{P}_1$, and if, for some $q \in P$, there is a positive time $T(q)$ such that $\Phi_t^{X_P}(q)$ is defined for $0 \leq t < T(q)$ but not for $t = T(q)$, but $\Phi_t^{X_P}(q)$ has a limit $\xi(q)$ as $t \rightarrow T(q)-$, and $\xi(q) \in S - P$, then the same is true for every $q \in P$ and, moreover, there is a $P' \in \mathcal{P}$ such that $\xi(q) \in P'$ for all $q \in P$, and the maps $T(\cdot): P \rightarrow \mathbb{R}$, $\xi(\cdot): P \rightarrow P'$ are of class C^k ;
- (3.IIvi) If $P \in \mathcal{P}_2$, then either $X_P \equiv 0$ or X_P never vanishes on P .

If $\mathcal{V} = (\mathcal{P}, \mathcal{P}_1, \mathcal{P}_2, \mathbf{X}, \mathbf{E}, \epsilon)$ is a piecewise C^k vector field on S , then those points $p \in S$ which belong to a $P \in \mathcal{P}_1$ such that $X_P \equiv 0$ are called *terminal points* of \mathcal{V} . The set of all such points is the *terminal set* of \mathcal{V} (notation: $\text{Term}(\mathcal{V})$). A *trajectory* of \mathcal{V} is a continuous curve $\gamma: I \rightarrow S$, where I is some interval, such that

- (3.IIIi) If $t \in I$, $t \neq \sup I$, and $\gamma(t) \in P \in \mathcal{P}_1$, then $\gamma(t) \notin \text{Term}(\mathcal{V})$, and there exists a $\delta > 0$ such that $\gamma \upharpoonright [t, t + \delta]$ is an integral curve of X_P ;
- (3.IIIii) If $t \in I$, $t \neq \sup I$, and $\gamma(t) \in P \in \mathcal{P}_2$, then there exists a $\delta > 0$ such that $\gamma(t + \tau) = E_P(\gamma(t), \tau)$ for $0 < \tau < \delta$.

It is clear from the definition of trajectories that there is local existence and global uniqueness of trajectories in the forward direction; i.e. (a) given $t_0 \in \mathbb{R}$, $q_0 \in S$, $q_0 \notin \text{Term}(\mathcal{V})$, there is a $\delta > 0$ and a \mathcal{V} -trajectory γ defined on $[t_0, t_0 + \delta[$, such that $\gamma(t_0) = q_0$ and (b) if γ_1, γ_2 are \mathcal{V} -trajectories such that $\gamma_1(t_0) = \gamma_2(t_0)$ for some t_0 , then $\gamma_1 \equiv \gamma_2$ on $\text{Dom}(\gamma_1) \cap \text{Dom}(\gamma_2) \cap [t_0, \infty[$. So, for any $q \in S - \text{Term}(\mathcal{V})$, there is a unique *maximal* forward trajectory from q , i.e., a trajectory $\gamma_q^{\mathcal{V}}$ of \mathcal{V} , such that $0 \in \text{Dom}(\gamma_q^{\mathcal{V}})$, that $\text{Dom}(\gamma_q^{\mathcal{V}}) \subseteq [0, \infty[$, that $\gamma_q^{\mathcal{V}}(0) = q$, and that $\gamma_q^{\mathcal{V}}$ cannot be extended to a \mathcal{V} -trajectory on an interval I such that

$$\text{Dom}(\gamma_q) \not\subseteq I \subseteq [0, \infty[.$$

A *piecewise C^k feedback control law* for the system (3.1) on a set S is a pair $\mathcal{V}' = (\mathcal{V}, \mathbf{u})$, where:

- (3.IV i) $\mathcal{V} = (\mathcal{P}, \mathcal{P}_1, \mathcal{P}_2, \mathbf{X}, \mathbf{E}, \epsilon)$ is a piecewise C^k vector field on S ;
- (3.IV ii) \mathbf{u} is a family $\{u_P: P \in \mathcal{P}_1\}$, such that each u_P is a map $u_P: P \rightarrow U$, of class C^k , such that

$$(3.2) \quad X_P(q) = F(q, u_P(q))$$

for all $q \in P$.

If $\mathcal{V}' = (\mathcal{V}, \mathbf{u})$ is a piecewise C^k feedback control law on S , then we can define an *admissible control*

$$u_q^{\mathcal{V}'}: \text{Dom}(\gamma_q^{\mathcal{V}}) \rightarrow U$$

by letting

$$(3.3) \quad u_q^{\mathcal{V}'}(t) = u_P(\gamma_q^{\mathcal{V}}(t))$$

if $\gamma_q^{\mathcal{V}}(t) \in P \in \mathcal{P}_1$, and by defining $u_q^{\mathcal{V}'}(t)$ in an arbitrary fashion if $\gamma_q^{\mathcal{V}}(t) \in P \in \mathcal{P}_2$. (It follows from the definition of a trajectory that the set of those $t \in \text{Dom}(\gamma_q^{\mathcal{V}})$ for which $\gamma_q^{\mathcal{V}}(t)$ is in some $P \in \mathcal{P}_2$ is finite or countable.) The pair $(u_q^{\mathcal{V}'}, \gamma_q^{\mathcal{V}}) = \Gamma_q^{\mathcal{V}'}$ is clearly admissible.

If \mathcal{V} is a piecewise C^k vector field on S (or if \mathcal{V}' is a piecewise C^k feedback control law on S), we say that \mathcal{V} (or \mathcal{V}') is *completely terminating* if $\text{Dom}(\gamma_q^{\mathcal{V}})$ is a compact interval for each $q \in S$. It is then clear that, if $q \in S$, the point $\gamma_q^{\mathcal{V}}(\max \text{Dom}(\gamma_q^{\mathcal{V}}))$ is in $\text{Term}(\mathcal{V})$. When \mathcal{V} is completely terminating, it is more convenient to define a new trajectory $\hat{\gamma}_q^{\mathcal{V}}$ by letting $\hat{\gamma}_q^{\mathcal{V}}(t) = \gamma_q^{\mathcal{V}}(t + \max \text{Dom}(\gamma_q^{\mathcal{V}}))$ for

$-\max \text{Dom}(\gamma_q^\mathcal{V}) \leq t \leq 0$. Then $\hat{\gamma}_q^\mathcal{V}$ terminates exactly at time 0. If $\mathcal{V}' = (\mathcal{V}, u)$ is a piecewise C^k completely terminating feedback control law, then we can also define $\hat{u}_q^{\mathcal{V}'}$ in a similar fashion. Now the pair

$$\hat{\Gamma}_q^{\mathcal{V}'} = (\hat{u}_q^{\mathcal{V}'}, \hat{\gamma}_q^{\mathcal{V}'})$$

is admissible for each $q \in S$ and, moreover, it is easy to see that the family $\hat{\Gamma}^{\mathcal{V}'} = \{\hat{\Gamma}_q^{\mathcal{V}'} : q \in S\}$ satisfies the compatibility property. Therefore $\hat{\Gamma}^{\mathcal{V}'}$ is a feedback control flow. We call $\hat{\Gamma}^{\mathcal{V}'}$ the *feedback control flow generated by \mathcal{V}'* .

If \mathcal{V} (or \mathcal{V}') is completely terminating, we say that \mathcal{V} (or \mathcal{V}') *terminates in a finite number of steps* if, for each $q \in S$, there exists a finite partition π of $\text{Dom}(\hat{\gamma}_q^\mathcal{V})$ into intervals such that, for each $I \in \pi$, $\hat{\gamma}_q^\mathcal{V}(I)$ is entirely contained in one member of \mathcal{P} .

Now, let $\mathcal{V}' = (\mathcal{V}, u)$ be completely terminating in a finite number of steps. For each $q \in S$, let $\tau^\mathcal{V}(q)$ denote the *termination time* of $\gamma_q^\mathcal{V}$, i.e., the length of $\text{Dom}(\gamma_q^\mathcal{V})$. It follows easily from our definitions that $\tau^\mathcal{V}$ is of class C^k on each $P \in \mathcal{P}_1$, and $P \not\subseteq \text{Term}(\mathcal{V})$, then X_P never vanishes on P , and

$$X_P(\tau^\mathcal{V} \upharpoonright P) \equiv -1.$$

Therefore, if $P \in \mathcal{P}_1$, $P \not\subseteq \text{Term}(\mathcal{V})$, the C^k function $\tau^\mathcal{V} \upharpoonright P$ has a nowhere vanishing gradient. So, if $q \in P$, the set $N(q)$ of those $q' \in P$ such that $\tau^\mathcal{V}(q') = \tau^\mathcal{V}(q)$ is a C^k submanifold of P , of codimension one. We say that $q \in P$ is a *Lipschitz point* if there is a neighborhood W of q in P and a constant $C > 0$, such that

$$(3.4) \quad \text{dist}(\hat{\gamma}_{q'}^\mathcal{V}(t), \hat{\gamma}_q^\mathcal{V}(t)) \leq C \text{dist}(q', q')$$

for all $q' \in W \cap N(q)$, and all $t \in [-\tau^\mathcal{V}(q), 0]$. (Here “dist” is the distance relative to some Riemannian metric on M .) We say that \mathcal{V}' has the *Lipschitz regularity property* if, for every $P \in \mathcal{P}_1$, the set of Lipschitz points of P is dense in P .

Now suppose that $\mathcal{V}' = (\mathcal{V}, u)$ is a piecewise C^k feedback control flow on S which completely terminates, and suppose that $\text{Term}(\mathcal{V}')$ consists of a single point p . Let L be a Lagrangian function as above. We define the *cost function* V , corresponding to \mathcal{V}' and L , by letting

$$(3.5) \quad V(q) = \int_{-\tau^{\mathcal{V}'}(q)}^0 L(\hat{\gamma}_q^{\mathcal{V}'}(t), \hat{u}_q^{\mathcal{V}'}(t)) dt.$$

We shall say that \mathcal{V}' satisfies the *near continuity property* if the following three conditions hold:

- (3.Vi) V is lower semicontinuous on S ;
- (3.Vii) For every $\delta > 0$ and every neighborhood W of p in M , there exists a submanifold W' of M such that $W' \subseteq W \cap S$, that all the vector fields $F(\cdot, u)$ are tangent to W' , and that $V(q) < \delta$ for all $q \in W'$ (it is not required that $p \in W'$);
- (3.Viii) Whenever $\gamma : I \rightarrow S$ is an integral trajectory of some vector field $F(\cdot, u)$, $u \in U$, then $t \rightarrow V(\gamma(t))$ is left continuous on I .

Notice that, if the feedback control flow $\hat{\Gamma}'$ actually were an L -optimal feedback control flow with target $p \in M$, then condition (3.Viii) would follow from (3.Vi). (Indeed, let γ be an integral curve of $F(\cdot, u)$ in $\text{Contr}(p)$, and let $t_n \rightarrow t$, $t_n \in \text{Dom}(\gamma)$, $t \in \text{Dom}(\gamma)$. Then $\liminf V(\gamma(t_n)) \geq V(\gamma(t))$ by the lower semicontinuity of V . On the other hand:

$$(3.6) \quad V(\gamma(t_n)) \leq \int_{t_n}^t L(\gamma(s), u) ds + V(\gamma(t)),$$

and so $\limsup V(\gamma(t_n)) \leq V(\gamma(t))$. Also, condition (3.Vii) follows automatically if, for instance, the system (3.1) is analytic. (This is quite easy to prove; see [Su 6].)

We are now ready to define regular synthesis. A C^k *regular synthesis* for the system (3.1), with Lagrangian L and target p , is an L -extremal feedback control flow Γ with target p , which is the flow $\hat{\Gamma}^{\mathcal{V}'}$ generated by a piecewise C^k feedback control law $\mathcal{V}' = (\mathcal{V}, u)$ which is completely terminating in finitely many steps, and satisfies the Lipschitz regularity and near continuity properties.

One of the main theorems of [Su 6] is that, if Γ is a C^k regular synthesis in the sense of the preceding definition, then Γ is optimal, if the control system satisfies a mild regularity condition. (Precisely the condition that the rank of the Lie algebra generated by the vector fields $F(\cdot, u)$ be constant along the integral curves of these vector fields. The condition is always satisfied when the $F(\cdot, u)$ are analytic.)

We are now ready to state the main theorem of this paper.

THEOREM 3.1. *Consider a control system Σ given by:*

$$\dot{x} = F(x) + uG(x), \quad |u| \leq 1, \quad x \in M,$$

where M is a two-dimensional analytic manifold, and F, G are analytic vector fields on M . Let $p \in M$ be such that the following "nonexplosion condition" holds:

(NE) For every $T > 0$ there exists a compact subset $K(T)$ of M such that, if $\gamma: [a, b] \rightarrow M$ is a trajectory of Σ , such that $\gamma(b) = p$ and $b - a \leq T$, then γ is entirely contained in $K(T)$.

Then the problem of reaching p in minimum time admits a C^ω regular synthesis Γ . Moreover, Γ can be chosen to be the feedback control flow generated by a $\mathcal{V}' = (\mathcal{V}, u)$ such that, if \mathcal{P} is the partition corresponding to \mathcal{V} , and if, for $T > 0$, K_T is the set of points that can be steered to p in time $\leq p$, then each K_T only meets finitely many members of \mathcal{P} .

Proof. We let M_0 denote the integral manifold through p of the Lie algebra L of vector fields generated by F and G . Clearly, the set $\text{Contr}(p)$ is contained in M_0 and so, in order to construct the synthesis, it suffices to work with the restriction of our system to M_0 . If $\dim M_0 < 2$ then the existence of the regular synthesis is a completely trivial matter. So we shall assume that $\dim M_0 = 2$, i.e. that Σ has the accessibility property at p . We now replace M by M_0 , i.e. we assume that M is itself the integral manifold of L through p . If L_0 is the ideal of L generated by G , then it follows from [Su J] that the dimension of $L_0(q)$ is the same for all $q \in M$. Obviously, the common value of this dimension is either two or one. In the former case, Σ has the SAP everywhere, and in the latter case Σ has the SAP nowhere.

The strategy of our proof will be to construct an *optimal* feedback control flow Γ with target p . Then we will invoke the results of [Su 6] to conclude that Γ is generated by a \mathcal{V}' that has all the desired properties. Before we construct Γ , let us observe that condition (NE) implies, in particular, that for every $q \in \text{Contr}(p)$ there exists a time-optimal trajectory γ going from q to p . So, if we define $V(q)$ to be the infimum of the times along all the trajectories γ from q to p , then $V(q)$ is actually attained by some γ . Moreover, it is easy to see (using (NE) again) that V is lower semicontinuous on $\text{Contr}(p)$. As explained earlier, the near-continuity property will then follow automatically.

So, what we really need is to construct an optimal Γ that is generated by a \mathcal{V}' which is a piecewise C^ω feedback control law which completely terminates in finitely many steps and satisfies the Lipschitz regularity condition. According to the main

theorem of [Su 6], such a generator \mathcal{V}' will exist if Γ is a *subanalytic feedback*. So all we need is to construct an optimal feedback Γ with target p , which is subanalytic.

First recall that a *subanalytic feedback* is a feedback control law $\Gamma = \{\Gamma_x\}$ with the property that each of the maps

$$(x, t) \rightarrow u_x(t)$$

and

$$(x, t) \rightarrow \gamma_x(t)$$

is subanalytic. (That is, the sets $\{(x, t, y): y = u_x(t)\}$ and $G_\Gamma = \{(x, t, z): z = \gamma_x(t)\}$ are subanalytic in $M \times \mathbb{R} \times \mathbb{R}$, $M \times \mathbb{R} \times M$, respectively.) Next observe that, if Γ is a feedback control flow with target p , and if $(x, t) \rightarrow \gamma_x(t)$ is subanalytic, then it automatically follows that, after a suitable modification, one can assume that $(x, t) \rightarrow u_x(t)$ is subanalytic. (Proof: let A be the set of those (x, t) such that $x \in \text{Contr}(p)$, $t \in \text{Dom}(\gamma_x)$. Recall that we are assuming that $\text{Dom}(\gamma_x)$ is of the form $[a, 0]$. Then A is the projection of G_Γ under the map $\pi: (x, t, y) \rightarrow (x, t)$, from $M \times \mathbb{R} \times M$ to $M \times \mathbb{R}$. Since π is proper on $\text{Clos } G_\Gamma$, because of (NE), it follows that A is subanalytic in $M \times \mathbb{R}$. Let A' be the set of those $(x, t) \in A$ for which $\lim_{h \rightarrow 0+} (\gamma_x(t+h) - \gamma_x(t))/h$ exists. Then it is easy to see that A' is subanalytic in $M \times \mathbb{R}$. Let A'' be the set of those $(x, t) \in A'$ such that $G(\gamma_x(t)) \neq 0$. Define $u_x(t) = 0$ if $(x, t) \in A - A''$. If $(x, t) \in A''$, let $u_x(t)$ be the unique u such that

$$\lim_{h \rightarrow 0+} \frac{\gamma_x(t+h) - \gamma_x(t)}{h} = F(\gamma_x(t)) + uG(\gamma_x(t)).$$

Then the map $(x, t) \rightarrow u_x(t)$ is subanalytic.)

So all we have to do is to find a way to select, for each $x \in \text{Contr}(p)$, a time-optimal trajectory $\gamma_x: [-V(x), 0] \rightarrow M$, so that the γ_x are compatible (i.e., whenever $\gamma_x(t) = \gamma_y(t)$ for some $t \leq 0$, then $\gamma_x(\tau) = \gamma_y(\tau)$ for $t \leq \tau \leq 0$) and that $(x, t) \rightarrow \gamma_x(t)$ is subanalytic.

If Σ has the SAP, let \mathcal{A} be the set whose elements are the symbols X , Y , and one symbol S for each maximal turnpike $S \in \mathcal{MT}(\Sigma)$. If Σ does not have the SAP, let \mathcal{A} consist of X and Y alone.

For each integer $n \geq 0$, let K_n be the subset of $\text{Contr}(p)$ which consists of those points $x \in \text{Contr}(p)$ which can be steered to p in time $\leq n$, i.e.,

$$(3.7) \quad K_n = \{x: x \in \text{Contr}(p) \wedge V(x) \leq n\}.$$

Then K_n is compact. Let \mathcal{A}_n be the subset of \mathcal{A} whose elements are X , Y , and—if Σ has the SAP—all the maximal turnpikes S of Σ such that $\text{Clos } S$ meets K_n . Since $\mathcal{MT}(\Sigma)$ is locally finite, it is clear that \mathcal{A}_n is a finite set. We claim that:

- (3.VI) For each n , there exists an integer $N(n) \geq 0$ such that, if $x \in K_n$, then x can be time-optimally steered to p by means of a trajectory γ which is a concatenation of at most $N(n)$ pieces, each of which is a ξ -trajectory for some $\xi \in \mathcal{A}_n$.

To prove this, we use our theorems on the structure of trajectories. If Δ has the SAP, then it follows from Theorem 3.18 of [B] that every $q \in K_n$ has a neighborhood $U(q)$ such that there is a $\nu(q)$ with the property that, if γ is a time-optimal trajectory in $U(q)$, then γ is a concatenation of at most $\nu(q)$ pieces, each of which is either an X -trajectory, or a Y -trajectory, or an S -trajectory for some $S \in \mathcal{MT}(\Sigma)$. (By definition, if $S \in \mathcal{MT}(\Sigma)$ an S -trajectory is a trajectory of Σ which is contained in $\text{Clos } S$.) Clearly,

we can choose $U(q)$ so that $U(q) \cap \text{Clos } S = \emptyset$ for all $S \in \mathcal{MT}(\Sigma)$ such that $(\text{Clos } S) \cap K_n = \emptyset$. Then, if γ is time-optimal in $U(q)$, γ is a concatenation of at most $\nu(q)$ pieces, each of which is a ξ -trajectory for a $\xi \in \mathcal{A}_n$. Next choose a finite covering of K_n , of the form $\mathcal{U} = \{U(q_1), \dots, U(q_m)\}$, and let $\bar{\nu} = \max(\nu(q_1), \dots, \nu(q_m))$. Let $\alpha > 0$ be a Lebesgue number for \mathcal{U} (i.e., an $\alpha > 0$ such that, whenever a set $J \subseteq K_n$ has diameter $\leq \alpha$, with respect to some fixed metric, then J is contained in one of the $U(q_i)$). Let $C > 0$ be such that $\|F(x)\| + \|G(x)\| \leq C$ for $x \in K_n$, and let $\beta > 0$ be such that $\beta C < \alpha$. Let μ be an integer such that $\mu\beta > n$. If $x \in K_n$, let $\gamma: [-V(x), C] \rightarrow M$ be a time-optimal trajectory from x to p . Then γ is a concatenation of at most μ pieces γ_j , such that $\text{Dom}(\gamma_j) = I_j$ is an interval of length $< \beta$. Therefore $\gamma_j(I_j)$ has diameter $< \alpha$. So each $\gamma_j(I_j)$ is contained in one of the $U(q_i)$. Therefore each γ_j is a concatenation of at most $\bar{\nu}$ pieces, each of which is a ξ -trajectory for some $\xi \in \mathcal{A}_n$. So γ is a concatenation of at most $N(n)$ pieces, each of which is a ξ -trajectory for some $\xi \in \mathcal{A}_n$, if we take $N(n) = \mu\bar{\nu}$. This completes the proof of (3.VI) when Σ has the SAP. If Σ does not have the SAP, i.e., if $\Delta_B \equiv 0$ on M , then the proof of (3.VI) is almost identical, except that we use Theorem 4.2 of [B] instead of Theorem 3.18 of [B]. For each $q \in K_n$ we choose $\nu(q)$, $U(q)$ such that, whenever it is possible to go time-optimally from an $r_1 \in U(q)$ to an $r_2 \in U(q)$ by means of a trajectory in $U(q)$, then one can go time-optimally from r_1 to r_2 by means of a trajectory which is bang-bang with at most $\nu(q) - 1$ switchings. The finite cover \mathcal{U} , and the numbers $\alpha, C, \beta, \bar{\nu}, \mu, N(n)$ are defined as before. If $x \in K_n$, one can find a time-optimal $\gamma: [-V(x), 0]$ from x to p , and then express γ as before as a concatenation of at most μ γ_j 's, each of which is contained in one $U(q_i)$. Therefore there is a γ_j^* that has the same endpoints as γ_j , is defined on the same domain as γ_j , and is bang-bang with at most $\bar{\nu} - 1$ switchings. If γ^* denotes the concatenation of the γ_j^* , then γ^* goes from x to p in the same time as γ , and γ^* is a concatenation of no more than $N(n)$ pieces, each of which is in $\text{Traj}(X \vee Y)$. So (3.VI) holds in this case as well.

If $\xi = (\xi_1, \dots, \xi_m)$ is a finite sequence of elements of \mathcal{A} , define a set

$$D(\xi) \subseteq M \times \mathbb{R}$$

to be the set of all pairs (x, t) such that x can be steered to p in time t by means of a trajectory γ which is a concatenation $\gamma_m * \dots * \gamma_1$, where each γ_i is a ξ_i trajectory. Then we have, if η denotes the sequence (ξ_2, \dots, ξ_m) :

$$(3.8) \quad (x, t) \in D(\xi) \Leftrightarrow (\exists s)(0 \leq t \leq s \wedge (\exists y)((x, y, s) \in \mathcal{F}^{\xi_1} \wedge (y, t-s) \in D(\eta))).$$

If (x, t) remains in a compact subset of $M \times \mathbb{R}$, then t is bounded by an n , and therefore the y which appears in the formula, if it exists at all, must actually be in K_n . So all the quantifiers in (3.8) are bounded. Therefore it follows, by induction, that all the $D(\xi)$ are subanalytic in $M \times \mathbb{R}$. Moreover, it is clear that they are closed.

We now define, by induction on n , a family Γ^n of subanalytic feedback flows on K_n , such that Γ^n steers all of K_n to p , and that each trajectory of Γ^n is time-optimal. Moreover, the Γ^n will be so constructed that Γ^n agrees with Γ^{n-1} on K_{n-1} .

The definition of Γ^0 is obvious. Suppose that $\Gamma^0, \dots, \Gamma^n$ have been constructed. Let $\Gamma_x^n = (u_x^n, \gamma_x^n)$, for $x \in K_n$. The set

$$G_n = \{(x, t, y): y = \gamma_x^n(t)\}$$

is subanalytic in $M \times M \times \mathbb{R}$, and is relatively compact (because it is contained in $K_n \times [-n, 0] \times K_n$). So its projection via $(x, t, y) \rightarrow x$ is subanalytic. Therefore K_n is subanalytic in M .

Let ξ_1, \dots, ξ_r be the elements of \mathcal{A}_{n+1} . For $\rho = 1, \dots, N(n+1)$, let E'_ρ denote the set of all elements $x \in K_{n+1} - K_n$ such that there is a time-optimal trajectory γ

which steers x to some element y of K_n in time $V(x) - V(y)$, and which is a concatenation of at most ρ pieces, each of which is a ξ -trajectory for some $\xi \in \mathcal{A}_{n+1}$. It is clear that the sets E'_ρ increase with ρ , and that $E'_{N(n+1)} = K_{n+1} - K_n$. Let $E_\rho = E'_\rho \cup K_n$.

We will construct Γ^{n+1} by constructing, for each $\rho \in \{0, \dots, N(n+1)\}$, a subanalytic feedback $\Gamma^{n+1, \rho}$ on E_ρ , which steers all of E_ρ to p , and which is such that each trajectory of $\Gamma^{n+1, \rho}$ is time-optimal. Moreover, the $\Gamma^{n+1, \rho}$ will be so constructed that $\Gamma^{n+1, 0} = \Gamma^n$, and that each $\Gamma^{n+1, \rho}$, $\rho > 0$, agrees with $\Gamma^{n+1, \rho-1}$ on $E_{\rho-1}$. Once this construction is carried out, then we can take Γ^{n+1} to be $\Gamma^{n+1, N(n+1)}$.

It is clear that $\Gamma^{n+1, 0}$ must be taken to be Γ^n . Assume that $0 \leq \rho < N(n+1)$, and that $\Gamma^{n+1, 0}, \dots, \Gamma^{n+1, \rho}$ have been constructed. We must construct $\Gamma^{n+1, \rho+1}$. First observe that, if $\Gamma^{n+1, \rho} = (u_x^{n+1, \rho}, \gamma_x^{n+1, \rho})$ for $x \in E_\rho$, then E_ρ is the projection via $(x, t, y) \rightarrow x$ of the relatively compact subanalytic subset

$$G_{n, \rho} = \{(x, t, y) : y = \gamma_x^{n+1, \rho}(t)\}$$

of $M \times M \times \mathbb{R}$. So E_ρ is subanalytic in M . Moreover, we have $t = -V(x)$, $x \in E_\rho$, if and only if $(x, t, p) \in G_{n, \rho}$. So the graph of $V|_{E_\rho}$ is subanalytic in $M \times \mathbb{R}$.

Now let

$$H_\rho = E_{\rho+1} - E_\rho.$$

A point x belongs to H_ρ iff it belongs to K_{n+1} and it can be steered in time $V(x) - V(y)$ to some $y \in K_n$ by a trajectory γ which is a concatenation of $\rho + 1$ pieces, each of which is a ξ -trajectory for some $\xi \in \mathcal{A}_{n+1}$, but it cannot be steered in time $V(x) - V(y)$ to any $y \in K_n$ by a γ which only involves ρ pieces. If γ is a trajectory with $\rho + 1$ pieces which steers x to $y \in K_n$ in time $V(x) - V(y)$, let \tilde{x} be the point where the first piece of γ ends. If δ steers y to p time-optimally (i.e., in time $V(y)$), then the concatenation $\delta * \gamma$ steers x to p in time $V(x)$; i.e., time-optimally. If $\tilde{\gamma}$ is the result of eliminating from γ its first piece, then $\delta * \tilde{\gamma}$ is time-optimal. Therefore $\tilde{\gamma}$ steers \tilde{x} to y in time $V(\tilde{x}) - V(y)$.

Since $\tilde{\gamma}$ is a concatenation of ρ pieces, each of which is a ξ -trajectory for a $\xi \in \mathcal{A}_{n+1}$, we conclude that $\tilde{x} \in E_\rho$. Therefore, if $x \in H_\rho$, then:

(3.VII) x can be steered to a $y \in E_\rho$ in time $V(x) - V(y)$ by means of a trajectory γ which is a ξ -trajectory for some $\xi \in \mathcal{A}_{n+1}$.

Conversely, it is easy to see that, if $x \in K_{n+1}$, $x \notin E_\rho$, then (3.VII) implies that $x \in H_\rho$.

For each $\sigma = 0, \dots, r$, let H_ρ^σ denote the set of those $x \in K_{n+1}$ that can be steered to some $y \in E_\rho$ in time $V(x) - V(y)$ by means of a γ which is a ξ_i -trajectory for some $i \in \{1, \dots, \sigma\}$. Then

$$\emptyset = H_\rho^0 \subseteq H_\rho^1 \subseteq \dots \subseteq H_\rho^r = H_\rho.$$

If we let $L_\rho^\sigma = H_\rho^\sigma \cup E_\rho$, then $E_\rho = L_\rho^0 \subseteq L_\rho^1 \subseteq \dots \subseteq L_\rho^r = E_{\rho+1}$.

We will construct $\Gamma^{n+1, \rho+1}$ by constructing, by induction on σ , subanalytic feedbacks $\Gamma^{n+1, \rho+1, \sigma}$ that steer all of L_ρ^σ to p and are time-optimal. Moreover, we will construct them in such a way that $\Gamma^{n+1, \rho+1, \sigma}$ agrees with $\Gamma^{n+1, \rho+1, \sigma-1}$ on $L_\rho^{\sigma-1}$, and that $\Gamma^{n+1, \rho+1, 0} = \Gamma^{n+1, \rho}$.

Assuming that $\sigma < r$, and that $\Gamma^{n+1, \rho+1, 0}, \dots, \Gamma^{n+1, \rho+1, \sigma}$ have been defined, let us define $\Gamma^{n+1, \rho+1, \sigma+1}$. If $x \in L_\rho^\sigma$, then we choose $\Gamma_x^{n+1, \rho+1, \sigma+1}$ to be $\Gamma_x^{n+1, \rho+1, \sigma}$. If $x \in L_\rho^{\sigma+1}$ but $x \notin L_\rho^\sigma$, then x can be steered to some $y \in E_\rho$ in time $V(x) - V(y)$ by means of a $\xi_{\sigma+1}$ -trajectory. Let δ be the maximal $\xi_{\sigma+1}$ -trajectory through x , parametrized in such a way that $\delta(-V(x)) = x$. Then there is a $\bar{t} > -V(x)$ such that $\delta(\bar{t}) = y \in E_\rho$, and that $\bar{t} = -V(y)$. Let $\bar{\delta} = \delta|_{[-V(x), \bar{t}]}$. If $\eta : [-V(y), 0] \rightarrow M$ steers y time-optimally to p , then $\eta * \bar{\delta}$ is time-optimal. Therefore $V(\delta(t)) = -t$ for $-V(x) \leq t \leq \bar{t}$.

It is clear that $\delta(\bar{t}) \in L_\rho^\sigma$. Let J be the set of those $t \in [-V(x), \bar{t}]$ such that $\delta(t) \in L_\rho^\sigma$. Let $t^* = \inf J$. We prove that $t^* \in J$. Let $t_m \searrow t^*$ be such that $t_m \in J$. Then $\delta(t_m) \in L_\rho^\sigma$, so that $\delta(t_m)$ can be steered to some $y_m \in E_\rho$ in time $\tau_m = V(\delta(t_m)) - V(y_m)$ by means of a trajectory γ_m that is a $\xi_{i(m)}$ -trajectory for some $i(m) \in \{1, \dots, \sigma\}$. By replacing $\{t_m\}$ by a subsequence, if necessary, we may assume that all the $i(m)$ are equal to a fixed $\bar{i} \in \{1, \dots, \sigma\}$.

Since $y_m \in E_\rho$, it is possible to steer y_m to a $z_m \in K_n$ in time $V(y_m) - V(z_m)$ by means of a trajectory γ'_m which is a concatenation of ρ pieces $\gamma'_{m,1}, \dots, \gamma'_{m,\rho}$, such that each $\gamma'_{m,j}$ is a $\xi_{i(m,j)}$ -trajectory for some $\xi_{i(m,j)} \in \mathcal{A}_{n+1}$. Since there is only a finite number of sequences $(i(1), \dots, i(\rho))$ of elements of $\{1, \dots, r\}$, we may assume (after passing to a subsequence, if necessary) that there is a fixed sequence $i^*(1), \dots, i^*(\rho)$, such that all the $\gamma'_{m,j}$ are $\xi_{i^*(j)}$ -trajectories for all m , and all j . Finally select, for each m , a time-optimal trajectory γ''_m that goes from z_m to p .

By passing to a subsequence, we may assume that the z_m, y_m, τ_m converge, as well as the times along all the $\gamma'_{m,j}$, and all the endpoints of $\gamma'_{m,j}$. Moreover, we can assume that the trajectories $\gamma''_m * \gamma'_m * \gamma_m$ converge uniformly. The limiting trajectory η is a concatenation $\gamma'' * \gamma' * \gamma$, where: (a) γ is a $\xi_{\bar{i}}$ -trajectory that steers $\delta(t^*)$ to $y = \lim y_m$, (b) γ' is a concatenation $\gamma'_1 * \dots * \gamma'_\rho$, where each γ'_j is a $\xi_{i^*(j)}$ -trajectory, and, moreover, γ' steers y to $z = \lim z_m$, (c) γ'' steers z to p .

Since $\gamma''_m * \gamma'_m * \gamma_m$ steers $\delta(t_m)$ to p in time $V(\delta(t_m))$ (i.e., $-t_m$), it follows that $\gamma'' * \gamma' * \gamma$ steers $\delta(t^*)$ to p in time $-t^*$, i.e., in time $V(\delta(t^*))$. So $\gamma'' * \gamma' * \gamma$ is time-optimal. Therefore γ' steers y to z in time $V(y) - V(z)$. Since $z_m \in K_n$, we conclude that $z \in K_n$, and so $y \in E_\rho$. Moreover, γ steers $\delta(t^*)$ to y in time $V(\delta(t^*)) - V(y)$. Since $\delta(t^*) \in K_{n+1}$, it follows that $\delta(t^*) \in L_\rho^\sigma$. Therefore $t^* \in J$, as asserted.

So, if $x \in L_\rho^{\sigma+1}$ but $x \notin L_\rho^\sigma$, the trajectory δ defined above is entirely contained in $L_\rho^{\sigma+1} - L_\rho^\sigma$, until a first time t^* is reached when $\delta(t^*) \in L_\rho^\sigma$. We define, for such an x , the trajectory $\gamma_x^{n+1, \rho+1, \sigma+1}$ as follows: we follow δ up to the time t^* , and then we follow $\gamma_{\delta(t^*)}^{n+1, \rho+1, \sigma}$. This defines a family of trajectories, one for each $x \in L_\rho^{\sigma+1}$, and it is clear that the trajectories obtained in this way constitute a compatible family and that the feedback control law so defined agrees with $\Gamma^{n+1, \rho+1, \sigma}$ on L_ρ^σ . This completes the definition of $\Gamma^{n+1, \rho+1, \sigma+1}$. The only remaining point is to show that $\Gamma^{n+1, \rho+1, \sigma+1}$ is subanalytic. Naturally, it is a part of our inductive hypothesis that $\Gamma^{n+1, \rho+1, \sigma}$ is subanalytic.

A triple (x, t, y) satisfies $y = \gamma_x^{n+1, \rho+1, \sigma+1}(t)$ if and only if one of the following conditions holds:

- (3.VIIIi) $y = \gamma_x^{n+1, \rho+1, \sigma}(t)$;
- (3.VIIIii) There exist τ, z, s such that:
 - (a) $z \in L_\rho^\sigma$ and $\tau > 0, \tau \leq n+1$,
 - (b) $(x, z, \tau) \in \mathcal{F}^{\xi_{\sigma+1}}$,
 - (c) there is no pair (z', τ') such that $0 < \tau' < \tau, z' \in L_\rho^\sigma$, and $(x, z, \tau') \in \mathcal{F}^{\xi_{\sigma+1}}$,
 - (d) $z = \gamma_x^{n+1, \rho+1, \sigma}(s)$ and $-n-1 \leq s \leq 0$,
 - (e) there exists no sequence ξ of elements of \mathcal{A}_{n+1} , of length $\leq N(n+1)$ such that there is an s' for which $0 \leq s' < \tau - s$ and $(x, s') \in D(\xi)$,
 - (f) either $t \geq s$ and $y = \gamma_z^{n+1, \rho+1, \sigma}(t)$, or $t < s$ and $(x, y, \tau - s + t) \in \mathcal{F}^{\xi_{\sigma+1}}$.

From this restatement of the definition of $\Gamma^{n+1, \rho+1, \sigma+1}$, together with the subanalyticity of all the sets \mathcal{F}^ξ , it follows that $\Gamma^{n+1, \rho+1, \sigma+1}$ is indeed a subanalytic feedback.

As explained above, this completes the inductive construction of the $\Gamma^{n+1,\rho+1,\sigma}$ for a given n, ρ . Then the inductive construction of the $\Gamma^{n+1,\rho}$ for a given n is also complete, and so the construction of the Γ^n is complete. We now define Γ by letting $\Gamma_x = \Gamma_x^n$ if $x \in K_n$. Then Γ is an optimal feedback control law on $\text{Contr}(p)$, with target p , and is subanalytic. As explained above, the theorem of [Su 6] then implies the existence of a regular synthesis.

As for the last assertion of our statement, it follows from the results of [Su 6]. Indeed, the partition \mathcal{P} constructed in the proof of the main theorem of [Su 6] is finite on each compact set K such that the time $\tau(x)$ along γ_x is bounded for $x \in K$, and that all the $\gamma_x, x \in K$, are Lipschitzian in t with a fixed constant. Clearly, the sets K_T satisfy this condition.

We conclude this section with a brief outline of an obvious extension of Theorem 3.1. If we have an analytic system Σ as has been considered here, and if we want to minimize an integral $\int L(x(t)) dt$, where L is analytic and $L(x) > 0$ for all x , then we can form the system Σ' given by

$$\dot{x} = \frac{F(x)}{L(x)} + u \frac{G(x)}{L(x)}.$$

The trajectories of Σ' are exactly the Σ -trajectories, reparametrized by cost. So, if the time-optimal problem for Σ' has a regular synthesis, we get a regular synthesis for the problem of minimizing $\int L$ for Σ . The precise hypotheses required are: (a) analyticity of F, G, L , (b) strict positivity of L and (c) a "nonexplosion" condition. The non-explosion condition is exactly that of Theorem 3.1, except that "time" has to be replaced by "cost" throughout. (That is, we must assume that, for each $T > 0$, there is a compact $K(T)$ such that, whenever γ steers an $x \in \text{Contr}(p)$ to p with cost less than T , then γ is entirely contained in $K(T)$.)

The conclusion then is the same as that of Theorem 3.1, except that the definition of regular synthesis has to be modified, so as to substitute "cost" for "time."

REFERENCES

- [A] H. J. SUSSMANN, *The structure of time-optimal trajectories for single-input systems in the plane: the C^∞ nonsingular case*, this Journal, 25 (1987), pp. 433-465.
- [B] ———, *The structure of time-optimal trajectories for single-input systems in the plane: the general real-analytic case*, this Journal, 25 (1987) pp. 868-904.
- [Su 3] ———, *Les semigroupes sousanalytiques et la régularité des commandes en boucle fermée*, Astérisque, Soc. Math. de France, 75-76 (1980), pp. 219-226.
- [Su 6] ———, *Subanalytic sets and regular synthesis*, to appear.
- [Ba] M. BAYTMAN, *The Optimal Synthesis of Trajectories in the Plane*, Zinatne, Riga (USSR), 1971. (In Russian.)
- [Bo] V. BOLTYANSKII, *Sufficient conditions for optimality and the justification of the dynamic programming method*, SIAM J. Control, 4 (1966), pp. 326-361.
- [Br 1] P. BRUNOVSKY, *Every normal linear system has a regular time-optimal synthesis*, Math. Slovaca, 28 (1978), pp. 81-100.
- [SuJ] H. J. SUSSMANN AND V. JURDJEVIC, *Controllability of nonlinear systems*, J. Differential Equations, 12 (1972), pp. 95-116.

OPTIMAL CONTROL FOR A BACKWARD PARABOLIC SYSTEM*

P. H. RIVERA†‡ AND C. F. VASCONCELLOS†

Abstract. In this paper the authors study two problems of optimal control for systems governed by a backward parabolic equation, when the \mathcal{U}_{ad} has nonempty interior.

Key words. optimal control, backward equation, nonwell-posed problem

AMS(MOS) subject classifications. 49A22, 35K99

1. Introduction. The objective of this paper is to present certain results on two problems of optimal control related to the partial differential operator $\partial z / \partial t + m(t) \Delta z$, in which $m: [0, T] \rightarrow \mathbb{R}$ is a smooth function changing the sign in $]0, T[$.

In J. L. Lions [4] we find the origin of the above problems, where he solved the control problem, that is, existence of optimal couple and optimality system for the backward parabolic problem:

$$\begin{aligned} (1.1) \quad & \frac{\partial z}{\partial t} + \Delta z = v \quad \text{in } Q, \\ & z = 0 \quad \text{on } \Sigma, \\ & z(0) = z_0 \quad \text{in } \Omega \end{aligned}$$

where Ω is a bounded open subset of \mathbb{R}^n with smooth boundary Γ and Q is the cylinder $\Omega \times]0, T[$, with lateral boundary $\Sigma = \Gamma \times]0, T[$. Lions considered $L^2(Q)$ as the admissible convex set for his problem.

In [4], Lions also proposed studying the same type of problem, where instead of constant coefficient $m = 1$, as in the case (1.1), we have the following state equation:

$$\begin{aligned} (1.2) \quad & \frac{\partial z}{\partial t} + m(t) \Delta z = v \quad \text{in } Q, \\ & z = 0 \quad \text{on } \Sigma, \\ & z(0) = z_0 \quad \text{in } \Omega \end{aligned}$$

where m is defined as above.

In Rivera [8] it is shown that the problem (1.1) is well posed when we consider some special functional space.

The functional space introduced by Rivera is defined as follows:

$$\text{Let } A = -\Delta \text{ with domain } D(A) = H_0^1(\Omega) \cap H^2(\Omega).$$

Let

$$W = \left\{ u \in \bigcap_{\nu=0}^{\infty} D(A^\nu) : \sum_{\nu=0}^{+\infty} \frac{1}{\nu!} |A^\nu u|_\Omega^2 < +\infty \right\}.$$

* Received by the editors December 20, 1983; accepted for publication (in revised form) June 5, 1986. This work was supported in part by Conselho de Ensino Para Graduados, Universidade Federal do Rio de Janeiro and Fundo Nacional Desenvolvimento Científico Tecnológico.

† Instituto de Matemática, Universidade Federal do Rio de Janeiro, C.P. 68530, Rio de Janeiro, RJ, Brazil.

‡ Deceased.

This is a Hilbert space with the norm

$$\|u\|_1^2 = \sum_{\nu=0}^{+\infty} \frac{1}{\nu!} |A^\nu u|_\Omega^2, \quad u \in W$$

where we represent by $|\cdot|_\Omega$ the norm of $L^2(\Omega)$.

Since A has a discrete spectrum, then W contains all the eigenfunctions of A ; hence W is dense in $L^2(\Omega)$. For related subjects see also Medeiros [7].

In this paper we prove, in § 2, the existence and uniqueness of a solution of the backward parabolic problem (1.2) when we consider the special space W which was defined above.

In § 3 we study the optimal control problems. First of all we consider a closed convex subset K of $H^{-1}(\Omega)$ and the following state equation:

$$\begin{aligned} \frac{\partial z}{\partial t} + m(t) \Delta z &= v \quad \text{in } Q, \\ z &= 0 \quad \text{on } \Sigma, \\ z(0) &\in K. \end{aligned} \tag{1.3}$$

The cost functional is defined by:

$$J(v, z) = |z - z_d|_Q^2 + N|v|_Q^2, \quad (v, z) \in L^2(Q) \times L^2(Q) \tag{1.4}$$

with $z_d \in L^2(Q)$ and $N > 0$.

To complete this paper, we study the final state optimal control which the state equation is given by:

$$\begin{aligned} \frac{\partial z}{\partial t} + m(t) \Delta z &= f \quad \text{in } Q, \\ z &= 0 \quad \text{on } \Sigma, \\ z(0) &= v \quad \text{in } \Omega. \end{aligned} \tag{1.5}$$

The cost functional is defined by:

$$J(v, z) = |z - z_d|_Q^2 + N\|v\|_{H^{-1}}^2, \quad (v, z) \in H^{-1}(\Omega) \times L^2(Q) \tag{1.6}$$

where z_d belongs to $L^2(Q)$ and $N > 0$.

In both problems we show the existence of a couple (u, y) such that $J(u, y) = \inf J(v, z)$, where $\{v, z\}$ are in suitable function spaces.

Moreover, we give necessary conditions for a couple $\{u, y\}$ to be an optimal couple; that is, we show that if a couple $\{u, y\}$ is an optimal couple, there exists an adjoint state p satisfying a system which will be defined in § 3 and is known by the optimality system.

We must note that the optimality system is crucial for the numerical approach to the optimal problem.

In this paper we denote $L^2(\Omega)$ by H and its scalar product by $(\cdot | \cdot)_\Omega$.

We denote by $H^1(\Omega)$ the Sobolev space of order one on Ω and by $H_0^1(\Omega)$ the closure in $H^1(\Omega)$ of the infinitely differentiable functions on Ω with compact support.

We represent $H_0^1(\Omega)$ by V and the norm in V shall be denoted by $\|\cdot\|$.

2. Existence of solution for the backward parabolic problem. We shall prove in this section that the system (1.2) is well posed when we pick up z_0 in W and v in $L^2(0, T; W)$.

For this we assume the following:

(2.1) Let $m(t)$ be a real function defined on $[0, T]$, $T > 0$, satisfying the following conditions:

- (i) m is continuously differentiable on $[0, T]$;
- (ii) m is monotonic decreasing in this interval;
- (iii) There exists $0 < T_0 < T$ such that $m(T_0) = 0$.

THEOREM 2.1. Suppose given $z_0 \in W$ and $v \in L^2(0, T; W)$. Then there exists a unique vector function $z: [0, T] \rightarrow H$ satisfying the following conditions:

$$(2.2) \quad z \in L^\infty(0, T; D(A)) \cap C([0, T]; V), \quad z' \in L^2(Q);$$

(2.3) There exists $c_0 > 0$ such that

$$\int_0^T |m(s)| \|z(s)\|^2 ds < c_0;$$

$$(2.4) \quad z' - m(t)Az = v \quad \text{in } L^2(Q);$$

$$(2.5) \quad z(0) = z_0.$$

Moreover the linear mapping $(z_0, v) \rightarrow (z, z')$ is continuous from $W \times L^2(0, T; W)$ into $L^\infty(0, T; D(A)) \times L^2(Q)$.

Before proving Theorem 2.1 we shall prove the following two propositions.

Let us represent by m_0, v_0 the restrictions, respectively, of m, v to the interval $[0, T_0]$.

PROPOSITION 2.1. There exists a unique vector function $x: [0, T_0] \rightarrow H$ such that

$$(2.6) \quad x \in C([0, T_0]; D(A)), \quad x' \in L^2(0, T_0; H).$$

(2.7) There exists $c_1 > 0$ such that

$$\int_0^{T_0} |m_0(s)| \|x(s)\|^2 ds \leq c_1,$$

$$(2.8) \quad x' - m_0(t)Ax = v_0 \quad \text{in } L^2(0, T_0; H),$$

$$x(0) = z_0.$$

Moreover, the linear mapping $(z_0, v_0) \rightarrow (x, x')$ is continuous from $W \times L^2(0, T; W)$ into $C([0, T_0]; D(A)) \times L^2(0, T_0; H)$.

Proof. Let w_1, \dots, w_ν, \dots be the eigenvectors of the operator A .

For each $j = 1, 2, \dots$, let f_j and g_j be the functions defined by:

$$(2.9) \quad (i) \quad f_j(t) = (z_0 | w_j)_\Omega \exp \left(\lambda_j \int_0^t m_0(s) ds \right), \quad j = 1, \dots, \nu,$$

$$(ii) \quad g_j(t) = \int_0^t (v_0(s) | w_j)_\Omega \exp \left(\lambda_j \int_s^t m_0(\sigma) d\sigma \right) ds, \quad j = 1, \dots, \nu.$$

Then, the function $h_j(t) = f_j(t) + g_j(t)$ is absolutely continuous on $[0, T_0]$ and satisfies, almost everywhere, the following system:

$$(2.10) \quad (i) \quad h_j'(t) - \lambda_j m_0(t) h_j(t) = (v_0(t) | w_j)_\Omega \quad \text{on }]0, T_0[,$$

$$(ii) \quad h_j(0) = (z_0 | w_j)_\Omega, \quad j = 1, \dots, \nu$$

where $\lambda_j, j = 1, 2, \dots, \nu$ are eigenvalues of A .

Hence, we deduce that $x_\nu(t) = \sum_{j=1}^\nu h_j(t)w_j$ is a solution of the following system:

$$(2.11) \quad (i) \quad (x'_\nu(t)|w_j)_\Omega - m_0(t)(Ax_\nu(t)|w_j)_\Omega = (v_0(t)|w_j)_\Omega,$$

$$(ii) \quad x_\nu(0) = \sum_{j=1}^\nu (z_0|w_j)_\Omega w_j.$$

We note that

$$(2.12) \quad (i) \quad |f_j(t)|^2 \leq \alpha_1 \sum_{k=0}^\infty (k!)^{-1} |(A^k z_0|w_j)_\Omega|^2 = S_j,$$

$$(ii) \quad |g_j(t)|^2 \leq \alpha_1 T_0 \sum_{k=0}^\infty (k!)^{-1} \int_0^{T_0} |(A^k v_0(s)|w_j)_\Omega|^2 ds = D_j,$$

$$j = 1, \dots, \nu \quad \text{and} \quad 0 \leq t \leq T_0$$

where $c = T_0 m_0(0)$ and $\alpha_1 = \exp(c^2)$.

Hence, we conclude that

$$(2.13) \quad |x_\nu(t)|_\Omega^2 \leq K_1 K_0, \quad \nu = 1, 2, \dots, \quad 0 \leq t \leq T_0 \quad \text{where} \quad K_0 = \|z_0\|_1^2 + \int_0^{T_0} \|v_0(s)\|_1^2 ds \\ \text{and} \quad K_1 = 2\alpha_1(1 + T_0).$$

Therefore by (2.13) we obtain:

$$(2.14) \quad \int_0^t m_0(s) \|x_\nu(s)\|^2 ds \leq K_2 K_0, \quad \nu = 1, 2, \dots, \quad 0 \leq t \leq T_0$$

where $K_2 = \frac{1}{2}(K_1 + 1 + K_1 T_0)$.

On the other hand, using (2.9), we have

$$(2.15) \quad (i) \quad |f'_j(t)| \leq m_0(0) \sum_{k=0}^\infty c^k (k!)^{-1} |(A^{k+1} z_0|w_j)_\Omega|,$$

$$(ii) \quad |g'_j(t)| \leq |(v_0(t)|w_j)_\Omega| \\ + m_0(0) \sum_{k=0}^\infty c^k (k!)^{-1} \int_0^{T_0} |(A^{k+1} v_0(s)|w_j)_\Omega| ds.$$

If we choose $\varepsilon > 0$ such that $0 < s = 1 + \varepsilon < 2$, there is $\alpha_0(\varepsilon) = \alpha_0 > 0$ which satisfies, for $u \in W$:

$$(2.16) \quad \sum_{k=0}^\infty (k!)^{-s} |A^{k+1} u|_\Omega^2 \leq \alpha_0 \|u\|_1^2.$$

Thus by (2.15) it follows that x'_ν belongs to $L^2(0, T_0; H)$ and

$$(2.17) \quad \|x'_\nu\|_{L^2(0, T_0; H)}^2 \leq K_3 K_0, \quad \nu = 1, 2, \dots$$

where $K_3 = 2T_0 \alpha_2 \alpha_0(1 + 2T_0) + 4$ and $\alpha_2 = m_0^2(0) \sum_{k=0}^\infty c^{2k} / (k!)^{s-2}$.

Now, since the series $\sum_{j=1}^\infty S_j$ and $\sum_{j=1}^\infty D_j$ converge, we obtain by (2.12) that

$$(2.18) \quad \lim_{\nu \rightarrow +\infty} x_\nu(t) = x(t) = \sum_{j=1}^\infty h_j(t)w_j \quad \text{in } H \text{ and uniformly in } t \in [0, T_0].$$

By the same argument we obtain:

$$(2.19) \quad \lim_{\nu \rightarrow +\infty} Ax_\nu(t) = Ax(t) \quad \text{in } H, \text{ uniformly in } t \in [0, T_0].$$

Now, existence of the solution of (2.8) follows using standard methods.

To prove the uniqueness we use the theorem of Lions and Malgrange [6].

PROPOSITION 2.2. *Let y_0 be in $D(A)$, v_1 in $L^2(0, T_1; D(A))$ and $m_1: [0, T_1] \rightarrow \mathbb{R}$, be a continuous monotonic increasing function such that $m_1(0) = 0 < m_1(T_1)$, where $T_1 > 0$.*

Then, there exists a unique vector function $y: [0, T_1] \rightarrow H$ such that

$$(2.20) \quad \begin{aligned} y &\in C([0, T_1]; V) \cap L^\infty(0, T_1; D(A)), \\ y' &\in L^2(0, T_1; V). \end{aligned}$$

$$(2.21) \quad \text{There exists } c_2 > 0 \text{ such that}$$

$$(2.22) \quad \begin{aligned} \int_0^{T_1} m_1(t) |Ay(t)|_\Omega^2 dt &\leq c_2, \\ y' + m_1(t)Ay &= v_1 \quad \text{in } L^2(0, T_1; H), \\ y(0) &= y_0. \end{aligned}$$

Moreover the linear mapping $(y_0, v_1) \rightarrow (y, y')$ is continuous from $D(A) \times L^2(0, T_1; D(A))$ into $L^\infty(0, T_1; D(A)) \times L^2(0, T_1; V)$.

Proof. We prove this proposition using the usual methods.

Proof of Theorem 2.1. Let $T_1 = T - T_0$ and let y_0 be $x(T_0)$, where $x: [0, T_0] \rightarrow H$ satisfies Proposition 2.1. We also consider $v_1(t) = v(t + T_0)$, $t \in]0, T_1[$ and $m_1(t) = -m(t + T_0)$, $t \in [0, T_1]$.

Under the above assumptions there is $y: [0, T_1] \rightarrow H$ which satisfies Proposition 2.2.

Then, if we define $z: [0, T] \rightarrow H$ by:

$$z(t) = \begin{cases} x(t) & \text{if } t \in [0, T_0], \\ y(t - T_0) & \text{if } t \in [T_0, T], \end{cases}$$

we can prove that z is the unique function which satisfies Theorem 2.1.

Remark 2.1. In Proposition 2.1, if we consider in $W \times L^2(0, T_0; W)$ the norm of $H \times L^2(0, T_0; H)$, the linear mapping $(z_0, v_0) \mapsto (x, x')$ is not continuous.

In fact let $x_\nu(t)$ be defined by:

$$x_\nu(t) = \frac{1}{\lambda_\nu} \left\{ \exp \left(\lambda_\nu \int_0^t m_0(s) ds \right) - 1 \right\} w_\nu, \quad t \in [0, T_0].$$

Then, for each $\nu = 1, 2, \dots$, x_ν satisfies

$$\begin{aligned} x'_\nu(t) - m_0(t)Ax_\nu &= m_0(t)w_\nu \\ x_\nu(0) &= 0 \end{aligned}$$

and

$$\int_0^{T_0} |m_0(t)w_\nu|_\Omega^2 dt \leq T_0 |m_0(0)|^2.$$

But $\lim_{\nu \rightarrow +\infty} \|Ax_\nu\|_\infty = +\infty$.

3. Optimal control problems. In this section the function $m(t)$ is that defined by the conditions (2.1) and we suppose that it is $C^\infty(0, T)$ instead of (i).

Optimal control for system (1.3). Let \mathcal{U}_{ad} be a closed convex subset of $L^2(Q)$ with nonempty interior. We also assume that the convex subset K has nonempty interior.

Then by Theorem 2.1, it follows that $X_{\text{ad}} = \{(v, z) \in \mathcal{U}_{\text{ad}} \times L^2(Q) : (v, z) \text{ satisfies (1.3)}\}$ is a nonempty closed convex subset of $L^2(Q) \times L^2(Q)$. Therefore it follows by [1] that the problem

$$(3.1) \quad \inf \{J(v, z) : (v, z) \in X_{\text{ad}}\}$$

has a unique solution $(u, y) \in X_{\text{ad}}$, where $J(v, z)$ is defined by (1.4).

THEOREM 3.1. *If \mathcal{U}_{ad} and K have nonempty interiors, then the optimal couple $(u, y) \in X_{\text{ad}}$ is characterized by the set (u, y, p) , the solution of the following system:*

$$(3.2) \quad \begin{aligned} \frac{\partial y}{\partial t} + m(t) \Delta y &= u \quad \text{in } Q, \\ -\frac{\partial p}{\partial t} + m(t) \Delta p &= y - z_d \quad \text{in } Q; \end{aligned}$$

$$(3.3) \quad \begin{aligned} y &= 0, \quad p = 0 \quad \text{on } \Sigma, \\ p(x, T) &= 0 \quad \text{in } \Omega, \\ y(0) &\in K, \quad p(0) \in H_0^1(\Omega); \end{aligned}$$

$$(3.4) \quad \begin{aligned} \langle k - y(0), p(0) \rangle_{H^{-1}H_0^1} &\geq 0 \quad \forall k \in K, \\ (p + Nu|v - u)_Q &\geq 0 \quad \forall v \in \mathcal{U}_{\text{ad}}. \end{aligned}$$

Remark 3.1. We consider the inner product in $H^{-1}(\Omega)$ defined by $(q|k)_{H^{-1}} = (\Lambda^{-1}q|\Lambda^{-1}k)_{H_0^1}$, where $\Lambda: H_0^1(\Omega) \rightarrow H^{-1}(\Omega)$ is the canonical isomorphism.

Moreover we observe that if p and q belong to $H^{-1}(\Omega)$ and $\langle p, \psi \rangle_{H^{-1}H_0^1} = (q|\psi)_{H^{-1}}$, $\forall \psi \in \mathcal{D}(\Omega)$, then $p \in H_0^1(\Omega)$ and $p = \Lambda^{-1}q$.

Remark 3.2. If $\mathcal{U}_{\text{ad}} = L^2(Q)$ and $K = H^{-1}(\Omega)$, then $(3.4)_1$ and $(3.4)_2$ are similar to

$$(3.5) \quad p(0) = 0, \quad p + Nu = 0.$$

Remark 3.3. If $\mathcal{U}_{\text{ad}} = \{v \in L^2(Q) : \|v - v_0\| \leq r\}$ then (cf. Lions [3]) $(3.4)_2$ is similar to

$$u = v_0 - \frac{p + Nv_0}{\|p + Nv_0\|_Q} r.$$

Proof of Theorem 3.1. Let Z be the set defined by:

$$(3.6) \quad Z = \left\{ z \in L^2(Q) : \frac{\partial z}{\partial t} + m(t) \Delta z \in L^2(Q), z = 0 \text{ on } \Sigma \right\}$$

and for each $\varepsilon > 0$ we define

$$J_\varepsilon(v, z, k) = J(v, z) + \frac{1}{\varepsilon} \left\| \frac{\partial z}{\partial t} + m(t) \Delta z - v \right\|_Q^2 + \frac{1}{\varepsilon} \|z(0) - k\|_{H^{-1}}^2$$

where $(v, z, k) \in \mathcal{U}_{\text{ad}} \times Z \times K$.

We can prove that there exists $(u_\varepsilon, y_\varepsilon, k_\varepsilon)$ belonging to $\mathcal{U}_{\text{ad}} \times Z \times K$, which is a unique solution of the penalized problem:

$$J_\varepsilon(u_\varepsilon, y_\varepsilon, k_\varepsilon) = \inf \{J_\varepsilon(v, z, k) : (v, z, k) \in \mathcal{U}_{\text{ad}} \times Z \times K\}.$$

Moreover, $J_\varepsilon(u_\varepsilon, y_\varepsilon, k_\varepsilon) \leq J_\varepsilon(u, y, y(0)) = J(u, y)$ for all $\varepsilon > 0$; hence there is $c_0 > 0$ such that

$$(3.7) \quad |y_\varepsilon|_Q \leq c_0, \quad |u_\varepsilon|_Q \leq c_0 \quad \forall \varepsilon > 0,$$

$$(3.8) \quad \left\| \frac{\partial y_\varepsilon}{\partial t} + m(t) \Delta y_\varepsilon - u_\varepsilon \right\|_Q \leq \sqrt{\varepsilon} c_0, \quad \|k_\varepsilon - y_\varepsilon(0)\|_{H^{-1}} \leq \sqrt{\varepsilon} c_0 \quad \forall \varepsilon > 0.$$

Therefore, we can prove that

$$\begin{aligned}(u_\varepsilon, y_\varepsilon) &\rightarrow (u, y) \quad \text{strong in } L^2(Q) \times L^2(Q), \\ \lim_{\varepsilon \rightarrow 0^+} k_\varepsilon &= y(0) \quad \text{strong in } H^{-1}(\Omega), \\ \lim_{\varepsilon \rightarrow 0^+} J_\varepsilon(u_\varepsilon, y_\varepsilon, k_\varepsilon) &= J(u, y).\end{aligned}$$

On the other hand, since

$$J'_\varepsilon(u_\varepsilon, y_\varepsilon, k_\varepsilon) \cdot (v - u_\varepsilon, z - y_\varepsilon, k - k_\varepsilon) \geq 0, \quad (v, z, k) \in \mathcal{U}_{\text{ad}} \times Z \times K,$$

we obtain:

$$(3.9) \quad (y_\varepsilon - z_d | z)_Q = \left(p_\varepsilon \left| \frac{\partial z}{\partial t} + m(t) \Delta z \right|_Q + (q_\varepsilon | z(0))_{H^{-1}}, \quad z \in Z,$$

$$(3.10) \quad (p_\varepsilon + Nu_\varepsilon | v - u_\varepsilon)_Q \geq 0 \quad \forall v \in \mathcal{U}_{\text{ad}},$$

$$(3.11) \quad (q_\varepsilon | k - k_\varepsilon)_{H^{-1}} \geq 0 \quad \forall k \in K$$

where $p_\varepsilon = -1/\varepsilon \{ \partial y_\varepsilon / \partial t + m(t) \Delta y_\varepsilon - u_\varepsilon \}$ and $q_\varepsilon = -1/\varepsilon \{ y_\varepsilon(0) - k_\varepsilon \}$.

Now, since \mathcal{U}_{ad} and K have nonempty interiors there are $r > 0, v_0 \in \mathcal{U}_{\text{ad}} \cap L^2(0, T; W)$ and $k_0 \in K \cap W$ such that

$$S_r(v_0) = \{v \in L^2(Q) : |v - v_0|_Q \leq r\} \subset \mathcal{U}_{\text{ad}},$$

$$S_r(k_0) = \{k \in H^{-1}(\Omega) : \|k - k_0\|_{H^{-1}} \leq r\} \subset K.$$

Moreover, there is $z_0 \in L^2(Q)$ such that $(v_0, z_0) \in X_{\text{ad}}$ (cf. Theorem 2.1).

Using (3.9), we obtain:

$$(y_\varepsilon - z_d | y_\varepsilon - z_0)_Q + \varepsilon |p_\varepsilon|_Q^2 + \varepsilon \|q_\varepsilon\|_{H^{-1}}^2 = (p_\varepsilon | u_\varepsilon - v_0)_Q + (q_\varepsilon | k_\varepsilon - k_0)_{H^{-1}}.$$

Now, adding $(Nu_\varepsilon | u_\varepsilon - v_0)_Q$ to both sides of the above expression, we have by (3.7) and (3.8)

$$|(Nu_\varepsilon + p_\varepsilon | u_\varepsilon - v_0)_Q + (q_\varepsilon | k_\varepsilon - k_0)_{H^{-1}}| \leq C_1 \quad \forall \varepsilon > 0;$$

hence, by (3.10) and (3.11) we obtain:

$$(3.12) \quad (p_\varepsilon + Nu_\varepsilon | v - v_0)_Q + (q_\varepsilon | k - k_0)_{H^{-1}} \geq -C_1 \quad \forall (v, k) \in \mathcal{U}_{\text{ad}} \times K.$$

On the other hand, if $w_\varepsilon = p_\varepsilon + Nu_\varepsilon$ and $w_\varepsilon \neq 0$ (if $w_\varepsilon = 0$ then p_ε is a bounded family) then $v = v_0 - rw_\varepsilon / |w_\varepsilon|_Q \in \mathcal{U}_{\text{ad}}$ and $k = k_0 - rq_\varepsilon / |q_\varepsilon|_{H^{-1}} \in K$.

Hence, replacing in (3.12) we have $C > 0$ such that:

$$|p_\varepsilon|_Q \leq C, \quad \|q_\varepsilon\|_{H^{-1}(\Omega)} \leq C \quad \forall \varepsilon > 0.$$

Therefore, there are p in $L^2(Q)$, q in $H^{-1}(\Omega)$ and subsequences also denoted by $\{p_\varepsilon\}$ and $\{q_\varepsilon\}$ such that

$$p_\varepsilon \rightarrow p \quad \text{weak in } L^2(Q),$$

$$q_\varepsilon \rightarrow q \quad \text{weak in } H^{-1}(\Omega).$$

Passing to the limit in (3.9) and (3.11), we obtain:

$$(3.13) \quad (y - z_d | z)_Q = \left(p \left| \frac{\partial z}{\partial t} + m(t) \Delta z \right|_Q + (q | z(0))_{H^{-1}} \quad \forall z \in Z,$$

$$(3.14) \quad (k - y(0) | q)_{H^{-1}} \geq 0 \quad \forall k \in K.$$

From (3.13) we have (3.2)₂, (3.3)₁ and (3.3)₂; moreover we also have that $\langle p(0), \psi \rangle_{H^{-1}H_0^1} = (q|\psi)_{H^{-1}}$, $\psi \in \mathcal{D}(\Omega)$. Hence by Remark 3.1 and by (3.14) we obtain (3.3)₃ and (3.4)₁.

Finally passing to the limit in (3.10), we obtain (3.4)₂.

Remark 3.4. If $K = \{0\}$, Theorem 3.1 is the same, except for (3.3)₃ and (3.4)₁ which are replaced by $y(0) = 0$.

Optimal control for system (1.5). Let \mathcal{U}_{ad} be a closed convex subset of $H^{-1}(\Omega)$ with nonempty interior; we also consider $f \in L^2(0, T; W) \subset L^2(Q)$ fixed.

Then, by Theorem 2.1 it follows that $X_{\text{ad}} = \{(v, z) \in \mathcal{U}_{\text{ad}} \times L^2(Q) : (v, z) \text{ satisfies (1.5)}\}$ is a nonempty closed convex subset of $H^{-1}(\Omega) \times L^2(Q)$. Therefore, it follows by Lions [1] that the problem

$$(3.15) \quad \inf \{J(v, z) : (v, z) \in X_{\text{ad}}\}$$

has a unique solution $(u, y) \in X_{\text{ad}}$ where $J(v, z)$ is defined by (1.6).

THEOREM 3.2. *If \mathcal{U}_{ad} has a nonempty interior, then the optimal couple (u, y) is characterized by the solution (u, y, p) of the following system:*

$$(3.16) \quad \begin{aligned} \frac{\partial y}{\partial t} + m(t) \Delta y &= f \quad \text{in } Q, \\ -\frac{\partial p}{\partial t} + m(t) \Delta p &= y - z_d \quad \text{in } Q, \end{aligned}$$

$$(3.17) \quad \begin{aligned} y &= p = 0 \quad \text{on } \Sigma, \\ p(x, T) &= 0 \quad \text{in } \Omega, \\ y(0) &= u, \quad p(0) \in H_0^1(\Omega), \end{aligned}$$

$$(3.18) \quad (\Lambda p(0) + Nu | v - u)_{H^{-1}} \geq 0 \quad \forall v \in \mathcal{U}_{\text{ad}}.$$

Remark 3.5. (i) If $\mathcal{U}_{\text{ad}} = H^{-1}(\Omega)$ then (3.18) is similar to

$$Nu + \Lambda p(0) = 0.$$

(ii) If $\mathcal{U}_{\text{ad}} = \{v \in H^{-1}(\Omega) : \|v - v_0\|_{H^{-1}} \leq r\}$ then

$$u = v_0 - \frac{\Lambda p(0) + Nv_0}{\|\Lambda p(0) + Nv_0\|_{H^{-1}}} r \quad (\text{cf. Lions [3]}).$$

Proof of Theorem 3.2. Let Z be the set defined by (3.6) and for each $\varepsilon > 0$ we define

$$J_\varepsilon(v, z) = J(v, z) + \frac{1}{\varepsilon} \left| \frac{\partial z}{\partial t} + m(t) \Delta z - f \right|_Q^2 + \frac{1}{\varepsilon} \|z(0) - v\|_{H^{-1}}^2$$

where $(v, z) \in \mathcal{U}_{\text{ad}} \times Z$.

We can prove that there exists $(u_\varepsilon, y_\varepsilon)$ in $\mathcal{U}_{\text{ad}} \times Z$ which is the unique solution of the penalized problem:

$$J_\varepsilon(u_\varepsilon, y_\varepsilon) = \inf \{J_\varepsilon(v, z) : (v, z) \in \mathcal{U}_{\text{ad}} \times Z\}.$$

Moreover, $J_\varepsilon(u_\varepsilon, y_\varepsilon) \leq J_\varepsilon(u, y) = J(u, y)$, for all $\varepsilon > 0$; hence there is $c_0 > 0$ such that

$$(3.19) \quad |y_\varepsilon|_Q \leq c_0, \quad \|u_\varepsilon\|_{H^{-1}} \leq c_0 \quad \forall \varepsilon > 0$$

$$(3.20) \quad \left| \frac{\partial y_\varepsilon}{\partial t} + m(t) \Delta y_\varepsilon - f \right|_Q \leq \sqrt{\varepsilon} c_0, \quad \|y_\varepsilon(0) - u_\varepsilon\|_{H^{-1}} \leq \sqrt{\varepsilon} c_0.$$

Therefore, we can prove that

$$(u_\varepsilon, y_\varepsilon) \rightarrow (u, y) \text{ strong in } H^{-1}(\Omega) \times L^2(Q),$$

$$\lim_{\varepsilon \rightarrow 0^+} J_\varepsilon(u_\varepsilon, y_\varepsilon) = J(u, y).$$

On the other hand, since

$$J'_\varepsilon(u_\varepsilon, y_\varepsilon) \cdot (v - u_\varepsilon, z - y_\varepsilon) \geq 0 \quad \forall (v, z) \in \mathcal{U}_{\text{ad}} \times Z,$$

we obtain:

$$(3.21) \quad (y_\varepsilon - z_d | z)_Q = \left(p_\varepsilon \left| \frac{\partial z}{\partial t} + m(t) \Delta z \right|_Q + (q_\varepsilon | z(0))_{H^{-1}} \right) \quad \forall z \in Z,$$

$$(3.22) \quad (q_\varepsilon + Nu_\varepsilon | v - u_\varepsilon)_{H^{-1}} \geq 0 \quad \forall v \in \mathcal{U}_{\text{ad}}$$

where

$$p_\varepsilon = -\frac{1}{\varepsilon} \left\{ \frac{\partial y_\varepsilon}{\partial t} + m(t) \Delta y_\varepsilon - f \right\} \quad \text{and} \quad q_\varepsilon = -\frac{1}{\varepsilon} \{ y_\varepsilon(0) - u_\varepsilon \}.$$

Now, since \mathcal{U}_{ad} has a nonempty interior, then there are $r > 0$, and $v_0 \in \mathcal{U}_{\text{ad}} \cap W$ such that

$$S_r(v_0) = \{v \in H^{-1}(\Omega) : \|v - v_0\|_{H^{-1}} \leq 0\} \subset \mathcal{U}_{\text{ad}}.$$

Moreover there is $z_0 \in L^2(Q)$ such that $(v_0, z_0) \in X_{\text{ad}}$ (cf. Theorem 2.1).

Using (3.21), we obtain:

$$(y_\varepsilon - z_d | y_\varepsilon - z_0)_Q + \varepsilon \|p_\varepsilon\|_Q^2 + \varepsilon \|q_\varepsilon\|_{H^{-1}}^2 = (q_\varepsilon | u_\varepsilon - v_0)_{H^{-1}}.$$

Adding $(Nu_\varepsilon | u_\varepsilon - v_0)_{H^{-1}}$ to both sides of the above expression, we have by (3.19) and (3.20):

$$|(q_\varepsilon + Nu_\varepsilon | u_\varepsilon - v_0)_{H^{-1}}| \leq C_1 \quad \forall \varepsilon > 0;$$

hence, by (3.22), we obtain

$$(3.23) \quad (q_\varepsilon + Nu_\varepsilon | v - v_0)_{H^{-1}} \geq -C_1 \quad \forall v \in \mathcal{U}_{\text{ad}}.$$

Since

$$v = v_0 - r \frac{Nu_\varepsilon + q_\varepsilon}{\|Nu_\varepsilon + q_\varepsilon\|_{H^{-1}}} \in \mathcal{U}_{\text{ad}}$$

we obtain by (3.23) $C > 0$ such that

$$(3.24) \quad \|q_\varepsilon\|_{H^{-1}} \leq C \quad \forall \varepsilon > 0.$$

Estimates for $\{p_\varepsilon\}$. By (3.21) and Remark 3.1 we obtain:

$$-\frac{\partial p_\varepsilon}{\partial t} + m(t) \Delta p_\varepsilon = y_\varepsilon - z_d \quad \text{in } \mathcal{D}'(Q),$$

$$p_\varepsilon = 0 \quad \text{on } \Sigma,$$

$$p_\varepsilon(T) = 0 \quad \text{in } \Omega,$$

$$p_\varepsilon(0) = \Lambda^{-1} q_\varepsilon \quad (\text{hence } p_\varepsilon(0) \in H_0^1(\Omega)).$$

We consider the interval $[0, T_0 - \delta]$ in which $m(t) > 0$; then (cf. Lions and Magenes [5]) p_ε belongs to $L^2(0, T_0 - \delta; H_0^1(\Omega))$, p'_ε belongs to $L^2(0, T_0 - \delta; L^2(\Omega))$. Moreover,

$$\frac{1}{2} \frac{d}{dt} \|p_\varepsilon(t)\|_\Omega^2 + m(t) \|p_\varepsilon(t)\|_{H_0^1} = (z_d - y_\varepsilon | p_\varepsilon)_Q, \quad 0 < t < T_0 - \delta;$$

hence by (3.19) and (3.24) there is $C > 0$ such that

$$(3.25) \quad \|p_\varepsilon\|_{L^2(0, T_0; L^2(\Omega))} \leq C \quad \forall \varepsilon > 0.$$

In the same way, if we consider the interval $[T_0 + \delta, T]$, since $m(t) < 0$, $T_0 + \delta \leq t \leq T$, and $p_\varepsilon(T) = 0$, then there exists $C > 0$ such that

$$(3.26) \quad \|p_\varepsilon\|_{L^2(T_0, T; L^2(\Omega))} \leq C \quad \forall \varepsilon > 0.$$

Therefore, by (3.24), (3.25) and (3.26), there are p in $L^2(Q)$, q in $H^{-1}(\Omega)$ and subsequences also denoted by $\{p_\varepsilon\}$ and $\{q_\varepsilon\}$ such that

$$\begin{aligned} p_\varepsilon &\rightarrow p \quad \text{weak in } L^2(Q), \\ q_\varepsilon &\rightarrow q \quad \text{weak in } H^{-1}(\Omega). \end{aligned}$$

Passing to the limit in (3.21) and (3.22), we have

$$(3.27) \quad (y - z_d | z)_Q = \left(p \left| \frac{\partial z}{\partial t} + m(t) \Delta z \right|_Q + (q | z(0))_{H^{-1}} \right) \quad \forall z \in Z,$$

$$(3.28) \quad (q + Nu | v - u)_{H^{-1}} \geq 0 \quad \forall v \in \mathcal{U}_{ad}.$$

By (3.27) we have (3.16)₂, (3.17)₁ and (3.17)₂; by Remark 3.1 and (3.28) we have (3.17)₃ and (3.18).

Acknowledgment. We wish to thank Professor L. A. Medeiros for his comments and suggestions.

Professor P. H. Rivera was the adviser of the second author and has been deceased since August 13, 1983.

REFERENCES

- [1] J. L. LIONS, *Contrôle optimal de systèmes gouvernés par des équations aux dérivées partielles*, Dunod, Paris, 1968.
- [2] ———, *Functions spaces and optimal control of distributed systems*, Instituto de Matemática-Universidade Federal do Rio de Janeiro, Rio de Janeiro, 1980.
- [3] ———, *Contrôle des systèmes distribués singuliers*, Dunod, Paris, 1983.
- [4] ———, *Some Methods in the Mathematical Analysis of Systems and Their Control*, Science Press, Beijing, People's Republic of China, 1981.
- [5] J. L. LIONS AND E. MAGENES, *Problèmes aux limites non homogènes*, vols. 1 et 2, Dunod, Paris, 1968.
- [6] J. L. LIONS AND B. MALGRANGE, *Sur l'unicité rétrograde dans les problèmes mixtes paraboliques*, Math. Scand., 8 (1960), pp. 277-286.
- [7] L. A. MEDEIROS, *Remarks on a non-well posed problem*, Proc. Roy. Soc. Edinburgh Sect. A, 102 (1986), pp. 131-140.
- [8] P. H. RIVERA, *On the optimal control of non-well posed linear evolution systems*, to be submitted for publication.
- [9] L. SCHWARTZ, *Théorie des distributions*, Hermann, Paris, 1966.

INVARIANCE OF THE REACHABLE SET UNDER NONLINEAR PERTURBATIONS*

THOMAS I. SEIDMAN†

Abstract. For the abstract controlled Volterra equation

$$(*) \quad x(t) = \bar{x}(t) + \int_0^t \Psi(t, s) [\phi(s, x(s)) + \mathbf{B}(s)v(s)] ds$$

on $[0, T]$, we consider the reachable set $\mathcal{K}_\phi := \{x(T): (*) \text{ for some } v \in \mathcal{V}\}$. Viewing $\phi(\cdot, \cdot)$ as a nonlinear perturbation of an otherwise linear control problem, conditions are obtained under which $\mathcal{K}_\phi = \mathcal{K}_0$ for a suitable class \mathcal{F} of such nonlinear perturbations ϕ whenever the linear problem is known to have a reachable set invariant under affine perturbations: $\phi(\cdot, \cdot) = g \in \mathcal{G}$. The results generalize those obtained by Naito [7] for control of the heat equation.

Key words. reachable, nonlinear, perturbation, control system, semigroup, fixed point

AMS(MOS) subject classifications. 93B05, 93C10, 93C25

1. Introduction. We consider a linear control system governed by an equation of the form

$$(1.1) \quad \dot{x} + \mathbf{A}x = \bar{f} + \mathbf{B}v, \quad x(0) = 0.$$

Here $x(\cdot)$ is to take values in a Banach space \mathcal{X} and v is \mathcal{V} -valued with, say, $\mathbf{B}: \mathcal{V} \rightarrow \mathcal{X}$. It is now standard that such abstract ODEs can be used to model distributed parameter systems corresponding to partial differential equations (e.g. wave and diffusion equations) and to equations with delay, etc. We do not require that \mathbf{A}, \mathbf{B} be independent of t so the “variation of parameters” formulation leads to an integral equation of the form

$$(1.2) \quad x(t) = \int_0^t \Psi(t, s) [\bar{f}(s) + \mathbf{B}(s)v(s)] ds + \Psi(t, 0)x_0,$$

involving the *transition operators* (impulse response function)

$$(1.3) \quad \Psi(t, s): \mathcal{X} \rightarrow \mathcal{X} \quad \text{for } 0 \leq s \leq t \leq T.$$

(Here we have arbitrarily fixed the *terminal time* T . Note that in certain circumstances one knows [10] that the reachable set is independent of T .) Without further reference (for awhile) to the abstract ODE (1.1) except for motivation, we take (1.2)—with suitable hypotheses regarding the family of operators in (1.3)—as defining the system. In this context $x(\cdot)$ is referred to as a *mild solution* of (1.1). The theory, especially for the autonomous case $\{\Psi(t, s) = \Psi_0(t - s)\} = C_0$ semigroup on \mathcal{X} , has been extensively treated; cf., e.g., [5] and references there.

As v ranges over a space \mathcal{V} of *admissible control functions*, we denote the corresponding solutions of (1.2) by $x_0(\cdot; v)$. The *reachable set* is then

$$(1.4) \quad \mathcal{K}_0 := \{x_0(T; v): v \in \mathcal{V}\}.$$

In writing (1.4) we have suppressed explicit indication of the dependence of \mathcal{K}_0 on the terminal time T and on the operators Ψ, \mathbf{B} . More to the point is that (1.4) does

* Received by the editors January 13, 1985; accepted for publication (in revised form) May 27, 1986. This work was supported in part by Air Force Office of Scientific Research grant AFOSR-82-0271.

† Department of Mathematics, University of Maryland Baltimore County, Catonsville, Maryland 21228.

not indicate any dependence of \mathcal{H}_0 on the (fixed) function \bar{f} appearing in (1.1), (1.2) since our present concern is with the effect of perturbing (1.1), (1.2) by replacing \bar{f} with $f = \bar{f} + g$. This will be an affine perturbation of the problem if g is a fixed function—say, drawn from a function space \mathfrak{Z} —but becomes more interesting if we consider nonlinear perturbations of the form

$$(1.5) \quad \begin{aligned} g(t) &:= \phi(t, x(t)), \\ f &= \bar{f} + g = \bar{f} + \phi(\cdot, x) \end{aligned}$$

so that (1.2) is now an integral equation of second kind:

$$(1.6) \quad x(t) = \bar{x}(t) + \int_0^t \Psi(t, s) [\phi(s, x(s)) + \mathbf{B}(s)v(s)] ds$$

where $\bar{x}(\cdot)$ is the (fixed) solution of (1.2) with $v = 0$.

We now take (1.6) as defining the dynamics under suitable conditions to ensure the existence of solutions $x = x_\phi(\cdot; v)$ of (1.6) for $v \in \mathfrak{V}$. The principal result of this paper is that the reachable set

$$(1.7) \quad \mathcal{H}_\phi := \{x_\phi(T; v) : v \in \mathfrak{V}\}$$

is invariant: $\mathcal{H}_\phi = \mathcal{H}_0$ for ϕ in a suitable class \mathfrak{F} of nonlinear perturbations, provided $\Psi, \mathbf{B}, \mathfrak{V}$ are such that the reachable set is known a priori to be invariant under affine perturbations $\phi = g \in \mathfrak{Z}$, already a fairly strong controllability hypothesis.

It is convenient at this point to introduce two relevant mappings, one linear and one nonlinear. We first define the linear mapping

$$(1.8) \quad \mathbf{T} : g \mapsto \int_0^T \Psi(T, s)g(s) ds$$

so that for the affine case one has

$$(1.9) \quad x_g(T; v) := \bar{x}(T) + \mathbf{T}(g + \mathbf{B}v).$$

(Note that, hopefully without introducing confusion, we have continued to use the letter \mathbf{B} also to denote the multiplication operator: $v(\cdot) \mapsto \mathbf{B}(\cdot)v(\cdot)$ acting on $v \in \mathfrak{V}$.) Next, supposing we have a unique solution $x_\phi(\cdot; v)$ of (1.6) for (every) $v \in \mathfrak{V}$, we define the *nonlinear* mappings Φ, \mathbf{G} by

$$(1.10) \quad \begin{aligned} \Phi : x &\mapsto \phi(\cdot, x(\cdot)), \\ \mathbf{G} = \mathbf{G}_\phi : v &\mapsto G_\phi(\cdot; v) := \phi(\cdot, x_\phi(\cdot; v)) = \Phi x_\phi. \end{aligned}$$

Inserting this in (1.6) gives

$$x_\phi(t; v) = \bar{x}(t) + \int_0^t \Psi(t, s) [G_\phi(s; v) + \mathbf{B}(s)v(s)] ds$$

so

$$(1.11) \quad x_\phi(T; v) = \bar{x}(T) + \mathbf{T}[\mathbf{G}_\phi v + \mathbf{B}v].$$

Let us consider now the controllability hypothesis that the reachable set is invariant under affine perturbations by $g \in \mathfrak{Z}$ —i.e., that $\mathcal{H}_g = \mathcal{H}_0$ for $g \in \mathfrak{Z}$. To have $\xi \in \mathcal{H}_0$ means that $\xi = \bar{x}(T) + \mathbf{T}\mathbf{B}v_0$ for some $v_0 \in \mathfrak{V}$. Comparing this with (1.9), we see that to have $\xi \in \mathcal{H}_g$ we must find $v \in \mathfrak{V}$ such that $\xi = \bar{x}(T) + \mathbf{T}g + \mathbf{T}\mathbf{B}v$ so, taking the difference and using the linearity of \mathbf{T}, \mathbf{B} , one must have

$$(1.12) \quad \mathbf{T}(g + \mathbf{B}w) = 0$$

for some $w = (v - v_0) \in \mathfrak{B}$. We impose the hypothesis

(H₀). (i) $\mathbf{T}: \mathfrak{B} \rightarrow \mathcal{X}$ and $\mathbf{TB}: \mathfrak{B} \rightarrow \mathcal{X}$ are continuous,

(ii) For each $g \in \mathfrak{B}$ there exists $w \in \mathfrak{B}$ such that $\mathbf{TB}w = -\mathbf{T}g$, i.e., (1.12).

Clearly this implies that $\mathcal{K}_g \supset \mathcal{K}_0$ for each $g \in \mathfrak{B}$ —and also conversely, since $\xi \in \mathcal{K}_g$ leads to the same condition (1.12) to have $\xi \in \mathcal{K}_0$. Hence (H₀)(ii) is equivalent to invariance of the reachable set under affine perturbations from \mathfrak{B} :

(H'₀). \mathcal{K}_g is independent of g for $g \in \mathfrak{B}$.

This condition (H₀) does not involve any nonlinearity but only the relation between \mathfrak{B} and $\mathbf{B}\mathfrak{B}$, given Ψ which defined \mathbf{T} . Indeed, in view of the linearity of \mathbf{T} we may equivalently write (H₀)(ii) in the form

(H''₀). $\mathbf{T}(\mathfrak{B}) \subset \mathbf{T}(\mathbf{B}\mathfrak{B}) = \mathcal{K}_0$.

Suppose now, that $\mathfrak{B}, \mathfrak{B}$ are given such that (H₀) holds. We wish to obtain a class \mathfrak{F} of nonlinearities such that

(H₁). $\mathbf{G} = \mathbf{G}_\phi: \mathfrak{B} \rightarrow \mathfrak{B}$ for each $\phi \in \mathfrak{F}$,

i.e., for each $\phi \in \mathfrak{F}$ we wish to know that the integral equation (1.6) has a solution $x_\phi(\cdot; v)$ for each $v \in \mathfrak{B}$ and that the substitution (1.10) then gives a function $g := \mathbf{G}_\phi(v)$ which is in \mathfrak{B} . Our object is an invariance result generalizing (H'₀): that

(*) \mathcal{K}_ϕ is independent of ϕ for $\phi \in \mathfrak{F}$

or, equivalently since we assume $0 \in \mathfrak{F}$, that $\mathcal{K}_\phi = \mathcal{K}_0$ for $\phi \in \mathfrak{F}$. Note that if we fix $\phi \in \mathfrak{F}$ and $\xi \in \mathcal{K}_0$ as earlier, then (H₁), (H₀) give a composed map¹

$$(1.13) \quad \tilde{w} \in \mathfrak{B} \mapsto g := \mathbf{G}_\phi(v_0 + \tilde{w}) \in \mathfrak{B} \mapsto w \in \mathfrak{B} \quad \text{giving (1.12).}$$

Existence of a fixpoint for the map (1.13) would show $\xi \in \mathcal{K}_\phi$. This, for each $\xi \in \mathcal{K}_0$, would show $\mathcal{K}_\phi \supset \mathcal{K}_0$; the converse, that $\mathcal{K}_\phi \subset \mathcal{K}_0$, follows easily from (H₀), (H₁). Our strategy, then, is to impose suitable hypotheses under which (1.13) can be shown to admit applicability of the Schauder Fixpoint theorem.

This paper is very much in the line of the “fixpoint theory approach to nonlinear controllability” for which see, e.g., [3] and the references therein. While most work along these lines has been intended to show (approximate) controllability to the entire state space, the considerations here have been stimulated by the original results [7] of Naito, introducing the problem of invariance of the exactly reachable set \mathcal{K}_0 under various perturbations. In [7], Naito introduced the hypothesis (H₀) in the context of a parabolic equation such as

$$(1.14) \quad \begin{aligned} \dot{x} &= \Delta x + \phi + \mathbf{B}v \quad \text{on } \mathcal{Q} := (0, T) \times \Omega, \Omega \subset \mathbb{R}^m, \\ x &= 0 \quad \text{on } \Sigma := (0, T) \times \partial\Omega, \end{aligned}$$

showing invariance of the reachable set for $\phi: \mathbb{R} \rightarrow \mathbb{R}$ uniformly bounded and uniformly Lipschitzian in (the pointwise values of) x on \mathcal{Q} . His argument immediately applies to (1.1) with $-\mathbf{A}$ generating a C_0 semigroup $\Psi(t-s)$ of compact operators on \mathcal{X} and autonomous bounded $\mathbf{B}: \mathcal{V} \rightarrow \mathcal{X}$ so also $\mathbf{B}: \mathfrak{B} := L^2([0, T] \rightarrow \mathcal{V}) \rightarrow L^2([0, T] \rightarrow \mathcal{X}) =: \mathfrak{B}$. The hypothesis (H₀) is then taken as a condition on \mathbf{B} .

Our objective here, accepting the formulation in terms of the controllability hypothesis (H₀), is to generalize the abstract setting and to weaken the hypotheses of [7]. In particular, we seek results such that

¹ At this point, of course, (1.13) is not a well-determined map—even accepting (H₀), (H₁)—since (1.12) does not, in general, uniquely determine w . However Lemma 2 will resolve this by asserting the existence of a continuous operator $\mathbf{C}: g \mapsto w: \mathfrak{B} \rightarrow \mathfrak{B}$.

A, B need not be autonomous (so the transition operator $\Psi(t, s)$ need not be a semigroup);

$\Psi(t, s)$ need not be compact: some compactness is needed for the general approach but we explore alternate ways of obtaining this which may, e.g., permit applications to hyperbolic equations;

B need not be bounded (so we may consider, e.g., boundary control);

The nonlinearity ϕ need not be bounded in range and may not be Lipschitzian (or may involve a Lipschitz condition with respect to a different norm so, e.g., in (1.14) one might take $\phi = \phi(t, x, \nabla x)$). One might also consider ϕ spatially nonlocal so one could treat

$$(1.15) \quad \phi : [0, T] \times \mathcal{X} \rightarrow \mathcal{Z},$$

with, for example, \mathcal{X}, \mathcal{Z} spaces of functions on Ω and $[\Phi x](t, w) := \phi(t, x(t, \cdot))(\omega)$; see § 5, Remark 3.

A summary of the set of (abstract) results obtained is given by Theorem 3. Some remarks as to possible contexts (in which various hypotheses might be verified) are presented in § 5 although, so far, little can be said about verifiability of (H_0) in the most interesting case: $\mathcal{H}_0 \neq \mathcal{X}$. For the moment, the hope for positive results (of some substantial generality) along these lines is a proposal for continued investigation. Another proposal would be for a comparable investigation of possible invariance of the approximately reachable set \mathcal{H}_0 (closure in \mathcal{X}) with (H_0'') replaced by $T(\mathfrak{J}) \subset \mathcal{H}_0$ and with suitable conditions imposed on the nonlinearity ϕ .

2. Continuity and the control map. Since we are defining a reachable set by evaluating the solution $x(\cdot)$ of (1.2) or (1.6) at $t = T$, we wish to consider as part of the definition of a solution that

$$(2.1) \quad x(\cdot) \in \mathfrak{X} := C([0, T] \rightarrow \mathcal{X})$$

so point evaluation makes sense. We assume directly that $\bar{x} \in \mathfrak{X}$, but must impose conditions on Ψ , **B** and \mathfrak{J} , \mathfrak{V} which ensure \mathcal{X} -continuity in t for the integral. For simplicity we will only consider \mathfrak{J} , \mathfrak{V} of the form:

$$(2.2) \quad \mathfrak{J} := L^p([0, T] \rightarrow \mathcal{Z}), \quad \mathfrak{V} := L^{p'}([0, T] \rightarrow \mathcal{V})$$

for suitable p, p' ($1 < p, p' < \infty$) and separable Banach spaces \mathcal{Z}, \mathcal{V} . (Obviously, for consistency one must take \mathcal{Z} and \mathcal{X} to be spaces of the same kind of objects. For example, one might have $\mathcal{Z} \subset \mathcal{X}$ or, perhaps, have $\mathcal{X} := L^2(\Omega)$ and $\mathcal{Z} = \text{span}\{z_1, \dots, z_n\}$ with each $z_k \in L^1(\Omega)$ for some $\Omega \subset \mathbb{R}^m$; see Remark 3.) We will assume

$$(2.3) \quad \begin{aligned} & \text{(i)} \quad \bar{x} \in \mathfrak{X}, \Psi(\cdot, \cdot), \mathbf{B}(\cdot) \text{ are suitably measurable,} \\ & \text{(ii)} \quad |\Psi(t, s)|_{\mathcal{X} \rightarrow \mathcal{X}} \leq \rho_1(t-s) \quad \text{with } \rho_1 \in L_+^q(0, T) (1/p + 1/q = 1), \\ & \text{(iii)} \quad |\Psi(t, s)\mathbf{B}(s)|_{\mathcal{V} \rightarrow \mathcal{X}} \leq \rho_2(t-s) \quad \text{with } \rho_2 \in L_+^{q'}(0, T) (1/p' + 1/q' = 1), \\ & \text{(iv)} \quad |\Psi(t, s) - \Psi(r, s)|_{\mathcal{X} \rightarrow \mathcal{X}} \leq \varepsilon, \quad |[\Psi(t, s) - \Psi(r, s)]\mathbf{B}(s)|_{\mathcal{V} \rightarrow \mathcal{X}} \leq \varepsilon \end{aligned}$$

for $0 \leq s \leq r - \varepsilon, r < t = r + h \leq T$ with $\varepsilon = \varepsilon(h) \rightarrow 0+$ as $h \rightarrow 0+$.

Under these assumptions we will have the desired continuity.

LEMMA 1. Assume (2.3) and the form (2.2) for $\mathfrak{J}, \mathfrak{V}$. Then the linear operator **S** given by

$$(2.4) \quad [\mathbf{S}g](t) := \int_0^t \Psi(t, s)g(s) ds$$

is well defined and continuous from \mathfrak{J} to \mathfrak{X} . Further, for g in a bounded subset of \mathfrak{J} one has a uniform modulus of continuity (depending only on (2.3) and the bound on $\|g\|_{\mathfrak{J}}$)

for **Sg**. Similarly, **SB** is continuous from \mathfrak{V} to \mathfrak{X} , again with a uniform modulus of continuity (depending only on (2.3) and a bound on $\|v\|_{\mathfrak{V}}$) for **SBv**.

Proof. Setting $y := \mathbf{S}g$ for $g \in \mathfrak{V}$, we have

$$\begin{aligned} |y(t) - y(r)|_{\mathfrak{X}} &\leq \int_{r-\varepsilon}^{r+h} |\Psi(t, s)|_{\mathfrak{X} \rightarrow \mathfrak{X}} |g(s)|_{\mathfrak{X}} ds \\ &\quad + \int_{r-\varepsilon}^r |\Psi(r, s)|_{\mathfrak{X} \rightarrow \mathfrak{X}} |g(s)|_{\mathfrak{X}} ds \\ &\quad + \int_0^{r-\varepsilon} |\Psi(t, s) - \Psi(r, s)|_{\mathfrak{X} \rightarrow \mathfrak{X}} |g(s)|_{\mathfrak{X}} ds \\ &\leq \left[\left(\int_{r-\varepsilon}^{r+h} \rho_1^q \right)^{1/q} + \left(\int_{r-\varepsilon}^r \rho_1^q \right)^{1/q} + \varepsilon(r-\varepsilon)^{1/q} \right] \|g\|_{\mathfrak{V}} \end{aligned}$$

with $\varepsilon = \varepsilon(h)$, ρ_1 as in (2.3). (Note: for $t < \varepsilon(h)$ one replaces $(r-\varepsilon)$ by 0.) Taking $r=0$ and $0 < t = h \leq T$, one obtains

$$\|\mathbf{S}\|_{\mathfrak{V} \rightarrow \mathfrak{X}} \leq \|\rho_1\|_{L^q}.$$

Since $\varepsilon \rightarrow 0$ as $h \rightarrow 0$, one has an estimate for the modulus of continuity as desired. The estimates for **SB**: $\mathfrak{V} \rightarrow \mathfrak{X}$ are essentially the same. \square

Note that **T**, as given by (1.8), is just **S** followed by evaluation at $t = T$ so continuity of **S**: $\mathfrak{V} \rightarrow \mathfrak{X}$ and of **SB**: $\mathfrak{V} \rightarrow \mathfrak{X}$ immediately give continuity of **T**: $\mathfrak{V} \rightarrow \mathfrak{X}$ and of **TB**: $\mathfrak{V} \rightarrow \mathfrak{X}$, i.e., $(H_0)(i)$. With this in hand, we make the following observation.

LEMMA 2. Assume (H_0) . Then there exists a continuous control map **C**: $\mathfrak{V} \rightarrow \mathfrak{V}$ of linear growth

$$(2.5) \quad \mathbf{TBC}g + \mathbf{T}g = 0, \quad \|Cg\|_{\mathfrak{V}} \leq \alpha \|g\|_{\mathfrak{V}} \quad \text{for } g \in \mathfrak{V}$$

for some fixed $\alpha > 0$. (If $\mathcal{N} := \mathcal{N}(\mathbf{TB})$ admits a closed complement in \mathfrak{V} , e.g., if \mathfrak{V} is a Hilbert space, then one may take **C** to be linear.)

Proof. As **TB** is continuous by $(H_0)(i)$, its nullspace \mathcal{N} is closed so we may consider the quotient space $\mathfrak{V}_0 := \mathfrak{V}/\mathcal{N}$. It is standard that **TB** factors through the canonical projection **P**: $\mathfrak{V} \rightarrow \mathfrak{V}_0$ so one has a continuous linear injection **T**₀: $\mathfrak{V}_0 \rightarrow \mathfrak{X}$ with **TB** = **T**₀**P**. In this context $(H_0)(ii)$ gives existence, for each $g \in \mathfrak{V}$, of $\tilde{w} := \mathbf{P}w \in \mathfrak{V}_0$ such that **Tg** = $-\mathbf{TB}w = -\mathbf{T}_0\tilde{w}$ and this \tilde{w} is unique by the injectivity of **T**₀. Thus, there is a well-defined map **C**₀: $\mathfrak{V} \rightarrow \mathfrak{V}_0$: $g \mapsto \tilde{w}$ such that **Tg** + **T**₀**C**₀ $g = 0$; clearly **C**₀ is linear. Introducing the continuous linear operator

$$\mathbf{L}: \mathfrak{V} \times \mathfrak{V}_0 \rightarrow \mathfrak{X}: [g, \tilde{w}] \mapsto (\mathbf{T}g + \mathbf{T}_0\tilde{w}),$$

we see that the closed subspace $\mathcal{N}(\mathbf{L}) \subset \mathfrak{V} \times \mathfrak{V}_0$ is just the graph of **C**₀. The Closed Graph theorem thus gives continuity of **C**₀: $\mathfrak{V} \rightarrow \mathfrak{V}_0$. (If \mathcal{N} admits a complement in \mathfrak{V} , we can identify \mathfrak{V}/\mathcal{N} with this complement—call it $\mathfrak{V}_0 \subset \mathfrak{V}$ —and so take **C** = **C**₀ linear.) At this point, the Michael Selection theorem (cf. [6, Thm. 7.2]) gives a continuous $\Gamma: \mathfrak{V}_0 \rightarrow \mathfrak{V}$ (a right inverse of the open surjection **P**) with $\|\Gamma(w)\|_{\mathfrak{V}} \leq \lambda \|w\|_{\mathfrak{V}_0}$, where $\lambda > 1$ was arbitrary. We take **C**: $\mathfrak{V} \rightarrow \mathfrak{V}$ to be the composition $\Gamma \circ \mathbf{C}_0$: $\mathfrak{V} \rightarrow \mathfrak{V}_0 \rightarrow \mathfrak{V}$ and have (2.5) with $\alpha = \lambda \|\mathbf{C}_0\|$. (Remark: When **C** is permitted to be nonlinear we note that uniform continuity is not asserted and, in particular, (2.5) gives linear growth but not a Lipschitz condition.) \square

3. Nonlinear perturbation. Our next goal is the introduction of a class \mathfrak{F} of nonlinearities for which (H_1) can be verified. Indeed, we will want to impose conditions on $\phi \in \mathfrak{G}$ enabling us also to verify the hypothesis

(H_2) . For each $v \in \mathfrak{V}$ consider the map **W**₀: $w \mapsto \mathbf{G}_\phi(v_0 + w)$; then there is a ball \mathfrak{B} (depending on $v_0 \in \mathfrak{V}$, $\phi \in \mathfrak{F}$, etc.) in \mathfrak{V} which is invariant under **W** := **CW**₀.

Note that we have separated out (H_2) for convenience of exposition but it is only meaningful in the context of (H_0) , (H_1) which, e.g., ensure the continuity of $\mathbf{W}_0: \mathfrak{B} \rightarrow \mathfrak{Z}$ and that \mathbf{C} is defined as in (2.5). The map \mathbf{W} is, of course, as in (1.13); compare footnote 1. We also state one final hypothesis:

(H_3) . The map $\mathbf{W}_0: w \mapsto \mathbf{G}_\phi(v_0 + w)$ is compact from \mathfrak{B} to \mathfrak{Z} for each $v_0 \in \mathfrak{B}$. The hypotheses (H_0) – (H_3) are, of course, to be used for application of the Schauder Fixpoint theorem in proving Theorem 1 below. Following that proof, we consider specific assumptions leading to verification of (H_1) – (H_2) ; discussion of alternative approaches to the compactness in (H_3) is the content of the next section.

THEOREM 1. *Let $\Psi(\cdot, \cdot)$, $\mathbf{B}(\cdot)$, \mathfrak{Z} , \mathfrak{B} , \mathfrak{F} be such that (H_0) – (H_3) hold for each $\phi \in \mathfrak{F}$. Then the reachable set $\mathcal{K}_\phi := \{x_\phi(T; v) : v \in \mathfrak{B}\}$, defined by (1.6), is independent of $\phi \in \mathfrak{F}$, i.e., $\mathcal{K}_\phi = \mathcal{K}_0$.*

Proof. The argument is as indicated in § 1. Given $\phi \in \mathfrak{F}$, for any $\xi \in \mathcal{K}_0$ we have $v_0 \in \mathfrak{B}$ such that $\xi = x_0(T; v_0) = \bar{x}(T) + \mathbf{TB}v_0$. With this ϕ , v_0 we have \mathbf{W}_0 , \mathbf{W} defined and continuous by Lemmas 1 and 2, using (H_0) , (H_1) . By (H_2) , (H_3) , then, the set

$$\mathfrak{B}_* := [\text{closed convex hull of } \mathbf{W}\mathfrak{B}; \mathfrak{B} \text{ as in } (H_2)]$$

is compact, convex, and invariant under \mathbf{W} ; we now redefine \mathbf{W} to be its restriction to \mathfrak{B}_* . The Schauder Fixpoint theorem then applies to give a fixpoint \hat{w} ; set $\hat{v} := v_0 + \hat{w}$, $\hat{x} := x_\phi(\cdot; \hat{v})$, and $\hat{g} := \phi\hat{x} = \mathbf{G}_\phi\hat{v}$. The definition gives

$$\begin{aligned}\hat{x}(T) &= \bar{x}(T) + \mathbf{T}\hat{g} + \mathbf{TB}(v_0 + \hat{w}) \\ &= (\bar{x}(T) + \mathbf{TB}v_0) + (\mathbf{T}\hat{g} + \mathbf{TB}\hat{w}).\end{aligned}$$

Since $\hat{w} = \mathbf{W}\hat{w} = \mathbf{C}\mathbf{G}_\phi(v_0 + \hat{w}) = \mathbf{C}\hat{g}$, the definition of \mathbf{C} gives $\mathbf{T}\hat{g} + \mathbf{TB}\hat{w} = 0$ so $x_\phi(T; \hat{v}) = \hat{x}(T) = \bar{x}(T) + \mathbf{TB}v_0 = \xi$. Thus, $\xi \in \mathcal{K}_0$ implies $\xi \in \mathcal{K}_\phi$ so $\mathcal{K}_0 \subset \mathcal{K}_\phi$. Conversely, for any $\xi \in \mathcal{K}_\phi$ we have $v \in \mathfrak{B}$ for which

$$\xi = x_\phi(T; v) = \bar{x}(T) + \mathbf{T}g + \mathbf{TB}v \quad \text{with } g := \mathbf{G}_\phi v.$$

Now let w_0 correspond to g as in (H_0) , e.g., $w_0 := \mathbf{C}g$. Then $\mathbf{T}g + \mathbf{TB}w_0 = 0$ so, setting $v_0 := v - w_0$, we have

$$\begin{aligned}x_0(T; v_0) &= \bar{x}(T) + \mathbf{TB}v_0 \\ &= (\bar{x}(T) + \mathbf{T}g + \mathbf{TB}v) - (\mathbf{T}g + \mathbf{TB}w_0) \\ &= \xi - 0 = \xi.\end{aligned}$$

Thus, $\xi \in \mathcal{K}_\phi$ implies $\xi \in \mathcal{K}_0$ so $\mathcal{K}_\phi = \mathcal{K}_0$. This, for each $\phi \in \mathfrak{F}$, shows the desired invariance of the reachable set under these quasilinear perturbations. \square

We will devote the remainder of this section to providing a set of conditions on ϕ (i.e. part of the specification of \mathfrak{F}) under which (H_1) , (H_2) can be verified. The next section will then consider possible approaches leading to verification of the compactness assumption of the theorem.

To obtain somewhat greater generality than otherwise, introduce yet another Banach space \mathcal{Y} and consider (1.6) as an integral equation for \mathcal{Y} -valued functions. (Obviously, for consistency \mathcal{X} and \mathcal{Y} must be spaces of the same kind of objects but, perhaps, with different norms. For example, one might consider $\mathcal{X} := L^2(\Omega)$ and $\mathcal{Y} := H^1(\Omega)$ for some region $\Omega \subset \mathbb{R}^m$.)

For each $\phi \in \mathfrak{F}$ we will assume

$$(3.1) \quad \begin{aligned} \phi: [0, T] \times \mathcal{Y} &\rightarrow \mathcal{X} \text{ satisfies Carathéodory conditions and a growth} \\ &\text{condition} \quad |\phi(s, \eta)|_{\mathcal{X}} \leq \alpha(s) + \beta|\eta|_{\mathcal{Y}} \quad (\eta \in \mathcal{Y}) \end{aligned}$$

where \mathcal{X} is as in (2.2) and $\alpha \in L^p_+(0, T)$, $0 \leq r < 1$. With this r we set

$$(3.2) \quad \mathcal{Y} := L^p([0, T] \rightarrow \mathcal{Y})$$

with $\bar{p} := rp$; with no loss of generality we may assume r in (3.1) gives $1 < \bar{p} < p < \infty$. We will use an exponentially weighted norm

$$(3.3) \quad \|y(\cdot)\|_{\mathcal{Y}} := \left[\int_0^T |e^{-\mu t} y(t)|_{\mathcal{Y}}^{\bar{p}} dt \right]^{1/\bar{p}}$$

with the value of the parameter $\mu \geq 0$ to be specified later. By Krasnoselski's theorem (cf., e.g., [2]) one then has

(3.4) $\phi: y \mapsto \phi(\cdot, y(\cdot))$ is continuous from \mathcal{Y} into \mathcal{Z} , taking bounded sets to bounded sets with

$$\|\phi(\cdot, y)\|_{\mathcal{Z}} \leq \|\alpha\|_{L^p} + \beta e^{\mu T} \|y\|_{\mathcal{Y}}^r.$$

We next consider conditions along the lines of (2.3):

- (3.5) (i) $\bar{x} \in \mathcal{X} \cap \mathcal{Y}$,
 (ii) $|\psi(t, s)|_{\mathcal{X} \rightarrow \mathcal{Y}} \leq \bar{\rho}_1(t-s)$ for some $\bar{\rho}_1 \in L_+^{\bar{q}}(0, T)$, where $1/\bar{q} + 1/p \leq 1 + 1/\bar{p}$ (e.g. $\bar{q} = 1$ since $\bar{p} < p$),
 (iii) $|\psi(t, s)\mathbf{B}(s)|_{\mathcal{V} \rightarrow \mathcal{Y}} \leq \rho_2(t-s)$ for some $\rho_2 \in L_+(0, T)$, where $1/\bar{q}' + 1/p' \leq 1 + 1/\bar{p}$

with a Lipschitz condition of the form

$$(3.6) \quad |\psi(t, s)[\phi(s, \eta) - \phi(s, \eta')]|_{\mathcal{Y}} \leq \bar{\rho}_3(t-s)|\eta - \eta'|_{\mathcal{W}} \quad \text{for some } \bar{\rho}_3 \in L_+^1(0, T).$$

THEOREM 2. Assume (H_0) , suppose ϕ satisfies (3.1) giving (3.4), and assume (3.5), (3.6). Then (1.6) has a (unique) solution $x_\phi(\cdot, v) \in \mathcal{X} \cap \mathcal{Y}$ for each $v \in \mathcal{Z}$ and (H_1) , (H_2) hold.

Proof. We choose $\mu \geq 0$ large enough so that

$$(3.7) \quad \|e^{-\mu \tau} \bar{\rho}_3\|_{L^1} =: \theta < 1$$

and use this to define the norm in (3.3). Define a map $F = F_v: y \mapsto F(y, v)$ by the right-hand side of (1.6):

$$(3.8) \quad \begin{aligned} Fy &:= \bar{x} + S\phi y + SBv =: F(y, v), \quad \text{i.e.,} \\ [F_v y](t) &:= \bar{x}(t) + \int_0^t \psi(t, s)[\phi(s, y(s)) + \mathbf{B}(s)v(s)] ds. \end{aligned}$$

Clearly a fixpoint of F_v is a solution of the integral equation (1.6). The form of the hypotheses permits us to use convolution estimates and we recall (cf., e.g., [8, p. 106]) that convolutions $f * g$ satisfy

$$(3.9) \quad \|f * g\|_{L^\pi} \leq C \|f\|_{L^{\pi'}} \|g\|_{L^{\pi''}} \quad (\text{note the fixed support } [0, T])$$

for $1 \leq \pi, \pi', \pi'' < \infty$ and $1/\pi \geq 1/\pi' + 1/\pi'' - 1$.

The estimates which show contractivity of F_v with respect to (3.3) and then estimate the fixpoint $x_\phi(\cdot; v)$ are fairly standard consequences of the hypotheses. We have

$$e^{-\mu t} |[Fy](t)|_{\mathcal{Y}} \leq |e^{-\mu t} \bar{x}(t)|_{\mathcal{Y}} + [\hat{\rho}_1 * \hat{\alpha} + \beta \hat{\rho}_1 * \hat{\eta}' + \hat{\rho}_2 * |v(\cdot)|_{\mathcal{V}}](t)$$

where

$$\begin{aligned} \hat{\rho}_1(\tau) &:= e^{-\mu \tau} \bar{\rho}_1(\tau), & \hat{\rho}_2(\tau) &:= e^{-\mu \tau} \bar{\rho}_2(\tau), \\ \hat{\alpha}(\tau) &:= e^{-\mu \tau} \alpha(\tau), & \hat{\eta}(\tau) &:= |e^{-\mu \tau} y(\tau)|_{\mathcal{Y}}. \end{aligned}$$

Hence

$$\|\mathbf{F}y\|_{\mathcal{Y}} \leq \|\bar{x}\|_{\mathcal{W}} + C\|\rho_1\|_{L^q}(\|\hat{\alpha}\|_{L^p} + \beta\|y\|_{\mathcal{Y}}^r + C\|\hat{\rho}_2\|_{L^q}\|v\|_{\mathcal{X}})$$

so $\mathbf{F}_v: \mathcal{Y} \rightarrow \mathcal{Y}$ and any fixpoint $y = x_\phi(\cdot; v)$ must satisfy

$$(3.10) \quad \|x_\phi(\cdot; v)\|_{\mathcal{Y}} = \mathcal{O}(\|v\|_{\mathcal{X}})$$

since $r < 1$. Using (3.6), (3.7), (3.9) gives

$$(3.11) \quad \|\mathbf{F}(y, v) - \mathbf{F}(y', v')\|_{\mathcal{Y}} \leq \theta\|y - y'\|_{\mathcal{Y}} + c\|v - v'\|_{\mathcal{X}}$$

where c is the L^q -norm of $\hat{\rho}_2$. With $v = v'$ this gives contractivity so $x_\phi(\cdot; v)$ is well-defined in \mathcal{Y} for each $v \in \mathcal{X}$. From (3.4) with $y := x_\phi(\cdot; v)$ we have $\mathbf{G}_\phi v := \phi(\cdot, y) \in \mathcal{Z}$ with

$$(3.12) \quad \|\mathbf{G}_\phi v\|_{\mathcal{Z}} = \mathcal{O}(\|v\|_{\mathcal{X}}^r) \text{ uniformly in } \phi \text{ satisfying (3.4)–(3.6) and, noting } (H_0)(i), \text{ we have } x_\phi(r, v) \in \mathcal{X} \text{ as well. Now } \mathcal{W}_0 \text{ factors as}$$

$$\mathcal{W}_0: w \mapsto x_\phi(\cdot; v_0 + w) =: y \mapsto \phi(\cdot; y): \mathcal{X} \rightarrow \mathcal{Y} \rightarrow \mathcal{Z},$$

with the first map (Lipschitz) continuous by (3.11) and the second map continuous by (3.4). This gives (H_1) . Using Lemma 2 and combining (3.12) with (2.7), we have

$$(3.13) \quad \|\mathbf{W}v\|_{\mathcal{X}} \leq \alpha\|\mathbf{G}_\phi v\|_{\mathcal{Z}} = \mathcal{O}(\|v\|_{\mathcal{X}}^r).$$

Since $r < 1$, this clearly gives (H_2) . \square

Remark 1. We note from Theorems 1 and 2 that the original space \mathcal{X} plays little role except, in terms of \mathcal{X} -continuity just at the terminal time T , to justify the consideration (and boundedness) of the maps \mathbf{T} , \mathbf{TB} in $(H_0)(i)$ so as to permit applicability of Lemma 2.

Let us now write $\hat{\mathbf{S}} = \hat{\mathbf{S}}_\phi$ for the solution map of (1.6) so

$$(3.14) \quad \hat{\mathbf{S}}: v \mapsto x_\phi(\cdot; v): \mathcal{X} \rightarrow \mathcal{Y}.$$

The principal concern of the proof of Theorem 2 was to show that $\hat{\mathbf{S}}_\phi$ is well defined and uniformly Lipschitz continuous for ϕ satisfying (3.4)–(3.6). As noted, \mathbf{W} factors as

$$(3.15) \quad \begin{array}{ccccccc} \mathcal{X} & \longrightarrow & \mathcal{X} & \xrightarrow{\hat{\mathbf{S}}_\phi} & \mathcal{Y} & \xrightarrow{\phi} & \mathcal{Z} \xrightarrow{c} \mathcal{X} \\ & & \downarrow \mathbf{SB} & & & & \downarrow \mathbf{s} \\ & & & & & & \mathcal{X}. \end{array}$$

Other than to validate our original notion of solution for (1.6), one could omit direct reference to $(H_0)(i)$ in the hypotheses for Theorem 2. \square

4. Compactness. Our object, in this section, is to indicate some approaches—i.e., alternative conditions supplementary to those considered in proving Theorems 2 and 3—leading to verification of the compactness hypothesis (H_3) . We will provide three such approaches: the first, using the Arzela–Ascoli theorem, imposes the strongest conditions on ϕ and returns to the interpretation of solutions of (1.6) as continuous \mathcal{X} -valued functions, while the second and third, using the Aubin Compactness theorem [1], return to the differential equation (1.1) with $f = g = \mathbf{G}_\phi(v)$. Each approach, then, limits the generality of the setting of the last section for existence/continuity.

We recall, before proceeding directly with the approaches to showing compactness of the map \mathbf{W}_0 , a compactness argument from [10].

LEMMA 3. *Let $F: \mathcal{Y} \times \mathcal{X} \rightarrow \mathcal{Y}$ (with \mathcal{Y} complete metric) and suppose*

$$(4.1) \quad (i) \quad F \text{ is uniformly contractive on } \mathcal{Y}, \text{ i.e., for some } \theta < 1 \\ d_{\mathcal{Y}}(F(y, w), F(y', w)) \leq \theta d_{\mathcal{Y}}(y, y') \quad (y, y' \in \mathcal{Y}, w \in \mathcal{X});$$

- (ii) For every compact set \mathfrak{A} in \mathfrak{Y} the set $F(\mathfrak{A}, \mathfrak{B}) := \{F(y, w) : y \in \mathfrak{A}, w \in \mathfrak{B}\}$ is precompact in \mathfrak{Y} .

Then the set $\mathcal{F}_{\mathfrak{B}}$ of fixpoints ($\mathcal{F}_{\mathfrak{B}} := \{y \in \mathfrak{Y} : y = F(y, w) \text{ for some } w \in \mathfrak{B}\}$) is precompact in \mathfrak{Y} .

Proof. See [10] for details, but we sketch the argument here. For any choice of $y_0 \in \mathfrak{Y}$, set $\mathfrak{A}_0 := \{y_0\}$ and, recursively, let \mathfrak{A}_{k+1} be the closure of the set $F(\mathfrak{A}_k, \mathfrak{B})$ so, by (ii), each \mathfrak{A}_k is compact. Using (i) to obtain the standard estimate of the Contractive Mapping Principle, one sees that $\mathcal{F}_{\mathfrak{B}}$ is uniformly approximable within $\varepsilon/2$ by some \mathfrak{A}_k and, by compactness, this \mathfrak{A}_k has a finite cover by $(\varepsilon/2)$ -balls. Hence $\mathcal{F}_{\mathfrak{B}}$ has a finite cover by ε -balls for each $\varepsilon > 0$ so $\mathcal{F}_{\mathfrak{B}}$ is precompact. \square

We now proceed to discuss approaches to the verification of the compactness hypothesis (H_3) used in Theorem 1.

Approach 1. In contrast to Remark 1, concerning existence, the space \mathfrak{X} plays a key role in this approach. We assume the continuity of \mathbf{S} , \mathbf{SB} as in Lemma 1 and that (1.6) has a unique solution $x_\phi(\cdot; v) \in \mathfrak{X}$ for each v in \mathfrak{B} with $\{x_\phi(\cdot; v)\}$ bounded in \mathfrak{X} for v bounded in \mathfrak{B} ; for this approach we may well be using the argument of Theorem 2 with, say, $\mathfrak{Y} = \mathfrak{X}$. Now impose an additional condition comparable to (3.1) but with a hypothesis of compact embedding:

- (4.2) For some space $\hat{\mathfrak{X}}$ such that $\mathfrak{X} \hookrightarrow \hat{\mathfrak{X}}$ is a compact embedding, assume $\phi : [0, T] \times \hat{\mathfrak{X}} \rightarrow \mathfrak{Z}$ satisfies Carathéodory conditions and a condition:

$$|\phi(s, \xi)|_{\mathfrak{Z}} \leq \alpha_M(s) \quad \text{for } \xi \in \hat{\mathfrak{X}} \quad \text{with } |\xi|_{\hat{\mathfrak{X}}} \leq M$$

where, for each $M \in \mathbb{R}^+$, one has $\alpha_M \in L^p_+(0, T)$.

This suffices to ensure compactness of \mathbf{W}_0 —we argue as follows: For v in a bounded subset of \mathfrak{B} we have $\{x_\phi(\cdot; v)\}$ in a bounded subset of $\mathfrak{X} = C([0, T] \rightarrow \mathfrak{X})$ by assumption and, from Lemma 1, equicontinuous from $[0, T]$ to \mathfrak{X} . By the Arzela–Ascoli theorem this gives $\{y = \{x_\phi(\cdot; v)\}\}$ in a compact subset of $\hat{\mathfrak{X}} := C([0, T] \rightarrow \hat{\mathfrak{X}})$ whence in a compact subset of, e.g., $\hat{\mathfrak{X}} := L^p([0, T] \rightarrow \hat{\mathfrak{X}})$. Boundedness in $\hat{\mathfrak{X}}$ gives some bound M on $|\eta|_{\hat{\mathfrak{X}}}$ for the relevant application of the inequality in (4.2) so Krasnoselski's theorem gives continuity of the map

$$y \mapsto \phi(\cdot, y) : \{y \in \hat{\mathfrak{X}}_p : |y(t)| \leq M \text{ on } [0, T]\} \rightarrow \mathfrak{Z}.$$

Thus, for v in a bounded subset of \mathfrak{B} one has $\{\phi(\cdot, x_\phi(\cdot; v)) = \mathbf{G}_\phi v\}$ in a compact subset of \mathfrak{Z} , i.e., \mathbf{W}_0 is compact. We note that this approach imposes no compactness condition involving the transition operators $\Psi(\cdot, \cdot)$. \square

Approach 2. For this approach we return to the differential equation so

$$(4.3) \quad \dot{x} = \mathbf{B}v + \phi(\cdot, x) - \mathbf{A}x, \quad x(0) = x_0.$$

While our earlier interpretation treated \mathbf{A} , \mathbf{B} as (possibly) unbounded operators, we now impose the extremely weak assumption that there is some space $\tilde{\mathfrak{Y}}$ such that (with $1 < \tilde{p}, \tilde{q} < \infty$; $1/\tilde{p} + 1/\tilde{q} = 1$)

$$(4.4) \quad \begin{aligned} \text{(i)} \quad & \mathfrak{Y} \hookrightarrow \tilde{\mathfrak{Y}}, \quad \mathfrak{X} \hookrightarrow \tilde{\mathfrak{Y}}, \\ \text{(ii)} \quad & |\mathbf{A}(t)|_{\mathfrak{Y} \rightarrow \tilde{\mathfrak{Y}}} \leq \alpha(t), \quad \alpha \in L^{\tilde{q}}(0, T) (1/\tilde{q} + 1/p < 1), \\ \text{(iii)} \quad & |\mathbf{B}(t)|_{\mathfrak{V} \rightarrow \tilde{\mathfrak{Y}}} \leq \beta(t), \quad \beta \in L^{\tilde{q}'}_+(0, T) (1/\tilde{q}' + 1/p' < 1). \end{aligned}$$

Then (4.3), (4.4) give

$$(4.5) \quad \text{Bounds on } x \text{ in } \mathfrak{Y} := L^{\tilde{p}}([0, T] \rightarrow \mathfrak{Y}), \text{ on } v \text{ in } \mathfrak{B}, \text{ and on } g := \phi(\cdot, x) \text{ in } \mathfrak{Z} \text{ give a bound on } \dot{x} \text{ in } L^{\tilde{p}}([0, T] \rightarrow \tilde{\mathfrak{Y}}).$$

Suppose, then, we were to assume, as for Theorem 2, that ϕ satisfies (3.1), (3.5), (3.6). The proof of Theorem 2 then gives, for v in any bounded subset of \mathfrak{B} ,

- (i) the existence of a solution $x = x_\phi(\cdot; v) \in \mathfrak{Y}$ for each v ,
- (ii) bounds, as for (4.5) on $x \in \mathfrak{Y}$, $g := \phi(\cdot, x) \in \mathfrak{Z}$; hence also on \dot{x} .

Now impose an additional condition supplementing (3.1) but with a hypothesis of compact embedding:

- (4.6) For some space $\hat{\mathfrak{Y}}$ such that $\mathfrak{Y} \hookrightarrow \hat{\mathfrak{Y}}$ is a compact embedding, assume $\phi: [0, T] \times \hat{\mathfrak{Y}} \rightarrow \mathfrak{Z}$ satisfies Carathéodory conditions and a growth condition

$$|\phi(s, \eta)|_{\mathfrak{Z}} \leq \alpha(s) + \beta |\eta|_{\hat{\mathfrak{Y}}}^r \quad (\eta \in \hat{\mathfrak{Y}})$$

with r as in (3.1) and $\alpha \in L^r_+(0, T)$.

By the Aubin Compactness theorem [1] we have, using (4.5) and the bound on $\{x = x_\phi(\cdot; v)\}$ in \mathfrak{Y} for v in a bounded subset $\mathfrak{B} \subset \mathfrak{B}$, that $\{x_\phi(\cdot; v)\}$ is in a compact subset of $\hat{\mathfrak{Y}} := L^p([0, T] \rightarrow \hat{\mathfrak{Y}})$. Using (4.6), Krasnoselski's theorem (compare (3.4)) gives continuity of $\phi: \hat{\mathfrak{Y}} \rightarrow \mathfrak{Z}$. Thus, combining these, we have $\{\phi(\cdot; x_\phi(\cdot, v)) = \mathbf{G}_\phi v\}$ in a compact subset of \mathfrak{Z} . Under these hypotheses also, then, we have shown that the map \mathbf{W}_0 is compact. \square

For this approach also, note that the supplementary hypotheses (4.4), (4.6) did not impose any compactness condition involving $\psi(\cdot, \cdot)$.

Approach 3. This really will split into two related approaches, using Lemma 3. For the first of these, as for the previous approach, we will assume (4.3), (4.4) to obtain (4.5) so that the Aubin theorem will be applicable. The supplementary hypothesis to be imposed will no longer involve ϕ but, rather, will be a strengthening of (3.5)(iii). Thus, we assume that (3.1), (3.5), (3.6) hold, with (3.5)(iii) replaced by

- (4.7) $|\psi(t, s)\mathbf{B}(s)|_{\mathfrak{Y} \rightarrow \hat{\mathfrak{Y}}} \leq \bar{\rho}_2(t-s)$ for some $\bar{\rho}_2 \in L^q_+(0, T)$ with $1/\bar{q}' + 1/p' \leq 1 + 1/\bar{p}$ and some $\hat{\mathfrak{Y}}$ such that $\hat{\mathfrak{Y}} \hookrightarrow \mathfrak{Y}$ is a compact embedding.

We wish to employ Lemma 3 to show that, for $v \in \mathfrak{B} = (\text{bounded in } \mathfrak{B})$, one has $\{x = x_\phi(\cdot; v)\} =: \mathcal{F}_{\mathfrak{B}}$ precompact in \mathfrak{Y} . The map \mathbf{F} under consideration here is, of course, that defined by (3.8), for which (4.1)(i) is already known from the proof of Theorem 2; we need only use (4.7) and the Aubin theorem to demonstrate (4.1)(ii).

Suppose, then, we were to have $v \in \mathfrak{B}$ and $y \in \mathfrak{A} = (\text{compact in } \mathfrak{Y})$. By (3.4) we would have $\phi \mathfrak{A}$ compact in \mathfrak{Z} so $\mathbf{S}\phi \mathfrak{A}$ compact in \mathfrak{Y} . Hence, it will suffice to show that (4.7) ensures compactness for the linear map $\mathbf{SB}: \mathfrak{B} \rightarrow \mathfrak{Y}$, whence we have precompactness of the set $\mathbf{SB}\mathfrak{B}$ and so also of $F(\mathfrak{A}, \mathfrak{B}) = \mathbf{S}\phi \mathfrak{A} + \mathbf{SB}\mathfrak{B} + \bar{x}$. We note that \mathbf{SB} is just the solution operator for the differential equation

$$(4.8) \quad \dot{y} = -\mathbf{A}y + \mathbf{B}v, \quad y(0) = 0,$$

i.e., (4.3) without ϕ and with $x_0 = 0$. We next show that (4.7) makes $\mathbf{SB}: \mathfrak{B} \rightarrow \mathfrak{Y}$ continuous. The estimate is like the one we used in the proof of Theorem 2 to obtain (3.10), although we do need to consider the exponential weight factor here:

$$y(t) := [\mathbf{SB}v](t) := \int_0^t \psi(t, s)\mathbf{B}(s)v(s) ds,$$

$$|y(t)|_{\hat{\mathfrak{Y}}} \leq \int_0^t \bar{\rho}_2(t-s)|v(s)|_{\mathfrak{Y}} ds = [\bar{\rho}_2 * |v|_{\mathfrak{Y}}](t),$$

$$\|y\|_{\hat{\mathfrak{Y}}} \leq C \|\bar{\rho}_2\|_{L^q(0, T)} \|v\|_{\mathfrak{B}}.$$

For $v \in \mathfrak{V}$ we have thus bounded $\{y \in \mathbf{SB}v\}$ in $\hat{\mathfrak{Y}}$ and, using (4.8), (4.4) as in (4.5), have a bound on $\{y: y \in \mathbf{SB}\mathfrak{V}\} =: \mathfrak{S}'$ in $\hat{\mathfrak{Y}}$. Thus, applying the Aubin theorem [1] gives precompactness of $\mathbf{SB}\mathfrak{V}$ in \mathfrak{Y} and so applicability of Lemma 3 to give precompactness in \mathfrak{Y} of $\mathcal{F}_{\mathfrak{V}} = \mathbf{S}_{\phi}\mathfrak{V}$. This shows the map \mathbf{S}_{ϕ} is compact and so $\mathbf{W}_0: w \mapsto \phi \mathbf{S}_{\phi}(w + v_0)$ is compact as desired.

Within the same context we also note the sufficiency of a somewhat different supplementary hypothesis: instead of replacing (3.5)(iii) by the stronger version (4.7), we now retain (3.5) unchanged and supplement it by a compactness condition on the transition operator $\Psi(\cdot, \cdot)$:

(4.9) For some $\hat{\mathcal{U}}$ such that $\mathcal{U} \hookrightarrow \hat{\mathcal{U}}$ is a compact embedding, assume that for $\delta > 0$ one has

$$|\psi(t, t - \delta)|_{\hat{\mathcal{U}} \rightarrow \mathcal{U}} \leq M_{\delta} (\delta \leq t \leq T).$$

We show that this too gives compactness of the operator $\mathbf{SB}: \mathfrak{V} \rightarrow \mathfrak{Y}$ and so gives applicability of Lemma 3 as above. Now, given (3.5)(iii) we have a bound on $\mathbf{SB}: \mathfrak{V} \rightarrow \mathfrak{Y}$. Thus, for v bounded in \mathfrak{V} , we have a bound on $\mathfrak{S} := \{y := \mathbf{SB}v\}$ in \mathfrak{Y} and, as in (4.5), a bound on $\mathfrak{S}' = \{y\}$ in $\hat{\mathfrak{Y}}$. Using the Aubin theorem again, this gives compactness in $\hat{\mathfrak{Y}} := L^p([0, T] \rightarrow \hat{\mathcal{U}})$ for \mathfrak{S} . Note that in the context of (4.3), (4.8) we expect the transition operator to satisfy the "causality condition"

$$(4.10) \quad \psi(t, r)\psi(r, s) = \psi(t, s) \quad (0 \leq s \leq r \leq t \leq T).$$

If we define $\mathbf{D}_{\delta}: y \mapsto y_{\delta}$ for $\delta > 0$ by setting

$$(4.11) \quad y_{\delta}(t) := \begin{cases} 0, & t \leq \delta, \\ \psi(t, t - \delta)y(t - \delta), & \delta \leq t \leq T, \end{cases}$$

then $|y_{\delta}(t)|_{\mathcal{U}} \leq M_{\delta}|y(t - \delta)|_{\hat{\mathcal{U}}}$ and $\mathbf{D}_{\delta}: \hat{\mathfrak{Y}} \rightarrow \mathfrak{Y}$ is continuous whence $\mathbf{D}_{\delta}\mathfrak{S}$ is compact in \mathfrak{Y} for each $\delta > 0$. Note that (4.10) gives, for $\delta \leq t \leq T$,

$$\begin{aligned} y_{\delta}(t) &= \psi(t, t - \delta) \int_0^{t - \delta} \psi(t - \delta, s) \mathbf{B}(s)v(s) ds \\ &= \int_0^{t - \delta} \psi(t, s) \mathbf{B}(s)v(s) ds, \quad \delta \leq t \leq T \end{aligned}$$

so using (3.5)(iii) gives

$$\begin{aligned} |y(t) - y_{\delta}(t)| &= \left| \int_{t - \delta}^t \psi(t, s) \mathbf{B}(s)v(s) ds \right|_{\mathcal{U}} \\ &\leq \int_{t - \delta}^t \bar{\rho}_2(t - s)|v(s)|_{\mathcal{V}} ds \\ &= [\bar{\rho}_{2,\delta} * |v|_{\mathcal{V}}](t) \end{aligned}$$

where $\bar{\rho}_{2,\delta}$ is the restriction of $\bar{\rho}_2$ to $[0, \delta]$. Thus, using (3.9),

$$(4.12) \quad \|y - y_{\delta}\|_{\mathfrak{Y}} \leq C \|\bar{\rho}_{2,\delta}\|_{L^{q'}(0,\delta)} \|v\|_{\mathfrak{V}}.$$

As in the argument for Lemma 3, this permits us to show precompactness of \mathfrak{S} in \mathfrak{Y} by finding a finite covering by ε -balls: choosing δ small enough, one can make $\|\bar{\rho}_{2,\delta}\|$

as small as desired so each $y \in \mathfrak{S}$ is within $\varepsilon/2$ of $y_\delta \in \mathbf{D}_\delta \tilde{\mathfrak{S}}$ for $0 < \delta \leq \delta(\varepsilon)$ while the compactness of $\mathbf{D}_\delta \tilde{\mathfrak{S}}$ permits finding a finite covering of $\mathbf{D}_\delta \tilde{\mathfrak{S}}$ by $(\varepsilon/2)$ -balls. Thus $\mathfrak{S} := \mathbf{S}\mathbf{B}\mathfrak{B}$ is precompact in \mathfrak{Y} for \mathfrak{B} bounded in \mathfrak{B} and, as above, this implies (4.1)(ii) for such \mathfrak{B} and applicability of Lemma 3, etc. Once again we have demonstrated the compactness of \mathbf{W}_0 . \square

Remark 2. In Approach 3 the Lipschitz condition (3.6) was used to obtain the contractivity condition (4.1)(i), which, in turn, was needed both to show existence in \mathfrak{Y} of solutions of (4.3) and also for application of Lemma 3 to obtain compactness. In the first two approaches the compactness was obtained without appeal to (4.1)(i) and, indeed, it is possible there to omit (3.6) as a hypothesis while still retaining the final conclusion: invariance of the reachable set ($\mathcal{H}_\phi = \mathcal{H}_0$).

Note that showing $\mathcal{H}_\phi \subset \mathcal{H}_0$ used no property of ϕ except that $\phi: \mathfrak{Y} \rightarrow \mathfrak{Z}$ but to prove $\mathcal{H}_0 \subset \mathcal{H}_\phi$ required an analysis of properties of \mathbf{W}_0 . Note that all the estimates used to show (3.13)—hence, to determine an invariant ball $\mathfrak{B} = \mathfrak{B}_{\mathfrak{B}}$ for \mathbf{W} —depend² on v_0 , on \bar{x} , on the α, β and r of (3.1) or (4.6), and on $\bar{\rho}_1, \bar{\rho}_2$ in (3.5). The choice of $\bar{\rho}_3$ in (3.6) did not enter into (3.13). Indeed, for Approaches 1 and 2 the Lipschitz condition (3.6) plays a role only in ensuring that (1.6) does have a solution in \mathfrak{Y} for each $v \in \mathfrak{B}$ (so \mathbf{S}_ϕ is well defined) and in showing that \mathbf{S}_ϕ is continuous. We now intend to dispense with a direct analysis of \mathbf{S}_ϕ .

Suppose, then, we were to be given (H_0) , (2.3) and (3.5) with $\mathfrak{Y} = \mathcal{X}$ together with a function ϕ satisfying (3.1) and (4.2) but not (3.6). Now let \mathcal{F}_* be the set of all $\tilde{\phi}$ satisfying (3.1), (4.2) (with the same data: $\alpha, \beta, r, \alpha_M$) and also satisfying (3.6) with some fixed choice of $\bar{\rho}_3$. Fixing v_0 , we can find fixed sets (i.e., independent of $\phi \in \mathcal{F}_*$) $\mathfrak{B}_{\mathfrak{B}}, \mathfrak{B}_{\mathfrak{Y}}, \mathfrak{B}_{\mathfrak{Z}}$ with $\mathfrak{B}_{\mathfrak{B}}$ bounded in \mathfrak{B} , $\mathfrak{B}_{\mathfrak{Y}}$ compact in $\mathcal{X} := C([0, T] \rightarrow \mathcal{X})$, and such that

$$(4.13) \quad \mathbf{S}_\phi(v_0 + \mathfrak{B}_{\mathfrak{B}}) \subset \mathfrak{B}_{\mathfrak{Y}}, \quad \tilde{\phi}\mathfrak{B}_{\mathfrak{Y}} \subset \mathfrak{B}_{\mathfrak{Z}}, \quad \mathbf{C}\mathfrak{B}_{\mathfrak{Z}} \subset \mathfrak{B}_{\mathfrak{B}}$$

for each $\tilde{\phi} \in \mathcal{F}_*$. Since we will only be interested in $\phi(\cdot, x)$ for x in the compact set $\mathfrak{B}_{\mathfrak{Y}}$, we can³ approximate ϕ by $\tilde{\phi}_k \in \mathcal{F}_*$ for which, say,

$$(4.14) \quad \|\tilde{\phi}_k(\cdot, x) - \phi(\cdot, x)\|_{\mathfrak{Z}} \leq \frac{1}{k} \quad (x \in \mathfrak{B}_{\mathfrak{Y}}).$$

Since each $\tilde{\phi}_k$ is in \mathcal{F}_* (so Theorem 2 applies and Approach 1 to compactness is available) we may apply Theorem 1 and have a fixpoint $w_k \in \mathfrak{B}_{\mathfrak{B}}$ of $\mathbf{W} = \mathbf{W}_k$ (i.e., defined as above using $\tilde{\phi}_k$ for ϕ); let x_k be the corresponding solution of (1.6) and let $g_k := \phi_k(\cdot, x_k)$. We have x_k in the compact set $\mathfrak{B}_{\mathfrak{Y}}$ so, extracting a subsequence if necessary, we have convergence: $x_k \rightarrow x_*$ in the sense of \mathcal{X} for some $x_* \in \mathfrak{B}_{\mathfrak{Y}}$. Now, setting $g_* := \phi(\cdot, x_*)$, we have

$$\begin{aligned} \|g_k - g_*\|_{\mathfrak{Z}} &= \|\phi_k(\cdot, x_k) - \phi(\cdot, x_*)\|_{\mathfrak{Z}} \\ &\leq \|\phi_k(\cdot, x_k) - \phi(\cdot, x_k)\|_{\mathfrak{Z}} + \|\phi(\cdot, x_k) - \phi(\cdot, x_*)\|_{\mathfrak{Z}}. \end{aligned}$$

The first term goes to 0 by (4.14) while $x_k \rightarrow x_*$ in \mathcal{X} clearly gives $x_k \rightarrow x_*$ in $\mathfrak{Y} := \bar{L}^p([0, T] \rightarrow \mathcal{X})$ and so, using (4.2) and Krasnoselski's theorem to obtain continuity of $\phi: \mathfrak{Y} \rightarrow \mathfrak{Z}$, the second term also goes to 0. Thus, $g_k \rightarrow g_*$ in \mathfrak{Z} so $w_k := \mathbf{C}g_k \rightarrow \mathbf{C}g_* =: w_*$

² Note that in using (4.2) in Approach 1 we did *not* suggest that it replace (3.1) which, as observed here, plays a key role in determining \mathfrak{B} .

³ In view of the assumption (2.3)(ii), the approximation can be accomplished relatively straightforwardly. The construction is rather cumbersome and we relegate it to the Appendix.

by the continuity of \mathbf{C} . But now we have

$$\begin{aligned} x_k &\rightarrow x_* \quad \text{in } \mathfrak{X}, \\ \bar{x}_k &= x + \mathbf{S}g_k + \mathbf{SB}(v_0 + w_k) \\ &\rightarrow \bar{x} + \mathbf{S}g_* + \mathbf{SB}(v_0 + w_*) \quad \text{in } \mathfrak{X} \quad (\text{by Lemma 1}) \\ &= \bar{x} + \mathbf{S}\phi(\cdot, x_*) + \mathbf{SB}(v_0 + w_*). \end{aligned}$$

Hence x_* does indeed satisfy⁴ the limit equation: (1.6) using ϕ itself. Finally, since $w_* = \mathbf{C}g_*$ it follows that $x_*(T) = \bar{x}(T) + [\mathbf{SB}v_0](T)$. As in the proof of Theorem 1, this suffices to show $\mathcal{H}_0 \subset \mathcal{H}_\phi$ and so the invariance.

The argument is almost the same for situations in which one might apply Approach 2. Suppose, then, we are given the conclusions of Lemma 2, (3.5), (4.4) in the context of a differential equation (4.3) with a function $\tilde{\phi}$ satisfying (4.6) but not (3.6). Now let $\tilde{\mathfrak{F}}_*$ be the set of all $\tilde{\phi}$ satisfying (4.6) with the same α, β, r and also satisfying (3.6) with some choice of \tilde{p}_3 . As above, we find $\mathfrak{B}_{\mathfrak{Z}}$ (bounded in \mathfrak{B}), $\mathfrak{B}_{\mathfrak{Y}}$, $\mathfrak{B}_{\mathfrak{Z}}$ such that (4.13) holds for each $\tilde{\phi}$ in $\tilde{\mathfrak{F}}_*$. Since the set $\mathfrak{B}_{\mathfrak{Y}}$ can be defined solely by the bounds obtained from (4.6), (4.4), its compactness in \mathfrak{Y} follows from Aubin's theorem without reference to the Lipschitz condition (3.6). Again, noting the compactness of $\mathfrak{B}_{\mathfrak{Y}}$ —hence the uniform continuity of $\tilde{\phi}$ on $\mathfrak{B}_{\mathfrak{Y}}$ —we expect to approximate ϕ on $\mathfrak{B}_{\mathfrak{Y}}$ as in (4.14) by a sequence $\{\tilde{\phi}_k\}$ in $\tilde{\mathfrak{F}}_*$. The remainder of the argument continues as earlier: Taking the fixpoints $w_k \in \mathfrak{B}_{\mathfrak{W}}$ of $\mathbf{W} = \mathbf{W}_k$ given by using Approach 2 with Theorem 2 and the proof of Theorem 1, let $y_k \in \mathfrak{B}_{\mathfrak{Y}}$ be the corresponding solutions and, noting compactness of $\mathfrak{B}_{\mathfrak{Y}}$ in \mathfrak{Y} , assume $y_k \rightarrow y_*$. Now $g_k := \tilde{\phi}_k(\cdot, y_k) \rightarrow g_* := \phi(\cdot, y_*)$ in \mathfrak{Z} and $w_k := \mathbf{C}g_k \rightarrow w_* := \mathbf{C}g_*$ as earlier. It is now the continuity of $\mathbf{S}: \mathfrak{Z} \rightarrow \mathfrak{Y}$ and $\mathbf{SB}: \mathfrak{B} \rightarrow \mathfrak{Y}$ given by (3.5) which gives

$$x_* \leftarrow x_k \rightarrow \bar{x} + \mathbf{S}\phi(\cdot, x_*) + \mathbf{SB}(v_0 + w_*)$$

and so the limit equation. Again this suffices to show $\mathcal{H}_0 \subset \mathcal{H}_\phi$ and so the desired invariance. \square

At this point it seems useful to gather together the results we have obtained. Recall that the basic setting is the abstract integral equation

$$(4.15) \quad x(t) = \bar{x}(t) + \int_0^t \Psi(t, s)[\phi(s, x(s)) + \mathbf{B}(s)v(s)] ds$$

or essentially equivalently, the differential equation

$$(4.16) \quad \dot{x}(t) = \phi(t, x(t)) + \mathbf{B}(t)v(t) + \mathbf{A}(t)x(t), \quad x(0) = x_0$$

for $0 \leq t \leq T$. Without seeking the most general possible version of what has been demonstrated in the previous sections, we consider:

- (4.17) (i) Banach spaces $\mathfrak{X}, \mathfrak{Y}, \mathfrak{Z}$ (of comparable elements) and \mathcal{V} ;
 (ii) Exponents $1 < p, p', \bar{p} < \infty$ with $\bar{p} = rp, r < 1$;
 (iii) $\mathfrak{X} := C([0, T] \rightarrow \mathfrak{X})$, $\mathfrak{Y} := L^{\bar{p}}([0, T] \rightarrow \mathfrak{Y})$,
 $\mathfrak{Z} := L^p([0, T] \rightarrow \mathfrak{Z})$, $\mathfrak{B} := L^{p'}([0, T] \rightarrow \mathcal{V})$;
 (iv) Operators $\Psi(t, s), \mathbf{B}(s)$, measurable for $0 \leq s \leq t \leq T$ and with
 $\Psi(t, \tau)\Psi(\tau, s) = \Psi(t, s)$;

⁴ Note that without the Lipschitz condition (3.6) we do not necessarily know the uniqueness of this solution.

- (v) A function $\phi: [0, T] \times \mathcal{Y} \rightarrow \mathcal{X}$ satisfying Carathéodory conditions and a function $\bar{x} \in \mathcal{X} \cap \mathcal{Y}$

and impose the conditions (2.3), (3.5), i.e.,

- (4.18) (i) $|\psi(t, s)|_{\mathcal{X} \rightarrow \mathcal{X}} \leq \rho_1(t-s)$ with $\rho_1 \in L_+^q(0, T)(1/q + 1/p = 1)$,
 (ii) $|\psi(t, s)|_{\mathcal{X} \rightarrow \mathcal{Y}} \leq \bar{\rho}_1(t-s)$ with $\bar{\rho}_1 \in L_+^{\bar{q}}(0, T)(1/\bar{q} + 1/p \leq 1 + 1/\bar{p})$,
 (iii) $|\psi(t, s)\mathbf{B}(s)|_{\mathcal{Y} \rightarrow \mathcal{X}} \leq \rho_2(t-s)$ with $\rho_2 \in L_+^q(0, T)(1/q' + 1/p' = 1)$,
 (iv) $|\psi(t, s)\mathbf{B}(s)|_{\mathcal{Y} \rightarrow \mathcal{Y}} \leq \bar{\rho}_2(t-s)$ with $\bar{\rho}_2 \in L_+^{\bar{q}}(0, T)$
 $(1/\bar{q}' + 1/p' \leq 1 + 1/\bar{p})$,
 (v) $|\psi(t, s) - \psi(t', s)|_{\mathcal{X} \rightarrow \mathcal{X}} \leq \varepsilon$, $|\psi(t, s) - \psi(t', s)]\mathbf{B}(s)|_{\mathcal{Y} \rightarrow \mathcal{X}} \leq \varepsilon$
 for $0 \leq s \leq t' - \varepsilon$, $t' < t = t' + h \leq T$ with $\varepsilon = \varepsilon(h) \rightarrow 0+$ as $h \rightarrow 0+$.

Under the conditions (4.18)(i), (4.18)(iii), (4.18)(v) we showed (Lemma 2) existence of continuous operators $\mathbf{S}: \mathcal{Y} \rightarrow \mathcal{X}$, $\mathbf{SB}: \mathcal{Y} \rightarrow \mathcal{X}$ such that the solution of $[\dot{x} + \mathbf{A}x = g + \mathbf{B}v$, $x(0) = 0]$ is given by $x = \mathbf{S}g + \mathbf{SB}v$, and of continuous operators $\mathbf{T}: \mathcal{Y} \rightarrow \mathcal{X}$, $\mathbf{TB}: \mathcal{Y} \rightarrow \mathcal{X}$ such that $\mathbf{T}g := [\mathbf{S}g](T)$, $\mathbf{TB}v := [\mathbf{SB}v](T)$. In terms of these we make the underlying assumption that

(H₀) $\mathcal{R}(T) \subset \mathcal{R}(\mathbf{TB})$, i.e., for each $f \in \mathcal{Y}$ there exists $v \in \mathcal{Y}$ such that $\mathbf{T}(g + \mathbf{B}v) = 0$. This is, of course, quite a strong controllability assumption, corresponding to the invariance of the reachable set

$$\mathcal{K}_g := \{x(T): x \text{ satisfies (4.15) for some } v \in \mathcal{Y}\}$$

as g ranges over \mathcal{Y} , i.e., taking

$$\phi(t, \xi) := g(t) \quad \text{for some } g \in \mathcal{Y}.$$

We have treated (H₀) here as given a priori and did not attempt to investigate specific situations for which it might be verifiable; see, however, Remark 3 in the next section.

THEOREM 3. Assume (4.17), (4.18) and assume that ϕ satisfies a growth condition:

$$(4.19) \quad |\phi(s, \eta)|_{\mathcal{X}} \leq \alpha(s) + \beta|\eta|_{\hat{\mathcal{Y}}}^r \quad \text{with } \alpha \in L_+^p(0, T), \quad r < 1,$$

and a Lipschitz condition:

$$(4.20) \quad |\psi(t, s)[\phi(s, \eta) - \phi(s, \eta')]|_{\mathcal{Y}} \leq \bar{\rho}_3(t-s)|\eta - \eta'|_{\mathcal{Y}} \quad \text{with } \bar{\rho}_3 \in L_+^1(0, T)$$

or, for Cases 1 and 2 below, can be approximated as in Remark 2, (4.14) by $\{\tilde{\phi}_k\}$ satisfying this condition. We consider four cases in the assumption regarding ϕ :

Case 1. $\hat{\mathcal{Y}} = \mathcal{Y} = \mathcal{X} \hookrightarrow \hat{\mathcal{X}}$ (compact embedding) and (4.19) is supplemented by

$$(4.21) \quad |\phi(s, \eta)|_{\mathcal{X}} \leq \alpha_M(s) \quad \text{for } |\eta|_{\hat{\mathcal{X}}} \leq M$$

with $\alpha_M \in L_+^p(0, T)$ for each $M > 0$.

Case 2. Consider (4.15) as equivalent to (4.16) with \mathbf{A} , \mathbf{B} as in (4.4) for some $\tilde{\mathcal{Y}}$ ($\mathcal{Y}, \mathcal{X} \hookrightarrow \tilde{\mathcal{Y}}$); for (4.20) assume $\mathcal{Y} \hookrightarrow \hat{\mathcal{Y}}$ is a compact embedding.

Case 3. Consider (4.15) equivalent to (4.16) with \mathbf{A} , \mathbf{B} as in (4.4); require (4.20) directly for ϕ ; taking $\hat{\mathcal{Y}} \hookrightarrow \mathcal{Y}$ (compact embedding) replace (4.18)(iv) by

$$(4.22) \quad |\psi(t, s)\mathbf{B}(s)|_{\mathcal{Y} \rightarrow \hat{\mathcal{Y}}} \leq \bar{\rho}_2(t-s) \quad \text{with } \bar{\rho}_2 \in L_+^{\bar{q}}(0, T) \quad (1/\bar{q}' + 1/p' \leq 1 + 1/\bar{p}).$$

Case 4. Consider (4.15) equivalent to (4.16) with \mathbf{A} , \mathbf{B} as in (4.4); require (4.20) directly for ϕ_k taking $\mathcal{V} \hookrightarrow \mathcal{Q}$ (compact embedding) and assume

$$(4.23) \quad |\psi(t, t - \varepsilon)|_{\mathcal{Q} \rightarrow \mathcal{Q}} \leq M_\varepsilon \quad (\pi \leq t \leq T)$$

for each $\varepsilon > 0$ and some M_ε .

Then if one has invariance of the reachable set under affine perturbations in \mathfrak{J} , i.e., if

(4.24) For each $g \in \mathfrak{J}$ there is some control $v \in \mathfrak{V}$ such that

$$\int_0^T \psi(T, s) \mathbf{B}(s) v(s) ds = \int_0^T \psi(T, s) g(s) ds;$$

then one also has invariance of the reachable set ($\mathcal{K}_\phi = \mathcal{K}_0$) for the nonlinear perturbation (4.15).

Proof. This merely summarizes the preceding discussions. \square

5. Further remarks. In this section we supplement the rather abstract discussion above by some remarks on possible settings in which various of the hypotheses can be verified.

Remark 3. The most difficult of the hypotheses made is (H_0) . Clearly, for any case in which \mathcal{K}_0 is the entire state space \mathcal{X} this is immediate. The results obtained are then very much along the main line of development of this fixed point approach to nonlinear control.

An equally immediate verification is available if $\mathfrak{J} \subset \mathbf{B}\mathfrak{V}$, say, if $p' \leq p$ in (2.2) and $\mathcal{X} \subset \mathbf{B}\mathcal{V}$. In this case, however, the final conclusion is equally immediate without the machinery of this paper: given $\xi \in \mathcal{K}_0$ with $x' + \mathbf{A}x = \mathbf{B}v_0$ giving $x(\tau) = \xi$, we need only take $g := \phi x \in \mathfrak{J}$ for this x and take w such that $\mathbf{B}w = -g$. Then $v := v_0 + w$ gives $x' + \mathbf{A}x = \phi(\cdot, x) + \mathbf{B}v$ with, of course, the same terminal value ξ since one has the same solution x .

The only other situation for which, *as of now*, we can show how to verify (H_0) involves ϕ of the form

$$(5.1) \quad \phi(t, \eta) := \sum_i^n \phi_k(t, \eta) z_k$$

where each $\phi_k(t, \cdot)$ is a scalar-valued functional, i.e., $\mathcal{X} = \text{span}\{z_1, \dots, z_n\}$ is finite-dimensional. If one were to have, e.g., an autonomous equation, so $\psi(\cdot, \cdot)$ is the semigroup generated by $-\mathbf{A}$, then requiring invariance of \mathcal{X} under \mathbf{A} (hence under ψ) and that $\mathcal{X} \subset \mathcal{K}_0$ would be sufficient to ensure (H_0) ; note that exact nullcontrollability for the linear problem would ensure $\mathcal{X} \subset \mathcal{K}_0$. In particular, this method of verification for (H_0) would apply for boundary control of the heat equation (see Remark 4 below) if each z_k were an eigenfunction of $[-\Delta; BC]$. Somewhat more generally, we may consider any closed (e.g. finite-dimensional) subspace $\mathcal{X}_* \subset \mathcal{K}_0$ and assume that \mathcal{X} is such that

$$(5.2) \quad \psi(T, s)z \in \mathcal{X}_* \quad \text{for } z \in \mathcal{X}, \quad s \in (0, T).$$

The point is that (1.8) shows that (5.2) gives $\mathbf{T}g \in \mathcal{X}_* \subset \mathcal{K}_0$ for each $g \in \mathfrak{J}$, assuming the integral converges (for which we need our other hypotheses), which gives (H_0'') . \square

Remark 4. We next consider settings for which the hypotheses for Lemma 1 can be verified.

The simplest case (compare [7]) would be to have $\mathcal{X} = \mathcal{X}$ and ψ continuous on the triangle $\{0 \leq s \leq t \leq T\}$ to be bounded operators on \mathcal{X} with $\mathbf{B}(\cdot)$ continuous on $[0, T]$ to bounded operators: $\mathcal{V} \rightarrow \mathcal{X}$.

The conditions (2.3) were designed to consider, e.g., such possibilities as the heat equation with \mathcal{L} "nastier" than $L^2(\Omega)$. Slightly more generally, take $\mathcal{X} := L^2(\Omega)$ and $-\mathbf{A}$ the generator of an analytic semigroup. If \mathbf{A} is second order, then $\mathcal{D}(\mathbf{A}^\sigma) = H^{2\sigma}(\Omega)$ to within consideration of boundary conditions. A standard estimation (cf., e.g., [5]) gives

$$(5.3) \quad |\psi(\tau)z|_{\mathcal{X}} \leq |\mathbf{A}^\sigma \psi(\tau)| |\mathbf{A}^{-\sigma} z|_{\mathcal{X}} \leq M\tau^{-\sigma} |z|_{\mathcal{X}}$$

for $\mathcal{X} := H^{-2\sigma}(\Omega)$. If we take $\sigma < 1/q := 1 - 1/p$, then setting $\rho_1(\tau) := M\tau^{-\sigma}$ in (2.3)(ii) gives $\rho_1 \in L^q$. Again, a standard semigroup estimation gives, with this choice of \mathcal{X} ,

$$(5.4) \quad \begin{aligned} |\psi(t+h-s) - \psi(t-s)|_{\mathcal{X} \rightarrow \mathcal{X}} &\approx \left| \mathbf{A} \int_{t-s}^{t-s+h} \psi(\tau) d\tau \right|_{\mathcal{X}} \\ &\leq M \int_{t-s}^{t-s+h} \tau^{-(1+\sigma)} d\tau \leq Mh\delta^{-(1+\sigma)} \end{aligned}$$

got $h > 0$, $t-s \geq \delta > p$. Thus, since $\sigma < 1$, we may take

$$\delta = \delta(h) := h^{1/2}, \quad \varepsilon := \max \{ \delta, Mh\delta^{-(1+\sigma)} \}$$

to have $\varepsilon = \varepsilon(h) \rightarrow 0$ as $h \rightarrow 0$, giving the first part of (2.3)(iv). Similar considerations apply to $\psi\mathbf{B}$.

Of particular interest, also, is the possibility of treating boundary control in this framework. While some more general situations could be treated, we take \mathbf{A} as above, so the equation is an autonomous parabolic equation of the form

$$(5.5) \quad \dot{x} + \mathbf{A}_0 x = \phi(\cdot, x), \quad \beta x = v.$$

Here β is a boundary operator and what corresponds to \mathbf{B} is the effect of the control v appearing in the boundary conditions on the evolution of x . We have written \mathbf{A}_0 in (5.5) to indicate that specification of the domain $\mathcal{D}(\mathbf{A})$ includes homogeneous boundary conditions [$\beta y = 0$]; \mathbf{A}_0 is the same (elliptic) differential operator "pointwise". We take $\mathcal{X} \subset \mathcal{X} = \mathcal{Y} := L^2(\Omega)$; \mathcal{V} will be a space of functions on $\partial\Omega$ —say, $L^2(\Gamma) = \{\text{functions in } L^2(\partial\Omega) \text{ with support in } \Gamma\}$ for some choice of $\Gamma \subset \partial\Omega$. Assuming 0 is not an eigenvalue of \mathbf{A} , we define a Green's operator \mathbf{D} by

$$(5.6) \quad \mathbf{D}: \zeta \rightarrow z: \mathcal{V} \rightarrow \mathcal{X} \quad \text{with } \mathbf{A}_0 z = 0, \quad \beta z = \zeta.$$

The important point now is that standard trace and regularity theory for these elliptic operators tells us (assuming $\partial\Omega$ is smooth enough) that

$$(5.7) \quad \mathbf{A}^\sigma \mathbf{D}: \mathcal{V} \rightarrow \mathcal{X} \quad \text{is bounded for } \sigma < \bar{\sigma}$$

where $\bar{\sigma} := \frac{3}{4}$ if β is the Dirichlet trace and $\bar{\sigma} = \frac{3}{4}$ if β is (uniformly) of first order, e.g., the Neumann conditions.

We now use the fact that there is a variation of parameters formula corresponding to (1.2), for this situation⁵:

$$(5.8) \quad x(t) = \psi(t, 0)x(0) + \int_0^t [\psi(t-s)\phi(s, x(s)) + \mathbf{A}^{1-\sigma}\psi(t-s)\mathbf{A}^\sigma \mathbf{D}v(s)] ds.$$

⁵ See, e.g., [2] for this, but we sketch an argument: In (5.5) write $x = y + z$ with $z := \mathbf{D}v$ so $\dot{y} + \mathbf{A}y = \phi - \dot{z}$. An integration by parts of the variation of parameters formula for y leads to (5.8) since $d\psi/dt = \mathbf{A}\psi = \mathbf{A}^{1-\sigma}\psi\mathbf{A}^\sigma$.

This essentially replaces the previous use of $[\Psi(t, s)\mathbf{B}(s)]$ by use of $[\mathbf{A}^{1-\sigma}\Psi(t-s)] \times [\mathbf{A}^\sigma \mathbf{D}]$. Noting (5.7), we can use the same kind of estimate $|\mathbf{A}^\lambda \Psi(\tau)| = \mathcal{O}(\tau^{-\lambda})$ as for (5.3) above to verify the hypotheses. We will take $\hat{\mathcal{Y}} := \mathcal{D}(\mathbf{A}^\varepsilon)$ in (4.7) = (4.22) for our “compactness hypothesis,” noting that $\mathcal{D}(\mathbf{A}^\varepsilon) = H^{2\varepsilon}(\Omega) \hookrightarrow L^2(\Omega) = \mathcal{Y}$ is a compact embedding for $\varepsilon > 0$.

Suppose we let β be the Dirichlet trace. If Γ is a large enough part of $\partial\Omega$, it is known that we have exact nullcontrollability whence, as in Remark 3, (H_0) is verifiable for ϕ as in (5.1) with z_k eigenfunctions of \mathbf{A} ; in any case, assume that (H_0) holds. Since $\mathcal{X} \subset \mathcal{X} = \mathcal{Y}$, we may take $\rho_1, \bar{\rho}_1$ to be constant in (4.18). We may take $\rho_2(\tau) := M\tau^{-(1-\sigma)}$ in $L^{q'}$ with $q' < \frac{4}{3}$ (as $\sigma < \bar{\sigma} = \frac{1}{4}$ here so $1 - \sigma > \frac{3}{4}$) corresponding to taking $p' > 4$ for (2.2). For (4.22) with $\hat{\mathcal{Y}} = H^{2\varepsilon}(\Omega)$, we may take $\bar{\rho}_2(\tau) := M\tau^{-(1-\sigma+\varepsilon)}$ and want this to be in $L^{\bar{q}'}$. Since we may choose any $\sigma < \frac{1}{4}$, $\varepsilon > 0$, this only requires $q' > \frac{4}{3}$ also which means that there are essentially no new restrictions on \bar{p} or p . While we cannot simply apply Theorem 3 directly to this boundary control setting, it is clear that the discussion here can be viewed either as an interpretation of the theorem in that setting or as indicating the requisite (minor) modifications of the arguments: either way we now may take the theorem as applicable. We conclude that if we can verify the Carathéodory and growth conditions on ϕ and (H_0) for

$$(5.9) \quad \begin{aligned} \mathcal{X} &:= L^2(\Omega) \supset \mathcal{X}, \\ \mathfrak{J} &:= L^p([0, T] \rightarrow \mathcal{X}), \quad 1 < p < \infty, \\ \mathfrak{B} &:= L^{p'}([0, T] \rightarrow L^2(\Gamma)), \quad 4 < p' < \infty, \end{aligned}$$

then the perturbation by ϕ appearing in (5.5) does not alter the exactly reachable set $\mathcal{H}_0 \subset \mathcal{X}$ associated with the linear boundary control problem.

The case for first order boundary control is similar. Now we have $\bar{\sigma} = \frac{3}{4}$ so we can have any $q' < 4$ (corresponding to $p' > \frac{4}{3}$) by again taking p_2 of the form $M\tau^{-(1-\sigma)}$, since now $1 - \sigma$ can be arbitrarily close to $\frac{1}{4}$. Similarly, we have $q' < 4$ and again there is no new restriction on p, \bar{p} . The situation is just as for (5.9) except that now we have $\frac{4}{3} < p' < \infty$ for \mathfrak{B} . \square

Remark 5. We turn finally to consideration of settings for which the hypotheses of Theorem 2 can be verified.

The conditions (3.1)–(3.6), specifically the introduction of the space \mathcal{Y} as distinguished from \mathcal{X} , were formulated so as to permit consideration of such problems as, e.g.,

$$(5.10) \quad \dot{x} - \Delta x = \phi(\cdot, x, \nabla x) + \mathbf{B}v, \quad x|_{\partial\Omega} = 0$$

on a suitable bounded region Ω in \mathbb{R}^m . Here, we might have $\mathcal{X} \subset \mathcal{X} := L^2(\Omega)$ but wish to take $\mathcal{Y} := H_0^1(\Omega) = \mathcal{D}(\mathbf{A}^{1/2})$ to have some suitable kind of (Lipschitz) continuity for the dependence of ϕ on the state x . The point is that if we think of $\phi = \phi(\cdot, x, \xi)$ with Lipschitz continuity and sublinear (power r) growth in the pair $[x, \xi]$, then this choice of \mathcal{Y} gives

$$(5.11) \quad \begin{aligned} |\phi(\cdot, x, \nabla x)|_{\mathcal{X}} &\leq \alpha + \beta(|x|_{\mathcal{X}} + |\nabla x|_{\mathcal{X}}^r) \\ &\leq \alpha + \beta'|x|_{\mathcal{Y}}^r \end{aligned}$$

and, as in Remark 4, we use the standard semigroup estimate (for such analytic semigroups)

$$(5.12) \quad |\Psi(\tau)|_{\mathcal{X} \rightarrow \mathcal{Y}} \leq M|\mathbf{A}^{1/2}\Psi(\tau)|_{\mathcal{X} \rightarrow \mathcal{X}} \leq M\tau^{-1/2}.$$

That is (5.12) corresponds to taking $\bar{\rho}_1(\tau) = M\tau^{-1/2}$ in (4.18)(ii) which is admissible since, as we noted at (3.5)(ii), we can always take $\bar{q} = 1$. Note that if, e.g., we were to take $p = p' = 2$ then we would expect, according to Theorem 2, a solution in

$$\mathfrak{X} \cap \mathfrak{Y} = C([0, T] \rightarrow L^2(\Omega)) \cap L^2([0, T] \rightarrow H_0^1(\Omega)),$$

which is exactly the standard estimate obtainable by energy methods.⁶

Note that in writing (5.10) we do *not* necessarily assume that $\phi(\cdot, \cdot)$ is defined as a Nemytsky operator *pointwise spatially* but admit possibilities along the lines of (5.1) with the functionals $\phi_k(t, \cdot)$ uniformly continuous from \mathfrak{Y} , e.g., integral functionals over Ω involving $x, \nabla x$ suitably. Thus, the discussion of Remark 3 can contribute to the verification of (H_0) for a problem of this sort. Alternatively, if one were to take $\mathfrak{X} = \mathcal{X}$ then one would require a very strong controllability assumption to have (H_0) . \square

Appendix. As promised in footnote 4, we wish to verify the possibility of approximating ϕ as in (4.14). In view of the assumption (2.3)(ii) and the setting $\mathfrak{Y} = \mathcal{X} \rightarrow \mathcal{X}$, our task is to approximate ϕ satisfying (3.1), (4.2) by $\tilde{\phi}$ such that

$$\begin{aligned} \text{(A.1)} \quad & \text{(i)} \quad \|\phi(\cdot, x) - \tilde{\phi}(\cdot, x)\|_3 \leq \bar{\varepsilon} \quad \text{for all } x \in \mathfrak{B}_\mathfrak{Y}, \\ & \text{(ii)} \quad |\tilde{\phi}(x, \eta) - \tilde{\phi}(s, \eta')|_{\mathcal{X}} \leq C|\eta - \eta'|_{\mathcal{X}} \quad \text{for all } \eta, \eta' \in \mathcal{R}. \end{aligned}$$

Here it is given that $\mathfrak{B}_\mathfrak{Y}$ is compact in $C([0, T] \rightarrow \mathcal{X})$ so we have $x(s) \in \mathcal{R}$ for $s \in [0, T]$, $x \in \mathfrak{B}_\mathfrak{Y}$ for some compact set $\mathcal{R} \subset \mathcal{X}$; with no loss of generality, take \mathcal{R} convex. The approximation is to be possible with $\bar{\varepsilon} > 0$ in (A.1)(i) arbitrarily small; the Lipschitz constant C in (A.1)(ii) will in general depend on $\tilde{\phi}$ (i.e., on $\bar{\varepsilon}, \dots$). Note that $\phi: [0, T] \times \mathcal{X} \rightarrow \mathcal{X}$ satisfies Carathéodory conditions by (4.2) so, in particular, $\phi(s, \cdot)$ is uniformly continuous on \mathcal{R} for a.e. s —there exists $\delta(\varepsilon, s) \geq 0$ with

$$\begin{aligned} \text{(A.2)} \quad & \text{(i)} \quad |\phi(s, \eta) - \phi(s, \eta')|_{\mathcal{X}} \leq \varepsilon \quad \text{for } \eta, \eta' \in \mathcal{R}, \quad |\eta - \eta'|_{\mathcal{X}} \leq \delta(\varepsilon, s), \\ & \text{(ii)} \quad \text{meas } \mathcal{S}(\bar{\delta}) \rightarrow 0 \quad \text{as } \bar{\delta} \rightarrow 0, \quad \text{where } \mathcal{S}(\bar{\delta}) = \mathcal{S}(\bar{\delta}, \varepsilon) := \{s: \delta(\varepsilon, s) < \bar{\delta}\}. \end{aligned}$$

Since $\mathcal{S}(\bar{\delta})$ decreases with $\bar{\delta}$, (4.2) permits us to choose $\bar{\delta} = \bar{\delta}(\varepsilon)$ such that the restriction of $\phi(\cdot, x)$ to $\mathcal{S}(\bar{\delta})$ has \mathfrak{B} -norm less than ε for all (measurable) $x: [0, T] \rightarrow \mathcal{R}$; set $\mathcal{S}^c(\bar{\delta}) = [0, T] \setminus \mathcal{S}(\bar{\delta})$. Now take a finite covering of \mathcal{R} by $\delta/4$ -balls with centers $\{\eta_1, \dots, \eta_N\}$ in \mathcal{R} and set $\mathcal{R}' := \text{hull}(\eta_1, \dots, \eta_N)$. Let \mathbf{P} be a Lipschitzian retraction of \mathcal{R} to \mathcal{R}' such that $|\eta - \mathbf{P}(\eta)| \leq \bar{\delta}/2$, possible as each point of \mathcal{R} is within $\bar{\delta}/4$ of \mathcal{R}' . We may assume \mathcal{R}' can be triangulated: partitioned into simplices $\{S_j\}$ with vertices which may be taken from among $\{\eta_1, \dots, \eta_N\}$, introducing more vertices from \mathcal{R}' if necessary, such that $\text{diam } \bar{S}_j \leq \delta/2$. Now set

$$\text{(A.3)} \quad \tilde{\phi}_0(s, \eta_n) := \phi(s, \eta_n) \quad (n = 1, \dots, N') \quad \text{and } \tilde{\phi}_0(s, \cdot) \text{ piecewise linear (the "pieces" being the simplices } S_j)$$

and then define

$$\text{(A.4)} \quad \tilde{\phi}(s, \eta) := \begin{cases} \phi(x, \bar{\eta}) & \text{for } s \in \mathcal{S}(\bar{\delta}, \varepsilon) \quad (\text{using any } \bar{\eta} \text{ fixed in } \mathcal{R}'), \\ \tilde{\phi}_0(s, \mathbf{P}(\eta)) & \text{for } s \notin \mathcal{S}(\bar{\delta}, \varepsilon), \quad \eta \in \mathcal{R}. \end{cases}$$

⁶ That is, multiply by x , use the Divergence Theorem, integrate over $[0, t]$ and then apply the Gronwall Inequality. (It would also be possible to take $\mathfrak{Y} := H^1(\Omega) \cap H^{2\sigma}(\Omega)$ for any $r < 1$ since this would give $\bar{\rho}(\tau) = M\tau^{-\sigma}$ in $L^1(0, T)$ but these methods do not give the result: $x \in L^2([0, T] \rightarrow H^2(\Omega))$ which is obtainable for (5.10) by a slightly different energy method, multiplying by \dot{x} , etc.)

One can uniformly bound the Lipschitz constants for $\tilde{\phi}_0(s, \cdot)$ on each simplex S_j in view of the bound $1/\bar{\delta}$ for $\{\phi(s, \eta_n)|_x\}$ and the minimum separation of the (finitely many) points $\{\eta_n\}$. Hence this $\tilde{\phi}$ is uniformly Lipschitzian in the sense of (A.1)(ii). To see (A.1)(i), suppose $\eta \in \mathcal{R}$ and $P(\eta) \in S_j (j = j(\eta))$ where S_j has vertices $\{\eta_k\}_{j(\eta)}$. As $|\eta - P(\eta)|_x \leq \bar{\delta}/2$ and $\text{diam } S_j \leq \bar{\delta}/2$ we have $|\eta - \eta_k|_x \leq \bar{\delta}$ so (A.2)(i) gives $|\phi(s, \eta) - \tilde{\phi}_0(s, \eta_k)|_x \leq \varepsilon$ for $s \in \mathcal{S}(\bar{\delta}, \varepsilon)$ and so $|\phi(s, \eta) - \tilde{\phi}(s, \eta)|_x \leq \varepsilon$ since $\tilde{\phi}(s, \eta) \in \text{hull } \{\phi_0(s, \eta_k)\}_{j(\eta)}$ by the piecewise linearity. Thus, by the choice of $\bar{\delta}$,

$$\begin{aligned} \|\phi(\cdot, x) - \tilde{\phi}(\cdot, x)\|_3 &\leq 2\varepsilon + \left[\int_{[0, T] \setminus \mathcal{S}(\bar{\delta}, \varepsilon)} |\phi(\cdot, x) - \tilde{\phi}(\cdot, x)|^r \right]^{1/p} \\ &\leq 2\varepsilon + T^{1/p} \varepsilon \end{aligned}$$

which gives (A.1)(i) as $\varepsilon > 0$ is arbitrary. \square

REFERENCES

- [1] J. P. AUBIN, *Un théorème de compacité*, C. R. Acad. Sci. Paris, 265 (1963), pp. 5042-5043.
- [2] A. V. BALAKRISHNAN, *Applied Functional Analysis*, Springer-Verlag, Berlin, New York, 1976, § 4.11.
- [3] N. CARMICHAEL AND M. D. QUINN, *Fixed point methods in nonlinear control*, in Distributed Parameter Systems, F. Kappel, K. Kunisch and W. Schapacher, eds., Springer-Verlag, Berlin, 1985, pp. 24-51.
- [4] J. A. GOLDSTEIN, *Semigroups of Linear Operators and Applications*, Oxford Univ. Press, Oxford, 1985.
- [5] D. HENRY, *Geometric Theory of Semilinear Parabolic Equations*, Lecture Notes in Math., 840, Springer-Verlag, Berlin, New York, 1981.
- [6] E. MICHAEL, *Continuous selections*, I, Ann. Math., 63 (1956), pp. 361-382.
- [7] K. NAITO, *Controllability of semilinear control systems*, I, this Journal, 25 (1987), pp. 715-722.
- [8] I. E. SEGAL AND R. A. KUNZE, *Integrals and Operators*, McGraw-Hill, New York, 1968.
- [9] T. I. SEIDMAN, *Time-invariance of the reachable set for linear control problems*, J. Math. Anal. Appl., 72 (1979), pp. 17-20.
- [10] ———, *Two compactness lemmas*, in Nonlinear Semigroups, Partial Differential Equations and Strange Attractors, T. L. Gill and W. W. Zachary, eds., Lecture Notes in Mathematics, Springer-Verlag, Berlin, New York, 1987.

DUALITY THEOREMS FOR AN OPTIMAL CONTROL PROBLEM WITH A LINEAR UNBOUNDED OPERATOR*

H. DIETRICH†

Abstract. The present paper applies the duality theory of Fenchel and Rockafellar to problems of optimal control with a linear operator-type equation and with a linear unbounded operator. If the linear operator is everywhere densely defined and closed, we can prove duality theorems and conditions for optimality. The results are applied to optimal control problems for systems with concentrated and distributed parameters.

Key words. duality theory, optimal control theory

AMOS(MOS) subject classification. 49A55, 49B27

1. Introduction and statement of the problem. Throughout this paper B_i , $i = 1, 2, 3$, will denote real reflexive Banach spaces with the dual spaces B_i^* . As usual, $\langle \cdot, \cdot \rangle_i$ will denote the canonical pairing between B_i and B_i^* , $i = 1, 2, 3$.

For the product space $B_1 \times B_2 \times B_3$ with the dual $B_1^* \times B_2^* \times B_3^*$ we define the canonical pairing by

$$\langle \cdot, \cdot \rangle_{1 \times 2 \times 3} = \langle \cdot, \cdot \rangle_1 + \langle \cdot, \cdot \rangle_2 + \langle \cdot, \cdot \rangle_3.$$

Furthermore let $\bar{R} := R \cup \{-\infty\} \cup \{+\infty\}$.

Many papers use the duality theory of Fenchel and Rockafellar for investigation of optimal control problems for distributed parameter systems (Lions [10], Ekeland and Temam [5], Barbu and Precupanu [2], Heins and Mitter [7], Mackenroth [11], [12], Mossino [13], Outrata [14]).

In the functional analytical description of such problems it is usual to choose the corresponding spaces of functions, for example as Sobolev spaces, so that the differential operators are continuous on the whole space. It is known that various differential operators can be unbounded. Under suitable assumptions on the unbounded operator we can use the duality theory of Fenchel and Rockafellar for investigation of such optimal control problems with an unbounded linear operator in the equation of state. We consider the following primal problem (P) of optimal control with a linear operator-type equation of state

$$(1.1) \quad \inf(P) := \inf \left\{ G(y, u) \left| \begin{array}{l} Ay + Bu + f = 0 \\ y \in D(A) \cap X, u \in U \end{array} \right. \right\}$$

where we let the cost function $G \in (B_1 \times B_2 \rightarrow \bar{R})$ be a real proper convex and lower semicontinuous function; let X or U , respectively, be a nonempty closed and convex set in B_1 or B_2 ; let $f \in B_3$ be a given element; let $B \in L(B_2, B_3)$ be a linear and continuous operator on B_2 and let the linear unbounded operator $A \in (B_1 \rightarrow B_3)$ in B_1 have an everywhere dense domain $D(A) \subset B_1$.

*Received by the editors February 2, 1984; accepted for publication (in revised form) June 25, 1986.

†Technische Hochschule Carl Schorlemmer, Leuna-Merseburg, Sektion Mathematik, Otto Nuschke Straße, DDR 4200 Merseburg, East Germany.

We assume that there exists an element $(y, u) \in (X \cap D(A)) \times U$ with $(y, u) \in \text{dom } G$ (the effective domain on the function G) and $Ay + Bu + f = 0$. Therefore for problem (1.1) $\inf(P) < +\infty$ is valid.

We interpret the variable y as variable of state, the variable u as variable of control.

The main objectives of this paper are the following:

- (i) To present a dual problem by duality theory according to Fenchel and Rockafellar's formalism;
- (ii) To show the duality relations and the necessary and sufficient conditions of optimality if certain conditions are given for G, A, B, X and U ;
- (iii) To show a lower bound for $\inf(P)$ by means of the dual problem;
- (iv) To show how the duality theory may be applied to certain control processes described by linear ordinary and partial differential equations and to quadratic and norm-minimal problems of optimal control.

This paper presents a new form of the dual problem different from works of Lions [10], Mossino [13], Mackenroth [11], [12] and Outrata [14] and is a generalization of known results on linear unbounded and closed operators A . This generalization is founded on the paper by Rockafellar [15].

2. The dual problem (D) and the relations between (P) and (D). If the primal minimization problem (1.1) is given, we consider a family of perturbed minimization problems

$$(2.1) \quad \varphi(x, v, w) = \inf_{y, u} F(y, u, x, v, w)$$

with

$$(2.2) \quad F(y, u, x, v, w) = G(y + x, u + v) + \delta((y, u, w) | K), \quad x \in B_1, v \in B_2, w \in B_3$$

and

$$K = \left\{ (y, u, w) \in B_1 \times B_2 \times B_3 \left| \begin{array}{l} Ay + Bu + f + w = 0 \\ y \in D(A) \cap X, u \in U \end{array} \right. \right\}$$

where $\delta(\cdot | K)$ is the indicator function of K .

The initial problem (P) corresponds to the value $(x, v, w) = (0, 0, 0)$. By means of the theory of Rockafellar in [15] (see also Ekeland and Temam [5]) we can derive the following dual problem (D) by introducing appropriate perturbations of (P) in (2.1) and (2.2). In the sense of convex analysis we introduce the support function of the sets $X \subset B_1$ and $U \subset B_2$ and we denote them by σ_X and σ_U , where for example

$$\sigma_X(y^*) = \sup_{y \in X} \langle y^*, y \rangle_1$$

holds.

THEOREM 1. *If A is a linear, everywhere densely defined and closed operator, the dual problem (D) of (P) is*

$$(2.3) \quad \sup(D) := \sup \left\{ \langle f, w^* \rangle_3 - G^*(x^*, v^*) - \sigma_X(-x^* - A^* w^*) - \sigma_U(-v^* - B^* w^*) \left| \begin{array}{l} x^* \in B_1^*, v^* \in B_2^* \\ w^* \in D(A^*) \end{array} \right. \right\}$$

and $\sup(D) \leq \inf(P)$ is valid.

In formula (2.3) G^* means the conjugate function of G , and A^* , or B^* , respectively, means the adjoint operator of A , or B .

Proof. A simple calculation shows that the conjugate function F^* of F from (2.2) is of the form

$$F^*(y^*, u^*, x^*, v^*, w^*) = -\langle f, w^* \rangle_3 + G^*(x^*, v^*) - \inf_{u \in U} \langle u, v^* - u^* + B^* w^* \rangle_2 \\ - \inf_{y \in D(A) \cap X} [\langle y, x^* - y^* \rangle_1 + \langle Ay, w^* \rangle_3].$$

Since the linear unbounded operator A is an everywhere densely defined and closed operator, it follows that

$$(2.4) \quad F^* = \begin{cases} -\langle f, w^* \rangle_3 + G^*(x^*, v^*) + \sigma_X(-x^* + y^* - A^* w^*) \\ + \sigma_U(-v^* + u^* - B^* w^*) & \text{if } w^* \in D(A^*), \\ +\infty & \text{otherwise,} \end{cases} \quad y^*, x^* \in B_1^*, \quad u^*, v^* \in B_2^*$$

and therefore we obtain (2.3) with $\sup(D) \leq \inf(P)$.

Remark 1. The case $A \in L(B_1, B_3)$ is contained in Theorem 1.

Remark 2. If the operator A is not closed, duality gaps can occur, independent of the properties of G, B, X and U . The following theorems show that for a closed operator A and certain conditions on G, B, X and U duality relations are valid.

The dual problem (2.3) is a problem of optimization without restrictions if the sets X and U are bounded. In contrast to recent papers, (2.3) does not contain the adjoint equation to $Ay + Bu + f = 0$ explicitly.

THEOREM 2. Let $G(y, u)$ be bounded below and continuous at (\bar{y}, \bar{u}) . If one of the following conditions (i) or (ii) is valid:

(i) There exists the inverse operator A^{-1} of A which is continuous on $R(A)$ (the range of A), for every $u \in U$ we have $Bu + f \in R(A)$; furthermore, there exist elements $\bar{y} \in D(A) \cap \text{int } X$ and $\bar{u} \in U$ such that $A\bar{y} + B\bar{u} + f = 0$ is valid.

(ii) The inverse operator B^{-1} of B is continuous on $R(B)$, for every $y \in D(A) \cap X$ we have $Ay + f \in R(B)$, and there exist elements $\bar{y} \in D(A) \cap X$ and $\bar{u} \in \text{int } U$ such that $A\bar{y} + B\bar{u} + f = 0$ is valid. Then the duality relation

$$\inf(P) = \max(D)$$

is true.

Proof. We prove (i); then the proof of (ii) is analogous. We show the subdifferentiability of $\varphi(x, v, w)$ at $(0, 0, 0)$ (see [15, Thms. 16, 17]). $\varphi(x, v, w)$ defined by (2.1) is proper convex. We have $\varphi(x, v, w) \leq G(A^{-1}w + A^{-1}(-B\bar{u} - f) + x, \bar{u} + v) =: g(x, v, w)$ for every $(x, v, w) \in V$ where V is a neighborhood of zero in $B_1 \times B_2 \times R(A)$. A is closed and A^{-1} is continuous, and this implies $R(A) \subset B_3$ is a closed subspace of B_3 .

Since $g(x, v, w)$ under the given assumptions is bounded above on a neighborhood of $(0, 0, 0)$ we have the subdifferentiability of $\varphi(x, v, w)$ at $(0, 0, 0)$.

LEMMA 1. If one of the following conditions (i)-(iv) is valid, the primal problem (P) possesses an optimal solution:

(i) X and U are bounded.

(ii) $G(y, u)$ is coercive on the set K_0 , where

$$K_0 = \left\{ (y, u) \in B_1 \times B_2 \left| \begin{array}{l} Ay + Bu + f = 0 \\ y \in D(A) \cap X, u \in U \end{array} \right. \right\}.$$

(iii) U is a bounded set and for every $u \in U$ we have $Bu + f \in R(A)$. The operator A^{-1} is continuous on $R(A)$.

(iv) X is a bounded set and for every $y \in D(A) \cap X$ we have $Ay + f \in R(B)$. The operator B^{-1} is continuous on $R(B)$.

The proof of this lemma follows from [5, Chap. II, Prop. 1.2]. In the following we show the duality relation:

$$(2.5) \quad \min (P) = \sup (D).$$

THEOREM 3. *If one of the following conditions is valid, the duality relation (2.5) holds:*

- (i) X and U are bounded sets.
- (ii) There exists an element $w^* \in D(A^*)$ with $(-A^*w^*, -B^*w^*) \in \text{int}(\text{dom } G^*)$.
- (iii) The set U is bounded and there exists an element $w^* \in D(A^*)$ such that $G^*(y^*, u^*)$ is continuous relative to the variable y^* in a neighborhood $V(-A^*w^*)$ of the element $-A^*w^*$ and for any fixed element $u^* = v^* \in B_2^*$, where $(y^*, v^*) \in \text{dom } G^*$ with $y^* \in V(-A^*w^*)$ holds.
- (iv) The set X is bounded and there exists an element $w^* \in D(A^*)$ such that $G^*(y^*, u^*)$ is continuous relative to the variable u^* in a neighborhood $V(-B^*w^*)$ of the element $-B^*w^*$ and for any fixed element $y^* = x^* \in B_1^*$, where $(x^*, u^*) \in \text{dom } G^*$ with $u^* \in V(-B^*w^*)$ holds.

To prove this theorem we apply [15, Thms. 16', 17']. A is a closed operator and this implies that the set K (2.2) is a closed set in $B_1 \times B_2 \times B_3$, and therefore the function F (2.2) is proper convex and lower semicontinuous. Analogous to the proof of Theorem 2, we can show the subdifferentiability of the perturbed function $\psi(y^*, u^*)$ of the dual problem

$$(2.6) \quad \Psi(y^*, u^*) = \sup \left\{ -F^*(y^*, u^*, x^*, v^*, w^*) \mid \begin{array}{l} x^* \in B_1^*, v^* \in B_2^* \\ w^* \in D(A^*) \end{array} \right\}$$

at $(y^*, u^*) = (0, 0)$ for the conditions (i)-(iv).

3. Conditions for optimality and the lower bound for $\inf (P)$. The duality relations $\inf (P) = \max (D)$, or $\min (P) = \sup (D)$, respectively, imply necessary and sufficient conditions for optimal solutions of primal or dual problems. An element (y, u) is called admissible for the primal problem if $(y, u) \in (D(A) \cap X) \times U$, $(y, u) \in \text{dom } G$ and $Ay + Bu + f = 0$ hold. An element $(x^*, v^*, w^*) \in \text{dom } F^*(0, 0, x^*, v^*, w^*)$ is called admissible for the dual problem.

With this notation we can show the following theorems.

THEOREM 4. *Let $\inf (P) = \max (D)$, and let (y_0, u_0) be admissible for the primal problem. Then (y_0, u_0) is an optimal solution of (P) if and only if there exists an element (x_0^*, v_0^*, w_0^*) admissible for (D) such that*

$$(3.1) \quad \begin{aligned} (x_0^*, v_0^*) &\in \partial G(y_0, u_0), \\ \langle y - y_0, x_0^* + A^*w_0^* \rangle_1 &\geq 0 \quad \forall y \in X, \\ \langle u - u_0, v_0^* + B^*w_0^* \rangle_2 &\geq 0 \quad \forall u \in U \end{aligned}$$

is valid.

THEOREM 5. *Let $\min (P) = \sup (D)$ and let (x_0^*, v_0^*, w_0^*) be admissible for the dual problem. Then (x_0^*, v_0^*, w_0^*) is an optimal solution of (D) if and only if there exists an admissible element (y_0, u_0) for (P) such that (3.1) is valid.*

The proofs of Theorem 4, or Theorem 5, respectively, follow from [15, Thm. 16 or Thm. 16']. For $X = B_1$ and $U = B_2$ the dual problem (2.3) is of the form

$$(3.2) \quad \sup (D) = \sup \{ \langle f, w^* \rangle_3 - G^*(-A^*w^*, -B^*w^*) \mid w^* \in D(A^*) \}.$$

Since the operator A is closed, it follows that the domain $D(A^*)$ of the adjoint operator A^* to A is everywhere densely defined in B_3^* . By

$$(3.3) \quad L(y, u) = -Ay - Bu$$

we define an everywhere densely defined linear and closed operator $L \in (B_1 \times B_2 \rightarrow B_3)$ with the domain $D(L) = D(A) \times B_2$. The adjoint operator $L^* \in (B_3^* \rightarrow B_1^* \times B_2^*)$ is

$$(3.4) \quad L^*w^* = (-A^*w^*, -B^*w^*), \quad w^* \in D(A^*) = D(L^*)$$

and we can write the dual problem (3.2) in the form

$$(3.5) \quad \sup (D) = \sup \{ \langle f, w^* \rangle_3 - G^*(L^*w^*) \mid w^* \in D(A^*) \}.$$

Let the function G^* be continuous in a point of image of the operator L^* . By [15, Thm. 19(i)] and Theorems 3 and 5 above, it follows that an element $w_0^* \in D(A^*)$ is an optimal solution of (D) if and only if

$$(3.6) \quad f \in L\partial G^*(L^*w_0^*)$$

is valid.

If we denote by $G^{*'}(y^*, u^*)$ the Gâteaux-derivative of the function $G^*(y^*, u^*)$, the following formula for a lower bound of $\inf (P)$ holds.

LEMMA 2. *Let G^* be continuous in a point of $R(L^*)$, and let $G^{*'}(y^*, u^*)$ be strongly monotone and Lipschitz continuous on $R(L^*)$. Let the dual problem (3.5) have an optimal solution. Then the formula*

$$(3.7) \quad \min (P) \geq \langle f, w^* \rangle_3 - G^*(L^*w^*) + \frac{m}{2M^2} \frac{\langle LG^{*'}(L^*w^*) - f, w^* \rangle_3^2}{\|L^*w^*\|_{1^* \times 2^*}^2}$$

holds for every $w^* \in D(A^*)$ with $L^*w^* \neq 0$ and $G^{*'}(L^*w^*) \in D(L)$, where $M > 0$ is the Lipschitz constant and $m > 0$ is the constant of the strong monotonicity of $G^{*'}$.

Proof. Analogous to [6] we can show that the following inequality holds

$$G^*(x^*) - G^*(y^*) \geq \langle G^{*'}(y^*), x^* - y^* \rangle_{1 \times 2} + \frac{m}{2} \|x^* - y^*\|_{1^* \times 2^*}^2.$$

For $x^* = L^*w^*$ and $y^* = L^*w_0^*$, where w_0^* is an optimal solution of (D), the lower bound (3.7) follows from this inequality and the Lipschitz continuity of $G^{*'}$.

In most cases lower bounds for optimal control problems are constructed for quadratic and norm-minimal problems of optimal control (Yavin [17], Benker [3], Chan and Ho [4]). The lower bound (3.7) complements the known results.

4. Application to quadratic and norm-minimal optimal control problems. We consider the following optimal control problem

$$(4.1) \quad \inf (P) = \inf \left\{ \frac{1}{2} \|y - R\|_1^2 + \frac{k}{2} \|u\|_2^2 \mid \begin{array}{l} Ay + Bu + f = 0 \\ y \in D(A) \cap X, u \in U \end{array} \right\}$$

with $k > 0$. Let $B_i = B_i^* = H_i$, $i = 1, 2, 3$, be Hilbert spaces. Let $R \in H_1$ be a given element and let the assumptions of § 1 be valid. Following § 2, we can immediately write a dual problem to (4.1). The form of the dual problem is, to a certain degree, dependent on the kind of perturbation in the primal problem. Through a suitable perturbation of the primal problem (4.1), which considers the specific structure of (4.1), in the following we obtain a convenient dual problem different from the dual problem of

(4.1) based on (2.3). We define the following perturbed function of the primal problem (4.1):

$$(4.2) \quad \varphi(w) = \inf_{y,u} \left[\frac{1}{2} \|y - R\|_1^2 + \frac{k}{2} \|u\|_2^2 + \delta((y, u, w) | K) \right]$$

with K as in (2.2).

If A is a linear everywhere densely defined and closed operator, the dual problem related to (4.1) is

$$(4.3) \quad \sup(D) = \sup_{w^* \in D(A^*)} g(w^*)$$

with

$$g(w^*) = \langle f, w^* \rangle_3 + \frac{1}{2} \|(I - P_X)(R - A^*w^*)\|_1^2 - \frac{1}{2} \|R - A^*w^*\|_1^2 + \frac{1}{2} \|R\|_1^2 \\ + \frac{k}{2} \left\| (I - P_U) \left(\frac{-1}{k} B^*w^* \right) \right\|_2^2 - \frac{1}{2k} \|B^*w^*\|_2^2$$

(I is the identic operator; P_X , or P_U , respectively, is the operator of the projection in the Hilbert space H_1 or H_2 on the set X or U). We can show the following duality theorem analogous to Theorem 3 using [15, Thm. 19(i)].

THEOREM 6. *It is always true that $\min(P) = \sup(D)$. The dual cost function $g(w^*)$ is subdifferentiable for every $w^* \in D(A^*)$ with $P_X(R - A^*w^*) \in D(A)$, with*

$$(4.4) \quad \partial g(w^*) = \left\{ AP_X(R - A^*w^*) + BP_U \left(\frac{-1}{k} B^*w^* \right) + f \right\}.$$

If for a $w^* \in D(A^*)$ we have $P_X(R - A^*w^*) \notin D(A)$, then we have $\partial g(w^*) = \emptyset$.

An element $w_0^* \in D(A^*)$ is an optimal solution of (D) if and only if

$$(4.5) \quad AP_X(R - A^*w_0^*) + BP_U \left(\frac{-1}{k} B^*w_0^* \right) + f = 0$$

holds. The element (y_0, u_0) with $y_0 = P_X(R - A^*w_0^*)$ and $u_0 = P_U(-1/k B^*w_0^*)$ is the optimal solution of (4.1).

Remark 3. If the operator $A \in L(H_1, H_3)$ is a linear and continuous operator on H_1 the cost function $g(w^*)$ of the dual problem (4.3) is Fréchet-differentiable for every $w^* \in D(A^*) = H_3$, since by Holmes [8] the function

$$f(\cdot) = \frac{1}{2} \|(I - P)(\cdot)\|^2,$$

possesses the Fréchet-derivative

$$f'(\cdot) = (I - P)(\cdot)$$

where P is the operator of projection on a convex and closed set in a Hilbert space.

The remaining results of §§ 2 and 3 are transferable to problem (4.1).

Next we consider the following problem of optimal control, which is equivalent to a norm-minimal control problem

$$(4.6) \quad \inf(P) = \inf \left\{ \frac{1}{2} \|Cy - R\|_4^2 \mid \begin{array}{l} Ay + Bu + f = 0 \\ y \in D(A) \cap X, u \in U \end{array} \right\}$$

where the assumptions of § 1 are valid and A is a linear everywhere densely defined and closed operator. Let B_4 be a real reflexive Banach space (a space of observation)

with the dual B_4^* , let $R \in B_4$ be a given element and let $C \in L(B_1, B_4)$. As a perturbed function of the problem (4.6), we define

$$(4.7) \quad \varphi(x, w) = \inf_{y, u} \left[\frac{1}{2} \|Cy + x - R\|_4^2 + \delta((y, u, w)|K) \right]$$

with K as in (2.2). Then we obtain for the dual problem to (4.6)

$$(4.8) \quad \sup(D) = \sup \left\{ \langle f, w^* \rangle_3 - \frac{1}{2} \|x^*\|_{4^*}^2 - \langle R, x^* \rangle_4 - \sigma_U(-B^*w^*) - \sigma_X(-C^*x^* - A^*w^*) \mid \begin{array}{l} x^* \in B_4^* \\ w^* \in D(A^*) \end{array} \right\}.$$

The results of §§ 2 and 3 are transferable to problem (4.6). For example, if $\inf(P) = \max(D)$ holds, we obtain the following as necessary and sufficient conditions for the optimality of an admissible element (y_0, u_0) for (P): (y_0, u_0) is the optimal solution of (P) if and only if there exists an element (x_0^*, w_0^*) admissible for (D) such that

$$(4.9) \quad \begin{aligned} -x_0^* &\in J(Cy_0 - R), \\ \langle y - y_0, C^*x_0^* + A^*w_0^* \rangle_1 &\geq 0 \quad \forall y \in X, \\ \langle u - u_0, B^*w_0^* \rangle_2 &\geq 0 \quad \forall u \in U \end{aligned}$$

is valid, where $J \in (B_4 \rightarrow 2^{B_4^*})$ is the duality mapping of B_4 .

5. Application to control processes described by linear ordinary and partial differential equations. First we consider an optimal control problem for a Cauchy problem with an ordinary differential operator of the second order: the operator $A \in L_2(0, 1) \rightarrow L_2(0, 1)$ given by

$$Ay(x) = p_0(x) \cdot y''(x) + p_1(x) \cdot y'(x) + p_2(x) \cdot y(x)$$

with the domain

$$(5.1) \quad D(A) = \left\{ y(x) \in L_2(0, 1) \mid \begin{array}{l} y' \text{ is absolutely continuous on } (0, 1) \\ y'' \in L_2(0, 1), y(0) = y'(0) = 0 \end{array} \right\}$$

where $p_i(x)$; $i = 0, 1, 2$ are real functions on $[0, 1]$ such that $p_0''(x)$, $p_1'(x)$ and $p_2(x)$ are continuous on $[0, 1]$ and $p_0(x) < 0$ for every $x \in [0, 1]$. A is a linear unbounded everywhere densely defined and closed operator in $L_2(0, 1)$ (see [9, Chap. III, §2.3]). The inverse operator A^{-1} exists on $R(A) = L_2(0, 1)$ and is continuous. Now we can apply the above results. In particular, we choose in (4.1): $H_1 = H_2 = H_3 = L_2(0, 1)$, $B = -I$, $f(x) = 0$, $R(x) = 1$, $k = 1$, $Ay(x) = -y''(x) + y'(x) + y(x)$ with $D(A)$ by (5.1). Then the dual problem related to this problem is of the form

$$(5.2) \quad \sup_{w^* \in D(A^*)} \left[\frac{1}{2} \int_0^1 [(I - P_X)(1 + w^{*''} + w^{*'} - w^*)]^2 dx + \frac{1}{2} \int_0^1 [(I - P_U)w^*]^2 dx + \frac{1}{2} \int_0^1 (1 - (1 + w^{*''} + w^{*'} - w^*)^2 - w^{*2}) dx \right]$$

where $A^*w^* = -w^{*''} - w^{*'} + w^*$, $A^* \in (L_2(0, 1) \rightarrow L_2(0, 1))$ with

$$D(A^*) = \left\{ w^* \in L_2(0, 1) \mid \begin{array}{l} w^{*'} \text{ is absolutely continuous on } (0, 1) \\ w^{*''} \in L_2(0, 1), w^*(1) = w^{*'}(1) = 0 \end{array} \right\}$$

for any convex and closed set X and U in $L_2(0, 1)$. Always

$$\min (P) = \sup (D).$$

For $X = U = L_2(0, 1)$ it follows from (4.5) for the optimal solution $w_0^*(x)$ of (D) that

$$w_0^{*(4)}(x) - 3 \cdot w_0^{*(2)}(x) + 2 \cdot w_0^*(x) = 1$$

with the conditions

$$\begin{aligned} w_0^*(1) = w_0^{*'}(1) = 0, \quad 1 + w_0^{*(2)}(0) + w_0^{*'}(0) - w_0^*(0) &= 0, \\ w_0^{*(3)}(0) + w_0^{*(2)}(0) - w_0^{*'}(0) &= 0. \end{aligned}$$

The system has a unique solution $w_0^*(x)$.

Next we consider an optimal control problem for a class of homogeneous boundary value problems of Dirichlet.

Let $G \subset R^n$ be a bounded open domain with boundary Γ . We consider the differential operator $E \in (L_2(G) \rightarrow L_2(G))$ given by

$$Ey(x) = - \sum_{i,j=1}^n \frac{\partial}{\partial x_i} \left[a_{ij}(x) \frac{\partial y(x)}{\partial x_j} \right] + c(x) \cdot y(x)$$

with $D(E) = C_0^\infty(G)$. Let $a_{ij}(x) = a_{ji}(x)$, $i, j = 1, 2, \dots, n$, and let $c(x) \geq 0$ be real functions from $\bar{C}^\infty(G)$ with

$$(5.3) \quad \sum_{i,j=1}^n a_{ij}(x) \cdot d_i \cdot d_j \geq \beta \sum_{i=1}^n d_i^2 \quad \forall x \in G,$$

$$\forall d = (d_1, d_2, \dots, d_n) \in R^n \quad \text{and} \quad \beta > 0.$$

The operator $A \in (L_2(G) \rightarrow L_2(G))$ with

$$(5.4) \quad D(A) = \left\{ y(x) \in \dot{W}_2^1(G) \left| - \sum_{i,j=1}^n \frac{\partial}{\partial x_i} \left[a_{ij}(x) \frac{\partial y(x)}{\partial x_j} \right] + c(x)y(x) \in L_2(G) \right. \right\}$$

is the extension by Friedrichs of the operator E (see [16, Satz 17.11, Satz 29.1]). A is a linear unbounded everywhere densely defined and closed operator, which is self adjoint, and because of (5.3) is strongly monotone. The operator A^{-1} is continuous on $R(A) = L_2(G)$. Therefore the above results are applicable.

To illustrate the application of the results, we consider a special example. We choose the following Dirichlet problem:

$$\begin{aligned} -\Delta y + y + u + f &= 0 \quad \text{in } G, \\ y &= 0 \quad \text{on } \Gamma \end{aligned}$$

with the operator $A = I - \Delta$, and $D(A)$ according to (5.4). Let $f \in L_2(G)$ be a given element, and, for example, consider the quadratic control problem (4.1).

Then we have $B = I$, $H_i = L_2(G)$, $i = 1, 2, 3$, $X \subset L_2(G)$ and $U \subset L_2(G)$. The dual problem (4.3) is of the form

$$\begin{aligned} \sup_{w^* \in D(A)} & \left[\int_G \left[f \cdot w^* + R(I - \Delta)w^* - \frac{1}{2}((I - \Delta)w^*)^2 - \frac{1}{2k} w^{*2} \right] dx \right. \\ & \left. + \frac{1}{2} \int_G \left(\left[(I - P_X)(R + \Delta w^* - w^*) \right]^2 + k \left[(I - P_U) \left(-\frac{w^*}{k} \right) \right]^2 \right) dx \right]. \end{aligned}$$

Finally we will apply the formula for the lower bound to this problem. In formula (3.7) we choose

$$G(y, u) = \frac{1}{2} \|y - R\|_1^2 + \frac{k}{2} \|u\|_2^2,$$

$k = 1$, $f = 0$, $R = 4$, the domain $G \subset \mathbb{R}^2$ as a circle of unit radius with the center at the origin and

$$w^*(x) = 1 - x_1^2 - x_2^2 \in D(I - \Delta).$$

Then we have $m = M = 1$, the condition

$$G^{*'}(L^* w^*) = \begin{pmatrix} -A^* w^* + R \\ -\frac{1}{k} B^* w^* \end{pmatrix} \in D(L)$$

is valid, and we obtain

$$25.133 > 8\pi = G(0, 0) \geq \min(P) \geq \frac{243\pi}{31} > 24.626.$$

Analogously we can solve other problems of optimal control, in particular for homogeneous and inhomogeneous problems of Dirichlet and Neumann or problems with parabolic equations.

The advantage of using a linear unbounded operator consists of the possibility that for a boundary value problem we can choose a suitable simple space, because the differential operator A is only defined on an everywhere dense subspace. On the other hand, the dual problem acquires a complicated form because of the unboundedness of the operator. This we can see in the dual of the quadratic control problem (4.3), where the Fréchet-differentiability of the cost function $g(w^*)$ for the dual problem disappears.

REFERENCES

- [1] E. ARONOFF AND C. T. LEONDES, *Lower bounds for a quadratic cost functional*, Internat. J. Systems Sci., 7 (1976), pp. 17-25.
- [2] V. BARBU AND TH. PRECUPANU, *Convexity and Optimization in Banach spaces*, Sijthoff and Noordhoff, Editura Academiei, Bucharest, Romania, 1978.
- [3] H. BENKER, *Upper and lower bounds for special optimal control problems described by operator equations*, Reihe Math., 5 (1981).
- [4] W. L. CHAN AND L. F. HO, *Lower bounds and duality in the optimal control of distributed systems*, J. Math. Anal. Appl., 70 (1979), pp. 530-545.
- [5] I. EKELAND AND R. TEMAM, *Convex Analysis and Variational Problems*, North-Holland, Amsterdam, Oxford, 1976.
- [6] K. GRÖGER, *Zur Lösung einer Klasse von Gleichungen mit einem monotonen Potentialoperator*, Math. Nachr., 81 (1978), pp. 7-24.
- [7] W. HEINS AND S. K. MITTER, *Conjugate convex functions, duality and optimal control. Problem I: Systems governed by ordinary differential equations*, Inform. Sciences, 2 (1970), pp. 211-243.
- [8] R. B. HOLMES, *A Course on Optimization and Best Approximation*, Springer-Verlag, Berlin-Heidelberg-New York, 1972.
- [9] T. KATO, *Perturbations Theory for Linear Operators*, Springer-Verlag, Berlin-Heidelberg-New York, 1966.
- [10] J. L. LIONS, *Remarks on the theory of distributed systems*, in Control Theory of Systems Governed by Partial Differential Equations, Academic Press, New York, 1977, pp. 1-103.
- [11] U. MACKENROTH, *Optimalitätsbedingungen und Dualität bei zustandsrestringierten parabolischen Kontrollproblemen*, Math. Operationsforsch. Statistik, Ser. Optim., 21 (1981), pp. 65-89.

- [12] ———, *Strong duality, weak duality and penalization for a state constrained parabolic control problem*, preprint, Universität Bayreuth, Bayreuth, West Germany, May 1980.
- [13] J. MOSSINO, *An application of duality to distributed optimal control problems with constraints of the control and the state*, J. Math. Anal. Appl., 50 (1975), pp. 223–243.
- [14] J. V. OUTRATA, *On the differentiability in dual optimal control problems*, Math. Operationsforsch. Statistik, Ser. Optim., 10 (1979), pp. 529–542.
- [15] R. T. ROCKAFELLAR, *Conjugate Duality and Optimization*, CBMS-NSF Regional Conference Series in Applied Mathematics, 16, Society for Industrial and Applied Mathematics, Philadelphia, PA, 1974.
- [16] H. TRIEBEL, *Höhere Analysis*, VEB Deutscher Verlag der Wissenschaften, Berlin, 1972.
- [17] Y. YAVIN, *Lower bounds on the cost functional for systems governed by partial differential equations*, J. Optim. Theory Appl., 11 (1973), pp. 605–612.

MODULAR FEEDBACK LOGIC FOR DISCRETE EVENT SYSTEMS*

P. J. RAMADGE† AND W. M. WONHAM‡

Abstract. We examine a modular approach to the synthesis of state feedback controls for the problem of maintaining a predicate on the state set of a discrete dynamic system invariant. Dynamical systems are modeled by automata together with a mechanism for enabling and disabling a subset of state transitions. The basic problem of interest is to ensure by appropriate control action that a given predicate on the state set of the process remains invariantly true whenever it is initially satisfied. Assuming the predicate can be decomposed into the conjunction or disjunction of component predicates, we determine conditions under which it is possible to synthesize the appropriate control in a modular fashion.

Key words. discrete event systems, modular control, supervisory control

AMS(MOS) subject classification. 93

1. Introduction. We examine a modular approach to the synthesis of state feedback controls for the problem of maintaining a predicate on the state set of a discrete dynamic system invariant. Our setting is the supervisory control framework developed in [6], [9].

The concept of modular control synthesis for discrete-event processes was first suggested and partially investigated in [5]. Some early work on this topic has also been reported in [10]. This paper is intended to be the first part of a comprehensive study of modular control synthesis in the supervisory control framework. Here we concentrate on nondynamic controls, i.e., state feedback. The basic control problem of interest is to ensure by appropriate control action that a given predicate P on the state set of the process remains invariantly true whenever it is initially satisfied. Assuming the predicate P can be decomposed into the conjunction or disjunction of component predicates, we determine conditions under which it is possible to synthesize the appropriate control in a modular fashion. Our investigation of modular synthesis is continued in a companion paper [5] where we examine the modular synthesis of supervisors, i.e., dynamic controllers, to achieve specified closed loop output behaviors.

The paper is organized as follows. In §§ 2–6 we review the basic supervisory control framework as well as introduce the concepts and definitions needed later in the paper. Section 7 contains the main results on the modular synthesis of feedback controls while § 8 deals with the modular determination of extremal control-invariant predicates. In § 9 we present two simple examples.

2. Controlled discrete event processes. We recall the basic definition of a controlled discrete-event process. For a more detailed discussion of this model the reader is referred to [7]. Descriptions of similar models for discrete-event systems can also be found in [3] and [1].

Let

$$G = (\Sigma, Q, \delta, q_0)$$

* Received by the editors January 20, 1986; accepted for publication (in revised form) August 11, 1986.

† Department of Electrical Engineering, Princeton University, Princeton, New Jersey 08544. The work of this author was supported by National Science Foundation grant ECS-8504584.

‡ Systems Control Group, Department of Electrical Engineering, University of Toronto, Toronto, Canada. The work of this author was supported by Natural Sciences and Engineering Research Council of Canada grant A-7399.

be an automaton. Here Q is the set of states, q_0 is the initial state, Σ is a finite set of output symbols, and $\delta: \Sigma \times Q \rightarrow Q$ (in general a partial function) is the state transition function.¹

We interpret G as a device that starts in the state q_0 and generates a sequence of events, i.e. state transitions, subject to the range of transitions permitted by the function δ . Each event is assumed to occur spontaneously, asynchronously and instantaneously, and to result in an output from the set Σ . Note that this is a nonstandard interpretation of the automaton structure.

To adjoin a control mechanism to G , we designate a subset $\Sigma_c \subseteq \Sigma$ of controllable events and set $\Sigma_u = \Sigma - \Sigma_c$. Control consists of specifying for each controllable event whether it is *enabled* (permitted to occur) or *disabled* (prevented from occurring). The set of *control patterns* for G is defined to be

$$\Gamma = \{\gamma: \gamma: \Sigma \rightarrow \{0, 1\} \text{ and } \gamma(\sigma) = 1 \text{ for each } \sigma \in \Sigma_u\}.$$

Events labeled by $\sigma \in \Sigma$ are enabled by γ if $\gamma(\sigma) = 1$, and disabled by γ if $\gamma(\sigma) = 0$.

The extended process

$$G_c = (\Gamma \times \Sigma, Q, \delta_c, q_0)$$

with

$$\delta_c(\gamma, \sigma, q) = \begin{cases} \delta(\sigma, q) & \text{if } \gamma(\sigma) = 1, \\ \text{undefined} & \text{otherwise} \end{cases}$$

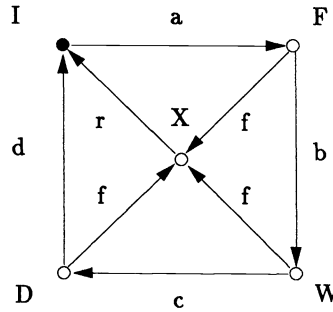
is called a *Controlled Discrete-Event Process* (CDEP). It is identical to the original process, and hence carries the same information, except that the control mechanism is now explicitly displayed.

In this paper we make extensive use of the algebraic structure of the set Γ . For each $\gamma_1, \gamma_2 \in \Gamma$ and each $\sigma \in \Sigma$ define

$$\begin{aligned} (\sim \gamma)(\sigma) &= \begin{cases} \sim \gamma(\sigma) & \text{if } \sigma \in \Sigma_c, \\ 1 & \text{otherwise,} \end{cases} \\ (\gamma_1 \wedge \gamma_2)(\sigma) &= \begin{cases} \gamma_1(\sigma) \wedge \gamma_2(\sigma) & \text{if } \sigma \in \Sigma_c, \\ 1 & \text{otherwise,} \end{cases} \\ \gamma_1 \vee \gamma_2 &= \sim((\sim \gamma_1) \wedge (\sim \gamma_2)). \end{aligned}$$

It is then readily shown that $(\Gamma; \sim, \wedge, \vee)$ is a Boolean algebra. Indeed it is clear that Γ is isomorphic to the family of subsets of Σ_c .

3. Example. A simple workstation in a manufacturing system is modeled by the following CDEP:



¹ In some situations it is convenient to add to the definition of G a subset $Q_m \subseteq Q$ of marker states, see e.g. [7]. However, this is not needed here.

In this diagram each state represents an activity, while transitions between states, i.e., events, represent the completion of one activity and the start of another. In the initial state I the machine is idle, in state F it is fetching a part from its input buffer, in state W it is working, and in state D it is depositing a part in its output buffer. These constitute the normal modes of machine operation. In state X the machine is broken down and awaits repair. Each state transition is labeled by its corresponding output.

In the notation of § 2 we have

$$Q = \{I, F, W, D, X\}, \quad \Sigma = \{a, b, c, d, f, r\}, \quad q_0 = I,$$

and the state transition function δ is displayed in the above graph.

Certain events are deemed controllable. Usually these are selected, subject to cost and feasibility constraints,² by the machine designer. If

$$\Sigma_c = \{a, c, r\}$$

then the machine's access to its input and output buffers and its repair can all be controlled. A control pattern simply consists of a specification of which events in Σ_c are enabled (i.e., permitted to occur), e.g., a and c but not r .

4. Predicates. Define a (unary) *predicate* on the nonempty set Q to be a function $P: Q \rightarrow \{0, 1\}$, i.e., a characteristic function on Q . Let \mathcal{P} denote the family of all predicates on Q , and define the operators \sim (*negation*), \wedge (*conjunction*), and \vee (*disjunction*) on \mathcal{P} by

$$\begin{aligned} (\sim P)(q) &= 1 \text{ iff } P(q) = 0, \\ (P_1 \wedge P_2)(q) &= 1 \text{ iff } P_1(q) = 1 \text{ and } P_2(q) = 1, \\ P_1 \vee P_2 &= \sim((\sim P_1) \wedge (\sim P_2)). \end{aligned}$$

\mathcal{P} together with the above operators forms a Boolean algebra. Indeed under the correspondence

$$P \leftrightarrow Q_P = \{q: q \in Q \text{ and } P(q) = 1\}$$

\mathcal{P} is isomorphic to the Boolean algebra of all subsets of Q . Thus in what follows the terms "predicate" and "subset" can be interchanged without essential loss.

Other operators on \mathcal{P} can be defined in terms of \wedge , \vee , and \sim . For example, *implication* is defined for each P_1 and $P_2 \in \mathcal{P}$ by

$$P_1 \Rightarrow P_2 = \sim P_1 \vee P_2.$$

We say that P is *true* at q if $P(q) = 1$, and *false* at q if $P(q) = 0$. Let 1 denote the predicate on Q which is true at all $q \in Q$, i.e., $1(q) = 1$, for each $q \in Q$, and 0 denote the predicate which is false at all $q \in Q$ (corresponding to the subsets Q and \emptyset , respectively).

Let P_1 , P_2 and P be predicates on Q . We say that P_1 and P_2 are equivalent *relative to* P , written $P_1 = P_2(\text{rel } P)$, if $P_1 \wedge P = P_2 \wedge P$. It is a straightforward matter to show that equivalence *rel* P is a congruence on the Boolean algebra \mathcal{P} .

The standard partial order on \mathcal{P} is defined by

$$P_1 \leq P_2 \quad \text{iff } P_1 \wedge P_2 = P_1.$$

Equivalently,

$$P_1 \leq P_2 \quad \text{iff } P_2 = 1(\text{rel } P_1) \quad \text{iff } (P_1 \Rightarrow P_2) = 1.$$

² For example, machine breakdown is always uncontrollable.

With respect to this partial order $P_1 \wedge P_2$ is the greatest lower bound and $P_1 \vee P_2$ the least upper bound of P_1 and P_2 . Each nonempty subset $U \subseteq \mathcal{P}$ has a unique greatest lower bound and a unique least upper bound in \mathcal{P} . These predicates are denoted $\inf U$ (or $\wedge U$) and $\sup U$ (or $\vee U$), respectively. Similarly each monotone sequence $\{P_j\}$ of predicates on Q has a limit in \mathcal{P} . For example, if $P_j \leq P_{j+1}$, for each index j , then we have

$$\lim_{j \rightarrow \infty} P_j = \inf \{P_j\} = \bigwedge_{j=0}^{\infty} P_j.$$

5. Predicate transforms. The transition function δ of the automaton G induces two useful transformations on the family of predicates $\mathcal{P} = \{0, 1\}^Q$.

First, for each $\sigma \in \Sigma$ we introduce a partial function

$$\delta_\sigma = \delta(\sigma, q)$$

and a predicate

$$D_\sigma(q) = \begin{cases} 1 & \text{if } \delta_\sigma(q) \text{ is defined,} \\ 0 & \text{otherwise.} \end{cases}$$

Then for each $\sigma \in \Sigma$ define the transformation $\text{wp}_\sigma: \mathcal{P} \rightarrow \mathcal{P}$ by

$$\text{wp}_\sigma(P)(q) = \begin{cases} 1 & \text{if } \delta_\sigma(q) \text{ is defined and } P(\delta_\sigma(q)) = 1, \\ 0 & \text{otherwise.} \end{cases}$$

When δ_σ is a total function on Q , $\text{wp}_\sigma(P)$ is simply the composition of the functions δ_σ and P . More generally $\text{wp}_\sigma(P)$ equals $P \cdot \delta_\sigma$ at those states q where $D_\sigma(q) = 1$ (i.e. $\delta_\sigma(q)$ is defined) and is zero elsewhere. From this it is evident that $\text{wp}_\sigma(P)(q) = 1$ is the weakest condition that guarantees that $\delta_\sigma(q)$ satisfies P . Hence, following Dijkstra [2], we call $\text{wp}_\sigma(P)$ the *weakest precondition* of P under σ .

As pointed out in [2] a variation of the above transform is useful when it is enough to guarantee that $\delta_\sigma(q)$ does not satisfy $\sim P$. For this define the *weakest liberal precondition* of P :

$$\text{wlp}_\sigma(P) = \text{wp}_\sigma(P) \vee \sim D_\sigma.$$

Note that if $\text{wlp}_\sigma(P)(q) = 1$, then either $\delta_\sigma(q)$ is undefined or $\delta_\sigma(q)$ satisfies P . In either case $\delta_\sigma(q)$ does not satisfy $\sim P$.

To illustrate these concepts, consider again the CDEP of § 3. If

$$P = (q = X)$$

then we have

$$\text{wp}_f(P) = (q \neq I)$$

and

$$\text{wlp}_f(P) = 1.$$

Notice that the weakest precondition specifies the set of all states from which an event with output f takes the process to the state X . On the other hand, the weakest liberal precondition appends to this set all states from which an event with output f is impossible (as defined by the process dynamics).

The second transformation induced on \mathcal{P} by δ is defined by

$$\text{sp}_\sigma(P)(q) = \begin{cases} 1 & \text{if } q = \delta_\sigma(q') \text{ for some } q' \in Q_P, \\ 0 & \text{otherwise.} \end{cases}$$

Clearly given that P is initially satisfied, $\text{sp}_\sigma(P)$ is the strongest condition whose truth can be inferred after a state transition under the map δ_σ . Hence we call $\text{sp}_\sigma(P)$ the *strongest postcondition* of P under σ .

The transforms wp_σ and sp_σ specify the action of G on predicates on the state set of G . As such they provide an alternative higher level description of the system dynamics which is particularly well suited, as we shall see, to the study of certain invariance problems.

It is readily shown that each of the above predicate transforms distributes over both a conjunction and a disjunction of predicates, see for example, Dijkstra [2, pp. 18–19]. We frequently use this fact in the following form. If $\{P_j, j \in J\} \subseteq \mathcal{P}$ is a set of predicates on Q , then

$$(5.1) \quad \text{wlp}_\sigma \left(\bigwedge_j P_j \right) = \bigwedge_j \text{wlp}_\sigma(P_j)$$

and

$$(5.2) \quad \text{wlp}_\sigma \left(\bigvee_j P_j \right) = \bigvee_j \text{wlp}_\sigma(P_j).$$

The proof of these equalities is straightforward.

6. State feedback. A *state feedback* for the CDEP G_c is a total function $f: Q \rightarrow \Gamma$. The application of f to G_c yields the closed loop process G_c^f defined by

$$G_c^f = (\Sigma, Q, \delta_c^f, q_0)$$

with

$$\delta_c^f(\sigma, q) = \delta_c(f(q), \sigma, q).$$

Note that G_c^f is again a process with output set Σ . Informally we regard G_c^f as a subsystem of G which has been constructed by the removal of certain state transitions.

For each $\sigma \in \Sigma$, the σ -component of the feedback $f \in \Gamma^Q$ is the map $f_\sigma: Q \rightarrow \{0, 1\}$ with

$$f_\sigma(q) = f(q)(\sigma).$$

The map f_σ gives precisely the conditions under which events associated with output σ are enabled under the control f . Clearly f is completely characterized by the set $\{f_\sigma: \sigma \in \Sigma\}$. This suggests that to synthesize a feedback map it is sufficient to synthesize each of its components individually. Of course for each $\sigma \in \Sigma_u$ we must have $f_\sigma = 1$. Hence it is only necessary to specify the components of f for $\sigma \in \Sigma_c$.

In addition to decomposing a feedback synthesis the component maps have the advantage of being predicates on the state set Q of G_c . This permits feedback to be analyzed within the algebra $\mathcal{P} = \{0, 1\}^Q$. For example, if wlp_σ is the predicate transform introduced in § 4 for G_c , and wlp_σ^f is the corresponding predicate transform for G_c^f , then it is readily verified that for each $\sigma \in \Sigma$, and $P \in \mathcal{P}$

$$(6.1) \quad \text{wlp}_\sigma^f(P) = \text{wlp}_\sigma(P) \vee \sim f_\sigma.$$

This relation will find frequent application in following sections.

Since Γ is a Boolean algebra, we can define a Boolean algebra on Γ^Q by point-wise definition of the algebraic operations. For example, if $f, g \in \Gamma^Q$, then $(f \wedge g)(q) = f(q) \wedge g(q)$.

This algebraic structure permits feedback controls to be constructed in a modular fashion, e.g., $h = f \wedge g$. It is precisely this aspect of control synthesis that we wish to investigate in the current paper.³

7. Control invariance. We consider the simple but fundamental problem of ensuring by control action that a given predicate on the state set of G_c remains invariantly true whenever it is initially satisfied.

This is equivalent to ensuring that a certain subset of states is invariant in the closed loop process, a problem that has been extensively studied in other contexts, see e.g. [4], [8]. Our primary purpose here, however, is to exploit the algebraic structure inherent in the supervisory control framework to synthesize modular solutions to the problem.

Throughout this section we let $G_c = (\Gamma \times \Sigma, Q, \delta_c, q_0)$ be a fixed CDEP and $\mathcal{P} = \{0, 1\}^Q$.

A predicate $P \in \mathcal{P}$ is said to be *control-invariant* (with respect to G_c) if for some feedback $f \in \Gamma^Q$ we have

$$(7.1) \quad P \leq \text{wlp}_\sigma^f(P) \quad \text{for each } \sigma \in \Sigma.$$

This ordering implies that if P is true at a state $q \in Q$, then P is true at all states reachable from q in the closed loop process G_c^f . Hence P will be invariantly true whenever it is initially satisfied.

It is a simple matter to characterize the control-invariant predicates in a “feedback independent” fashion. Say that the predicate $P \in \mathcal{P}$ is Σ_u -invariant (with respect to G_c) if

$$P \leq \text{wlp}_\sigma(P) \quad \text{for each } \sigma \in \Sigma_u.$$

We then have the following proposition:

PROPOSITION 7.1. *A predicate P is control-invariant iff P is Σ_u -invariant.*

Proof.

IF. Define the feedback f component-wise by

$$f_\sigma(q) = \begin{cases} 1 & \text{if } \sigma \in \Sigma_u, \\ 0 & \text{if } \sigma \in \Sigma_c. \end{cases}$$

Then for $\sigma \in \Sigma_u$

$$\text{wlp}_\sigma^f(P) = \text{wlp}_\sigma(P) \vee (\sim f_\sigma) = \text{wlp}_\sigma(P).$$

Thus $P \leq \text{wlp}_\sigma(P) = \text{wlp}_\sigma^f(P)$.

Alternatively, for $\sigma \in \Sigma_c$

$$\text{wlp}_\sigma^f(P) = \text{wlp}_\sigma(P) \vee \sim f_\sigma = \text{wlp}_\sigma(P) \vee 1 = 1.$$

Thus $P \leq 1 = \text{wlp}_\sigma^f(P)$.

ONLY IF. For $\sigma \in \Sigma_u$ we have $\text{wlp}_\sigma^f = \text{wlp}_\sigma(P)$. Hence if $P \leq \text{wlp}_\sigma^f(P)$, then $P \leq \text{wlp}_\sigma(P)$. \square

³ It is readily verified that modular constructions carry over to each component of the maps f and g , e.g., $(f \wedge g)_\sigma = f_\sigma \wedge g_\sigma$.

A predicate P' on Q satisfying the relation

$$P \leq (P' \Rightarrow \text{wlp}_\sigma(P))$$

is said to be a σ -friend of P .

By applying relation (6.1) to (7.1) we see that a predicate P is control-invariant in G_c iff for some $f \in \Gamma^Q$

$$P \leq \text{wlp}_\sigma(P) \vee \sim f_\sigma \quad \text{for each } \sigma \in \Sigma$$

or equivalently iff for some $f \in \Gamma^Q$

$$P \leq (f_\sigma \Rightarrow \text{wlp}_\sigma(P)) \quad \text{for each } \sigma \in \Sigma.$$

Thus f ensures that P is invariant in G_c^f iff for each $\sigma \in \Sigma_c$ the σ -component of f is a σ -friend of P . It follows that if P is control-invariant, then we can synthesize a feedback to ensure the invariance of P by selecting for the σ -component of f any σ -friend of P .

Let

$$F_\sigma(P) = \{P' : P' \text{ is a } \sigma\text{-friend of } P\}.$$

This set has the following properties:

PROPOSITION 7.2. *Let P be a control-invariant predicate and $\sigma \in \Sigma$. Then $F_\sigma(P)$ is nonempty and is closed under arbitrary conjunctions and disjunctions. In particular, $F_\sigma(P)$ has a unique maximal and a unique minimal element.*

Proof. It follows from the definition of control-invariance that $F_\sigma(P)$ is nonempty. Let $\{P_j, j \in J\} \subseteq F_\sigma(P)$. Now for each $j \in J$ we have

$$P \leq \text{wlp}_\sigma(P) \vee \sim P_j.$$

Hence

$$P \leq \bigwedge_j (\text{wlp}_\sigma(P) \vee \sim P_j) = \text{wlp}_\sigma(P) \vee \sim \left(\bigvee_j P_j \right).$$

The proof for conjunction is similar. \square

We can think of $\sup F_\sigma(P)$ as the least restrictive, or alternatively the most permissive control among the σ -friends of P . The dual predicate $\inf F_\sigma(P)$ is somewhat less interesting; although it ensures that P is invariant under the event σ it generally does so in an unnecessarily restrictive fashion.

It is clear that whether or not P_1 is a σ -friend of P depends only on the structure of P_1 on the set of states where P is true. Thus for the purposes of control-invariance we may regard potential feedback components P_1 and P_2 as equivalent, with respect to the predicate P , if $P_1 \wedge P = P_2 \wedge P$, i.e., if $P_1 = P_2(\text{rel } P)$.

We end our preliminary results with the following characterization of the class of predicates $F_\sigma(P)$.

PROPOSITION 7.3. *Let P be a control-invariant predicate and $\sigma \in \Sigma_c$. Then*

$$P' \in F_\sigma(P) \quad \text{iff } P' \leq \text{wlp}_\sigma(P)(\text{rel } P)$$

i.e., $P' \leq \text{wlp}_\sigma(P)$ at those q where $P(q) = 1$.

COROLLARY 7.1. *Let $P \in \mathcal{P}$ be control-invariant and $\sigma \in \Sigma_c$. Then*

$$\sup F_\sigma(P) = \text{wlp}_\sigma(P)(\text{rel } P).$$

Proof. We have

$$P \leq (P' \Rightarrow \text{wlp}_\sigma(P)) \quad \text{iff } (P' \Rightarrow \text{wlp}_\sigma(P)) = 1(\text{rel } P),$$

$$\text{iff } P' \leq \text{wlp}_\sigma(P)(\text{rel } P). \quad \square$$

Armed with Propositions 7.1 and 7.3, we are now ready to determine whether or not control-invariance problems admit modular solutions.

Let P_1, \dots, P_k be predicates on the state set Q and

$$(7.2) \quad P = g(P_1, \dots, P_k)$$

be constructed from the P_i by a finite number of conjunctions and disjunctions. We think of (7.2) as a modular specification of P . Our aim is to exploit this structure to analyze the control-invariance of P in terms of the component predicates P_1, \dots, P_k . It will be sufficient to consider the case $k = 2$.

It is a simple matter to show that the control-invariance of P is implied by the control-invariance of the predicates P_i . For this we use the fact that conjunction and disjunction are monotone operators, i.e., $P_1 \leq P_2$ and $T_1 \leq T_2$ together imply that $P_1 \wedge T_1 \leq P_2 \wedge T_2$.

PROPOSITION 7.4. $P_1 \wedge P_2$ and $P_1 \vee P_2$ are control-invariant whenever P_1 and P_2 are control-invariant.

Proof. For each $\sigma \in \Sigma_u$ we have $P_1 \leq \text{wlp}_\sigma(P_1)$ and $P_2 \leq \text{wlp}_\sigma(P_2)$. Hence

$$P_1 \wedge P_2 \leq \text{wlp}_\sigma(P_1) \wedge \text{wlp}_\sigma(P_2) = \text{wlp}_\sigma(P_1 \wedge P_2) \quad \text{by (5.1).}$$

The proof for disjunction is similar. \square

The more interesting question concerns modular feedback synthesis. The problem is to determine when we can synthesize a σ -friend of P from the σ -friends of the P_i . Our first result deals with conjunction.

PROPOSITION 7.5. Let $P_1, P_2 \in \mathcal{P}$ and $\sigma \in \Sigma_c$. Then $f_\sigma \in F_\sigma(P_1)$ and $g_\sigma \in F_\sigma(P_2)$ together imply that

$$f_\sigma \wedge g_\sigma \in F_\sigma(P_1 \wedge P_2).$$

Proof. Let $f_\sigma \in F_\sigma(P_2)$ and $g_\sigma \in F_\sigma(P_2)$. Then

$$\begin{aligned} (f_\sigma \wedge g_\sigma) \wedge (P_1 \wedge P_2) &= (f_\sigma \wedge P_1) \wedge (g_\sigma \wedge P_2) \\ &\leq (\text{wlp}_\sigma(P_1) \wedge P_1) \wedge (\text{wlp}_\sigma(P_2) \wedge P_2) \quad \text{by Proposition 7.3} \\ &= (P_1 \wedge P_2) \wedge (\text{wlp}_\sigma(P_1) \wedge \text{wlp}_\sigma(P_2)) \\ &= (P_1 \wedge P_2) \wedge \text{wlp}_\sigma(P_1 \wedge P_2) \quad \text{by (5.1).} \end{aligned}$$

Thus by Proposition 7.3 $f_\sigma \wedge g_\sigma \in F_\sigma(P_1 \wedge P_2)$. \square

Proposition 7.5 shows that the conjunction of any pair of σ -friends of P_1 and P_2 yields a σ -friend of P . Thus it is always possible to synthesize a σ -friend of $P = P_1 \wedge P_2$ in a modular fashion.

For a similar construction to be viable in the case of the disjunction operator it is necessary to place additional restrictions on the choice of f_σ and g_σ .

PROPOSITION 7.6. Let $P_1, P_2 \in \mathcal{P}$ and $\sigma \in \Sigma_c$. Then $f_\sigma \in F_\sigma(P_1)$, $g_\sigma \in F_\sigma(P_2)$, $f_\sigma \leq (\text{wlp}_\sigma(P_1) \vee \text{wlp}_\sigma(P_2))(\text{rel } P_2)$, and $g_\sigma \leq (\text{wlp}_\sigma(P_1) \vee \text{wlp}_\sigma(P_2))(\text{rel } P_1)$ together imply that

$$f_\sigma \vee g_\sigma \in F_\sigma(P_1 \vee P_2).$$

Proof. We have

$$\begin{aligned}
 (f_\sigma \vee g_\sigma) \wedge (P_1 \vee P_2) &= (f_\sigma \wedge P_1) \vee (f_\sigma \wedge P_2) \vee (g_\sigma \wedge P_1) \vee (g_\sigma \wedge P_2) \\
 &\leq (\text{wlp}_\sigma(P_1) \wedge P_1) \vee (\text{wlp}_\sigma(P_1) \wedge P_2) \\
 &\quad \vee (\text{wlp}_\sigma(P_2) \wedge P_1) \vee (\text{wlp}_\sigma(P_2) \wedge P_2) \\
 &= (P_1 \vee P_2) \wedge (\text{wlp}_\sigma(P_1) \vee \text{wlp}_\sigma(P_2)) \\
 &= (P_1 \vee P_2) \wedge \text{wlp}_\sigma(P_1 \vee P_2) \quad \text{by (5.2).}
 \end{aligned}$$

Hence $f_\sigma \vee g_\sigma \in F_\sigma(P_1 \vee P_2)$. \square

The additional assumptions in Proposition 7.6 are easily interpreted. The standard assumption $f_\sigma \in F_\sigma(P_1)$ requires that on the set Q_1 where P_1 is true σ is enabled only if the corresponding state transition maps into Q_1 . The additional assumption $f_\sigma \leq (\text{wlp}_\sigma(P_1) \vee \text{wlp}_\sigma(P_2))(\text{rel } P_2)$ requires that on the set Q_2 where P_2 is true σ is enabled only if the corresponding state transition maps into Q_1 or Q_2 . Thus the second condition restricts f_σ on the set $Q_2 - Q_1$ with the purpose of ensuring compatibility with the predicate P_2 .⁴

We end this section by determining conditions under which a modular control synthesis using the conjunction and disjunction operators preserves the maximality of the control.

PROPOSITION 7.7. *Let $P_1, P_2 \in \mathcal{P}$ be control-invariant, $\sigma \in \Sigma_c, f_\sigma = \sup F_\sigma(P_1)(\text{rel } P_1)$ and $g_\sigma = \sup F_\sigma(P_2)(\text{rel } P_2)$. Then*

$$f_\sigma \wedge g_\sigma = \sup F_\sigma(P_1 \wedge P_2)(\text{rel } P_1 \wedge P_2).$$

Proof. Let $h_\sigma = \sup F_\sigma(P_1 \wedge P_2)$. Then

$$\begin{aligned}
 (P_1 \wedge P_2) \wedge h_\sigma &= (P_1 \wedge P_2) \wedge \text{wlp}_\sigma(P_1 \wedge P_2) \\
 &= (P_1 \wedge \text{wlp}_\sigma(P_1)) \wedge (P_2 \wedge \text{wlp}_\sigma(P_2)) \\
 &= (P_1 \wedge f_\sigma) \wedge (P_2 \wedge g_\sigma) \\
 &= (P_1 \wedge P_2) \wedge (f_\sigma \wedge g_\sigma). \quad \square
 \end{aligned}$$

Proposition 7.7 shows that the least restrictive σ -friend of the modular predicate $P_1 \wedge P_2$ can be synthesized by forming the conjunction of the least restrictive controls for the component predicates P_1 and P_2 . Thus no loss of maximality need result from a modular synthesis over the conjunction operator.

To prove an analogous result to Proposition 7.7 for the disjunction operator, it is necessary to place further assumptions on the maps f_σ and g_σ .

PROPOSITION 7.8. *Let $P_1, P_2 \in \mathcal{P}$ be control-invariant, $\sigma \in \Sigma_c, f_\sigma = \text{wlp}_\sigma(P_1)(\text{rel } P_1 \vee P_2)$ and $g_\sigma = \text{wlp}_\sigma(P_2)(\text{rel } P_1 \vee P_2)$. Then*

$$f_\sigma \vee g_\sigma = \sup F_\sigma(P_1 \vee P_2)(\text{rel } P_1 \vee P_2).$$

Proof. Similar to the proof of Proposition 7.7. \square

8. Extremal control invariant predicates. We continue with the notation and setting of § 7. Since in general not every predicate $P \in \mathcal{P}$ will be control-invariant, it is of interest to exploit the order structure of \mathcal{P} to find good control-invariant approximations to P .

⁴ A simple condition which ensures that $f_\sigma \leq (\text{wlp}_\sigma(P_1) \vee \text{wlp}_\sigma(P_2))(\text{rel } P_2)$ for all predicates P_2 is $f_\sigma \leq \text{wlp}_\sigma(P_1)$. However, this may be restrictive in problems involving more than just invariance.

We begin with a quick review of two (essentially) standard results. First, there are two optimal control-invariant approximations to $P \in \mathcal{P}$: the weakest control-invariant predicate stronger than P (denoted P^\uparrow) and the strongest control-invariant predicate weaker than P (denoted P^\downarrow). Second, each of these predicates can be characterized as an extremal fixpoint of an appropriate monotone lattice map. We then turn our attention to the modular computation of P^\uparrow and P^\downarrow . Our main result is that the transform \uparrow is an automorphism of the algebra (\mathcal{P}, \wedge) while the transform \downarrow is an automorphism of the algebra (\mathcal{P}, \vee) .

For $P \in \mathcal{P}$ define

$$CI_{<}(P) = \{P' : P' \in \mathcal{P}, P' \leq P \text{ and } P' \text{ is control-invariant}\}.$$

$CI_{<}(P)$ is simply the set of control-invariant predicates on Q that are stronger than P .

Since 0 is control-invariant, $CI_{<}(P)$ is nonempty. By a straightforward extension of Proposition 7.4 it can be shown that $CI_{<}(P)$ is closed under arbitrary disjunctions. Then since $CI_{<}(P)$ is bounded above by P , it must have a unique maximal element. Denote this predicate by P^\uparrow . We can think of P^\uparrow as the best control-invariant approximation to P among the predicates which logically imply P . Of course we may have $P^\uparrow = 0$.

Similarly we define

$$CI_{>}(P) = \{P' : P' \in \mathcal{P}, P \leq P' \text{ and } P' \text{ is control-invariant}\}.$$

$CI_{>}(P)$ is the set of all control-invariant predicates on Q that are weaker than P .

Since 1 is control-invariant, $CI_{>}(P)$ is nonempty. Again by a straightforward extension of Proposition 7.4 it can be shown that $CI_{>}(P)$ is closed under arbitrary conjunctions. Hence since $CI_{>}(P)$ is bounded below by P , it has a unique minimal element. Denote this predicate by P^\downarrow . We can think of P^\downarrow as the tightest boundary we can place around P by control action. Of course, it is possible that $P^\downarrow = 1$.

To develop the fixpoint characterization of the predicate P^\uparrow we bring in the map $H : \mathcal{P} \rightarrow \mathcal{P}$ with

$$H(P') = P \wedge \left(\bigwedge_{\Sigma_u} \text{wlp}_\sigma(P') \right).$$

A predicate P' such that $H(P') = P'$ is said to be a *fixpoint* of the map H . It is not difficult to show that P^\uparrow is the unique maximal fixpoint of the map H . However, for our current purposes it will be sufficient to show that P^\uparrow is simply a fixpoint of H . This follows by noting that $P^\uparrow \leq \text{wlp}_\sigma(P^\uparrow)$ for each $\sigma \in \Sigma_u$, and hence

$$P^\uparrow = P \wedge P^\uparrow \leq P \wedge \left(\bigwedge_{\Sigma_u} \text{wlp}_\sigma(P^\uparrow) \right) = H(P^\uparrow).$$

But $P \wedge \left(\bigwedge_{\Sigma_u} \text{wlp}_\sigma(P^\uparrow) \right)$ is clearly Σ_u -invariant and less than P . Hence by the maximality of P^\uparrow we must have $H(P^\uparrow) \leq P^\uparrow$. To complete our characterization of P^\uparrow we introduce the sequence of predicates:

$$(8.1) \quad P_0 = P, \quad P_{j+1} = H(P_j).$$

This leads to our desired result:

PROPOSITION 8.1. *The sequence defined by (8.1) is monotone decreasing and $P^\uparrow = \bigwedge_{j=0}^{\infty} P_j$.*

Proof. We use the fact that H is monotone, i.e., $P_1 \leq P_2$ implies that $H(P_1) \leq H(P_2)$, and the fact that P^\uparrow is a fixpoint of H .

To begin

$$P_1 = H(P_0) \leq P = P_0$$

and if $P_j \leq P_{j-1}$, then

$$P_{j+1} = H(P_j) \leq H(P_{j-1}) = P_j.$$

Hence $P_0 \geq P_1 \geq P_2 \geq \dots$.

Now $P^\uparrow \leq P = P_0$ and if $P^\uparrow \leq P_j$, then

$$P^\uparrow = H(P^\uparrow) \leq H(P_j) = P_{j+1}.$$

Thus $P^\uparrow \leq \bigwedge_{j=0}^{\infty} P_j$.

For the reverse ordering it will be sufficient to show that $\bigwedge_{j=0}^{\infty} P_j$ is control-invariant. Fix $\sigma \in \Sigma_u$. Then by Proposition 7.1 we must show that

$$\bigwedge_{j=0}^{\infty} P_j \leq \text{wlp}_\sigma \left(\bigwedge_{j=0}^{\infty} P_j \right).$$

Now for each j

$$P_{j+1} = P \wedge \left(\bigwedge_{\omega \in \Sigma_u} \text{wlp}_\omega (P_j) \right) \leq \text{wlp}_\sigma (P_j).$$

Hence

$$\bigwedge_{j=0}^{\infty} P_j \leq \bigwedge_{j=0}^{\infty} \text{wlp}_\sigma (P_j) = \text{wlp}_\sigma \left(\bigwedge_{j=0}^{\infty} P_j \right). \quad \square$$

It is possible to provide an analogous fixpoint characterization of the predicate P^\downarrow . For this we introduce the map $G: \mathcal{P} \rightarrow \mathcal{P}$ with

$$G(P') = P \vee \left(\bigvee_{\Sigma_u} \text{sp}_\sigma (P') \right).$$

It can then be shown that P^\downarrow is the unique minimal fixpoint of G . Proceeding as before we define

$$(8.2) \quad R_0 = P, \quad R_{j+1} = G(R_j), \quad j \geq 0.$$

This leads to the following proposition:

PROPOSITION 8.2. *The sequence defined by (8.2) is monotone increasing and $P^\downarrow = \bigvee_{j=0}^{\infty} R_j$. \square*

We are now ready for the main results of this section. Our objective is to determine the relationship between the transforms \uparrow and \downarrow and the algebraic operations of conjunction and disjunction on \mathcal{P} . We first consider the transform \uparrow .

THEOREM 8.1. *For each $P_1, P_2 \in \mathcal{P}$*

$$P_1^\uparrow \wedge P_2^\uparrow = (P_1 \wedge P_2)^\uparrow$$

and

$$P_1^\uparrow \vee P_2^\uparrow \leq (P_1 \vee P_2)^\uparrow.$$

Proof.

(a) It is clear from the definition of \uparrow that if $P_1 \leq P_2$, then $P_1^\uparrow \leq P_2^\uparrow$, i.e., \uparrow is a monotone operator.

Set $P = P_1 \wedge P_2$. Since $P \leq P_1$ and $P \leq P_2$ it follows that $P^\uparrow \leq P_1^\uparrow$ and $P^\uparrow \leq P_2^\uparrow$. Thus $P^\uparrow \leq P_1^\uparrow \wedge P_2^\uparrow$. For the reverse ordering we note that $P_1^\uparrow \wedge P_2^\uparrow \leq P$ and hence that $P_1^\uparrow \wedge P_2^\uparrow \leq P^\uparrow$.

(b) The assertion follows immediately from Proposition 5.4 and the definition of $(P_1 \wedge P_2)^\uparrow$. \square

Not surprisingly a dual result holds for the \downarrow transform.

THEOREM 8.2. *For each $P_1, P_2 \in \mathcal{P}$*

$$P_1^\downarrow \wedge P_2^\downarrow \cong (P_1 \wedge P_2)^\downarrow$$

and

$$P_1^\downarrow \vee P_2^\downarrow = (P_1 \vee P_2)^\downarrow.$$

Proof. Similar to the proof of Theorem 8.1. \square

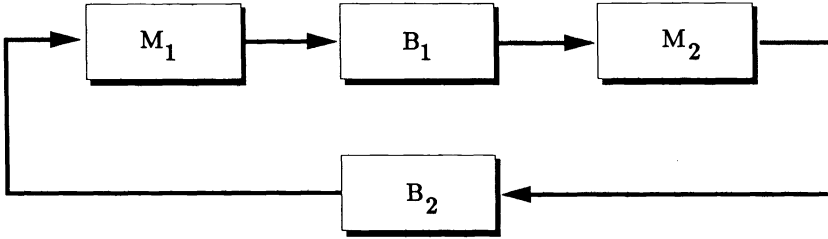
Theorem 8.1 indicates that the operation of forming the supremal control-invariant predicate (\uparrow) is an automorphism of the algebra (\mathcal{P}, \wedge) . This is a pleasing result from the point of view of modular synthesis since it implies that no loss of optimality is incurred in computing \uparrow by modularizing over a conjunction of predicates, i.e., to compute $(P_1 \wedge P_2)^\uparrow$ we can first compute P_1^\uparrow and P_2^\uparrow , then form their conjunction.

The same is not true of disjunction. In computing \uparrow , a modularization over a disjunction of predicates need not yield the maximal solution.

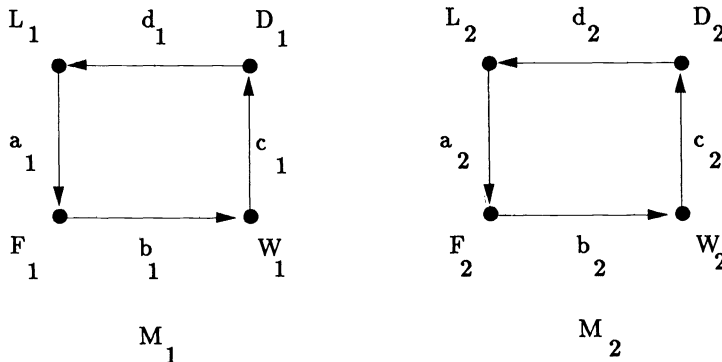
The corresponding results for the operation of forming the infimal control-invariant predicate (\downarrow) are dual to the above. Namely, the \downarrow transform is an automorphism of the algebra (\mathcal{P}, \vee) but not in general of the algebra (\mathcal{P}, \wedge) .

9. Examples.

Example 9.1. In a simple manufacturing system we consider two machines and two buffers connected in the configuration shown below.



The M_i are CDEPs with the following state diagrams:



As in the previous example each state represents an activity, while transitions between states, i.e. events, mark the completion of one activity and the start of another. In detail: in state I_i , M_i is “idle”; in state F_i , M_i is “fetching” a part from its input buffer (B_2 in the case of M_1 , and B_1 in the case of M_2); in state W_i , M_i is “working”; in state D_i , M_i is “depositing” a part in its output buffer (B_1 in the case of M_1 and B_2 in the case of M_2). (There may also be states and events associated with breakdown

and repair, and with the addition or removal of parts, but we assume that these are not relevant to the task at hand.)

Each buffer is modeled as an event driven automaton that records the number of workpieces in the buffer, i.e., each buffer is an automaton with state set \mathbf{N} (the natural numbers) and with the state transitions

$$B_1: \quad c_1: n \rightarrow n + 1,$$

$$a_2: n \rightarrow n - 1,$$

$$B_2: \quad c_2: n \rightarrow n + 1,$$

$$a_1: n \rightarrow n - 1.$$

Let

$$\Sigma = \{a_i, b_i, c_i, d_i; i = 1, 2\}, \quad Q_i = \{I_i, F_i, W_i, D_i\}, \quad i = 1, 2$$

and for the state space of the machine/buffer system we take $Q = Q_1 \times Q_2 \times \mathbf{N} \times \mathbf{N}$. For $(q_1, q_2, i, j) \in Q$: q_1 is the state of M_1 , q_2 is the state of M_2 , i is the state of B_1 , and j is the state of B_2 . Let

$$\Sigma_c = \{a_i, c_i; i = 1, 2\}$$

be the set of controllable events and assume that we can model the concurrent operation of M_1 and M_2 by "shuffling" their state transitions (see e.g. [6]).

Our aim is to ensure that the conjunction P of the following predicates is invariant:

- (1) Mutually exclusive use of B_1 : $P_1 = [q \neq (D_1, F_2)]$,
- (2) Mutually exclusive use of B_2 : $P_2 = [q \neq (F_1, D_2)]$,
- (3) Capacity of B_1 : $P_3 = (0 \leq i) \wedge (i \leq N)$,
- (4) Capacity of B_2 : $P_4 = (0 \leq j) \wedge (j \leq N)$.

It is readily shown that each of these predicates is control-invariant. Hence by Proposition 7.4 $P = \bigwedge_{i=1}^4 P_i$ is control-invariant.

Some elementary computation yields

$$\begin{aligned} \text{wlp}_\sigma(P_1) &= \begin{cases} (q_1, q_2) \neq (D_1, I_2) & \text{if } \sigma = a_2, \\ (q_1, q_2) \neq (W_1, F_2) & \text{if } \sigma = c_1, \\ 1 & \text{otherwise.} \end{cases} \\ \text{wlp}_\sigma(P_2) &= \begin{cases} (q_1, q_2) \neq (F_1, W_2) & \text{if } \sigma = c_2, \\ (q_1, q_2) \neq (I_1, D_2) & \text{if } \sigma = a_1, \\ 1 & \text{otherwise.} \end{cases} \\ \text{wlp}_\sigma(P_3) &= \begin{cases} i \neq N & \text{if } \sigma = c_1, \\ i \neq 0 & \text{if } \sigma = a_2, \\ 1 & \text{otherwise.} \end{cases} \\ \text{wlp}_\sigma(P_4) &= \begin{cases} j \neq N & \text{if } \sigma = c_2, \\ j \neq 0 & \text{if } \sigma = a_1, \\ 1 & \text{otherwise.} \end{cases} \end{aligned}$$

Now by Corollary 7.1 we know that $f_{i\sigma} = \text{wlp}_\sigma(P_i)$ is a σ -friend of P_i , $i = 1, \dots, 4$ (in fact it is equivalent $\text{rel}(P_i)$ to the supremal σ -friend of P_i). These component maps can be combined in a variety of ways to yield valid solutions to our problem. For

example, to obtain a feedback that ensures P_1 and P_3 are invariant (i.e., a controller for the buffer B_1) we form $g_\sigma = f_{1\sigma} \wedge f_{3\sigma}$, viz:

$$\begin{aligned} g_{a_1} &= g_{c_2} = 1, \\ g_{a_2} &= [(q_1, q_2) \neq (D_1, I_2) \wedge (i \neq 0)], \\ g_{c_1} &= [(q_1, q_2) \neq (W_1, F_2) \wedge (i \neq N)]. \end{aligned}$$

By Proposition 7.5 we know that g_σ is a σ -friend of $P_1 \wedge P_3$ (in fact by Proposition 7.7 it is equivalent $\text{rel}(P_1 \wedge P_3)$ to the supremal σ -friend of $P_1 \wedge P_3$). A similar procedure can be used to form a control for the buffer B_2 . Alternatively we can specify a single global control for each controlled event by forming $f_\sigma = \bigwedge_{i=1}^4 f_{i\sigma}$. This yields a single (centralized if you like) feedback that ensures P is invariant.

Example 9.2. Consider a simplified version of the example of ([9, § 7]) to which the reader is referred for any explained terminology below. A cat and a mouse are placed in the maze shown in Fig. 9.1. Each doorway in the maze must be traversed in the direction indicated and is either for the exclusive use of the cat (displayed as $-|\uparrow|$) or for the exclusive use of the mouse (displayed as $-|\downarrow|$). In addition each door, with the exception of c_7 , can be opened or closed as required in order to control the movement of the cat and the mouse. Our objective is to ensure that the cat and the mouse never occupy the same room simultaneously.

As our model of the system we adopt the CDEP $G_c = (Q, \Gamma \times \Sigma, \delta, q_0)$ shown in Fig. 9.2 (see [9, § 7] for further details). Here

$$\begin{aligned} \Sigma &= \{c_i, m_j : 1 \leq i \leq 7, 1 \leq j \leq 6\}, \\ Q &= \{(i, j) : 0 \leq i \leq 4, 0 \leq j \leq 4\}, \\ q_0 &= (2, 4), \\ \Sigma_u &= \{c_7\}. \end{aligned}$$

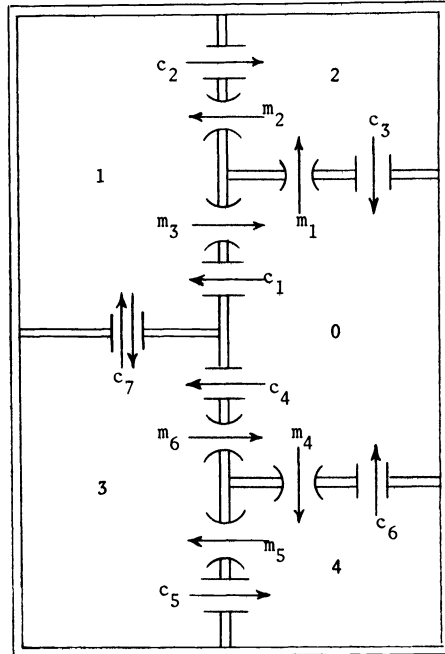


FIG. 9.1. Example 9.2: Maze for cat and mouse.

	00	01	02	03	04	10	11	12	13	14	20	21	22	23	24	30	31	32	33	34	40	41	42	43	44
00	.	.	m ₁	.	m ₄	c ₁	c ₄
01	m ₃	c ₁	c ₄
02	.	.	m ₂	.	.	.	c ₁	c ₄
03	m ₆	c ₁	c ₄
04	.	.	.	m ₅	c ₁	c ₄
10	m ₁	.	m ₄	c ₂	c ₇
11	m ₃	.	.	.	c ₂	c ₇
12	m ₂	.	.	.	c ₂	c ₇
13	m ₆	c ₂	c ₇
14	m ₅	c ₂	c ₇
20	c ₃	m ₁	.	m ₄
21	.	c ₃	m ₃
22	.	.	c ₃	m ₂
23	.	.	.	c ₃	m ₆
24	c ₃	m ₅
30	c ₇	m ₁	.	m ₄	c ₅
31	c ₇	m ₃	.	.	.	c ₅
32	c ₇	m ₂	.	.	.	c ₅	.	.	.
33	c ₇	m ₆	c ₅	.	.
34	c ₇	m ₅	c ₅	.
40	c ₆	m ₁	.	m ₄	.	.	.
41	.	c ₆	m ₃
42	.	.	c ₆	m ₂	.	.	.
43	.	.	.	c ₆	m ₆	.	.	.
44	c ₆	m ₅	.

FIG. 9.2. Example 9.2: Model for cat and mouse.

We want to find a feedback $f \in \Gamma^Q$ such that $P = (i \neq j)$ is an invariant of the closed loop process.

We begin by computing P^\dagger :

$$P_0 = (i \neq j),$$

$$P_1 = H(P_0)$$

$$= P_0 \wedge \text{wlp}_{c_7}(P_0)$$

$$= (i \neq j) \wedge [(i, j) \neq (1, 3)] \wedge [(i, j) \neq (3, 1)],$$

$$P_2 = H(P_1)$$

$$= (i \neq j) \wedge [(i, j) \neq (1, 3)] \wedge [(i, j) \neq (3, 1)]$$

$$= P_1.$$

Hence

$$P^\dagger = (i \neq j) \wedge [(i, j) \neq (1, 3)] \wedge [(i, j) \neq (3, 1)].$$

To find an appropriate feedback solution, we use Corollary 7.1 and set $f_\sigma = \text{wlp}_\sigma(P^\dagger)$. This yields

$$\begin{aligned} f_{c_1} &= [q \neq (0, 1) \wedge q \neq (0, 3)], & f_{m_1} &= [q \neq (2, 0)], \\ f_{c_2} &= [q \neq (1, 2)], & f_{m_2} &= [q \neq (1, 2) \wedge q \neq (3, 2)], \\ f_{c_3} &= [q \neq (2, 0)], & f_{m_3} &= [q \neq (0, 1)], \\ f_{c_4} &= [q \neq (1, 2) \wedge q \neq (0, 1)], & f_{m_4} &= [q \neq (4, 0)], \\ f_{c_5} &= [q \neq (3, 4)], & f_{m_5} &= [q \neq (3, 4) \wedge q \neq (1, 4)], \\ f_{c_6} &= [q \neq (4, 0)], & f_{m_6} &= [q \neq (0, 3)]. \end{aligned}$$

This control implements the following policy. A door that does not access rooms 1 or 3 is closed iff the cat and the mouse are occupying the rooms connected by that door. However, the doors m_2 , m_5 , c_1 , and c_4 (i.e. the doors accessing rooms 1 or 3) require special attention. For example, door c_1 is closed iff the cat is in room 0 and the mouse is in either of rooms 1 or 3. The condition for closing each of the remaining doors follows a similar pattern.

10. Conclusion. We have shown that for the invariant predicate problem it is possible to construct a feedback solution in a modular fashion. Specifically, if the predicate of interest is specified by a conjunction of predicates, e.g., $P = P_1 \wedge P_2$, then it is possible to construct a state feedback to ensure the invariance of the supremal control-invariant predicate stronger than P (i.e. P^\dagger) by first solving the equivalent problem for P_1 and P_2 and forming the conjunction of the solutions. Further, this procedure does not result in any loss of optimality in either the achievable invariant or the maximality of the control.

Although we have restricted attention to state feedback control, all is not lost if the state of the process cannot be measured. Provided the initial state of the process is known it is possible to track the process state by recording the sequence of generated outputs. In practice this is achieved by driving an automaton model G of the process by the output sequence. The pair $S = (G, f)$, where f is the desired feedback map, yields a valid supervisor for the process which can be further simplified if desired using the results of [6].

REFERENCES

- [1] A. ARNOLD AND M. NIVAT, *Controlling behaviours of systems: some basic concepts and some applications*, MFCS 1980 (Lecture Notes in Comput. Sci., 88), Springer-Verlag, New York, 1980, pp. 113–122.
- [2] E. W. DIJKSTRA, *A Discipline of Programming*, Prentice-Hall, Englewood Cliffs, NJ, 1976.
- [3] C. A. R. HOARE, *Communicating Sequential Processes*, Prentice-Hall, Englewood Cliffs, NJ, 1985.
- [4] P. E. LIEPA AND W. M. WONHAM, *Feedback systems in a general algebraic setting*, IEEE Trans. Circuits and Systems, CAS-25 (1978), pp. 728–741.
- [5] P. J. RAMADGE, *Control and Supervision of Discrete Event Processes*, Ph.D. thesis, Dept. of Electrical Engineering, Univ. of Toronto, 1983.
- [6] P. J. RAMADGE AND W. M. WONHAM, *Supervisory control of a class of discrete-event processes*, this Journal, 25 (1987), pp. 206–230; see also Proc. Sixth Internat. Conf. Analysis and Optimization of Systems, Nice, France, June 1984.
- [7] ———, *Modular supervisory control for discrete event systems*, in preparation, 1986; see also Proc. Seventh Internat. Conf. Analysis and Optimization of Systems, Nice, France, June 1986.

- [8] W. M. WONHAM, *Linear Multivariable Control: A Geometric Approach*, Springer-Verlag, New York, 1979.
- [9] W. M. WONHAM AND P. J. RAMADGE, *On the supremal controllable sublanguage of a given language*, this Journal, 25 (1987), pp. 637-659; see also Proc. 23rd IEEE Conf. on Decision and Control, Las Vegas, Nevada, December 1984.
- [10] ———, *On modular synthesis of supervisory controls for discrete-event processes*, International Conference on Computers, Systems and Signal Processing, Bangalore, India, December 1984.

RELAXATION METHODS FOR NETWORK FLOW PROBLEMS WITH CONVEX ARC COSTS*

DIMITRI P. BERTSEKAS[†], PATRICK A. HOSEIN[†] AND PAUL TSENG[†]

Abstract. We consider the standard single commodity network flow problem with both linear and strictly convex possibly nondifferentiable arc costs. For the case where all arc costs are strictly convex we study the convergence of a dual Gauss-Seidel type relaxation method that is well suited for parallel computation. We then extend this method to the case where some of the arc costs are linear. As a special case we recover a relaxation method for the linear minimum cost network flow problem proposed in Bertsekas [1] and Bertsekas and Tseng [2].

Key words. network flow, relaxation methods, parallel computation

1. Introduction. Consider a directed graph with set of nodes N and set of arcs A . We will write $j \sim (i, k)$ to denote that the start and end nodes of arc j are i and k , respectively. The network incidence matrix is denoted by E and has elements e_{ij} given by

$$(1) \quad e_{ij} = \begin{cases} 1 & \text{if } i \text{ is the start node of arc } j, \\ -1 & \text{if } i \text{ is the end node of arc } j, \\ 0 & \text{otherwise.} \end{cases}$$

We denote by x_j the *flow* of arc j , and by d_i the *deficit* of node i which is defined by

$$(2) \quad d_i = \sum_{j \in A} e_{ij} x_j \quad \forall i \in N.$$

In words d_i is the balance of flow outgoing from i and flow coming into i . The vectors with coordinates x_j and d_i are denoted x and d respectively. Thus (2) is written as

$$(3) \quad d = Ex.$$

In what follows the association of particular deficit vectors and flow vectors via (3) should be clear from the context.

Each arc j has associated with it a cost function $f_j: R \rightarrow (-\infty, +\infty]$. We consider the problem of minimizing total cost subject to a conservation of flow constraint at each node:

$$(4) \quad \begin{aligned} &\text{minimize} \quad f(x) = \sum_{j \in A} f_j(x_j) \\ &\text{subject to} \quad x \in C \end{aligned}$$

where C is the *circulation subspace*:

$$(5) \quad C = \{x | d_i = 0, i \in N\} = \{x | Ex = 0\}.$$

We make the following assumptions on f_j .

Assumption A. Each function f_j is convex, lower semicontinuous, and there exists at least one feasible solution for problem (4), i.e., the effective domain of f

$$\text{dom}(f) = \{x | f(x) < +\infty\}$$

and the circulation subspace C have a nonempty intersection.

* Received by the editors January 6, 1986; accepted for publication (in revised form) September 3, 1986. This work was supported by the National Science Foundation under grant NSF-ECS-8217668.

[†] Laboratory for Information and Decision Systems, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139.

Assumption B. The conjugate convex function of each f_j defined by

$$(6) \quad g_j(t_j) = \sup_{x_j} \{t_j x_j - f_j(x_j)\}$$

is real valued, i.e., $-\infty < g_j(t_j) < +\infty$ for all $t_j \in \mathbb{R}$.

Assumption B implies that $f_j(x_j) > -\infty$ for all x_j and j . It follows that the set of points where f_j is real valued, denoted $\text{dom}(f_j)$, is a nonempty interval the right and left endpoints of which (possibly $+\infty$ or $-\infty$) we denote by c_j and l_j , respectively, i.e.,

$$c_j = \sup \{x_j | f_j(x_j) < +\infty\}, \quad l_j = \inf \{x_j | f_j(x_j) < +\infty\}.$$

We call c_j and l_j the *upper and lower capacity bounds* of f_j respectively. It is easily seen that Assumptions A and B imply that for every t_j there is some $x_j \in \text{dom}(f_j)$ attaining the supremum in (6), and furthermore

$$\lim_{|x_j| \rightarrow +\infty} f_j(x_j) = +\infty.$$

It follows that the cost function of (4) has bounded level sets, and therefore (using also the lower semicontinuity of f) *there exists at least one optimal flow vector*.

Assumptions A and B are satisfied, if, for example, f_j is of the form

$$(7) \quad f_j(x_j) = \begin{cases} \hat{f}_j(x_j) & \text{if } x_j \in [l_j, c_j], \\ \infty & \text{otherwise} \end{cases}$$

where l_j , c_j are given upper and lower bounds on the flow of arc j , and \hat{f}_j is a real valued convex function on the real line \mathbb{R} . In this case $g_j(t_j)$ is linear for $|t_j|$ large enough with slopes l_j and c_j as t_j approaches $-\infty$ and $+\infty$, respectively (see Fig. 1.1).

Problem (4) is called the *optimal distribution problem* in Rockafellar [3]. The same reference develops in detail a duality theory (a refinement of what can be obtained from Fenchel's duality theorem) involving the dual problem

$$(8) \quad \begin{aligned} &\text{minimize} && g(t) \triangleq \sum_{j \in A} g_j(t_j) \\ &\text{subject to} && t \in C^\perp \end{aligned}$$

where t is the vector with coordinates t_j , $j \in A$, and C^\perp is the orthogonal complement of C . We call t_j the *tension* of the arc j and C^\perp the *tension subspace*. From (1)–(3) and

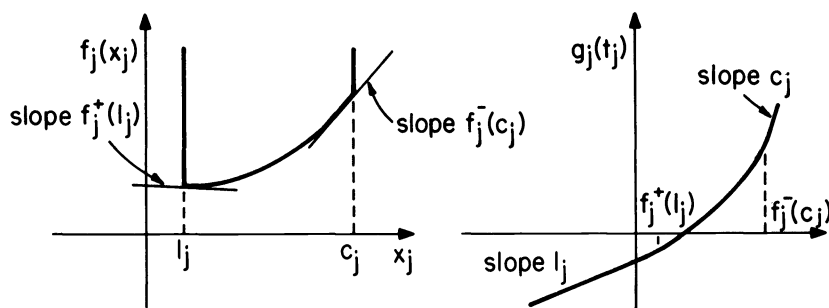


FIG. 1.1. Primal cost function of the form (7) and its dual.

(5) we have that $t \in C^\perp$ if and only if there exist scalars p_i , $i \in N$, called *prices*, such that

$$(9) \quad t_j = p_i - p_k \quad \forall j \in A \text{ with } j \sim (i, k),$$

or equivalently

$$(10) \quad t = E^T p$$

where E^T is the transpose of the network incidence matrix E and p is the vector with coordinates p_i , $i \in N$. Therefore the dual problem (8) can also be written as

$$(11) \quad \begin{array}{ll} \text{minimize} & q(p) \\ \text{subject to} & \text{no constraints on } p \end{array}$$

where q is the *dual functional*

$$(12) \quad q(p) = \sum_{\substack{j \in A \\ j \sim (i, k)}} g_j(p_i - p_k).$$

As shown in [3, p. 349], Assumption A guarantees that there is no duality gap in the sense that the primal and dual optimal costs are opposites of each other.

An important fact for the purposes of the present paper is that (in view of Assumption B above) *the dual problem (11) is an unconstrained optimization problem*. If each function f_j is strictly convex, the dual functional is also differentiable ([4, p. 253]) and as a result unconstrained smooth optimization methods can be applied for solution. This is particularly so since the gradient of the dual cost can be easily calculated. Indeed, when f_j is strictly convex, for every tension vector t there exists a unique flow vector x such that

$$(13) \quad x_j = \arg \max_{z_j} \{t_j z_j - f_j(z_j)\} \quad \forall j \in A$$

and it can be shown [4, p. 218] that x_j is the gradient of g_j at t_j

$$(14) \quad x_j = \nabla g_j(t_j) \quad \forall j \in A.$$

From (1) and (12) we see that for a given price vector p the partial derivatives of the dual functional q are given by

$$(15) \quad \frac{\partial q(p)}{\partial p_i} = \sum_{j \in A} e_{ij} \nabla g_j(t_j) \quad \forall i \in N.$$

Equivalently (cf. (2)), *the partial derivative $\partial q(p)/\partial p_i$ equals the deficit of node i when the arc flows x_j are the unique scalars defined by (13)*.

The differentiability of the dual cost when the primal cost is strictly convex motivates a Gauss-Seidel type of algorithm whereby, given a price vector p , one calculates the corresponding flows $x_j = \nabla g_j(t_j)$, $j \in A$, chooses a node i with positive (negative) deficit and decreases (increases) p_i up to the point where the corresponding partial derivative $\partial q/\partial p_i$ becomes zero. (This amounts to minimizing the dual functional q along the coordinate p_i .) One then repeats the procedure iteratively. The algorithm above is attractive not only because of its simplicity but also because *it lends itself naturally to distributed computation*, whereby minimization along different price coordinates is carried out simultaneously by several processors. Indeed, this can be done

in an asynchronous format as described and analyzed in Bertsekas and El Baz [5]. Simulations of a synchronous parallel method of this type [19] have shown remarkable speedup in computation time.

Gauss-Seidel relaxation methods for unconstrained optimization have been studied extensively [6]–[10]. However they typically require for convergence something like a strict convexity assumption on the cost minimized as well as boundedness of its level sets (see [10] for a counterexample). Unfortunately the dual cost (12) always has unbounded level sets since adding the same constant to all node prices leaves the cost unchanged. Even if we remove this degree of freedom by restricting the price of some special node to be zero (i.e. passing to a quotient space), the dual cost may still have unbounded levels sets and is not strictly convex when the functions f_j are nondifferentiable as in the important special case (7) where they imply capacity constraints. One contribution of the present paper (§ 2) is to show convergence of a flow sequence generated by the Gauss-Seidel method to the unique optimal solution of the primal problem (4). Convergence of the corresponding price vector sequence to some optimal solution of the dual problem (11) is also shown assuming the dual has an optimal solution. For this we actually require that the minimization along coordinates be done only approximately. Furthermore nodes can be relaxed in arbitrary order. The only requirement is that each node is relaxed infinitely often. This result is new and is remarkable in that it requires a rather unconventional method of proof. It improves on a result by Pang [11] (see also an earlier paper by Cottle and Pang [12]) which asserts convergence of the flow vector sequence under the assumption that g_j is of the form (7) with \hat{f}_j differentiable, and strongly convex (rather than just strictly convex as we assume). Pang's result requires exact minimization along each coordinate and contains no assertion on convergence of the price vector sequence; however it applies to a more general problem where the primal cost function need not be separable and the linear constraints need not have a network structure. The paper by Cottle and Pang [12] asserts subsequence convergence to a dual optimal solution for a transportation problem with quadratic arc costs but also uses a nondegeneracy assumption and places a restriction in the way relaxation is carried out. This result is strengthened in our analysis as described above.

When some of the arc cost functions f_j are not strictly convex, the dual cost is not differentiable, and the Gauss-Seidel method breaks down. However Bertsekas [1] and Bertsekas and Tseng [2] have proposed methods that are conceptually related to Gauss-Seidel and work with linear arc costs. They allow line minimization along directions involving several coordinates to cope with situations where minimizing along a single coordinate is not possible. Computational experimentation with standard benchmark problems and a code named RELAX [1], [2] shows that these methods are very promising and outperform, in terms of computation time, some of the best primal simplex and primal dual codes currently available. The second objective of this paper is to propose in § 3 a new relaxation method that in some sense bridges the gap between the strictly convex arc cost Gauss-Seidel method described earlier and the Bertsekas-Tseng linear arc cost version. We show that this method works with both linear and nonlinear (convex) arc costs and contains as special cases both relaxation methods described above. To our knowledge the only other known algorithm for network problems with both linear and nonlinear, possibly nondifferentiable, arc costs is Rockafellar's fortified descent method [3, Chap. 9]. Our algorithm relates in roughly the same way to the Bertsekas-Tseng relaxation method, as Rockafellar's relates to the classical primal-dual method. We note that the methods considered here for linear costs and, more generally, not strictly convex costs are not easily parallelizable. Related

synchronous and asynchronous relaxation methods that admit massive parallelization have been proposed recently in [20], [21].

The last section of the paper provides results of computational experimentation with codes implementing both of the relaxation algorithms proposed.

2. The relaxation method for strictly convex arc costs. In this section in addition to Assumptions A and B, there will be a standing assumption that each f_j is *strictly convex*. Two important consequences of this assumption are that *the optimal flow vector is unique* and *the conjugate functions g_j are differentiable* (in addition to being real valued by Assumption B). Indeed it is easily verified (see also [3] and [4, p. 218]) that we have for all t_j

$$(16) \quad \nabla g_j(t_j) = \arg \max_{x_j} \{t_j x_j - f_j(x_j)\}.$$

Furthermore $\nabla g_j(t_j)$ is the unique scalar x_j satisfying together with t_j the *Complementary Slackness (CS) condition*

$$(17) \quad f_j^-(x_j) \leq t_j \leq f_j^+(x_j)$$

where $f_j^-(x_j)$ and $f_j^+(x_j)$ denote the left and right derivatives of f_j at x_j (see Fig. 2.1). These derivatives are defined in the usual way for x_j in the interior of $\text{dom}(f_j)$. When $-\infty < l_j < c_j$ we define

$$f_j^+(l_j) = \lim_{\xi \downarrow l_j} f_j^+(\xi), \quad f_j^-(l_j) = -\infty.$$

When $l_j < c_j < +\infty$ we define

$$f_j^-(c_j) = \lim_{\xi \uparrow c_j} f_j^-(\xi), \quad f_j^+(c_j) = +\infty.$$

Finally when $l_j = c_j$ we define $f_j^-(l_j) = -\infty$, $f_j^+(c_j) = +\infty$. Note that $\nabla g_j(t_j)$ is continuous and monotonically nondecreasing. We define the *deficit functions* d_i by

$$d_i(p) = \sum_{j \in A} e_{ij} \nabla g_j(t_j) \quad \forall i \in N$$

where $t = E^T p$, and denote by $d(p)$ the vector with coordinates $d_i(p)$. Note that the definition of d is identical to that given in (2), except that here we have used the strict convexity of f_j to express flow and deficit as functions of the dual price vector. In view of the form of the dual functional, the relation above yields

$$d_i(p) = \frac{\partial q(p)}{\partial p_i} \quad \forall i \in N.$$

Since $d_i(p)$ is a partial derivative of a differentiable convex function, we have that $d_i(p)$ is *continuous and monotonically nondecreasing in the coordinate p_i* .

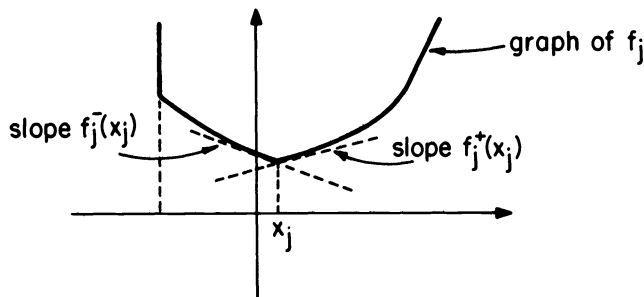


FIG. 2.1. The left and right derivatives of f_j .

We now define a Gauss-Seidel type of algorithm similar to the one sketched in § 1 whereby at each iteration a node s with positive (negative) deficit $d_s(p)$ is chosen and p_s is decreased (increased) with the aim of decreasing the dual cost $q(p)$. More formally, we initially choose a price vector p and a fixed scalar δ in the interval $(0, 1)$. Then we execute repeatedly the relaxation iteration described below.

Relaxation iteration for strictly convex arc costs.

If $d_i(p) = 0 \forall i \in N$ then STOP.

Else

Choose any node s . Set $\beta = d_s(p)$.

If $\beta = 0$, do nothing.

If $\beta > 0$, then decrease p_s so that $0 \leq d_s(p) \leq \delta\beta$.

If $\beta < 0$, then increase p_s so that $0 \geq d_s(p) \geq \delta\beta$.

The only assumption we make regarding the order in which nodes are chosen for relaxation is the following.

Assumption C. Every node in N is chosen as the node s in the relaxation iteration an infinite number of times.

The relaxation iteration is well defined, in the sense that every step in the iteration is executable. To see this suppose that $\beta > 0$ and there does not exist a $\Delta < 0$ such that $d_s(p + \Delta e_s) \leq \delta\beta$, where e_s denotes the s th coordinate vector. Then using the definition of d , l_j , and c_j , it is easily seen that

$$\lim_{\Delta \rightarrow -\infty} d_s(p + \Delta e_s) = \sum_{e_{sj} > 0} e_{sj} l_j + \sum_{e_{sj} < 0} e_{sj} c_j \geq \delta\beta > 0,$$

which implies that the flow deficit of node s is positive for any flow x within the upper and lower arc capacity bounds and contradicts the existence of a feasible flow (Assumption A). An analogous argument can be made for the case where $\beta < 0$.

In order to obtain our convergence result we must show that the sequence of flow vectors generated by the relaxation algorithm approaches the circulation subspace C (given by (5)). The line of argument that we will use is as follows: We will lower bound the amount of improvement in the dual functional q per iteration by a positive quantity. We will then show that if the sequence of flow vectors do not approach the circulation subspace, the quantity itself can be lower bounded by a positive constant which implies that the optimal dual functional has a value of $-\infty$. This will contradict the finiteness of the optimal primal cost.

We will denote the price vector generated at the r th iteration by p^r , $r = 0, 1, 2, \dots$ and the node operated on at the r th iteration by s^r , $r = 0, 1, 2, \dots$. To simplify notation we will denote

$$t^r = E^T p^r, \quad x_j^r = \nabla g_j(t_j^r).$$

We denote by x^r the vector with coordinates x_j^r , $j \in A$. Note the symmetry following from the CS condition (16) or (17): x_j^r is the gradient of the dual cost g_j at t_j^r , while t_j^r is a subgradient of the primal cost f_j at x_j^r . For any directed cycle Y of the network we will use Y^+ to denote the set of arcs $\{j \in A \mid j \text{ is positively oriented in } Y\}$, and Y^- to denote $Y \setminus Y^+$. We first show three preliminary results.

PROPOSITION 2.1. *We have for all r such that $p^{r+1} \neq p^r$ [i.e. $d_{s^r}(p^r) \neq 0$]*

$$(18) \quad q(p^r) - q(p^{r+1}) \geq \sum_{j \in A} [f_j(x_j^{r+1}) - f_j(x_j^r) - (x_j^{r+1} - x_j^r) t_j^r] > 0,$$

with equality holding if line minimization is used [$d_{s^r}(p^{r+1}) = 0$].

Proof. Fix an index $r \geq 0$. Denote $s = s^r$ and $\Delta = p_s^{r+1} - p_s^r$. From (6), (12) and (16) we have

$$q(p^r) = \sum_{j \in A} [t_j^r x_j^r - f_j(x_j^r)] \quad \forall r \geq 0.$$

Therefore

$$\begin{aligned} q(p^r) - q(p^{r+1}) &= \sum_{j \in A} [t_j^r x_j^r - f_j(x_j^r)] - \sum_{j \in A} [t_j^{r+1} x_j^{r+1} - f_j(x_j^{r+1})] \\ &= \sum_{j \in A} [t_j^r x_j^r - f_j(x_j^r)] - \sum_{j \in A} [(t_j^r + e_{sj} \Delta) x_j^{r+1} - f_j(x_j^{r+1})] \\ &= \sum_{j \in A} [f_j(x_j^{r+1}) - f_j(x_j^r) - (x_j^{r+1} - x_j^r) t_j^r - e_{sj} \Delta x_j^{r+1}] \\ &= \sum_{j \in A} [f_j(x_j^{r+1}) - f_j(x_j^r) - (x_j^{r+1} - x_j^r) t_j^r] - \Delta \sum_{j \in A} e_{sj} x_j^{r+1} \\ &= \sum_{j \in A} [f_j(x_j^{r+1}) - f_j(x_j^r) - (x_j^{r+1} - x_j^r) t_j^r] - \Delta d_s(p^{r+1}). \end{aligned}$$

Since $\Delta d_s(p^{r+1}) \leq 0$ (and $d_s(p^{r+1}) = 0$ if we use line minimization) the left side of (18) follows. The right side of (18) follows from the strict convexity of f_j and the fact $x_j^{r+1} \neq x_j^r$. QED

PROPOSITION 2.2. *The sequence $\{x^r\}$ is bounded.*

Proof. We first note that at every iteration the total deficit does not increase, i.e.,

$$\sum_{i \in N} |d_i(p^{r+1})| \leq \sum_{i \in N} |d_i(p^r)|.$$

(This follows from the fact that a flow change on an arc reflects itself in a change of the deficit of its start node and an opposite change in the deficit of its end node. Furthermore the deficit of node s^r chosen for relaxation at the r th iteration cannot increase in absolute value or change sign during that iteration.) It follows that $\{d(p^r)\}$ is bounded. We now argue by contradiction. Suppose $\{x^r\}$ is unbounded. Then there must exist an arc j and a subsequence R such that $|x_j^r| \rightarrow +\infty$ as $r \rightarrow \infty$, $r \in R$. Since $\{d(p^r)\}$ is bounded it follows (passing into another subsequence if necessary) that there exists a directed cycle Y such that $x_j^r \rightarrow +\infty$ for all $j \in Y^+$, and $x_j^r \rightarrow -\infty$ for all $j \in Y^-$ as $r \rightarrow \infty$, $r \in R$. Since by the CS condition (17)

$$f_j^-(x_j^r) \leq t_j^r \leq f_j^+(x_j^r),$$

and also

$$\sum_{j \in Y^+} t_j^r - \sum_{j \in Y^-} t_j^r = 0,$$

we have for all r

$$\sum_{j \in Y^+} f_j^-(x_j^r) - \sum_{j \in Y^-} f_j^+(x_j^r) \leq 0.$$

This is a contradiction since $x_j^r \rightarrow +\infty$ implies $f_j^-(x_j^r) \rightarrow +\infty$ while $x_j^r \rightarrow -\infty$ implies $f_j^+(x_j^r) \rightarrow -\infty$. QED

The next result is remarkable in that it shows that under a mild restriction on the way the relaxation iteration is carried out (which is typically very easy to satisfy in practice), the sequence of price vectors approaches the dual optimal set in an unusual manner. The result depends on the monotonicity of the functions ∇g_j .

PROPOSITION 2.3. *Given $p \in R^{|N|}$, let s be a node and let \bar{p} denote a dual price vector obtained by applying the relaxation iteration to p using node s . Assume in addition that \bar{p} is chosen so that*

$$(19a) \quad \text{if } d_s(p) > 0 \text{ then } d_s[\bar{p} + \alpha(p - \bar{p})] > 0 \quad \forall \alpha > 0,$$

$$(19b) \quad \text{if } d_s(p) < 0 \text{ then } d_s[\bar{p} + \alpha(p - \bar{p})] < 0 \quad \forall \alpha > 0.$$

Then for all $k \in N$, and all optimal dual price vectors p^ we have*

$$(20) \quad \min \{p_i - p_i^* | i \in N\} \leq \bar{p}_k - p_k^* \leq \max \{p_i - p_i^* | i \in N\}.$$

Note. Assumption (19) when $d_s(p) > 0$ [$d_s(p) < 0$] is equivalent to assuming that \bar{p}_s is chosen greater (less) or equal to the largest (smallest) minimizing point of the dual cost along the s th coordinate starting from p . It is automatically satisfied if the dual cost has a unique minimizing point along the line $\{p + \alpha e_s | \alpha \in R\}$.

Proof. Fix an optimal dual price vector p^* and consider an arbitrary price vector \tilde{p} . Let k be such that $\tilde{p}_k - p_k^* = \max \{\tilde{p}_i - p_i^* | i \in N\}$. We have

$$\tilde{p}_k - p_k^* \geq \tilde{p}_i - p_i^* \quad \forall i \neq k$$

so that

$$\tilde{p}_k - \tilde{p}_i \geq p_k^* - p_i^* \quad \forall j \sim (k, i), \quad \tilde{p}_i - \tilde{p}_k \leq p_i^* - p_k^* \quad \forall j \sim (i, k).$$

Since ∇g_j is a nondecreasing function, we have that

$$\nabla g_j(\tilde{p}_k - \tilde{p}_i) \geq \nabla g_j(p_k^* - p_i^*) \quad \forall j \sim (k, i), \quad \nabla g_j(\tilde{p}_i - \tilde{p}_k) \leq \nabla g_j(p_i^* - p_k^*) \quad \forall j \sim (i, k).$$

Thus $d_k(\tilde{p}) \geq d_k(p^*) = 0$.

The desired assertion (20) holds if $d_s(p) = 0$ since then we have $\bar{p} = p$. Assume that $d_s(p) < 0$. Consider the vector \tilde{p} defined by

$$\tilde{p}_i = \begin{cases} p_i & \text{if } i \neq s, \\ p_s^* + \max \{p_j - p_j^* | j \in N\} & \text{if } i = s. \end{cases}$$

Then we have $\tilde{p}_s - p_s^* = \max \{\tilde{p}_i - p_i^* | i \in N\} = \max \{p_i - p_i^* | i \in N\}$ and by the preceding argument we have $d_s(\tilde{p}) \geq 0$. Therefore, using assumption (19), we have $\bar{p}_s \leq \tilde{p}_s$ while at the same time $p_s < \bar{p}_s$, and $p_i = \bar{p}_i$ for all $i \neq s$. The assertion (20) follows. The proof is similar when $d_s(p) > 0$. Q.E.D.

Note that Proposition 2.3 implies among other things that, if (19) is satisfied at all iterations, the sequence $\{p^r\}$ generated by the relaxation method is bounded. Furthermore if we can show that $\{p^r\}$ accumulates at an optimal price vector, the proposition implies that $\{p^r\}$ must converge to that vector. We are now ready to show our main result.

PROPOSITION 2.4. *Let $\{p^r, x^r\}$ be a sequence generated by the relaxation method for strictly convex arc costs. Then*

$$(21) \quad (a) \quad \lim_{r \rightarrow \infty} d(p^r) = 0.$$

$$(22) \quad (b) \quad \lim_{r \rightarrow \infty} x^r = x^*$$

where x^* is the unique optimal flow vector.

$$(c) \quad \lim_{r \rightarrow \infty} q(p^r) = -f(x^*) = \inf_p q(p).$$

(d) If condition (19) is satisfied at each iteration, and the dual problem has an optimal solution, then

$$(23) \quad \lim_{r \rightarrow \infty} p^r \rightarrow p^*$$

where p^* is some optimal price vector.

Proof. (a) We first show that

$$(24) \quad \lim_{r \rightarrow \infty} d_{s^r}(p^r) = 0.$$

Indeed, if this is not so there must exist an $\varepsilon > 0$ and a subsequence R such that $|d_{s^r}(p^r)| \geq \varepsilon$ for all $r \in R$. Without loss of generality we assume that $d_{s^r}(p^r) \geq \varepsilon$ for all $r \in R$. Since $\delta |d_{s^r}(p^r)| \geq |d_{s^r}(p^{r+1})|$ we have that at the r th iteration some arc incident to node s^r must change its flow by at least Δ where $\Delta = (1 - \delta)\varepsilon/|A|$. By passing to a subsequence if necessary we assume that this happens for the same arc j^* for all $r \in R$, and that $x_{j^*}^{r+1} - x_{j^*}^r \geq \Delta$, for all $r \in R$. Using the boundedness of $\{x^r\}$ (Proposition 2.2) we may also assume that the subsequence $\{x_{j^*}^r\}_{r \in R}$ converges to some x_{j^*} . Using the convexity of f_j and Proposition 2.1 we have

$$\begin{aligned} q(p^r) - q(p^{r+1}) &\geq f_{j^*}(x_{j^*}^{r+1}) - f_{j^*}(x_{j^*}^r) - (x_{j^*}^{r+1} - x_{j^*}^r) t_{j^*}^r \\ &\geq f_{j^*}(x_{j^*}^r + \Delta) - f_{j^*}(x_{j^*}^r) - \Delta t_{j^*}^r \\ &\geq f_{j^*}(x_{j^*}^r + \Delta) - f_{j^*}(x_{j^*}^r) - \Delta f_{j^*}^+(x_{j^*}^r). \end{aligned}$$

Taking the limit as $r \rightarrow \infty$, $r \in R$ and using the facts $x_{j^*}^r \rightarrow x_{j^*}$ and $\lim_{r \rightarrow \infty} f_{j^*}^+(x_{j^*}^r) \leq f_{j^*}^+(x_{j^*})$ (in view of the upper semicontinuity of $f_{j^*}^+$) we obtain

$$\liminf_{\substack{r \rightarrow \infty \\ r \in R}} [q(p^r) - q(p^{r+1})] \geq f_{j^*}(x_{j^*} + \Delta) - f_{j^*}(x_{j^*}) - \Delta f_{j^*}^+(x_{j^*}) > 0.$$

This implies that $\lim_{r \rightarrow \infty} q(p^r) = -\infty$. But this is not possible because from (6) and (12) we have $q(p) \geq -\sum_{j \in A} f_j(x_j)$ for all p and $x \in C$. Therefore (24) is proved by contradiction.

We now show (21). Choose any $i \in N$. Take any $\varepsilon > 0$ and let R be the set of indices r such that $d_i(p^r) > 2\varepsilon$. Assume without loss of generality that $d_i(p^r) < \varepsilon$ for all r with $i = s^r$ (cf. (24)). For every $r \in R$ let r' be the first index with $r' > r$ such that $i = s^{r'}$. Then during iterations $r, r+1, \dots, r'-1$ node i is not chosen for relaxation while its deficit decreases from greater than 2ε to lower than ε . We claim that during these iterations the total deficit $\sum_{k \in N} |d_k(p)|$ is decreased by an amount of more than 2ε . To see this, note that the total absolute deficit cannot increase at any iteration as noted earlier in the proof of Proposition 2.2. Next observe that for any of the iterations $r, r+1, \dots, r'-1$, say \bar{r} , for which the deficit of i is decreased by an amount $\xi > 0$ from a positive value $d_i(p^{\bar{r}}) > 0$, it must be that the node s chosen for relaxation is a neighbor of i and has a negative deficit $d_s(p^{\bar{r}}) < 0$. Since all increase in $d_s(p^r)$ during the iteration must be matched by decreases of the deficits of the neighbor nodes of s , and the deficit of s will remain nonpositive after the iteration, it follows that the total absolute deficit will be decreased by at least $2 \min \{\xi, d_i(p^{\bar{r}})\}$ during the iteration. This shows that during iterations $r, r+1, \dots, r'-1$ the total absolute deficit must decrease by more than 2ε . It follows that the set R of indices r for which $d_i(p^r) > 2\varepsilon$ cannot be infinite. Since $\varepsilon > 0$ is arbitrary we obtain $\limsup_{r \rightarrow \infty} d_i(p^r) \leq 0$. Similarly we can show that $\liminf_{r \rightarrow \infty} d_i(p^r) \geq 0$ and therefore $d_i(p^r) \rightarrow 0$.

(b) For all r and arcs j we have the CS condition

$$(25) \quad f_j^-(x_j^r) \leq t_j^r \leq f_j^+(x_j^r).$$

If Y is any cycle we have

$$\sum_{j \in Y^+} t_j^r - \sum_{j \in Y^-} t_j^r = 0,$$

so from (25) we obtain

$$(26) \quad \sum_{j \in Y^+} f_j^-(x_j^r) - \sum_{j \in Y^-} f_j^+(x_j^r) \leq 0 \leq \sum_{j \in Y^+} f_j^+(x_j^r) - \sum_{j \in Y^-} f_j^-(x_j^r).$$

Let $\{x^r\}_{r \in R}$ be a subsequence converging to some \bar{x} (cf. Proposition 2.2). Then from (26) and the lower (upper) semicontinuity of $f_j^-(f_j^+)$, we have for all cycles Y

$$\sum_{j \in Y^+} f_j^-(\bar{x}_j) - \sum_{j \in Y^-} f_j^+(\bar{x}_j) \leq 0 \leq \sum_{j \in Y^+} f_j^+(\bar{x}_j) - \sum_{j \in Y^-} f_j^-(\bar{x}_j),$$

while from part (a) we have $\bar{x} \in C$. This implies that \bar{x} is an optimal flow ([3, Chap. 8]) and therefore must be equal to the unique optimal flow x^* . Since, by Proposition 2.2, $\{x^r\}$ is bounded we obtain $x^r \rightarrow x^*$.

(c) For every arc j for which $l_j < c_j$ there are three possibilities:

- (1) $\{t_j^r\}$ is bounded.
- (2) $x_j^* = c_j < +\infty$, $x_j^r \leq x_j^*$ and $-\infty < \liminf_{r \rightarrow \infty} t_j^r \leq \limsup_{r \rightarrow \infty} t_j^r = +\infty$.
- (3) $x_j^* = l_j > -\infty$, $x_j^r \geq x_j^*$ and $-\infty = \liminf_{r \rightarrow \infty} t_j^r \leq \limsup_{r \rightarrow \infty} t_j^r < +\infty$,

while for an arc j with $l_j = c_j$ we must have $x_j^* = x_j^r$ for all r . Using this fact we can easily see that we can construct a subsequence R such that

$$\sum_{j \in A} t_j^r(x_j^r - x_j^*) \leq \sum_{j \in B} t_j^r(x_j^r - x_j^*) \quad \forall r \in R$$

where B is a set of arcs j such that $\{t_j^r\}_R$ is bounded. We have (since $t^r \in C^\perp$, $x^* \in C$, and therefore $\sum_{j \in A} t_j^r x_j^* = 0$)

$$f(x^r) + q(p^r) = \sum_{j \in A} t_j^r x_j^r = \sum_{j \in A} t_j^r(x_j^r - x_j^*) \leq \sum_{j \in B} t_j^r(x_j^r - x_j^*).$$

Since $x_j^r \rightarrow x_j^*$ and $\{t_j^r\}_R$ is bounded for $j \in B$ we obtain by taking the limit above $f(x^*) + \lim_{r \rightarrow \infty} q(p^r) \leq 0$. On the other hand we have for all p using (6) and (12) $f(x^*) + q(p) \geq 0$. This together with the preceding relation show the desired result.

(d) By Proposition 2.3, $\{p^r\}$ is bounded. Let $\{p^r\}_{r \in R}$ be a subsequence converging to a vector p^* and let $t^* = E^T p^*$. We have for all $j \in A$

$$f_j^-(x_j^r) \leq t_j^r \leq f_j^+(x_j^r) \quad \forall r \in R.$$

It follows using part (b) and the lower (upper) semicontinuity of $f_j^-(f_j^+)$ that for all $j \in A$, $f_j^-(x_j^*) \leq t_j^* \leq f_j^+(x_j^*)$ where x^* is the optimal flow vector. Therefore t^* satisfies together with x^* the complementary slackness conditions and must be dual optimal. Proposition 2.3 shows that $\{p^r\}$ cannot have two different dual optimal price vectors as limit points and the conclusion follows. QED

3. The relaxation method for mixed linear and strictly convex arc costs. We first introduce some terminology. We will say that a point $b \in \text{dom}(f_j)$ is a *breakpoint* of f_j if $f_j^-(b) < f_j^+(b)$. Note that the dual functional q , as given by (12), is separable and is piecewise either linear or strictly convex. Roughly speaking each linear piece (breakpoint) of the primal cost function f_j corresponds to a breakpoint (linear piece) of the dual cost function g_j (see Fig. 3.1).

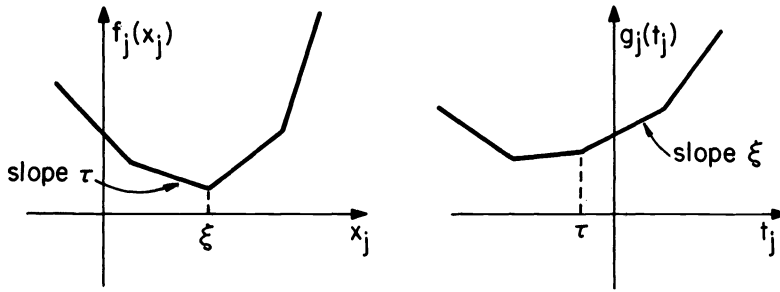


FIG. 3.1. Correspondence between the breakpoints of f_j and the linear pieces of g_j (and vice versa).

Assumption D. $f_j^+(x_j) > -\infty$ and $f_j^-(x_j) < +\infty$ for all $x_j \in \text{dom}(f_j)$.

In the terminology of ([3, Chap. 8]), Assumption D implies that every feasible primal solution is regularly feasible and guarantees, together with Assumptions A and B, that the dual problem has an optimal solution ([3, p. 360]). For a given $\varepsilon > 0$, we say that $x \in R^{|A|}$ and $p \in R^{|N|}$ satisfy ε -Complementary Slackness (ε -CS for short) if

$$(27) \quad f_j^-(x_j) - \varepsilon \leq t_j \leq f_j^+(x_j) + \varepsilon \quad \forall j \in A$$

where $t = E^T p$. For a given p , (27) defines upper and lower bounds, called ε -bounds, on the flow vector:

$$(28) \quad l_j^\varepsilon = \min \{ \xi | f_j^+(\xi) \geq t_j - \varepsilon \}, \quad c_j^\varepsilon = \max \{ \xi | f_j^-(\xi) \leq t_j + \varepsilon \} \quad \forall j \in A.$$

Then x and p satisfying ε -CS is equivalent to

$$(29) \quad x_j \in [l_j^\varepsilon, c_j^\varepsilon] \quad \forall j \in A$$

where $t = E^T p$. For a given t_j , we can obtain l_j^ε and c_j^ε from the graph of the subdifferential mapping of f_j as shown in Figs. 3.2-3.3. Intuition suggests that if x is in the

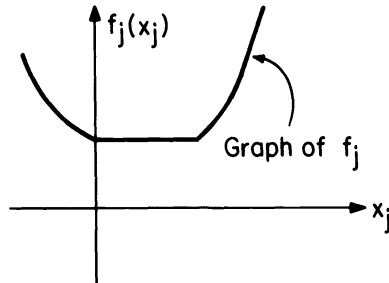


FIG. 3.2. Graph of f_j .

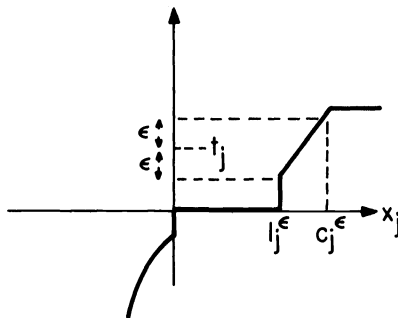


FIG. 3.3. Graph of ∂f_j and ε -bounds corresponding to t_j .

circulation subspace C , x and p satisfy ε -CS and ε is small, then both x and p should be near optimal. This idea will be made precise later when we explore the near optimality properties of the solution generated by a relaxation algorithm that uses the notion of ε -CS. The definition of ε -CS is related to the ε -subgradient idea introduced in nondifferentiable optimization in [13] as well as to the fortified descent method of Rockafellar [3]. The latter method, however, for a given p and $t = E^T p$, uses different lower and upper bounds on x_j given by

$$\inf_{\Delta > 0} \frac{g_j(t_j + \Delta) - g_j(t_j) + \varepsilon}{\Delta} \quad \text{and} \quad \sup_{\Delta > 0} \frac{g_j(t_j) - g_j(t_j - \Delta) - \varepsilon}{\Delta}.$$

Our bounds of (28) seem simpler for implementation purposes particularly when some of the cost functions f_j are linear within their effective domain.

For a given x within the ε -bounds, we define the *deficit* of node i as in (2) and say that a sequence of nodes $\{n_1 \cdots n_k\}$ forms a *flow augmenting path* if

$$d_{n_1} < 0, d_{n_k} > 0 \quad \text{and} \quad \begin{cases} x_j < c_j^\varepsilon & \text{if } j \sim (n_m, n_{m+1}), \\ x_j > l_j^\varepsilon & \text{if } j \sim (n_{m+1}, n_m), \end{cases} \quad m \in \{1, \dots, k-1\}.$$

Let

$$\mu_m = \begin{cases} c_j^\varepsilon - x_j & \text{if } j \sim (n_m, n_{m+1}), \\ x_j - l_j^\varepsilon & \text{if } j \sim (n_{m+1}, n_m), \end{cases} \quad m \in \{1, \dots, k-1\}.$$

We will call

$$\mu = \min \{-d_{n_1}, d_{n_k}, \mu_1, \dots, \mu_{k-1}\}$$

the *capacity* of the path. The relaxation algorithm of this section uses the labeling method of Ford and Fulkerson [14] for finding flow augmenting paths and for augmenting flow along them.

For a given tension vector $t \in C$ and any subset of nodes S , we define $C_\varepsilon(S, t)$ by

$$(30) \quad C_\varepsilon(S, t) = \sum_{j \in [S, N \setminus S]} l_j^\varepsilon - \sum_{j \in [N \setminus S, S]} c_j^\varepsilon$$

where we use the notation

$$[S, N \setminus S] = \{j | j \sim (i, k), i \in S, k \notin S\}, \quad [N \setminus S, S] = \{j | j \sim (i, k), i \notin S, k \in S\}.$$

We also define the $|N|$ -vector $u(S)$ by

$$u_i(S) = \begin{cases} -1 & \text{if } i \in S, \\ 0 & \text{if } i \notin S. \end{cases}$$

The importance of these notions is due to the fact that for any $\varepsilon \geq 0$, $C_\varepsilon(S, t) > 0$ implies that $u(S)$ is a dual descent direction at p , where p is any price vector satisfying $E^T p = t$. This follows from the fact that the directional derivative of q at p in the direction $u(S)$ defined by

$$q'(p; u(S)) = \lim_{\Delta \downarrow 0} \frac{q(p + \Delta u(S)) - q(p)}{\Delta}$$

is easily verified to be

$$q'(p; u(S)) = \sum_{j \in [N \setminus S, S]} c_j^0 - \sum_{j \in [S, N \setminus S]} l_j^0 \leq -C_\varepsilon(S, t)$$

where c_j^0, l_j^0 are the ε -bounds corresponding to $\varepsilon = 0$, and we are making use of the fact $c_j^0 \leq c_j^\varepsilon, l_j^0 \geq l_j^\varepsilon$, for all $\varepsilon \geq 0$.

We now describe the relaxation algorithm. The algorithm is iterative and uses the ε -CS idea. The scalar ε is kept fixed throughout the algorithm. At the beginning of

each iteration we have a dual price vector p and a flow vector x satisfying $l_j^\varepsilon \leq x_j \leq c_j^\varepsilon$ for all $j \in A$. If $x \in C$ then we terminate. Otherwise we use labeling to either find a flow augmenting path, in which case a flow augmentation is performed to bring x "closer" to C ; or to find a dual descent direction, in which case a dual descent along this direction is performed. When each f_j is linear within its effective domain, $\varepsilon = 0$, and all problem data is integer, the algorithm coincides with the relaxation method of [1], [2]. When each f_j is strictly convex and $\varepsilon = 0$ the algorithm coincides with the exact line minimization version of the algorithm of the previous section ($\delta = 0$).

Relaxation iteration.

Step 0. Given p and x satisfying $l_j^\varepsilon \leq x_j \leq c_j^\varepsilon$ for all j , let t and d be the corresponding tension and deficit vectors.

Step 1. Pick a node s such that $d_s > 0$. If no such node exists terminate. Else set all nodes to be unlabeled and unscanned. Give the label 0 to node s . Set $S = \emptyset$ and go to Step 2.

Step 2. Choose a labeled but unscanned node k . Set $S \leftarrow S \cup \{k\}$ and go to Step 3.

Step 3. Scan the label of the node k as follows: Give the label k to all unlabeled nodes m such that $x_j < c_j^\varepsilon$ for $j \sim (m, k)$ and to all unlabeled nodes m such that $x_j > l_j^\varepsilon$ for $j \sim (k, m)$. If $C_\varepsilon(S, t) > 0$ then go to Step 5. Else if for any of the nodes m labeled from k we have $d_m < 0$ then go to Step 4. Else go to Step 2.

Step 4. (Flow Augmentation Step). A flow augmenting path has been found which starts at the node m (with $d_m < 0$) identified in Step 3 and ends at the node s . The path can be constructed by tracing labels starting from m . Let μ be the capacity of the path. Increase by μ the flow of all arcs on the path oriented in the direction from m to s , decrease by μ the flow of all other arcs on the path. Update the deficit vector d and return.

Step 5. (Dual Descent Step). Determine λ^* such that

$$q(p + \lambda^* u(S)) = \min_{\lambda > 0} \{q(p + \lambda u(S))\}.$$

Set $p \leftarrow p + \lambda^* u(S)$ and update the bounds l_j^ε and c_j^ε . Update x to maintain the ε -CS condition $l_j^\varepsilon \leq x_j \leq c_j^\varepsilon$ and return.

Validity and finite termination of the relaxation iteration. We will show that, under Assumption D, all steps in the Relaxation Iteration are executable and that the iteration terminates in a finite number of operations.

Steps 0, 1, and 3 are trivially executable. Step 2 is certainly executable on its first pass since the node s is labeled but unscanned. To show that it remains executable on subsequent passes we only need to verify that each time we go to Step 2 from Step 3 there always exists a labeled but unscanned node. In Step 2, if all labeled nodes are also scanned we have

$$\begin{aligned} \sum_{i \in S} d_i &= \sum_{j \in [S, N \setminus S]} x_j - \sum_{j \in [N \setminus S, S]} x_j \\ &= \sum_{j \in [S, N \setminus S]} l_j^\varepsilon - \sum_{j \in [N \setminus S, S]} c_j^\varepsilon = C_\varepsilon(S, t). \end{aligned}$$

Since node s has positive deficit and all other labeled nodes have nonnegative deficits we obtain that $C_\varepsilon(S, t) > 0$ and therefore in the previous pass through Step 3 we would have branched to Step 5 rather than to Step 2. Step 4 is executable since the rule for

labeling ensures that a flow augmenting path exists from node m to node s , so a flow augmentation is possible. Step 5 is executable since $C_\varepsilon(S, t) > 0$ implies that $u(S)$ is a dual descent direction at p , and we can show that there exists a minimizing stepsize λ^* . To see this assume the contrary, i.e., that there does not exist a stepsize λ^* achieving the minimum along the direction $u(S)$. In that case the convexity of q implies that

$$q'(p + \lambda u(S); u(S)) < 0 \quad \forall \lambda > 0,$$

$$\lim_{\lambda \rightarrow +\infty} q'(p + \lambda u(S); u(S)) \leq 0.$$

Then it can be easily seen that either

$$\sum_{j \in [S, N \setminus S]} l_j - \sum_{j \in [N \setminus S, S]} c_j > 0,$$

in which case Assumption A is violated [$\text{dom}(f) \cap C$ is empty], or

$$\sum_{j \in [S, N \setminus S]} l_j - \sum_{j \in [N \setminus S, S]} c_j = 0$$

and either $f_j^+(l_j) = -\infty$ for some $j \in [S, N \setminus S]$ or $f_j^-(c_j) = +\infty$ for some $j \in [N \setminus S, S]$, in which case Assumption D is violated. To complete the proof that the relaxation iteration terminates in a finite number of operations we note that we cannot loop between Step 2 and Step 3 infinitely often since the number of scanned nodes is increased by one each time we visit Step 3.

We next show that the relaxation algorithm, when applied in conjunction with an easily implementable labeling rule, terminates in a finite number of iterations. The proof may be divided into two separate parts. The first part involves showing that the number of dual descent steps is not infinite. This is done by arguing that the optimal dual cost is necessarily $-\infty$ if the number of dual descent steps is infinite. The second part involves showing that the number of flow augmentations between successive dual descent steps is finite. This is done by choosing an appropriate labeling scheme for the relaxation algorithm and showing that the number of flow augmentations is finite under the chosen scheme. For this purpose we will propose two schemes: *breadth-first search* and *arc discrimination*.

We first show that the stepsize in each dual descent step is bounded from below by ε . Indeed our definition of ε -CS was motivated primarily by this fact.

PROPOSITION 3.1. *The stepsize in each dual descent step is greater than ε .*

Proof. Under Assumption B, $q(p)$ is subdifferentiable everywhere. Let S denote the subset of nodes corresponding to the dual descent direction generated by the relaxation iteration. In other words, the dual descent direction u is given by

$$(31) \quad u_i = \begin{cases} -1 & \text{if } i \in S, \\ 0 & \text{if } i \notin S, \end{cases}$$

and S satisfies $C_\varepsilon(S, t) > 0$. Now consider p' given by $p' = p + \varepsilon u$ and $t' = E^T p'$. Then

$$t'_j = t_j - \varepsilon \quad \text{if } j \in [S, N \setminus S],$$

$$t'_j = t_j + \varepsilon \quad \text{if } j \in [N \setminus S, S],$$

$$t'_j = t_j \quad \text{otherwise,}$$

so that

$$l_j^\varepsilon = \min \{ \xi | f_j^+(\xi) \geq t_j - \varepsilon \} = \min \{ \xi | f_j^+(\xi) \geq t_j' \} \quad \text{for all } j \in [S, N \setminus S],$$

$$c_j^\varepsilon = \max \{ \xi | f_j^-(\xi) \leq t_j + \varepsilon \} = \max \{ \xi | f_j^-(\xi) \leq t_j' \} \quad \text{for all } j \in [N \setminus S, S].$$

Therefore

$$q'(p'; u) = - \sum_{j \in [S, N \setminus S]} l_j^\varepsilon + \sum_{j \in [N \setminus S, S]} c_j^\varepsilon = -C_\varepsilon(S, t) < 0.$$

Since q is convex, $q'(p; u) < 0$ and $q'(p + \varepsilon u; u) < 0$ imply that $q'(p + \alpha u; u) < 0$ for all $\alpha \in [0, \varepsilon]$. Therefore the stepsize in a dual descent step is greater than ε . QED

We will now use Proposition 3.1 to prove that the number of dual descent steps is necessarily finite. The following result is a first step in this direction.

PROPOSITION 3.2. *Let p^r denote the price vector generated by the relaxation algorithm just before the r th dual descent step. Then for each $r \in \{0, 1, 2, \dots\}$*

$$(32) \quad q(p^r) - q(p^{r+1}) > \sum_{\substack{j \in [S', N \setminus S'] \text{ or} \\ j \in [N \setminus S', S']}} [f_j(\Psi_j^r) - f_j(X_j^r) - (\Psi_j^r - X_j^r)t_j'] \geq 0$$

where we define

$$\Psi_j^r = \begin{cases} g_j^+(t_j' - \varepsilon) & \text{if } j \in [S', N \setminus S'], \\ g_j^-(t_j' + \varepsilon) & \text{if } j \in [N \setminus S', S'], \end{cases} \quad X_j^r = \begin{cases} g_j^-(t_j') & \text{if } j \in [S', N \setminus S'], \\ g_j^+(t_j') & \text{if } j \in [N \setminus S', S'], \end{cases}$$

and S^r denotes the node subset corresponding to the descent direction at the r th dual descent step.

Proof. From the definition of Ψ_j^r and X_j^r we have that

$$g_j(t_j') = X_j^r t_j' - f_j(X_j^r), \quad g_j(t_j' - \varepsilon) = \Psi_j^r(t_j' - \varepsilon) - f_j(\Psi_j^r) \quad \forall j \in [S', N \setminus S'],$$

$$g_j(t_j') = X_j^r t_j' - f_j(X_j^r), \quad g_j(t_j' + \varepsilon) = \Psi_j^r(t_j' + \varepsilon) - f_j(\Psi_j^r) \quad \forall j \in [N \setminus S', S'].$$

From the definition of q , S^r and $u(S^r)$ we have that

$$q(p^r + \varepsilon u(S^r)) = q(p^r) + \sum_{j \in [S', N \setminus S']} [g_j(t_j' - \varepsilon) - g_j(t_j')] \\ + \sum_{j \in [N \setminus S', S']} [g_j(t_j' + \varepsilon) - g_j(t_j')]$$

and from Proposition 3.1 we have that

$$q(p^r) - q(p^{r+1}) \geq q(p^r) - q(p^r + \varepsilon u(S^r)).$$

Combining the above three sets of equalities and inequalities we obtain that

$$q(p^r) - q(p^{r+1}) \geq \sum_{j \in [S', N \setminus S']} [[X_j^r t_j' - f_j(X_j^r)] - [(t_j' - \varepsilon)\Psi_j^r - f_j(\Psi_j^r)]] \\ + \sum_{j \in [N \setminus S', S']} [[X_j^r t_j' - f_j(X_j^r)] - [(t_j' + \varepsilon)\Psi_j^r - f_j(\Psi_j^r)]] \\ = \sum_{\substack{j \in [S', N \setminus S'] \text{ or} \\ j \in [N \setminus S', S']}} [f_j(\Psi_j^r) - f_j(X_j^r) - (\Psi_j^r - X_j^r)t_j'] \\ + \varepsilon \left[\sum_{j \in [S', N \setminus S']} \Psi_j^r - \sum_{j \in [N \setminus S', S']} \Psi_j^r \right].$$

Since

$$\begin{aligned} \sum_{j \in [S^r, N \setminus S^r]} \Psi_j^r - \sum_{j \in [N \setminus S^r, S^r]} \Psi_j^r &= \sum_{j \in [S^r, N \setminus S^r]} g_j^+(t_j^r - \varepsilon) - \sum_{j \in [N \setminus S^r, S^r]} g_j^-(t_j^r + \varepsilon) \\ &\geq \sum_{j \in [S^r, N \setminus S^r]} g_j^-(t_j^r - \varepsilon) - \sum_{j \in [N \setminus S^r, S^r]} g_j^+(t_j^r + \varepsilon) \\ &= -q'(p^r + \varepsilon u(S^r); u(S^r)) > 0 \end{aligned}$$

(where the last strict inequality is obtained from Proposition 3.1) the left side of (32) follows. The right side of (32) follows from the convexity of f_j . QED

PROPOSITION 3.3. *Under Assumption D the number of dual descent steps is finite.*

Proof. We will argue by contradiction. Suppose that the number of dual descent steps is infinite. We denote the price vector, the tension vector and the flow vector generated by the relaxation algorithm at the r th dual descent step by p^r , t^r , and x^r , respectively. First we show the following property of the sequence $\{t^r\}$:

For each j

$$(33) \quad \{t_j^r\}_R \rightarrow +\infty \text{ for some subsequence } R \Rightarrow c_j < +\infty, \quad f_j(c_j) < \infty,$$

$$(34) \quad \{t_j^r\}_R \rightarrow -\infty \text{ for some subsequence } R \Rightarrow l_j > -\infty, \quad f_j(l_j) > -\infty.$$

If $\{t^r\}$ is bounded then (33), (34) trivially hold. Consider a subsequence R such that $\{t^r\}_R$ is unbounded. Without loss of generality suppose that, for each arc $j \in A$, $\{t_j^r\}$ is either bounded, or tends to ∞ , or tends to $-\infty$. We now partition N into a collection of nonempty subsets N_0, N_1, \dots, N_L ($L \geq 1$) such that

$$\{(p_i^r - p_k^r)\} \rightarrow \infty \quad \text{if } \alpha > \beta \text{ and } i \in N_\alpha, k \in N_\beta.$$

(One way to construct such a collection is to consider a graph identical to the original except that all arcs j such that $\{t_j^r\}_R$ is bounded are discarded, and all arcs j such that $\{t_j^r\}_R \rightarrow -\infty$ are reversed in their orientation. Since the sum of tensions along a directed cycle is zero we see that this graph is acyclic. The set N_0 is the set of nodes of this acyclic graph having no outgoing arcs. The set N_1 is obtained similarly after all arcs incident to N_0 in the acyclic graph have been discarded, etc.)

For $a = 1, 2, \dots, L$, we define the following arc sets:

$$\begin{aligned} A_a^+ &= \left\{ j \sim (i, k) \mid i \in \bigcup_{\tau \geq a} N_\tau, k \in \bigcup_{\tau < a} N_\tau \right\}, \\ A_a^- &= \left\{ j \sim (i, k) \mid i \in \bigcup_{\tau < a} N_\tau, k \in \bigcup_{\tau \geq a} N_\tau \right\}. \end{aligned}$$

Then each set $A_a^+ \cup A_a^-$ is a cut in the network and

$$(35) \quad \begin{aligned} t_j^r &\rightarrow +\infty, r \in R \quad \text{if and only if } j \text{ belongs to some } A_a^+, \\ t_j^r &\rightarrow -\infty, r \in R \quad \text{if and only if } j \text{ belongs to some } A_a^-. \end{aligned}$$

Consider any fixed positive scalar Δ . Equation (35) implies that for all a

$$(36) \quad \lim_{r \rightarrow \infty, r \in R} g_j^-(t_j^r - \Delta) = c_j \quad \forall j \in A_a^+, \quad \lim_{r \rightarrow \infty, r \in R} g_j^+(t_j^r + \Delta) = l_j \quad \forall j \in A_a^-.$$

Since

$$q'(p^r + \Delta u; u) = - \sum_{j \in A_a^+} g_j^-(t_j^r - \Delta) + \sum_{j \in A_a^-} g_j^+(t_j^r + \Delta)$$

where u is given by

$$u_i = \begin{cases} -1 & \text{if } i \in \bigcup_{\tau \cong a} N_\tau, \\ 0 & \text{otherwise,} \end{cases}$$

it follows from (36) that

$$(37) \quad \lim_{r \rightarrow \infty, r \in R} q'(p^r + \Delta u; u) = - \sum_{j \in A_a^+} c_j + \sum_{j \in A_a^-} l_j.$$

Let Θ denote the right-hand side quantity in (37). We will argue that $\Theta = 0$. Clearly we cannot have $\Theta > 0$ since this would imply that there does not exist a primal feasible solution. We also cannot have $\Theta < 0$ since then (37) implies that for r sufficiently large

$$q(p^r + \Delta u) \leq q(p^r) + \Delta \Theta.$$

This is not possible since Δ can be chosen arbitrarily large while $q(p^r)$ is nonincreasing with r . This leaves the only possibility that $\Theta = 0$ or that

$$\sum_{j \in A_a^+} c_j = \sum_{j \in A_a^-} l_j, \quad a = 1, \dots, L.$$

It follows that for every feasible flow vector we have

$$x_j = c_j \quad \forall j \in A_a^+, \quad x_j = l_j \quad \forall j \in A_a^-, \quad a = 1, \dots, L.$$

This implies (33) and (34).

Now we will bound from below the amount of improvement in the dual cost per dual descent step by a positive constant. Proposition 3.1 assures us that at each dual descent step the step length is more than ε . Consider the interval $[\frac{1}{4}\varepsilon, \frac{3}{4}\varepsilon]$, which we denote by I . Also let u^r denote the dual descent direction at the r th dual descent step. We have that the dual cost is decreasing on the line segment connecting t^r and t^{r+1} . It follows from (33), (34) and Assumption D that there exists a subsequence R such that for r sufficiently large, $r \in R$, we have for all $\Delta \in I$

$$\begin{aligned} q'(p^r + \Delta u^r; u^r) &= \sum_{j \in J^+} c_j v_j^r + \sum_{j \in J^-} l_j v_j^r + \sum_{\substack{j \in J^0 \\ v_j^r > 0}} g_j^+(t_j^r + \Delta v_j^r) v_j^r \\ &\quad + \sum_{\substack{j \in J^0 \\ v_j^r < 0}} g_j^-(t_j^r + \Delta v_j^r) v_j^r < 0 \end{aligned}$$

where we define

$$\begin{aligned} J^+ &= \{j | \{t_j^r\}_{r \in R} \rightarrow \infty\}, & J^- &= \{j | \{t_j^r\}_{r \in R} \rightarrow -\infty\}, \\ J^0 &= \{j | \{t_j^r\}_{r \in R} \text{ is bounded} \} \end{aligned}$$

and $v^r = E^T u^r$. Consider a fixed $r \in R$. Define $\Theta: R \rightarrow R$ by

$$\Theta(\Delta) = q(p^r + \Delta u^r).$$

We consider two cases. In case (i) the right derivative of $\Theta(\Delta)$ assumes at most $2|A|$ distinct values in the interval I . In case (ii) the right derivative of $\Theta(\Delta)$ assumes more than $2|A|$ distinct values in the interval I . In case (i) $q(p^r + \Delta u^r)$ is linear for Δ in some subinterval I' of I of length at least $\varepsilon/4|A|$ and it follows that $q'(p^r + \Delta u^r; u^r)$ over I' is linear of the form

$$(38) \quad q'(p^r + \Delta u^r; u^r) = \sum_{j \in J^+} c_j v_j^r + \sum_{j \in J^-} l_j v_j^r + \sum_{j \in J^0} b_j v_j^r$$

where $v^r = E^T u^r$ and b_j denotes some breakpoint of f_j . This implies that, for each $j \in J^0$ such that $v_j^r \neq 0$, the dual functional $g_j(t_j^r + \Delta v_j^r)$ is linear with slope b_j for Δ in I^r . For each $j \in J^0$, $\{t_j^r\}_{r \in R}$ is bounded, and therefore the number of distinct linear pieces of g_j of length $\geq \varepsilon/4|A|$ encountered during the course of the algorithm is finite. This together with the fact that v^r is chosen from a finite set imply that $q'(p^r + \Delta u^r; u^r)$ (cf. (38)) can only assume one of a finite set of values over the subinterval I^r . It follows that in case (i) we can bound the amount of dual cost improvement from below by $\delta\varepsilon/4|A|$ where δ is some positive scalar. This implies that case (i) can occur for only a finite set of indexes r (for otherwise the dual cost tends to $-\infty$) and we need only to consider case (ii). In case (ii) for each $r \in R$ there must exist a $j \in J^0$ such that $v_j^r \neq 0$ and the right derivative of the function $h(\Delta)$ defined by $h(\Delta) = g_j(t_j^r + \Delta v_j^r)$ assumes at least three distinct values in the interval I . Since v_j^r equals either 1 or -1 it follows that either $t_j^{r+1} \geq t_j^r + \varepsilon$ and $g_j^+(t_j^r + \Delta_1) < g_j^-(t_j^r + \Delta_2)$ for at least two points $\Delta_1 < \Delta_2$ in I or $t_j^{r+1} \leq t_j^r - \varepsilon$ and $g_j^+(t_j^r - \Delta_2) < g_j^-(t_j^r - \Delta_1)$ for at least two points $\Delta_1 < \Delta_2$ in I . Passing to a subsequence if necessary, we can assume that it is the same j and either $t_j^{r+1} \geq t_j^r + \varepsilon$ or $t_j^{r+1} \leq t_j^r - \varepsilon$ for all $r \in R$ that are sufficiently large. Without loss of generality we will assume that $t_j^{r+1} \geq t_j^r + \varepsilon$ for all $r \in R$ that are sufficiently large. Since $j \in J^0$ the subsequence $\{t_j^r\}_{r \in R}$ is bounded and therefore has a limit point t_j^* . Passing to a subsequence if necessary, we assume that $\{t_j^r\}$ converges to t_j^* . Then it follows that there exists a fixed interval L such that

$$(39) \quad L \subset [t_j^r, t_j^r + \varepsilon] \quad \forall r \in R, r \text{ sufficiently large}$$

and

$$\eta_1 < \eta_2 \quad \text{and} \quad g_j^+(\eta_1) < g_j^-(\eta_2)$$

for at least two distinct points η_1 and η_2 in L . We then define

$$\xi_1 = g_j^-(\eta_1), \quad \xi_2 = g_j^+(\eta_2).$$

Then ξ_1 and ξ_2 belong to the interval

$$[g_j^-(a), g_j^+(b)]$$

where a, b are the left and the right endpoints of L , respectively, and they satisfy

$$(40) \quad \xi_1 < \xi_2 \quad \text{and} \quad f_j^+(\xi_1) < f_j^-(\xi_2).$$

Then for r sufficiently large, $r \in R$, we obtain (cf. (39)) that

$$(41) \quad g_j^+(t_j^r) \leq \xi_1 < \xi_2 \leq g_j^-(t_j^r + \varepsilon).$$

It follows from Proposition 3.2 that for all sufficiently large $r \in R$

$$(42) \quad \begin{aligned} q(p^r) - q(p^{r+1}) &\geq f_j(g_j^-(t_j^r + \varepsilon)) - f_j(g_j^+(t_j^r)) - f_j^+(g_j^+(t_j^r))(g_j^-(t_j^r + \varepsilon) - g_j^+(t_j^r)) \\ &\geq f_j(\xi_2) - f_j(\xi_1) - f_j^+(\xi_1)(\xi_2 - \xi_1) \end{aligned}$$

where the second inequality follows from (41) and the convexity of f_j . From (40) and the convexity of f_j we obtain that the right-hand side of (42) is positive. Therefore the dual cost improvement per dual descent is bounded from below by a positive constant, and the dual cost tends to $-\infty$, contradicting Assumption A. QED

The second part of our finite termination proof involves showing that the number of flow augmentations between successive dual descent steps is finite. Since the ε -bounds remain unchanged between successive dual descent steps, the issue in effect is whether the labeling algorithm used will solve finitely the max flow problem with the given ε -bounds taken as capacity constraints. It was shown by Ford and Fulkerson

([14, p. 126]) that, when the data is irrational, an arbitrary choice of labeled nodes may result in an infinite number of flow augmentations, so a more specific scheme for labeling is necessary to deal with irrational data. Here we propose two such schemes: breadth-first search and arc discrimination. In practice, the data is always rational, being stored on a finite precision machine, and therefore finite convergence is assured even if labeling is done arbitrarily.

Breadth-first search is a well-known scheme used in labeling. It can be easily implemented using a FIFO queue. In [15, Chap. 9.3, 16] it was shown that, under breadth-first search, the number of flow augmentations is finite if *all* nodes with positive deficit are labeled initially. We now show that the same conclusion holds if a *single* node with positive deficit is labeled initially, as is the case for the relaxation iteration. This fact requires a nontrivial proof and, to our knowledge, is not reported in the literature.

PROPOSITION 3.4. *When labeling is done by breadth-first search, the number of flow augmentations between successive dual descent steps is finite.*

Proof. We will assume that the number of flow augmentations is infinite and obtain a contradiction. For simplicity, we call a node with negative deficit a *source* and a node with positive deficit a *sink*. Since the number of flow augmentations is infinite, after a while the set of sources and sinks must become fixed (since a source cannot become a sink or vice versa) and the set of flow augmenting paths must repeat (since all flow augmenting paths are simple and therefore there are only a finite number of them). Let P be the set of flow augmenting paths that repeat infinitely often. We say that an arc belonging to a path $p \in P$ is *saturated in the direction of p* if the flow of the arc is at the upper (lower) bound and the arc is oriented from the source (sink) of p to the sink (source) of p .

Consider a path $p \in P$. After a flow augmentation using p as the path, some arc of p will become saturated in the direction of p . Let A_p denote the set of arcs on p that become saturated in the direction of p infinitely often. A_p is clearly nonempty. We will show, by induction, that A_p is empty when breadth-first search is done and thus obtain a contradiction.

Initialization. For all $p \in P$, every $a \in A_p$ is at least one arc away from the sink t of p .

Proof. This is true since if the arc on p incident to t is saturated, it must remain saturated from then on.

kth inductive step. Suppose that, for all $p \in P$, every $a \in A_p$ is at least k arcs away from the sink of p . We will show that, for all $p \in P$, every $a \in A_p$ is $k+1$ arcs away from the sink of p . Suppose the contrary. Then there exists a $p \in P$, whose source and sink we denote by s and t respectively, and an arc a in A_p such that a is k arcs away from t . After a becomes saturated, there must be a flow augmenting path p' to unsaturate it (see Fig. 3.4). From the inductive hypothesis, the arcs on p between a and t are unsaturated in the direction of p . Since the labeling is done by breadth-first search,

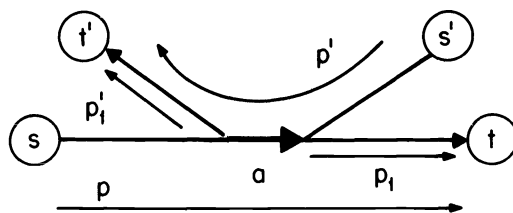


FIG. 3.4. p' unsaturating the arc a .

this implies that the number of arcs on the subpath p'_1 (see Fig. 3.4) must be *strictly* less than that of the subpath p_1 (otherwise during the iteration that generated p' as the flow augmenting path node t would have been labeled before node t'). It follows that, just before the iteration that generated path p , some arc of the subpath p'_1 must be saturated in the direction of p' (otherwise during the iteration that generated p node t' would be labeled before node t). This arc must then belong to $A_{p'}$. Since the number of arcs on p'_1 is strictly less than k , the inductive hypothesis is contradicted.

Since the inductive hypothesis holds for all k and the number of arcs on each flow augmenting path is at most $|N| - 1$, it follows that A_p is empty for all p and the desired contradiction is obtained. QED

In the arc discrimination scheme, the order in which nodes are labeled and scanned is given by the following simple rule:

Each labeled but unscanned node records whether it is connected to an unlabeled neighbor by an arc whose flow is strictly between the lower and upper bounds. A node with such a neighbor is scanned first.

The proof of finite convergence under this scheme is given in [17]. The implementation of the arc discrimination scheme requires more global information than breadth-first search. However, when the relaxation algorithm is extended to operate on both positive and negative deficit nodes between successive dual descent steps, which had been shown to be computationally beneficial in the case of linear cost problems, arc discrimination can still be shown to yield finite convergence. It is not known if this is also true of breadth-first search.

Propositions 3.3 and 3.4 show that the relaxation algorithm of this section terminates after a finite number of iterations. Since the algorithm only terminates when all the node deficits have zero value, the final flow vector x must belong to C . Since ε -CS is maintained at all iterations of the algorithm, it follows that x and the final dual price vector must satisfy ε -CS also.

We next show that we can bring the cost of the solution generated by the relaxation algorithm arbitrarily close to the optimal cost by taking ε sufficiently small. The main part of the argument is embodied in the next proposition.

PROPOSITION 3.5. *Let x and p satisfy ε -CS, and let ξ and p satisfy CS. If $x \in C$ then*

$$0 \leq f(x) + q(p) \leq \varepsilon \sum_{j \in A} |x_j - \xi_j|.$$

Proof. Let $t = E^T p$. Since ξ and p satisfy CS we have

$$f_j(\xi_j) = \xi_j t_j - g_j(t_j) \quad \forall j \in A.$$

Take an arc j such that $x_j \geq \xi_j$. Then by convexity of f_j

$$f_j(x_j) + (\xi_j - x_j)f_j^-(x_j) \leq f_j(\xi_j) = \xi_j t_j - g_j(t_j).$$

Hence

$$\begin{aligned} f_j(x_j) + g_j(t_j) &\leq (x_j - \xi_j)(f_j^-(x_j) - t_j) + x_j t_j \\ &\leq |x_j - \xi_j| \varepsilon + x_j t_j \end{aligned}$$

where the second inequality follows from ε -CS. This inequality is similarly obtained when $x_j \leq \xi_j$, so we have

$$f_j(x_j) + g_j(t_j) \leq |x_j - \xi_j| \varepsilon + x_j t_j \quad \forall j \in A.$$

From the definition of g_j we also have

$$x_j t_j \leq f_j(x_j) + g_j(t_j) \quad \forall j \in A.$$

By combining these two inequalities and adding over all $j \in A$, we obtain

$$\sum_{j \in A} x_j t_j \leq \sum_{j \in A} [f_j(x_j) + g_j(t_j)] \leq \varepsilon \sum_{j \in A} |x_j - \xi_j| + \sum_{j \in A} x_j t_j.$$

Since $x \in C$ we have $\sum_{j \in A} x_j t_j = 0$ and the result follows. QED

From Proposition 3.5 we can obtain a simple bound on the suboptimality of the solution in the special case where $l_j > -\infty$ and $c_j < +\infty$ for all $j \in A$.

COROLLARY 3.1. *Let x and p satisfy ε -CS. If $x \in C$, and $-\infty < l_j \leq c_j < +\infty$ for all $j \in A$, then*

$$0 \leq f(x) + q(p) \leq \varepsilon \sum_{j \in A} (c_j - l_j).$$

For the general case we have the following.

PROPOSITION 3.6. *Let $x(\varepsilon)$ and $p(\varepsilon)$ denote any flow and price vector pair such that $x(\varepsilon)$ and $p(\varepsilon)$ satisfy ε -CS and $x(\varepsilon) \in C$. Then $f(x(\varepsilon)) + q(p(\varepsilon)) \rightarrow 0$ as $\varepsilon \rightarrow 0$.*

Proof. First we show that $x(\varepsilon)$ remains bounded as $\varepsilon \rightarrow 0$. If $x(\varepsilon)$ is not bounded as $\varepsilon \rightarrow 0$, then since $x(\varepsilon) \in C$ for all $\varepsilon > 0$ there exists a directed cycle Y and a sequence $\{\varepsilon_n\} \rightarrow 0$ such that $c_j = +\infty$, $x_j(\varepsilon_n) \rightarrow +\infty$ for all $j \in Y^+$ and $l_j = -\infty$, $x_j(\varepsilon_n) \rightarrow -\infty$ for all $j \in Y^-$. By Assumption B

$$\lim_{\xi \rightarrow +\infty} f_j^-(\xi) = +\infty \quad \text{for all } j \in Y^+, \quad \lim_{\xi \rightarrow -\infty} f_j^+(\xi) = -\infty \quad \text{for all } j \in Y^-.$$

This implies that for n sufficiently large,

$$(43) \quad t_j(\varepsilon_n) > t_j(\varepsilon_0) \quad \text{for all } j \in Y^+ \quad \text{and} \quad t_j(\varepsilon_n) < t_j(\varepsilon_0) \quad \text{for all } j \in Y^-$$

where $t(\varepsilon_n) = E^T p(\varepsilon_n)$. Since $t(\varepsilon_n) = E^T p(\varepsilon_n)$ we have

$$\sum_{j \in Y^+} t_j(\varepsilon_n) - \sum_{j \in Y^-} t_j(\varepsilon_n) = 0 \quad \text{for all } n,$$

which contradicts (43). Therefore $x(\varepsilon)$ is bounded as $\varepsilon \rightarrow 0$.

Now we will show that $\xi_j(\varepsilon) - x_j(\varepsilon)$ is bounded for all $j \in A$ as $\varepsilon \rightarrow 0$, where $\xi(\varepsilon)$ is some vector satisfying $f_j^-(\xi_j(\varepsilon)) \leq t_j(\varepsilon) \leq f_j^+(\xi_j(\varepsilon))$, for all $j \in A$. If $c_j < \infty$ then $\xi_j(\varepsilon)$ is trivially bounded from above. If $c_j = +\infty$ then by Assumption B we have $f_j^-(\xi) \rightarrow +\infty$ as $\xi \rightarrow +\infty$. Since $x_j(\varepsilon)$ is bounded we have that $t_j(\varepsilon)$ is bounded from above which in turn implies that $\xi_j(\varepsilon)$ is bounded from above. Similarly, we can argue that $\xi_j(\varepsilon)$ is bounded from below. Therefore $|\xi_j(\varepsilon) - x_j(\varepsilon)|$ is bounded for all $j \in A$ as $\varepsilon \rightarrow 0$. This then completes our proof in view of Proposition 3.5. QED

Unfortunately Proposition 3.6 does not tell us how small ε must be to achieve a certain degree of near optimality. We need to solve the problem first for some guess ε to obtain $x(\varepsilon)$ and $\xi(\varepsilon)$, evaluate the quality of the solution on the basis of the gap $f(x(\varepsilon)) + q(p(\varepsilon))$ between primal and dual solution, and then decide whether ε needs to be decreased. If however the bounds l_j and c_j are finite, we can, by Corollary 3.1, obtain an a priori estimate on ε .

4. Computational experimentation. Two experimental codes implementing the methods of the paper were developed and tested on linear benchmark problems and nonlinear variations.

The first code, named NRELAX, implements the relaxation method for strictly convex problems of § 2. The second code, named MNRELAX, implements the method for mixed linear and strictly convex problems of § 3. Both codes were written in Fortran on a VAX 11-750 and were compiled and run under the VMS version 3.7 operating system.

TABLE 1

Times for NETGEN benchmark problems. MNRELAX uses $\varepsilon = 0$. Times in this and subsequent tables are in secs on a VAX 11-750. All codes are written in Fortran and compiled under VMS version 3.7.

Problem number	Number of nodes	Number of arcs	MNRELAX $\varepsilon = 0$	RELAX-II
1	200	1300	5.13	2.07
2	200	1500	6.33	2.12
3	200	2000	4.86	1.92
4	200	2200	7.74	2.52
5	200	2900	6.83	2.97
6	300	3150	12.85	4.37
7	300	4500	13.46	5.46
8	300	5155	14.54	5.39
9	300	6075	17.38	6.38
10	300	6300	14.39	4.12
11	400	1500	4.71	1.23
12	400	2250	5.81	1.38
13	400	3000	6.27	1.68
14	400	3750	7.79	2.43
15	400	4500	9.64	2.79
16	400	1306	9.06	2.79
17	400	2443	8.87	2.67
18	400	1306	8.98	2.56
19	400	2443	8.81	2.73
20	400	1416	9.82	2.85
21	400	2836	10.36	3.80
22	400	1416	9.08	2.56
23	400	2836	13.80	4.91
24	400	1382	4.73	1.27
25	400	2676	7.15	2.01
26	400	1382	3.73	1.79
27	400	2676	6.41	2.15
28	1000	2900	20.17	4.90
29	1000	3400	19.15	5.57
30	1000	4400	25.62	7.31

The test problems were generated using the public domain code NETGEN [18]. There are 40 "standard" benchmark linear cost problems that can be obtained using this code. We tested our codes with some of these problems either in their standard (linear cost) form or in a modified form whereby a quadratic cost was added to the linear cost of some or all of the arcs as discussed below. In order to test coding efficiency we tested MNRELAX with $\varepsilon = 0$ against the very efficient linear cost code RELAX-II (see [1], [2]) under identical conditions on the first 30 NETGEN benchmark problems. The two codes are close to being mathematically equivalent on linear cost problems but MNRELAX uses floating point arithmetic. The results shown in Table 1 appear to indicate that MNRELAX is coded fairly efficiently.

There were two issues that we wanted to clarify through the experimentation:

- (a) The effect of the parameter ε on the performance of MNRELAX;
- (b) The relative efficiency of NRELAX versus MNRELAX with optimal choice of the parameter ε on strictly convex problems.

A large number of experiments some of which are presented in Tables 2 and 3 showed that for all except some very "difficult" problems it is best to operate

TABLE 2

Times for NETGEN benchmark problems modified so that 50% of the arcs have an additional quadratic cost with coefficient from the range [5, 10]. Numbers in parentheses where present indicate significant digits of accuracy of the answer. In MNRELAX ε is kept constant during the solution of each problem.

Problem number	Number of nodes	Number of arcs	MNRELAX $\varepsilon > 0$	MNRELAX $\varepsilon = 0$	Sign. digits of accuracy
1	200	1300	30.79	11.49	6
2	200	1500	40.66	11.73	6
3	200	2000	34.48	9.31	7
4	200	2200	32.05	11.60	6
5	200	2900	50.43	28.14	5
6	300	3150	74.10	26.01	6
7	300	4500	140.70	48.64	6
8	300	5155	116.06	76.35	6
9	300	6075	96.59	49.25	6
10	300	6300	94.71	36.43	6
11	400	1500	263.14	26.35	4
12	400	2250	180.93	31.86	4
13	400	3000	240.76	(5)36.25	4
14	400	3750	436.80	79.88	4
15	400	4500	146.69	(3)42.23	2
16	400	1306	144.08	(7)86.40	3
17	400	2443	261.91	(7)47.31	5
18	400	1306	294.88	53.71	4
19	400	2443	108.04	37.54	5
20	400	1416	214.99	(7)68.17	3
21	400	2836	37.24	(7)18.43	4
22	400	1416	366.85	(7)53.37	4
23	400	2836	34.56	(7)18.09	4
24	400	1382	66.58	(5)45.87	3
25	400	2676	167.53	(6)22.73	5

MNRELAX with $\varepsilon = 0$ and terminate the iterations when the deficit of all nodes becomes sufficiently close to zero. Indeed it appears that for such problems the time required for MNRELAX to terminate increases with ε . The reason is probably that with large ε the intervals defined by the ε -bounds become larger and, as a result, a large number of flow augmentations are needed before a descent direction can be found. Given that a large value of ε leads also to inaccurate solutions (see Proposition 3.5), it appears that for most problems the best way to operate MNRELAX is with $\varepsilon = 0$ or with ε very small.

When $\varepsilon = 0$ and all arc costs are strictly convex, MNRELAX and NRELAX are mathematically equivalent. However NRELAX is somewhat faster because of more efficient coding as shown in Table 3.

Finally in Table 4 we show results obtained on some "difficult" problems with strictly convex arc costs. These problems were constructed by choosing the quadratic cost coefficients of some arcs to be very small relative to others as described in Table 4. This is similar to a situation in nonlinear unconstrained minimization where the Hessian matrix of the cost function has some eigenvalues that are very small relative to other eigenvalues. For this class of problems MNRELAX with nonzero ε can outperform both NRELAX and MNRELAX with $\varepsilon = 0$. This is not surprising in view of the coordinate descent interpretation of NRELAX. The version of MNRELAX that we found most efficient for these problems is one whereby we start with a moderate value

TABLE 3

Times for NETGEN benchmark problems modified so that all arcs have an additional quadratic cost with coefficient from the range [5, 10]. Numbers in parentheses where present indicate significant digits of accuracy of the answer. In MNRELAX ϵ is kept constant during solution of each problem.

Problem number	Number of nodes	Number of arcs	MNRELAX $\epsilon > 0$	MNRELAX $\epsilon = 0$	NRELAX	Sign. digits of accuracy
1	200	1300	22.63	17.97	10.80	7
2	200	1500	23.37	19.12	11.51	7
3	200	2000	19.33	20.77	12.04	8
4	200	2200	37.87	25.38	17.22	7
5	200	2900	34.82	32.37	21.44	6
6	300	3150	113.75	57.95	40.23	7
7	300	4500	85.26	50.49	37.77	6
8	300	5155	95.11	70.08	49.38	8
9	300	6075	70.48	69.44	48.04	7
10	300	6300	(6)99.69	69.33	41.41	5
11	400	1500	(7)43.19	34.37	14.67	6
12	400	2250	(6)39.56	33.31	12.98	5
13	400	3000	(7)34.62	32.66	18.34	5
14	400	3750	(6)34.97	35.32	20.86	5
15	400	4500	64.90	42.53	24.95	5
16	400	1306	65.86	54.19	21.54	6
17	400	2443	60.62	46.20	21.89	6
18	400	1306	84.26	72.41	48.97	7
19	400	2443	60.56	46.18	20.80	6
20	400	1416	108.49	72.11	38.70	7
21	400	2836	62.78	38.79	38.69	7
22	400	1416	95.91	55.25	42.03	6
23	400	2836	43.41	33.21	20.70	6
24	400	1382	59.83	65.35	42.47	7
25	400	2676	53.57	42.06	37.88	7

of ϵ , operate MNRELAX to termination, then reduce ϵ by a factor of 10 and repeat the process up to the point where primal and dual values differ by a specified accuracy. Still, the proper starting value for ϵ was not easy to determine and it was necessary to do some initial experimentation with several of these difficult problems. The conclusion is that the methods of this paper may not be successful for such problems. We do not know, however, of a better alternative.

TABLE 4

Times for NETGEN benchmark problems modified so that all arcs have an additional quadratic term. In 50% of the arcs the quadratic cost coefficient was small as indicated. In the other 50% of the arcs the quadratic cost coefficient was from the range [5, 10]. Numbers in parentheses where present indicate significant digits of accuracy of the answer. In MNRELAX ϵ is progressively decreased during solution of each problem.

Problem number	Number of nodes	Number of arcs	Small quad coeff.	MNRELAX $\epsilon > 0$	NRELAX
1	200	1300	.0001	(4)52.98	(3)58.00
5	200	2900	.001	(5)227.93	(2)50.15
7	300	4500	.0001	(3)301.03	(2)131.10
11	400	1500	.0001	(4)111.36	(2)218.24
15	400	4500	.01	(3)361.71	—
16	400	1306	.001	(3)287.13	(3)1957.70
18	400	1306	.001	(3)188.78	(2)350.97
24	400	1382	.001	(4)417.83	(2)321.76
				(2)46.21	(2)134.51

REFERENCES

- [1] D. P. BERTSEKAS, *A unified framework for primal-dual methods in minimum cost network flow problems*, Math. Programming, 32 (1985), pp. 125–145.
- [2] D. P. BERTSEKAS AND P. TSENG, *Relaxation methods for minimum cost ordinary and generalized network flow problems*, LIDS Report P-1462, Massachusetts Inst. of Technology, May 1985.
- [3] R. T. ROCKAFELLAR, *Network Flows and Monotropic Programming*, Wiley-Interscience, New York, 1983.
- [4] ———, *Convex Analysis*, Princeton Univ. Press, Princeton, NJ, 1970.
- [5] D. P. BERTSEKAS AND D. EL BAZ, *Distributed asynchronous relaxation methods for convex network flow problems*, LIDS Report P-1417, Massachusetts Inst. of Technology, October 1984, this Journal, 25 (1987), pp. 74–85.
- [6] W. I. ZANGWILL, *Nonlinear Programming*, Prentice-Hall, Englewood Cliffs, NJ, 1969.
- [7] R. W. H. SARGENT AND D. J. SEBASTIAN, *On the convergence of sequential minimization algorithms*, J. Optim. Theory Appl., 12 (1973), pp. 567–575.
- [8] E. POLAK, *Computational Methods in Optimization: A Unified Approach*, Academic Press, New York, 1971.
- [9] D. G. LUENBERGER, *Linear and Nonlinear Programming*, Addison-Wesley, Reading, MA, 1984.
- [10] M. J. D. POWELL, *On search directions for minimization algorithms*, Math. Programming, 4 (1973), pp. 193–201.
- [11] J. S. PANG, *On the convergence of dual ascent methods for large-scale linearly constrained optimization problems*, Univ. of Texas at Dallas, unpublished manuscript, 1984.
- [12] R. W. COTTLE AND J. S. PANG, *On the convergence of a block successive over-relaxation method for a class of linear complementarity problems*, Math. Programming Stud., 17 (1982), pp. 126–138.
- [13] D. P. BERTSEKAS AND S. K. MITTER, *A descent numerical method for optimization problems with nondifferentiable cost functionals*, this Journal, 11 (1973), pp. 637–652.
- [14] L. R. FORD, JR. AND D. R. FULKERSON, *Flows in Networks*, Princeton Univ. Press, Princeton, NJ, 1962.
- [15] C. H. PAPADIMITRIOU AND K. STEIGLITZ, *Combinatorial Optimization: Algorithms and Complexity*, Prentice-Hall, Englewood Cliffs, NJ, 1982.
- [16] J. EDMONDS AND R. KARP, *Theoretical improvements in algorithmic efficiency for network flow problems*, J. Assoc. Comput. Mach., 19 (1972), pp. 248–264.
- [17] P. TSENG, *Relaxation methods for monotropic programming problems*, Ph.D. thesis, Operations Research Center, Massachusetts Inst. of Technology, 1986.
- [18] D. KLINGMAN, A. NAPIER AND J. STUTZ, NETGEN—*A program for generation large scale (un)capacitated assignment, transportation and minimum cost network problems*, Management Sci., 20 (1974), pp. 814–822.
- [19] S. ZENIOS AND J. M. MULVEY, *Simulating a distributed synchronous relaxation method for convex network problems*, Working Paper, Department of Civil Engineering, Princeton Univ., January 1985.
- [20] D. P. BERTSEKAS, *Distributed relaxation methods for linear network flow problems*, Proc. of 25th Conf. on Decision and Control, Athens, Greece, December 1986.
- [21] ———, *Distributed asynchronous relaxation methods for linear network flow problems*, LIDS Report P-1606, Massachusetts Inst. of Technology, Cambridge, MA, September 1986.

SOME PROPERTIES OF CONSTRAINED VISCOSITY SOLUTIONS OF HAMILTON-JACOBI-BELLMAN EQUATIONS*

PAOLA LORETI†

Abstract. We consider an optimal control problem with space constraints and we show some properties of the optimal cost function. Our main result is the equality between the value functions of four optimal control problems and the constrained viscosity solution of the Hamilton-Jacobi-Bellman equation.

Key words. optimal control, Hamilton-Jacobi equation, relaxed control, viscosity solution, space constraints

AMS(MOS) subject classification. 49C20

1. Introduction. In this paper we deal with an optimal control problem with state constraints: our goal is to show how to use the existing knowledge on solution of Hamilton-Jacobi-Bellman equations in order to prove various properties of the optimal cost function.

We want to explain briefly what we mean by optimal control problems with state constraints. Let Ω be a smooth open set of \mathbb{R}^n . We consider the optimal control of the solution of an ordinary differential equation (the state process) where the control acts on the coefficients of the equation. The problem is to minimize a certain cost function over all controls such that the state process remains in $\bar{\Omega}$ (the closure of Ω) for all $t \geq 0$.

To explain our results, let us first recall a few facts. As is well known, a certain nonlinear partial equation called the Hamilton-Jacobi-Bellman equation is associated with general optimal control problems via the dynamic programming principle. This equation is of the form

$$(H-J) \quad u(x) + H(x, Du(x)) = 0 \quad \text{in } \Omega$$

where u is a scalar function on $\bar{\Omega}$; where Du is the gradient of u ; and H , called the Hamiltonian, is a continuous function on $\bar{\Omega} \times \mathbb{R}^n$. Our results will use the notion of the viscosity solution of (H-J) equations introduced in [3], [4]. In [6] it is proved that the usual dynamic programming argument shows that the optimal cost function (the value function) is a viscosity solution for the associated Hamilton-Jacobi-Bellman equation. On the other hand, the knowledge that there exists a unique viscosity solution of the Hamilton-Jacobi-Bellman equation, without knowing a priori whether the value function is continuous, implies that this solution is the value function [7], [8].

The optimal control problems we consider are different from those treated in [7], [8] by the state constraints we are imposing which should yield a particular boundary condition on $\partial\Omega$ for the value function. M. H. Soner [9] introduced a notion of constrained viscosity solution which combines the usual viscosity formulation in Ω and a related formulation on $\partial\Omega$.

We recall the definition in § 2. It is easy (see [9], [2]) to check that the value function is such a constrained viscosity solution if we know it is continuous on $\bar{\Omega}$. In addition uniqueness of constrained viscosity solutions was proved by Soner [9] and general existence results are obtain in [2] by PDE techniques.

Therefore, to characterize the value function as the unique constrained viscosity solution two strategies are possible. The first consists in proving that the value function

* Received by the editors May 19, 1986; accepted for publication (in revised form) September 3, 1986. This work was done while the author was visiting C  remade-Universit   Paris IX and was sponsored by Consiglio Nazionale Delle Ricerche under grant 203.01.36.

† Dipartimento di Matematica, Universit   di Rome, P. le Aldo Moro 5, 00100 Rome, Italy.

of an optimal control problem with state constraints is a continuous function. Then one concludes using the above results. This is what was proven by M. H. Soner [9]. In [9] it is proven that an appropriate geometric assumption implies that the value function is continuous on $\bar{\Omega}$. The second strategy uses the existence of a unique constrained viscosity solution of the Hamilton–Jacobi–Bellman associated with our control problem which is

$$(H-J-B) \quad u(x) + \max_{a \in A} \{-b(x, a) \cdot Du(x) - f(x, a)\} = 0.$$

Our goal is to show that this existence result (with some additional information on the approximation of the solution) enables us to check directly that the unique constrained viscosity solution is the value function. Even if one of the main results we show is contained in [9] our proof is entirely different and is probably more flexible, allowing us to solve similar problems like optimal control problems with space constraints and different boundary assumptions, and an optimal control problem with a final target. We will come back to these questions in future publications.

In addition, in the course of checking that the constrained viscosity solution is the value function we will show some other results of independent interest. More precisely, we show the equivalence of the value functions obtained by considering three other problems. Two of these, say (\hat{P}) , (\hat{P}_0) , are relaxed optimal control problems with state constraints, respectively, in $\bar{\Omega}$ or Ω . The third one is an optimal control problem with state constraints in Ω , say (P_0) , and we will denote by (P) our original problem. We call \hat{u} , \hat{u}_0 , u_0 , u the value functions of (\hat{P}) , (\hat{P}_0) , (P_0) , (P) , respectively, and \bar{u} the unique constrained viscosity solution of the (H–J–B) equation under the same geometric assumption as in Soner [9] on the controls at the boundary. Then, our main result is

$$(1.1) \quad \bar{u}(x) = \hat{u}(x) = \hat{u}_0(x) = u_0(x) = u(x) \quad \forall x \in \bar{\Omega}.$$

Let us briefly sketch the proof of this result. Our method will rely on [2] where the theory of existence, uniqueness and approximation of constrained viscosity solutions of the (H–J) equation is developed. Our result is achieved in two steps. First, we consider a method to approximate the constrained viscosity solution, introduced in [2]. This method of penalty type consists in extending the problem in \mathbb{R}^n and in introducing a function that penalizes the distance to the set $\bar{\Omega}$. In [2] it is proved that the sequence obtained by this method, which is a sequence of viscosity solutions for an appropriate (H–J) equation in \mathbb{R}^n , converges to the constrained viscosity solution of the (H–J–B) equation. For this approximated problem, the relation between viscosity solutions and both standard and relaxed value functions are well known and we check that passing to the limit they converge to the relaxed value functions \hat{u} .

The second step consists in showing just that

$$(1.2) \quad \hat{u} = \hat{u}_0 \quad \forall x \in \Omega.$$

This is achieved by the use of another approximation method which consists in restricting the (H–J–B) equation in certain appropriate subset Ω , say Ω_δ . We show that the sequence of constrained viscosity solution of the (H–J–B) on $\bar{\Omega}_\delta$ converges to $\hat{u}_0 \in C(\bar{\Omega})$, a viscosity solution of the (H–J–B) equation. Using the first step, we show (1.2). Then we show

$$(1.3) \quad \begin{aligned} \hat{u}_0(x) &= u_0(x) = u(x) & \forall x \in \Omega, \\ u_0(x) &= \bar{u}(x) & \forall x \in \partial\Omega \end{aligned}$$

by a very simple argument and it is easy to check (1.1).

We conclude this introduction with a summary of the paper: in § 2 we recall some basic definitions and notation, and we describe the optimal control problem with state constraints we are considering. In § 3, we present the main result of the paper, which is proved in §§ 4 and 5.

2. Statement of the problem and some known properties.

2.1. Statement of the problem. Let Ω be a smooth open set of \mathbb{R}^n , that we assume to be bounded, for simplicity. We denote by $\bar{\Omega}$, $\partial\Omega$, respectively, the closure and the boundary of Ω . Let A be a compact metric space. We consider functions $b \in C(\mathbb{R}^n \times A; \mathbb{R}^n)$, $f \in C(\mathbb{R}^n \times A; \mathbb{R})$ satisfying

$$(2.1) \quad |b(x, a) - b(x', a)| \leq C|x - x'|, \quad |b(x, a)| \leq C,$$

$$(2.2) \quad |f(x, a) - f(x', a)| \leq w(|x - x'|), \quad |f(x, a)| \leq C,$$

for all $(x, x', a) \in \mathbb{R}^n \times \mathbb{R}^n \times A$ for some constant $C > 0$ and where $w(t) \rightarrow 0$, $t \rightarrow 0^+$.

We introduce the controlled process

$$(2.3) \quad \dot{x}_t = b(x_t, a_t), \quad x_0 = x \in \bar{\Omega}$$

where the function a_t is the control (process) that we choose as follows:

$$(2.4) \quad \mathcal{A}_x = \{a \in L^\infty(\mathbb{R}^+; A) / \forall t \in \mathbb{R}^+, x_t \in \bar{\Omega}\}.$$

In other words we consider controls $a \in L^\infty(\mathbb{R}^+; A)$ such that the corresponding state of the system (state process) remains in $\bar{\Omega}$ for all $t \geq 0$.

Finally, we introduce the cost function

$$(2.5) \quad J(x, a) = \int_0^{+\infty} f(x_t, a_t) e^{-t} dt.$$

We want to minimize the cost function over all controls $a \in \mathcal{A}_x$, i.e., we want to determine the value function

$$(2.6) \quad u(x) = \inf_{a \in \mathcal{A}_x} J(x, a).$$

Here, in order to have a meaningful infimum, we assume that $\mathcal{A}_x \neq \emptyset \forall x \in \bar{\Omega}$ and this is the case in particular if we assume (see [9] where such an assumption was introduced):

$$(2.7) \quad \exists \nu > 0 / \forall x \in \partial\Omega, \quad \exists a \in A / b(x, a) \cdot n(x) \leq -\nu < 0.$$

We assume throughout the paper (2.1), (2.2) and we will not recall these assumptions.

2.2. Viscosity solutions. Let $u: \bar{\Omega} \rightarrow \mathbb{R}$ and $H: \bar{\Omega} \times \mathbb{R}^n \rightarrow \mathbb{R}$ be continuous functions.

DEFINITION 2.1. u is a viscosity subsolution (resp. supersolution) on Ω (or $\bar{\Omega}$) of the Hamilton–Jacobi equation

$$(2.8) \quad u(x) + H(x, Du(x)) = 0 \quad \text{in } \Omega$$

if, for all $\varphi \in C^1(\bar{\Omega})$ such that $u - \varphi$ has a local maximum (resp. minimum) in $x \in \Omega$ (or $\bar{\Omega}$) we have

$$u(x) + H(x, D\varphi(x)) \leq 0 \quad (\text{resp. } \geq 0).$$

Finally, u is a constrained viscosity solution of the equation (2.8) if it is a viscosity subsolution on Ω and a viscosity supersolution on $\bar{\Omega}$ of (2.8).

(See [3], [4] and [9] for the definition of viscosity solution and constrained viscosity solution, respectively.)

2.3. Known results. We consider the equation (H-J-B)

$$(2.9) \quad u(x) + \max_{a \in A} \{-b(x, a) \cdot Du(x) - f(x, a)\} = 0 \quad \text{in } \Omega.$$

By [2] it is known that under the assumptions (2.1), (2.2) and (2.7) that (2.9) admits a unique constrained viscosity solution \bar{u} in $C(\bar{\Omega})$. The uniqueness part is due to Soner and it is slightly improved in [2].

Furthermore, such a solution can be obtained as the limit of viscosity solutions of various approximated problems (see § 4). In [2] it is proved that $u \in C^{0,a}(\bar{\Omega})$ for some $a > 0$ but we will not use this result here.

3. The main result.

3.1. Relaxed controls. Let $M(A)$ be the dual space of $C(A)$ endowed with the weak-star topology of $C(A)^*$, i.e., the space of bounded measures on A .

We want to define the class of admissible relaxed controls for our problem. To this end we introduce the class

$$\hat{\mathcal{A}} = \{\hat{a} \in L^\infty(\mathbb{R}^+, M(A)) / \hat{a}_s \text{ is a probability measure for almost every } s \in \mathbb{R}^+\}.$$

We consider the state of the system corresponding to the control $\hat{a}_t \in \hat{\mathcal{A}}$

$$(3.1) \quad x(t) = x + \int_0^t \int_A b(x_s, a) d\hat{a}_s(a) ds \quad \forall t \in \mathbb{R}^+, \quad \forall x \in \mathbb{R}^n.$$

We say that \hat{a} is admissible for the point $x \in \bar{\Omega}$ if $\hat{a} \in \hat{\mathcal{A}}$ and \hat{a} constrains the relaxed state to lie in $\bar{\Omega}$ and we write $\hat{a} \in \hat{\mathcal{A}}_x$. In other words, we have for all $t \geq 0$

$$(3.2) \quad \hat{\mathcal{A}}_x = \{\hat{a}_t \in \hat{\mathcal{A}} / x_t \in \bar{\Omega}, \forall t \in \mathbb{R}^+\}.$$

3.2. Four different problems.

3.2.1. We begin with relaxed problems. We introduce the relaxation of the problem (2.6), i.e., we consider

$$(3.3) \quad \hat{u}(x) = \inf_{\hat{a} \in \hat{\mathcal{A}}_x} J(x, \hat{a})$$

where $x \in \bar{\Omega}$, $\hat{\mathcal{A}}_x$ is given by (4.2) and $J(x, \hat{a})$ is given by

$$(3.4) \quad J(x, \hat{a}) = \int_0^{+\infty} \int_A f(x_t, a) e^{-t} d\hat{a}_t(a) dt.$$

3.2.2. We also introduce the class of controls

$$(3.5) \quad \hat{\mathcal{A}}_x^0 = \{\hat{a} \in \hat{\mathcal{A}} / x_t \in \Omega, \forall t > 0\},$$

and we consider the associated value function

$$(3.6) \quad \hat{u}_0(x) = \inf_{\hat{a} \in \hat{\mathcal{A}}_x^0} J(x, \hat{a})$$

where $x \in \bar{\Omega}$ and $\hat{\mathcal{A}}_x^0$ is given by (3.5).

3.2.3. We finally consider the class of controls given by

$$(3.7) \quad \mathcal{A}_x^0 = \{a \in A / x_t \in \Omega, \forall t > 0\}.$$

We want to determine

$$(3.8) \quad u_0(x) = \inf_{a \in \mathcal{A}_x^0} J(x, a)$$

where $x \in \bar{\Omega}$, \mathcal{A}_x^0 , $J(x, a)$ are given, respectively, by (3.7) and (2.5).

3.2.4. Finally, let us recall that our original corresponds to the following class of problems:

$$\mathcal{A}_x = \{a \in \mathcal{A} / x_t \in \bar{\Omega}, \forall t > 0\}$$

and the value function is given by

$$u(x) = \inf_{a \in \mathcal{A}_x} J(x, a).$$

3.3. The main result.

THEOREM 3.1. *We assume (2.7). Then all the preceding value functions coincide with the unique constrained viscosity solution \bar{u} of (2.9), i.e., we have*

$$\hat{u} = \hat{u}_0 = u_0 = u = \bar{u} \quad \text{on } \bar{\Omega}.$$

The proof is given in §§ 4 and 5. In § 4 we prove that $\bar{u} = \hat{u}$ for all $x \in \bar{\Omega}$ and in § 5 we show the other equalities.

Remark. For our problem, by the definition of constrained viscosity solution, the result above implies the continuity of the value function.

4. Constrained viscosity solutions and the representation of the solution.

4.1. Representation of the solution.

PROPOSITION 4.1. *The unique constrained solution of the equation (2.8) is the relaxed value function, i.e.,*

$$\bar{u}(x) = \hat{u}(x) \quad \forall x \in \bar{\Omega}$$

where \hat{u} is given by (3.3).

Proof. Following [2], we consider the approximated problem

$$(4.1) \quad u_\varepsilon(x) + \max_{a \in A} \{-b(x, a) \cdot Du_\varepsilon - f(x, a)\} = \frac{p(x)}{\varepsilon} \quad \text{in } \mathbb{R}^n.$$

In (4.1) p is a function in $BUC(\mathbb{R}^n)$ such that

$$(4.2) \quad \forall \varepsilon > 0, \delta > 0, p(x) \geq \delta \text{ if } d(x, \bar{\Omega}) \geq \varepsilon, p(x) = 0 \text{ in } \bar{\Omega}.$$

By [6] there exists, for every $\varepsilon > 0$, a unique viscosity solution u_ε in $BUC(\mathbb{R}^n)$ of (4.1). By [2] the sequence u_ε , which is nondecreasing with respect to $\varepsilon > 0$, converges uniformly on $\bar{\Omega}$ to the constrained viscosity solution of (2.8).

On the other hand (see [6]) u^ε is the following value function:

$$(4.3) \quad u_\varepsilon(x) = \inf_{a \in \mathcal{A}} \int_0^{+\infty} \left[f(x_t, a_t) + \frac{p(x_t)}{\varepsilon} \right] e^{-t} dt.$$

We write

$$(4.4) \quad J^\varepsilon(x, a) = \int_0^{+\infty} \left[f(x_t, a_t) + \frac{p(x_t)}{\varepsilon} \right] e^{-t} dt.$$

But, in this case, from [1] and [10] we deduce

$$(4.5) \quad u_\varepsilon(x) = \inf_{\hat{a} \in \hat{\mathcal{A}}} \int_0^{+\infty} \int_A \left[f(x_t, a) + \frac{p(x_t)}{\varepsilon} \right] e^{-t} d\hat{a}_t dt$$

denoting by

$$(4.6) \quad J^\varepsilon(x, \hat{a}) = \int_0^{+\infty} \int_A \left[f(x_t, a) + \frac{p(x_t)}{\varepsilon} \right] e^{-t} d\hat{a}_t(a) dt.$$

Indeed, we observe that $\hat{\mathcal{A}}_x \subseteq \hat{A}$ and thus

$$(4.7) \quad u_\varepsilon(x) \leq \inf_{\hat{a} \in \hat{\mathcal{A}}_x} J(x, \hat{a}).$$

Passing to the limit, we find

$$\bar{u}(x) \leq \hat{u}(x) \quad \forall x \in \bar{\Omega}.$$

We want to show the converse. We consider an optimal relaxed control \hat{a}_t^ε ; without loss of generality we may assume that \hat{a}_t^ε (considering a subsequence if necessary) converges weakly star in $L^\infty([0, T]; M(A))$ to $\hat{a}_t \in \hat{A}$ as ε goes to 0. This is a consequence of the fact that the convex set

$$\hat{\mathcal{A}}_T = \{\hat{a}|_{[0, T]} / \hat{a} \in \hat{\mathcal{A}}\}$$

is sequentially compact in $L^\infty([0, T], M(A))$ endowed with the weak-star topology [1].

Next, we observe that $\{x_t^\varepsilon\}$ is a uniform and equicontinuous class of functions on $[0, T]$ for each $T > 0$. By the Ascoli-Arzelà Theorem, choosing a subsequence if necessary, we have

$$(4.8) \quad x_t^\varepsilon \xrightarrow{\varepsilon \rightarrow 0} x_t \quad \text{uniformly on } [0, T] \quad \forall T < +\infty.$$

Thus, the weak-star convergence of the sequence \hat{a}_t^ε implies

$$(4.9) \quad \int_0^{+\infty} \int_A (f(x_t^\varepsilon, a) d\hat{a}_t^\varepsilon(a)) e^{-t} dt \xrightarrow{\varepsilon \rightarrow 0} J(x, \hat{a}).$$

If $\hat{a}_t \in \hat{\mathcal{A}}_x$ then

$$(4.10) \quad u_\varepsilon(x) = J^\varepsilon(x, \hat{a}_t^\varepsilon) \geq J(x, \hat{a}) - \varepsilon \geq \inf_{\hat{a} \in \hat{\mathcal{A}}_x} J(x, \hat{a}) - \varepsilon$$

and we conclude the proof letting ε go to 0.

To conclude, we just have to show that $\hat{a}_t \in \hat{\mathcal{A}}_x$. Now, since $u^\varepsilon(x) \leq c$ for all $x \in \mathbb{R}^n$, by the definition (3.3),

$$\int_0^{+\infty} \int_A f(x_t^\varepsilon, a) d\hat{a}_t^\varepsilon(a) e^{-t} dt + \frac{1}{\varepsilon} \int_0^{+\infty} p(x_t^\varepsilon) e^{-t} dt \leq C.$$

By (1.2),

$$\int_0^{+\infty} \int_A f(x_t^\varepsilon, a) d\hat{a}_t^\varepsilon(a) e^{-t} dt \geq -C.$$

Thus

$$\int_0^{+\infty} \frac{p(x_t^\varepsilon)}{\varepsilon} e^{-t} dt \leq 2C$$

and passing to the limit for all $T < +\infty$

$$\int_0^T p(x_t) dt = 0.$$

Therefore $x(t) \in \bar{\Omega}$, for all $t \leq T$, for all $T < +\infty$ and we conclude the proof since $\hat{a}_t \in \hat{A}_x$. \square

5. Properties of the value functions.

5.1. Internal approximation.

LEMMA 5.1. For all $x \in \Omega$, $\hat{u}(x) = \hat{u}_0(x)$.

Proof. Let $\Omega_\delta = \{x \in \Omega / d(x, \bar{\Omega}) \leq \delta\}$ be with $\delta > 0$. We notice that, for δ small enough, the following assumption holds:

$$(5.1) \quad \exists \nu < 0, \quad \forall x \in \partial\Omega_\delta, \quad \exists a \in A \quad b(x, a) \cdot n(x) \leq -\nu < 0$$

where ν does not depend on δ . In fact if (2.7) holds we may consider $y \in \partial\Omega_\delta$ with $y = x - \delta n(x)$; then

$$b(y, a) \cdot n(y) \leq -\frac{\nu}{2} < 0$$

for δ small enough.

By (5.1) and [2] there exists a unique constrained viscosity solution $u_\delta \in C(\bar{\Omega}_\delta)$ of the equation

$$(5.2) \quad u_\delta(x) + \max_{a \in A} \{-b(x, a) \cdot Du_\delta(x) - f(x, a)\} = 0 \quad \text{in } \Omega_\delta.$$

By Proposition 4.1

$$(5.3) \quad u_\delta(x) = \inf_{\hat{a}_t \in \hat{\mathcal{A}}_x^\delta} J(x, \hat{a}) \quad \forall x \in \Omega_\delta$$

where

$$(5.4) \quad \hat{\mathcal{A}}_x^\delta = \{\hat{a} \in \hat{\mathcal{A}} / x_t \in \bar{\Omega}_\delta, \forall t \in \mathbb{R}^+\}.$$

Obviously, (5.3) yields

$$(5.5) \quad |u_\delta(x)| \leq C$$

and by [2] we know that the following a priori estimate holds:

$$(5.6) \quad |u_\delta(x) - u_\delta(y)| \leq \mu(d_\delta(x, y)) \quad \forall \delta > 0$$

where

$$d_\delta(x, y) = |x - y| + k|d_\delta(x) - d_\delta(y)|, \quad x, y \in \bar{\Omega}_\delta$$

and

$$d_\delta(x) = \text{dist}(x, \partial\Omega_\delta)$$

where μ is uniformly continuous modulus independent by δ . In fact we observe, using (5.1) and the proofs on [2], that k does not depend on δ . Therefore

$$(5.7) \quad |u_\delta(x) - u_\delta(y)| \leq \mu(|x - y|).$$

Thus, by the Ascoli-Arzelà Theorem, there exists $\underline{u} \in C(\bar{\Omega})$ and

$$(5.8) \quad \sup_{\bar{\Omega}_\delta} |u_\delta - \underline{u}| \xrightarrow{\delta \rightarrow 0} 0.$$

And, by standard stability results on viscosity solutions [6], $\underline{u} \in C(\bar{\Omega})$ is the viscosity solution of the problem (2.8).

By their definitions we observe that the following relations between the various classes of controls we consider hold:

$$(5.9) \quad \hat{\mathcal{A}}_x \supseteq \hat{\mathcal{A}}_x^0 \supseteq \mathcal{A}_x^\delta.$$

Hence, by (2.2), (3.3) and (3.6) we obtain

$$(5.10) \quad \hat{u} \leq \hat{u}_0 \leq u_\delta \quad \forall x \in \bar{\Omega}_\delta$$

and letting δ go to 0, we find

$$(5.11) \quad \hat{u} \leq \hat{u}_0 \leq \underline{u}.$$

On the other hand, $\hat{u}(x)$ is the maximal viscosity subsolution of (2.8) (see [1], [2]) and thus

$$(5.12) \quad \underline{u} \leq \hat{u} \quad \forall x \in \bar{\Omega}$$

and because of (5.10) and (5.12), the lemma is proved.

5.2. Relations with the function u_0 .

LEMMA 5.2. For all $x \in \Omega$, $\hat{u}_0(x) = u_0(x)$.

Proof. Let \hat{a}_t be such that $x_t \in \Omega$ and

$$(5.13) \quad \hat{u}_0(x) \geq J(x, \hat{a}_t) - \varepsilon.$$

By the Lyapunov Theorem [10] there exists $a_t^n \in \mathcal{A}$

$$(5.14) \quad a_t^n \xrightarrow{n \rightarrow +\infty} \hat{a}_t \quad \text{weak star in } L^\infty([0, T], M(A)).$$

Therefore, for all $T < +\infty$ and for all $t \in [0, T]$

$$(5.15) \quad x_t^n \rightarrow x_t \quad \text{uniformly on compact subsets of } \Omega.$$

Hence, for all $T < +\infty$, $x_t^n \in \Omega$, for all $t \in [0, T]$, for n large enough. Then, we consider the sequence

$$(5.16) \quad \bar{a}_t^n = a_t^n 1_{[0, T]} + \tilde{a}_t^n 1_{[T, +\infty]}$$

where \tilde{a}_{t-T}^n is any admissible control for the point x_T^n . ($\tilde{a}_{t-T}^n \in \mathcal{A}_{x_T^n}$) where x_T^n is the solution at time T of (2.3) corresponding to a_t^n .

We observe that \bar{a}_t^n is any admissible sequence for $x \in \Omega$. Hence, we have

$$(5.17) \quad J(x, \bar{a}_t^n) \geq \inf_{a_t \in \mathcal{A}_x^0} J(x, a_t)$$

and we pass to the limit as n, T go to ∞

$$(5.18) \quad \hat{u}_0(x) + \varepsilon \geq J(x, \hat{a}_t) \geq \inf_{a_t \in \mathcal{A}_x^0} J(x, a_t).$$

Letting ε go to 0, we deduce

$$(5.19) \quad \hat{u}_0 \geq u_0 \quad \forall x \in \Omega.$$

We just have to prove the converse. This is an easy consequence of the classical fact that every classical control can be written as a relaxed control with Dirac measure. Hence

$$(5.20) \quad u_0 \leq \hat{u}_0 \quad \forall x \in \bar{\Omega}$$

and we conclude.

5.3. Conclusion.

LEMMA 5.3. For all $x \in \bar{\Omega}$, $u_0(x) = \bar{u}(x)$.

Proof. We just have to prove the lemma for $x \in \partial\Omega$. By standard dynamic programming

$$u_0(x) = \inf_{a_t \in \mathcal{A}_x^0} \int_0^h f(x_s, a_s) e^{-s} ds + u_0(x_h) e^{-h}$$

and since $x_h \in \Omega$, $u_0 = \bar{u}$ on Ω , we deduce that

$$u_0(x) = \inf_{a_t \in \mathcal{A}_x^0} \int_0^h f(x_s, a_s) e^{-s} ds + \bar{u}(x_h) e^{-h}.$$

Next, we pass to limit to the limit as h goes to 0 and we obtain

$$(5.21) \quad u_0(x) = \bar{u}(x) \quad \forall x \in \bar{\Omega}.$$

We conclude the proof of Theorem 3.1. Indeed we observe that

$$(5.22) \quad \mathcal{A}_x^0 \subseteq \mathcal{A}_x \subseteq \hat{\mathcal{A}}_x,$$

$$(5.23) \quad \mathcal{A}_x^0 \subseteq \hat{\mathcal{A}}_x^0 \subseteq \hat{\mathcal{A}}_x.$$

Hence we have

$$(5.24) \quad u_0 \geq u \geq \hat{u} \quad \text{in } \bar{\Omega},$$

$$(5.25) \quad u_0 \geq \hat{u}_0 \geq \hat{u} \quad \text{in } \bar{\Omega}.$$

By Lemma 5.3 we have

$$(5.26) \quad u_0 = \bar{u} \quad \text{in } \bar{\Omega}.$$

By Proposition 4.1 we have

$$(5.27) \quad \bar{u} = \hat{u} \quad \text{in } \bar{\Omega}.$$

Then, by (5.24), (5.25), we obtain

$$u = u_0 = \hat{u} \quad \text{in } \bar{\Omega}, \quad u_0 = \hat{u}_0 = \hat{u} \quad \text{in } \bar{\Omega}.$$

Hence, we have

$$\bar{u} = \hat{u} = \hat{u}_0 = u_0 = u \quad \text{in } \bar{\Omega}$$

and we conclude. \square

Remark. The results utilized here can be generalized to a set Ω not bounded (see [2], [9]), so our results, which do not utilize the boundedness of Ω , hold also if Ω is not bounded.

Acknowledgment. The author wishes to thank Professor Pierre-Louis Lions for useful discussions and suggestions.

REFERENCES

- [1] I. CAPUZZO-DOLCETTA AND H. ISHII, *Approximate solutions of the Bellman equation of deterministic control theory*, Appl. Math. Optim., to appear.
- [2] I. CAPUZZO-DOLCETTA AND P. L. LIONS, *Hamilton-Jacobi equations and state-constraints problem*, to appear.
- [3] M. G. CRANDALL, L. C. EVANS AND P. L. LIONS, *Some properties of viscosity solutions of Hamilton-Jacobi equations*, Trans. Amer. Math. Soc., 282 (1984), pp. 487-502.
- [4] M. G. CRANDALL AND P. L. LIONS, *Viscosity solution of Hamilton-Jacobi equations*, Trans. Amer. Math. Soc., 277 (1983), pp. 1-42; announced in C.R. Acad. Sci. Paris, 292 (1981), pp. 183-186.
- [5] E. B. LEE AND L. MARKUS, *Foundations of Optimal Control Theory*, John Wiley, New York, 1967.
- [6] P. L. LIONS, *Generalized Solutions of Hamilton-Jacobi Equations*, Pitman, London, 1982.
- [7] ———, *Optimal control of diffusion processes and Hamilton-Jacobi-Bellman equations, Part 2*; Comm. Partial Differential Equations, 8 (1983), pp. 1229-1276.
- [8] P. L. LIONS AND P. E. SOUGANIDIS, *Differential games, optimal control and directional derivatives of viscosity solutions of Bellman's and Isaacs' equations*, this Journal, 23 (1985), pp. 566-583.
- [9] M. H. SONER, *Optimal Control Problem with State-Space Constraint*, to appear.
- [10] L. TARTAR, *Compensated compactness and applications to partial differential equations*, in Nonlinear Analysis and Mechanics, Heriot-Watt Symposium, Vol. IV, R. J. Knops, ed., Pitman, London, 1979.

REACHING ZERO RAPIDLY*

STEVEN OREY[†], VICTOR PESTIEN[‡] AND WILLIAM SUDDERTHS

Abstract. The controlled process $X(t)$ with values in $(-\infty, 0]$ is given by a stochastic differential equation:

$$dX(t) = \mu(t) dt + \sigma(t) dW_t, \quad X(0) = x$$

where the nonanticipative controls μ and σ are to be chosen so that $(\mu(t), \sigma(t))$ remains in a given set \mathcal{S} . The object is to maximize (minimize) the expectation of β^T where $0 < \beta < 1$ ($\beta > 1$) and T is the hitting time of zero. A complete solution is given for any \mathcal{S} , and an application is made to continuous-time red-and-black.

Key words. stochastic control, gambling theory

AMS(MOS) subject classifications. 60G40, 60J60, 93E20

Introduction. Consider a process $\{X_1(t)\}$ on $(-\infty, 0]$ given by a stochastic differential equation:

$$dX_1(t) = \mu(t) dt + \sigma(t) dW_t, \quad X_1(0) = x_1$$

where $\{W_t\}$ is standard Brownian motion and $\{\mu(t)\}$ and $\{\sigma(t)\}$ are nonanticipative controls to be chosen so that $(\mu(t), \sigma(t))$ remains in a specified set \mathcal{S} . Let T be the first time X_1 reaches the origin. In [2] the problems of minimizing or maximizing the expected value of T were solved for any \mathcal{S} ; the value function and optimal strategies were given explicitly in terms of \mathcal{S} . Minimizing the expected value of T is a natural criterion for getting to the origin rapidly. However, there are other criteria, two of which are considered here.

Our first problem is to maximize the expected value of β^T , where β is a positive constant less than one. The second problem is to minimize the expected value of β^T where now the constant β is chosen greater than one. In both cases we obtain a complete solution, for arbitrary \mathcal{S} .

The problems are of unequal difficulty. The first problem is actually quite easy. The reason the second problem is harder is that in the solution one must distinguish between the two cases $I < \infty$ and $I = \infty$, where the quantity I is defined by

$$I = \inf_{\varepsilon > 0} \sup \{ \mu + \varepsilon \sigma^2 : (\mu, \sigma) \in \mathcal{S} \}.$$

As is common in control theory problems, we obtain a natural candidate $Q(x_1)$ for a value function and try to prove it correct by applying an appropriate verification theorem. As it turns out, Q is indeed the correct value function if and only if $I < \infty$. Hence there must be something in the application of the verification theorem distinguishing between $I < \infty$ and $I = \infty$, and this indicates we should expect technical difficulties. The $I < \infty$, $I = \infty$ dichotomy also appeared in [2], but in the present work more delicate constructions for overcoming the obstacles are required.

* Received by the editors June 9, 1986; accepted for publication October 1, 1986.

[†] School of Mathematics, University of Minnesota, Minneapolis, Minnesota 55455. The work of this author was supported by National Science Foundation grant MCS83-01080.

[‡] Department of Mathematics and Computer Science, University of Miami, Coral Gables, Florida 33124.

[§] School of Statistics, University of Minnesota, Minneapolis, Minnesota 55455. The work of this author was supported by National Science Foundation grant DMS-8421208.

The solutions of the two general problems make it possible for us to solve the particular problems of discounted ($0 < \beta < 1$) and inflated ($\beta > 1$) red-and-black. It is shown here for the continuous-time problem, as it was by Klugman [3] in discrete time, that bold play is optimal for subfair, discounted red-and-black. The superfair case is also explicitly solved here for the continuous-time problem although it remains open in discrete time.

1. Preliminaries. We begin by explaining the continuous time gambling set-up of [4] and [2].

A *continuous-time gambling problem* is a triple (F, Σ, u) where

- (1.1) the *state space* F is a Borel subset of the Euclidean space \mathbb{R}^d having nonempty interior,
- (1.2) the *gambling house* Σ is a mapping which assigns to each $x \in F$ a nonempty collection $\Sigma(x)$ of processes $X = \{X_t, t \geq 0\}$ with state space F such that $X_0 = x$ and X has right-continuous paths with left-limits,
- (1.3) the *utility function* u is a Borel function from F to the real line.

A process $X \in \Sigma(x)$ is said to be *available* at x . Each available X is defined on some probability space (Ω, \mathcal{F}, P) and is adapted to an increasing filtration $(\mathcal{F}_t, t \geq 0)$ of complete subsigma fields of \mathcal{F} . The probability space and filtration may depend on X .

A player, starting at position $x \in F$, selects a process $X \in \Sigma(x)$ and receives payoff $u(X)$ defined by

$$(1.4) \quad u(X) = E[\limsup_{t \rightarrow \infty} u(X_t)].$$

The expectation occurring on the right is assumed to be well defined for every available process X .

The *value function* V is defined by

$$V(x) = \sup \{u(X) : X \in \Sigma(x)\}$$

for every $x \in F$. A process $X \in \Sigma(x)$ is *optimal* at x if

$$u(X) = V(x).$$

From now on, each process $X = \{X_t\}$ under consideration will be an *Ito process* of the form

$$(1.5) \quad X_t = x + \int_0^t \alpha(s) ds + \int_0^t \beta(s) dW_s,$$

where $W = \{W_t\}$ is a standard m -dimensional Brownian motion process on (Ω, \mathcal{F}, P) adapted to increasing, right-continuous σ -fields $\{\mathcal{F}_t\}$ and \mathcal{F}_t is independent of $\{W_{t+s} - W_t, s \geq 0\}$. The function $\alpha = \alpha(t, \omega)$ is to be \mathbb{R}^d -valued, progressively measurable, and such that

$$(1.6) \quad \int_0^t |\alpha(s)| ds < \infty \quad \text{a.s. for all } t.$$

The function $\beta = \beta(t, \omega)$ has as values real $d \times m$ matrices, is progressively measurable, and satisfies

$$(1.7) \quad \int_0^t |\beta(s)|^2 ds < \infty \quad \text{a.s. for all } t.$$

For each pair (a, b) , where $a \in \mathbb{R}^d$ is a $d \times 1$ vector and b is a $d \times m$ real-valued matrix, define the differential operator $D(a, b)$ for sufficiently smooth functions $Q: \mathbb{R}^d \rightarrow \mathbb{R}$ by the following:

$$D(a, b)Q(y) = Q_x(y)a + \frac{1}{2} \sum_{i=1}^d \sum_{j=1}^d Q_{x_i x_j}(y)(bb')_{ij}$$

where

$$Q_x(y) = \left(\frac{\partial Q}{\partial x_1}, \dots, \frac{\partial Q}{\partial x_d} \right), \quad Q_{x_i x_j} = \frac{\partial^2 Q}{\partial x_i \partial x_j},$$

and b' is the transpose of b .

We now specify $\Sigma(x)$ by specifying the possible values of α and β . To this end, let $C(x)$ be, for each $x \in F$, a nonempty set of pairs (a, b) , where $a \in \mathbb{R}^d$ and b is a real $d \times m$ matrix. (The idea is that $C(x)$ is the set from which a player at state x may choose the value of (α, β) .) Assume also that every available process X is absorbed at the time T_X of its first exit from F^0 , the interior of F . These conditions define a function Σ_C on F where $\Sigma_C(x)$ is the collection of all processes X having paths in F and satisfying (1.5), (1.6), and (1.7) together with

$$(1.8) \quad (\alpha(t, \omega), \beta(t, \omega)) \in C(X_t(\omega)) \quad \text{for all } (t, \omega),$$

$$(1.9) \quad (\alpha(t, \omega), \beta(t, \omega)) = (0, 0) \quad \text{for all } t \geq T_X(\omega),$$

$$(1.10) \quad C(x) = \{(0, 0)\} \quad \text{for } x \in F - F^0.$$

Let Σ be a gambling house such that $\Sigma(x) \subset \Sigma_C(x)$ for every $x \in F$. The following proposition is related to other verification lemmas in [2] and [4]. Since the hypotheses here differ slightly from those in [2] and [4], we provide a proof.

PROPOSITION 1.1. *Let G be an open subset of \mathbb{R}^d which contains F . Suppose $Q: G \rightarrow \mathbb{R}$ has continuous second order derivatives on G and that for every $x \in F^0$ and every $X \in \Sigma(x)$,*

- (i) $E[\limsup_{t \rightarrow \infty} Q(X_t)] \geq E[\limsup_{t \rightarrow \infty} u(X_t)],$
- (ii) $P[D(\alpha(t), \beta(t))Q(x_t) \leq 0 \text{ for all } t \geq 0] = 1, \text{ where } \alpha \text{ and } \beta \text{ are related to } X \text{ as in (1.5),}$
- (iii) *there exists an integrable random variable Y such that for all $t \geq 0$, $Q(X_t) \geq Y$.*

Then $Q \geq V$.

Proof. Let $x_0 \in F$ and $X \in \Sigma(x_0)$. For each $t \geq 0$, Ito's Lemma gives

$$(1.11) \quad Q(X_t) = Q(x_0) - A_t + M_t$$

where

$$A_t = - \int_0^t D(\alpha(s), \beta(s))Q(X_s) ds$$

and

$$M_t = \int_0^t Q_x(X_s)\beta(s) dW_s.$$

Notice that by conditions (ii) and (iii),

$$M_t = Q(X_t) - Q(x_0) + A_t \geq Y - Q(x_0)$$

holds for all $t \geq 0$ with probability one. Therefore the local martingale M_t is a supermartingale (see [1, VI.29]), and if τ is any almost-surely finite stopping time,

$$(1.12) \quad EM_\tau \leq EM_0 = 0.$$

From (1.11), (1.12) and condition (ii),

$$EQ(X_\tau) \leq Q(x_0).$$

By condition (i) and by Lemma 1 of [4], it follows that $Q \geq V$. \square

2. Maximizing $E[\beta^T]$, $0 < \beta < 1$. We formalize the first problem from the introduction as a continuous-time gambling problem in \mathbb{R}^2 . The first coordinate x_1 will be constrained to $(-\infty, 0]$ and indicates the player's position, while the second coordinate x_2 in $(-\infty, \infty)$ increases at a constant rate one and merely keeps track of the time. Define $F = \{x \in \mathbb{R}^2: x = (x_1, x_2), -\infty < x_1 \leq 0\}$. By our conventions each process X available at x will be absorbed at $T = T_X = \inf\{t: X_1(t) = 0\}$.

Let $\mathcal{S} \subseteq \mathbb{R} \times [0, \infty)$ and

$$(2.1) \quad C_0 = \left\{ \begin{pmatrix} \mu \\ 1 \end{pmatrix}, \begin{pmatrix} \sigma \\ 0 \end{pmatrix} : (\mu, 0) \in \mathcal{S} \right\}$$

and for every x in the interior of F , $C(x) = C_0$. Every $X \in \Sigma_C(x)$ can be specified by stochastic differential equations

$$(2.2) \quad \begin{aligned} dX_1(t) &= \mu(t) dt + \sigma(t) dW_t, \\ dX_2(t) &= dt, \\ X_1(0) &= x_1, \quad X_2(0) = x_2, \end{aligned}$$

where μ and σ are progressively measurable and $(\mu(t), \sigma(t)) \in \mathcal{S}$, $t < T$, $X(t) = X(T)$ for $t \geq T$.

Our utility function will be

$$(2.3) \quad u(x) = \beta^{x_2}.$$

Note that for $T < \infty$,

$$X_1(T) = 0, \quad X_2(T) = T + X_2(0)$$

and so

$$(2.4) \quad u(X) = E(\beta^{T+x_2}).$$

Here we interpret β^{T+x_2} as 0 if $T = \infty$. Now let $\Sigma(x) = \Sigma_C(x)$. The value function $V(x_1, x_2)$ must clearly have the form

$$(2.5) \quad V(x_1, x_2) = \beta^{x_2} V(x_1).$$

The form of $V(x_1)$ is also easily obtained. Given $X \in \Sigma(x)$ and $y \in [x, 0]$, let T_{xy} be the first time X attains y . Consider the problem of maximizing $E[\beta^{T_{xy}}]$ and denote the corresponding value function by V_{xy} . One easily sees $V_{xy} = V(x-y)$. It now appears that $T_{x0} = T_{xy} + T_{y0}$ and hence a stopping time argument gives $V(x) = V(x-y) \cdot V(y)$, leading to

$$(2.6) \quad V(x_1) = e^{\lambda x_1}, \quad \lambda \geq 0.$$

Also, since the problem of minimizing $E[\beta^{T_{xy}}]$ looks the same as that of minimizing $E[\beta^{T_{x-y}}]$, one expects to obtain optimal strategies of the form $\mu(x) \equiv \mu$ and $a(x) \equiv a$,

where we write $a(x) = \sigma^2(x)$ and μ and a will be constants depending on \mathcal{S} . If $a = 0$ and $\mu \leq 0$, then the process reaches zero with probability zero. So assume the maximum of a and μ is positive. For such a constant strategy the expected payoff is $W(x_1, x_2) = \beta^{x_2} W(x_1)$, with $W(x_1)$ satisfying

$$\frac{1}{2}aW'' + \mu W' + (\log \beta) W = 0, \quad W(0) = 1.$$

But again W should be an exponential $e^{\lambda x_1}$ with $\lambda \geq 0$, and this implies

$$(2.7) \quad W(x_1) = e^{\lambda x_1},$$

where

$$(2.8) \quad \lambda = \lambda(\mu, a) = \begin{cases} \frac{-\mu + \sqrt{\mu^2 - 2a \log \beta}}{a}, & a > 0, \\ \frac{-(\log \beta)}{\mu}, & a = 0, \quad \mu > 0. \end{cases}$$

Notice that in the case $a > 0$, λ is the positive root of

$$(2.9) \quad \frac{1}{2}a\lambda^2 + \mu\lambda + \log \beta = 0.$$

For fixed λ , the relation between λ , μ , and a in (2.8) defines a line in the (μ, a) -plane, namely

$$(2.10) \quad l_\lambda: a = \frac{-2}{\lambda}\mu - \frac{2 \log \beta}{\lambda^2}.$$

For our purposes, only

$$(2.11) \quad \bar{l}_\lambda = l_\lambda \cap \{(\mu, a): a \geq 0\}$$

is relevant. It now appears that to optimize the expected payoff in (2.7) we want to use strategies whose (μ, a) -pairs lie on \bar{l}_{λ^*} , where

$$(2.12) \quad \lambda^* = \inf \{\lambda(\mu, \sigma^2): (\mu, \sigma) \in \mathcal{S}, \max(\sigma, \mu) > 0\}.$$

THEOREM 2.1. *If $\mathcal{S} \subseteq (-\infty, 0] \times \{0\}$, then $V(x_1, x_2) = 0$. Otherwise,*

$$(2.13) \quad V(x_1, x_2) = \beta^{x_2} e^{\lambda^* x_1}$$

with λ^* as defined in (2.12).

Proof. The first assertion of the theorem is obvious. Assume then that \mathcal{S} contains a point (μ, σ) with $\max(\mu, \sigma^2) > 0$. Let $Q(x_1, x_2) = \beta^{x_2} e^{\lambda^* x_1}$. By (2.5) and (2.7) one can realize expected payoffs arbitrarily close to $Q(x_1, x_2)$, so that $Q(x_1, x_2) \leq V(x_1, x_2)$. For the opposite inequality use Proposition 1.1. It must be checked that Q satisfies conditions (i)–(iii). Condition (iii) is immediate. Condition (i) follows from the fact that for every available process $X = \{X_t\}$,

$$\limsup_{t \rightarrow \infty} Q(X_t) = \limsup_{t \rightarrow \infty} u(X_t) = \begin{cases} \beta^{T+x_2} & \text{if } T < \infty, \\ 0 & \text{if } T = \infty. \end{cases}$$

Condition (ii) follows once we have checked

$$(2.14) \quad \frac{1}{2}\sigma^2(\lambda^*)^2 + \mu\lambda^* + \log \beta \leq 0, \quad (\mu, \sigma) \in \mathcal{S}.$$

In the case $\sigma^2 = 0$, $\mu = 0$, this is immediate from the definitions of $\lambda(\mu, \sigma^2)$ and λ^* . In the case $\sigma^2 > 0$, (2.14) follows again from these definitions together with the fact that λ^* lies between 0 and $\lambda = \lambda(\mu, \sigma^2)$, which is the positive root of the equation (2.9). \square

Example. Discounted, continuous-time red-and-black. Suppose an investor has initial fortune y , $0 < y < 1$, and seeks to attain a fortune of 1. If $Y(t)$ is the player's fortune at time $t \geq 0$, he can invest $s(t)Y(t)$, $0 \leq s(t) \leq 1$, in a venture with rate of return μ_0 and standard deviation $\sigma_0 > 0$. More formally, the process $y(t)$ is given by a stochastic differential equation:

$$dY(t) = s(t)Y(t)[\mu_0 dt + \sigma_0 dW_t], \quad Y(0) = y$$

where $s(t)$ is a nonanticipative function such that $0 \leq s(t) \leq 1$. If the object is to maximize the probability of reaching 1, then (cf. [4], [5]) *bold play*, for which $s(t) \equiv 1$, is optimal in the *subfair case* ($\mu_0 \leq 0$), and *proportional play*, for which $s(t) \equiv c$ ($0 < c < 2\mu_0\sigma_0^{-2}$), is optimal in the *superfair case* ($\mu_0 > 0$). Here the value of the goal is discounted at rate β and the object is to maximize $E\beta^T$, where $T = \inf\{t \geq 0: Y(t) = 1\}$. The problem can be reduced to a special case of the problem of this section by the change of coordinates

$$(2.15) \quad X_1(t) = \log Y(t).$$

By Ito's formula,

$$dX_1(t) = (s(t)\mu_0 - \frac{1}{2}s(t)^2\sigma_0^2) dt + s(t)\sigma_0 dW_t.$$

The control set \mathcal{S} is now a one-parameter family

$$(2.16) \quad \mathcal{S} = \{(s\mu_0 - \frac{1}{2}s^2\sigma_0^2, s\sigma_0): 0 \leq s \leq 1\},$$

and formula (2.8) can be written in the form

$$(2.17) \quad \begin{aligned} \lambda(\mu, \sigma^2) &= \lambda(s) \\ &= \frac{-(\mu_0 - \frac{1}{2}s\sigma_0^2) + \sqrt{q(s)}}{s\sigma_0^2} \quad \text{if } s > 0 \end{aligned}$$

where $q(s) = (\mu_0 - \frac{1}{2}s\sigma_0^2)^2 - 2\sigma_0^2 \log \beta$. After some algebra it follows from (2.17) that, for $\mu_0 \leq 0$, $\lambda'(s) \leq 0$ on $(0, 1]$ and, hence, the infimum $\lambda^* = \lambda(1)$. So bold play is again optimal in the subfair case. Consider next the superfair case $\mu_0 > 0$.

Define the number

$$c = c(\mu_0, \sigma_0, \beta) = \frac{\mu_0}{\sigma_0^2} - \frac{2(\log \beta)}{\mu_0}.$$

More algebra shows that $\lambda'(s) > 0$ if and only if $s > \text{maximum}(c, 0)$. Thus if $c \geq 1$, λ is decreasing on $(0, 1]$ and bold play is optimal yet again. If $0 < c < 1$, an optimal strategy is given by $s(t) \equiv c$. If $c \leq 0$, there is no optimal strategy, but $s(t) \equiv \varepsilon$ for small positive ε will be almost optimal. An analogous problem in discrete time was solved by Klugman [3] in the subfair case, but the discrete-time superfair problem remains open.

3. Minimizing $E[\beta^T]$, $1 < \beta$. Again we fit the problem into the set-up of § 1. Since we discussed maximization problems there, we will seek to maximize $-\beta^T$. That is, we proceed exactly as in § 2, assuming (2.1), (2.2) but now defining

$$(3.1) \quad u(x) = -\beta^{x_2},$$

so that $u(X) = -E(\beta^{T+x_2})$. Let

$$\Sigma(x) = \{X \in \Sigma_C(x): u(X) > -\infty\} = \{X \in \Sigma_C(x): E\beta^T < \infty\},$$

and we have

$$(3.2) \quad V(x_1, x_2) = \beta^{x_2} V(x_1).$$

The arguments that lead to (2.6) now give

$$(3.3) \quad V(x_1) = -e^{\lambda x_1}, \quad \lambda \leq 0.$$

Again let us consider the expected payoff $W(x_1, x_2)$ for a constant strategy $\mu(t) \equiv \mu$ and $\sigma(t) \equiv \sigma$, $\sigma^2 = a$. If $a = 0$ and $\mu > 0$, obviously

$$(3.4) \quad W(x_1, x_2) = -\beta^{x_2} e^{\lambda x_1}$$

where $\lambda = -\log \beta / \mu$. If $a > 0$, $\mu > 0$, the expected payoff is of the form $W(x_1, x_2) = -\beta^{x_2} W(x_1)$, where $W(x_1)$ satisfies

$$(3.5) \quad \frac{1}{2}a W'' + \mu W' + (\log \beta) W = 0$$

and $W(0) = -1$. To see what condition to impose at $-\infty$ consider the problem as a limit of problems on the interval $[-M, 0]$, $M \rightarrow \infty$. So we consider (3.5) with $W(0) = -1$ and $W(-M) = -k_M$, with k_M chosen appropriately. Since in the limiting problem the process will hit zero with probability one and receive no help from the boundary on the left, it is appropriate to choose $k_M = 0$ (actually all choices of $k_M \geq 0$ which do not grow too rapidly with M will lead to the same limiting result). This limiting procedure gives us

$$(3.6) \quad W(x_1) = -e^{\lambda x_1},$$

for $\mu \geq \sqrt{2a \log \beta}$, where $\lambda = (-\mu + \sqrt{\mu^2 - 2a \log \beta})/a$ is the maximal root of the equation

$$(3.7) \quad f(\nu) = \frac{1}{2}a\nu^2 + \mu\nu + \log \beta = 0.$$

If

$$(3.8) \quad \mu < \sqrt{2a \log \beta},$$

the roots of (3.7) are not real and the expected payoff from the constant strategy $\mu(t) \equiv \mu$ and $\sigma(t) \equiv \sigma$ is $-\infty$. In every case, the expected payoff $W(x_1, x_2)$ is given by (3.4), where

$$(3.9) \quad \lambda = \lambda(\mu, a, \beta) = \begin{cases} \frac{-\mu + \sqrt{\mu^2 - 2a \log \beta}}{a}, & \mu \geq \sqrt{2a \log \beta}, \\ \frac{-\log \beta}{\mu}, & a = 0, \quad \mu > 0, \\ -\infty & \text{otherwise.} \end{cases}$$

Condition (3.8) holds if and only if (μ, a) lies on the left of the semi-parabola

$$(3.10) \quad p: \mu^2 = 2a \log \beta, \quad \mu > 0$$

in the (μ, a) -plane. Furthermore all points on the same line segment

$$l_\lambda: a = \frac{-2}{\lambda} \mu - \frac{2 \log \beta}{\lambda^2}, \quad 0 \leq a \leq \frac{\mu^2}{2 \log \beta}$$

give rise to constant strategies with the same expected payoff $-\beta^{x_2} e^{-\lambda x_1}$, $-\infty < \lambda < 0$. Observe that l_λ is the line segment connecting the point $(-\log \beta / \lambda, 0)$ on the μ -axis to the point $(-2 \log \beta / \lambda, 2 \log \beta / \lambda^2)$ on p ; at the latter point p and l_λ are tangential. Now set

$$(3.11) \quad \lambda^* = \lambda^*(\beta) = \sup \{ \lambda(\mu, \sigma^2, \beta) : (\mu, \sigma) \in \mathcal{S} \}.$$

Notice that, for $\lambda = \lambda(\mu, a, \beta) > -\infty$, the function $f(\nu)$ given in (3.7) is increasing to the right of $\nu = \lambda$. Also, $f(\lambda) = 0$ and $\lambda^* \geq \lambda$. Hence

$$(3.12) \quad \frac{1}{2}\sigma^2\nu^2 + \mu\nu + \log \beta \geq 0, \quad (\mu, \sigma) \in \mathcal{S}, \quad \lambda^* \leq \nu.$$

We define

$$(3.13) \quad I = \inf_{\varepsilon > 0} \sup \{ \mu + \varepsilon\sigma^2 : (\mu, \sigma) \in \mathcal{S} \}$$

and observe that the condition

$$(3.14) \quad I < \infty$$

holds if and only if \mathcal{S} is contained below some line with negative slope.

LEMMA 3.1. Assume $I = \infty$. Then $V(x_1, x_2) = -\beta^{x_2}$.

Proof. The assumption allows one to specify, for any $x_1 < 0$ and $\varepsilon > 0$ an $X \in \Sigma(x_1)$ with $ET < \varepsilon$. This was shown in [2]. This means that for $x_1 < 0$ and $\varepsilon > 0$ it is possible to find $X^{(1)} \in \Sigma(x_1)$ with $P[T > \varepsilon] < \frac{1}{2}$. Next, we find $X^{(2)} \in \Sigma(X^{(1)}(\varepsilon))$ such that $P[T > \varepsilon/2] < 2^{-2}$. Then $X \in \Sigma(x_1)$ is constructed as follows: X agrees with $X^{(1)}$ up to time ε ; if $X^{(1)} \neq 0$, then $X_{\varepsilon+t} = X_t^{(2)}$ for $0 \leq t \leq \varepsilon/2$; etc. Then for the process X , $E\beta^T \leq \beta^{2\varepsilon}$. Since ε is arbitrary the lemma follows by (3.2). \square

In view of Lemma 3.1 and the considerations preceding it, one might hope that if $I < \infty$,

$$V(x_1, x_2) = -\beta^{x_2} e^{\lambda^* x_1}.$$

We will now prove that this is indeed the case.

Some difficulties occur where points on the semi-parabola p introduced in (3.10) lie on the boundary of \mathcal{S} without belonging to \mathcal{S} . To deal with these we choose a sequence (β_n) of positive reals increasing up to β . Use (3.9) to define

$$(3.15) \quad \lambda_n(\mu, \sigma^2) = \lambda(\mu, \sigma^2, \beta_n)$$

and set

$$(3.16) \quad \lambda_n^* = \lambda^*(\beta_n) = \sup \{ \lambda_n(\mu, \sigma^2) : (\mu, \sigma) \in \mathcal{S} \}.$$

Note $\lambda_n^* \leq 0$ and λ_n^* decreases as a function of n . Finally let

$$(3.17) \quad \lambda^* = \inf_n \lambda_n^*$$

which is consistent with (3.11).

THEOREM 3.2(a). If $I < \infty$ and $\lambda^* > -\infty$ then

$$V(x_1, x_2) = -\beta^{x_2} e^{\lambda^* x_1}.$$

(b) If $I < \infty$ and $\lambda^* = -\infty$ then

$$V(x_1, x_2) = -\infty.$$

(c) If $I = \infty$ then

$$V(x_1, x_2) = -\beta^{x_2}$$

(that is, $E[\beta^T]$ can be made arbitrarily close to 1).

The proof of the theorem will involve the use of Proposition 1.1. We will construct a sequence of functions $Q_n(x_1, x_2) = \beta^{x_2} Q_n(x_1)$. Let us concentrate on the second factor and write simply x for x_1 . Consider the inequality

$$(3.18) \quad \frac{1}{2}\sigma^2 Q_n''(x) + \mu Q_n'(x) + (\log \beta) Q_n(x) \leq 0.$$

Let

$$(3.19) \quad U_n(x) = \frac{Q'_n(x)}{\lambda_n^* Q_n(x)}.$$

Then

$$U'_n(x) = \frac{1}{\lambda_n^*} \left[\frac{Q''_n(x)}{Q_n(x)} - \left[\frac{Q'_n(x)}{Q_n(x)} \right]^2 \right]$$

and so

$$U'_n(x) + \lambda_n^* (U_n(x))^2 = \frac{1}{\lambda_n^*} \frac{Q''_n(x)}{Q_n(x)}.$$

If Q_n is negative, the inequality (3.18) holds if and only if

$$(3.20) \quad \frac{1}{2} \sigma^2 \lambda_n^{*2} (U_n(x))^2 + \mu \lambda_n^* U_n(x) + \log \beta + \frac{1}{2} \sigma^2 \lambda_n^* U'_n(x) \geq 0.$$

Our plan is to define an appropriate sequence of functions U_n so that (3.20) holds for all (μ, σ) in \mathcal{S} . Then, using the relation (3.19), we transform U_n into Q_n and obtain inequality (3.18) for all $(\mu, \sigma) \in \mathcal{S}$.

The next lemma involves a construction of the functions U_n .

LEMMA 3.3. *Let $\{k_n : n \geq 1\}$ be positive constants. There exists a sequence $\{U_n : n \geq 1\}$ of real functions with domain $(-\infty, 1)$ such that*

- (i) U_n is continuously differentiable, $n \geq 1$;
- (ii) $0 < U_n(x) \leq 1$, $x < 1$, $n \geq 1$;
- (iii) $\lim_{n \rightarrow \infty} U_n(x) = 1$, $x \leq 0$;
- (iv) $\int_{-\infty}^0 U_n(x) dx < \infty$;
- (v) $0 < U'_n(x) \leq k_n U_n(x)$, $n \geq 1$, $x \leq 0$.

Proof. For each n define

$$U_n(x) = \begin{cases} 1, & -n < x < 1, \\ c_n \int_{-\infty}^{x+n} y e^{k_n y} dy, & x < -n \end{cases}$$

where $c_n = [\int_{-\infty}^0 y e^{k_n y} dy]^{-1}$. Then (i)–(iii) are easily checked. Integration by parts gives

$$U_n(x) = \frac{c_n}{k_n} e^{k_n(x+n)} \left(x+n - \frac{1}{k_n} \right), \quad x < -n.$$

The last identity implies (iv) and also

$$U_n(x) \geq \frac{c_n}{k_n} e^{k_n(x+n)} (x+n), \quad x < -n.$$

That is,

$$U_n(x) \geq \frac{1}{k_n} U'_n(x),$$

establishing property (v). \square

Assume now that $I < \infty$ (see (3.13)), and $\lambda^* > -\infty$. Then there exist positive constants ρ and M such that

$$(3.21) \quad \text{for all } (\mu, \sigma) \text{ in } \mathcal{S}, \quad \mu + \rho \sigma^2 \leq M.$$

For each positive integer n let

$$(3.22) \quad k_n = \begin{cases} \min \left\{ 2\rho, \frac{-2\rho \log(\beta/\beta_n)}{M\lambda_n^*} \right\}, & \lambda_n^* < 0, \\ 2\rho, & \lambda_n^* = 0. \end{cases}$$

With these constants k_n , let U_n be functions satisfying the conditions of Lemma 3.3.

LEMMA 3.4. Assume $\lambda^* > -\infty$, (3.21), and let $\{U_n: n \geq 1\}$ be the sequence of functions specified above. Then for $n \geq 1$, $x < 0$ and $(\mu, \sigma) \in \mathcal{S}$, inequality (3.20) holds.

Proof. The assertion is clear for $\lambda^* = 0$. So we suppose $\lambda^* < 0$. By properties (ii) and (v) of Lemma 3.3 and the definition of k_n in (3.22) we have

$$(3.23) \quad U'_n(x) \leq \frac{-2\rho \log(\beta/\beta_n)}{M\lambda_n^*}.$$

Let $g_n(x)$ denote the expression on the left of (3.20).

Case 1: $\mu \geq 0$, $\lambda_n^* < 0$. Use (3.21) to obtain

$$(3.24) \quad \rho\sigma^2 \leq \mu + \rho\sigma^2 \leq M.$$

Now write

$$(3.25) \quad g_n(x) = \left[\frac{1}{2} \sigma^2 \lambda_n^{*2} (U_n(x))^2 + \mu \lambda_n^* U_n(x) + \log \beta_n \right] + \log \frac{\beta}{\beta_n} + \frac{1}{2} \sigma^2 \lambda_n^* U'_n(x).$$

Since $\lambda_n^* \leq \lambda_n^* U_n(x) \leq 0$, the expression in brackets is greater than or equal to zero by (3.12). Then (3.24) and (3.23) show that the sum of the last two terms in (3.25) is nonnegative. Hence $g_n(x) \geq 0$, as desired.

Case 2: $\mu < 0$, $\lambda_n^* < 0$. Use (3.21) to obtain

$$(3.26) \quad \begin{aligned} g_n(x) &\geq \frac{1}{2} \sigma^2 (\lambda_n^* U_n(x))^2 + \mu \lambda_n^* U_n(x) + \log \beta + \frac{\lambda_n^* (M - \mu)}{2\rho} U'_n(x) \\ &\geq \mu \lambda_n^* \left[U_n(x) - \frac{1}{2\rho} U'_n(x) \right] + \frac{M\lambda_n^*}{2\rho} U'_n(x) + \log \beta. \end{aligned}$$

By properties (ii) and (v) of Lemma 3.3 and the definition of k_n in (3.22),

$$U_n(x) - \frac{1}{2\rho} U'_n(x) \geq 0 \quad \text{and} \quad U'_n(x) \leq \frac{-2\rho \log \beta}{M\lambda_n^*}.$$

Using these inequalities in (3.26) and recalling $\mu < 0$, $\lambda_n^* < 0$ one concludes $g_n(x) \geq 0$. \square

The next lemma will be used to create processes whose payoffs approach $-\beta^{x_2} e^{\lambda x_1}$.

LEMMA 3.5. If $I < \infty$, $\lambda^* > -\infty$, then there exists a sequence $\{(\mu_k, \sigma_k): k \geq 1\}$ of elements of \mathcal{S} such that the sequence converges in \mathbb{R}^2 to a limit $(\bar{\mu}, \bar{\sigma})$ and

$$(3.27) \quad \frac{1}{2} (\lambda^* \bar{\sigma})^2 + \lambda^* \bar{\mu} + \log \beta = 0.$$

(Note $(\bar{\mu}, \bar{\sigma})$ might not lie in \mathcal{S} .)

Proof. Using the definition of λ^* , there is a strictly increasing sequence $\{n_k; k \geq 1\}$ of positive integers and a sequence $\{(\mu_k, \sigma_k): k \geq 1\}$ of elements of \mathcal{S} such that

$$\lim_{k \rightarrow \infty} \lambda_{n_k}(\mu_k, \sigma_k^2) = \lambda^*.$$

By passing to a subsequence if necessary, we can assume that either (Case 1) $\sigma_k > 0$ and $\mu_k \geq \sigma_k \sqrt{2 \log \beta_{n_k}}$ for each k or (Case 2) $\sigma_k = 0$ and $\mu_k > 0$ for each k . Since $I < \infty$,

the set $\{(\mu, \sigma): (\mu, \sigma) \in \mathcal{S}, \mu \geq 0\}$ is a bounded subset of \mathbb{R}^2 . Thus by passing to a further subsequence, we can assume the sequence converges in \mathbb{R}^2 .

Case 1. $\sigma_k > 0$ and $\mu_k \geq \sigma_k \sqrt{2 \log \beta_{n_k}}$ for each k . By definition,

$$\lambda_{n_k}(\mu_k, \sigma_k) = \frac{-\mu_k + \sqrt{\mu_k^2 - 2\sigma_k^2 \log \beta_{n_k}}}{\sigma_k^2},$$

and so for each k ,

$$\frac{1}{2}[\lambda_{n_k}(\mu_k, \sigma_k^2)]^2 \sigma_k^2 + [\lambda_{n_k}(\mu_k, \sigma_k^2)]\mu_k + \log \beta_{n_k} = 0.$$

Taking the limit as $k \rightarrow \infty$, we get (3.27).

Case 2. $\sigma_k = 0$ and $\mu_k > 0$ for each k . By definition,

$$\lambda_{n_k}(\mu_k, \sigma_k^2) = \frac{-\log \beta_{n_k}}{\mu_k}.$$

Take limits to get $\lambda^* \bar{\mu} = -\log \beta$ and $\bar{\sigma} = 0$, and (3.27) follows. \square

Proof of the theorem. (a) Define $Q: (-\infty, 0] \times \mathbb{R} \rightarrow \mathbb{R}$ by

$$Q(x_1, x_2) = -\beta^{x_2} e^{\lambda^* x_1}.$$

To show $Q \leq V$, begin by fixing $x_1 \leq 0$ and $x_2 \in \mathbb{R}$. By Lemma 3.5, there is a sequence $\{(\mu_k, \sigma_k): k \geq 1\}$ in \mathcal{S} satisfying relation (3.27). Let $\varepsilon > 0$ and use (3.27) to create two functions $\hat{\mu}: [0, \infty) \rightarrow \mathbb{R}$ and $\hat{\sigma}: [0, \infty) \rightarrow \mathbb{R}$ such that

$$(\hat{\mu}(s), \hat{\sigma}(s)) \in \{(\mu_k, \sigma_k): k \geq 1\}$$

for each $s \geq 0$,

$$(3.28) \quad \int_0^\infty \left| \frac{1}{2} \lambda^{*2} [\hat{\sigma}(s)]^2 + \lambda^* \hat{\mu}(s) + \log \beta \right| ds < \log(1 + \varepsilon),$$

and

$$(3.29) \quad \inf \{\hat{\mu}(s): s \geq 0\} > 0.$$

Let X be the process given by

$$X_1(t) = x_1 + \int_0^t \hat{\mu}(s) ds + \int_0^t \hat{\sigma}(s) dW_s, \quad X_2(t) = x_2 + t$$

for $t \leq T$, where $T = \inf \{t: X_1(t) = 0\}$, and $X(t) = X(T)$ for $t \geq T$. Notice that (3.29) guarantees that the stopping time T is finite almost surely.

The aim is to show that X has payoff near $-\beta^{x_2} e^{\lambda^* x_1}$. Define the process Y by

$$Y_t = \exp \left[\lambda^* \left[X_1(t) - \int_0^t \hat{\mu}(s) ds \right] - \frac{\lambda^{*2}}{2} \int_0^t \hat{\sigma}^2(s) ds \right].$$

That is,

$$Y_t = \exp \left[\lambda^* x_1 + \lambda^* \int_0^t \hat{\sigma}(s) dW_s - \frac{\lambda^{*2}}{2} \int_0^t \hat{\sigma}^2(s) ds \right].$$

By Ito's formula,

$$Y_t = \exp [\lambda^* x_1] + \lambda^* \int_0^t Y_s \hat{\sigma}(s) dW_s,$$

and so $\{Y_t\}$ is a local martingale. Further, since $\{Y_t\}$ is nonnegative, it follows (see Dellacherie and Meyer [1, VI.29]) that $EY_T \leq EY_0$. That is,

$$(3.30) \quad E \left\{ \exp \left[- \int_0^T \left\{ \lambda^* \hat{\mu}(s) + \frac{1}{2} \lambda^{*2} \hat{\sigma}^2(s) \right\} ds \right] \right\} \leq \exp [\lambda^* x_1].$$

Using (3.28), (3.30), and the fact that $X_1(T) = 0$ a.s.,

$$E\beta^T = E[\exp[(\log \beta)T]] \leq (1 + \varepsilon) \exp[\lambda^* x_1].$$

Thus X is available at (x_1, x_2) because $E\beta^T < \infty$, and X has payoff at least $-\beta^{x_2}(1 + \varepsilon) \exp[\lambda^* x_1] = (1 + \varepsilon)Q(x_1, x_2)$. We conclude $Q \leq V$.

To show $Q \geq V$ we would like to apply Proposition 1.1. However, Q does not satisfy (iii) of that proposition. So we construct a sequence $\{Q_n\}$ converging pointwise to Q such that Proposition 1.1 applies to each Q_n .

For each $n \geq 1$, define $Q_n: (-\infty, 1) \times \mathbb{R} \rightarrow \mathbb{R}$ by

$$(3.31) \quad Q_n(x_1, x_2) = -\beta^{x_2} \exp \left[- \int_{x_1}^0 \lambda_n^* U_n(y) dy \right].$$

Notice that $\lim_{n \rightarrow \infty} Q_n(x) = Q(x)$ for each x in $F = (-\infty, 0] \times \mathbb{R}$ because of properties (ii) and (iii) of Lemma 3.3 and the dominated convergence theorem. Also, each Q_n is twice continuously differentiable because of property (i) of Lemma 3.3.

Now verify the conditions of Proposition 1.1 with Q_n ($n \geq 1$) in place of Q . Condition (i) is immediate. For condition (ii), let $x = (x_1, x_2) \in F^0$ and check that for each (μ, σ) in \mathcal{S} ,

$$(3.32) \quad \left[\frac{\partial}{\partial x_2} + \frac{1}{2} \sigma^2 \frac{\partial^2}{\partial x_1^2} \right] Q_n(x_1, x_2) \\ = Q_n(x_1, x_2) \left(\frac{1}{2} \sigma^2 \lambda_n^{*2} [U_n(x_1)]^2 + \frac{1}{2} \sigma^2 \lambda_n^* U_n'(x_1) + \mu \lambda_n^* U(x_1) + \log \beta \right).$$

Use Lemma 3.4 and the fact that $Q_n(x_1, x_2) \leq 0$ to show that the expression (3.32) is nonpositive. For condition (iii), let $x \in \Sigma_C(x)$ and use (3.31) and (ii) and (iv) of Lemma 3.3 to see that

$$(3.33) \quad Q_n(X_1(t), X_2(t)) \geq -C\beta^{X_2(t)} \geq -C\beta^{x_2+T}$$

where C is a constant satisfying $0 < C < \infty$. Now $E(\beta^T) < \infty$ for each $X \in \Sigma_C(x)$, and hence condition (iii) follows from (3.33). Thus Proposition 1.1 shows that $Q_n \geq V$ for each n and hence that $Q \geq V$. This completes the proof of (a) of the theorem.

(b) We reduce the result to (a). The hypothesis for (b) is that $I < \infty$ and $\lambda^* = -\infty$. Let $\varepsilon > 0$ and consider a new problem based on the set

$$\mathcal{S}_\varepsilon = \mathcal{S} \cup \{(\varepsilon, 0)\}.$$

The quantity corresponding to λ^* for the new problem is

$$\lambda_\varepsilon^* = \frac{-\log \beta}{\varepsilon}.$$

Thus part (a) can be applied to obtain the value function

$$V_\varepsilon(x_1, x_2) = -\beta^{x_2 - x_1/\varepsilon} \quad \text{for } x_1 \leq 0, \quad x_2 \in \mathbb{R}.$$

Clearly $V(x_1, x_2) \geq V_\varepsilon(x_1, x_2) \rightarrow -\infty$ as $\varepsilon \rightarrow 0$, and so the proof of (b) is finished.

(c) This was proved in Lemma 3.1. \square

Example. Inflated, continuous-time red-and-black. The problem considered here is the same as that in the example of § 2 except that $\beta > 1$ and the player seeks to minimize $E\beta^T$. (Imagine a borrower of \$1.00 who must pay back β^T if the loan is repaid at time T .) After the change of coordinates (2.15), the control set \mathcal{S} is given by (2.16) and is obviously bounded so that $I < \infty$. The quantity $\lambda(\mu, \sigma^2) = \lambda(s)$ is given by (2.17) if $s > 0$ and $\mu > \sqrt{2\sigma^2 \log \beta}$. Substitute $\mu = s\mu_0 - \frac{1}{2}s^2\sigma_0^2$, $\sigma = s\sigma_0$ and the latter condition reduces to $s \leq M$, $M = 2\mu_0/\sigma_0^2 - 2\sqrt{2(\log \beta)}/\sigma_0$. To reiterate, $\lambda(s)$ is given by (2.17) if $0 < s \leq M$ and $\lambda(s) = -\infty$ if not. Notice that, in the subfair case ($\mu_0 \leq 0$), $M < 0$, and consequently $\lambda^* = -\infty$ and, by Theorem 3.2, $V = -\infty$. In the superfair case ($\mu_0 > 0$), one shows that $\lambda'(s) > 0$ if and only if $0 < s < M \wedge c$ where $c = \mu_0/\sigma_0^2 - 2(\log \beta)/\mu_0$. Thus $\lambda^* = \lambda(s^*)$ where $s^* = (M \wedge c \wedge 1) \vee 0$, and V is given by Theorem 3.2.

REFERENCES

- [1] C. DELLACHERIE AND P.-A. MEYER, *Probabilities and Potential* B, North Holland, Amsterdam, 1982.
- [2] D. HEATH, S. OREY, V. PESTIEN AND W. SUDDERTH, *Minimizing or maximizing the expected time to reach zero*, this Journal, 25 (1987), pp. 195–205.
- [3] S. KLUGMAN, *Discounted and rapid subfair red and black*, Ann. Statist., 5 (1977), pp. 734–745.
- [4] V. PESTIEN AND W. SUDDERTH, *Continuous-time red and black: how to control a diffusion to a goal*, Math. Oper. Res., 10 (1985), pp. 599–611.
- [5] ———, *Continuous-time casino problems*, Math. Oper. Res. to appear.

ASYMPTOTIC PROPERTIES OF DISTRIBUTED AND COMMUNICATING STOCHASTIC APPROXIMATION ALGORITHMS*

HAROLD J. KUSHNER† AND G. YIN‡

Abstract. The asymptotic properties of extensions of the type of distributed or decentralized stochastic approximation proposed in [1] are developed. Such algorithms have numerous potential applications in decentralized estimation, detection and adaptive control, or in decentralized Monte Carlo simulation for system optimization (where they can exploit the possibilities of parallel processing). The structure involves several isolated processors (recursive algorithms) that communicate to each other asynchronously and at random intervals. The asymptotic (small gain) properties are derived. The communication intervals need not be strictly bounded, and they and the system noise can depend on the (communicating) system state. State space constraints are also handled. In many applications, the dynamical terms are merely indicator functions, or have other types of discontinuities. The “typical” such case is also treated, as is the case where there is noise in the communication. The linear stochastic differential equation satisfied by the (interpolated) asymptotic normalized error sequence is derived, and issued to compare alternative algorithms and communication strategies. Weak convergence methods provide the basic tools.

Key words. stochastic approximation, distributed stochastic approximation, weak convergence, asymptotic properties of recursive algorithms, communicating recursive systems, distributed stochastic computation

AMS(MOS) subject classifications. 60F05, 60F17, 62L20, 93E10, 93E12, 93E25

1. Introduction. Tsitsiklis [1] and Tsitsiklis, Bertsekas and Athens [2] proposed a very interesting model for a decentralized (distributed) recursive algorithm of the stochastic approximation (SA) type, with only asynchronous communications between the separate processors, and developed a scheme for proving w.p.1 (with probability 1) convergence. That work appears to be the first of its type for the decentralized SA problem. Such distributed algorithms are of rapidly growing interest. Various potential applications in adaptive control, estimation and in communication networks were proposed; e.g., several processors might do an identification of the parameters of an identical linear system (but with different inputs) and occasionally (asynchronously) share their latest estimates, or several processors might do Monte Carlo simulations of the SA type to locate the minimum of a regression function, and occasionally share their estimates. There are two main purposes for algorithms of the type discussed here and in [1]: exploiting the opportunities provided by parallel processing for Monte Carlo methods of system optimization or evaluation, situations in which there are physically separate systems (estimators, trackers, controllers) acting on or following essentially the same physical system; and occasionally communicating to take advantage of the “others” information.

The assumptions in [1] were fairly strong with respect to the great variety of potential applications, and the method of analysis required numerous detailed estimates. We analyze essentially the same algorithm here. In addition to getting the basic convergence results, our methods can handle the constrained (projected) algorithm, the case where the noise and the communication intervals depend on the state, the

* Received by the editors March 12, 1986; accepted for publication (in revised form) October 28, 1986.

† Lefschetz Center for Dynamical Systems, Division of Applied Mathematics, Brown University, Providence, Rhode Island 02912. The research of this author was supported in part by the Air Force Office of Scientific Research under grant AFOSR-81-0116, and Office of Naval Research grant N00014-83-K-0542.

‡ Lefschetz Center for Dynamical Systems, Division of Applied Mathematics, Brown University, Providence, Rhode Island 02912. The research of this author was supported in part by the Army Research Office under grant DAAG-29-84-K-0082, and the Office of Naval Research under grant N00014-85-K-0607.

general rate of convergence problem, and the case where there is communication noise. Instead of letting the “gain” parameter go to zero as $n \rightarrow \infty$ (as frequently done in classical SA) we keep it a constant, and work with convergence in the sense of weak convergence. There are several reasons for this. First, when we work with practical systems the chosen gains almost never go to zero, since one usually wants an algorithm that can track slow changes and is robust with respect to large bursts of noise. Our method can be adapted to get weak and even w.p.1 convergence when the gains do go to zero, and we comment on this in § 8. Even if the gains do go to zero, w.p.1 convergence is not much more useful or interesting than weak convergence. Weak convergence methods locate the points where the process spends most time (asymptotically), and as time goes to ∞ , an increasing (to one) proportion of time is spent arbitrarily close to such points. Then, one can often use the powerful “large deviations” methods to show, under very broad conditions, that ultimate escape from a small neighborhood of such points is impossible (when the gains go to zero) [3], [4]. Alternatively, once the weak convergence methods have located the “stable points” perturbed Lyapunov methods such as that in [10] can often be used to get w.p.1 convergence. One of the key questions in the analysis of any algorithm is the rate of convergence (the asymptotic normalized variance), and the analysis of the “rate” is almost always done via weak convergence methods. General background and applications in many areas are in [6]–[8]. Weak convergence methods are also *much easier* to use than the standard w.p.1 oriented methods; in many cases, a valid result can be obtained almost by inspection. This and the wide variety of problems which can be handled make it a more widely useful tool than “w.p.1” methods. The symbol \Rightarrow is used to denote weak convergence, and some definitions and properties of this convergence are stated in Appendix A.

The methods used here are quite efficient. Problems with potentially unbounded intercommunication intervals (e.g., where the interval is geometrically distributed) can be handled. We can also treat important cases where the dynamics are discontinuous or where the communication intervals and system noise depend on the system state, or where there are state space constraints. The case of discontinuous dynamics is of considerable importance in applications: often an estimate increases or decreases by a fixed amount ε —depending simply on whether a certain event occurred or not. Similarly, for state dependent communication times, a processor might want to communicate either if a given amount of time has passed since the last communication or if the state of the processor has changed by more than a given amount. In many applications (e.g., the decentralized form of the automata routing problem in [5]) the noise is naturally state dependent.

A theory of “rate of convergence” is also developed, which allows an objective comparison among alternative algorithms. Using this, in §§ 6 and 7, we comment on and compare the behavior of the algorithm with the centralized and various “deterministically” decentralized forms, in order to get a better understanding of its behavior and to see what the preferable communication strategies are. We can also allow “noise” in the communication, such as might be the case if the processors were physically separated and communicated via a noisy radio link (see § 7).

The basic algorithm will be described next. Section 2 contains a “technical” estimate which will be useful in the sequel. Section 3 deals with the basic weak convergence result in the function spaces $D[0, \infty)$ or $C[0, \infty)$ (see the Appendix for the definitions) and shows that a suitable continuous time interpolation $X^\varepsilon(\cdot)$ of the iterates $\{X_n\}$ converges weakly to the solution of a certain ODE as the gain parameter $\varepsilon \rightarrow 0$. The state dependent noise/intercommunication time case and the discontinuous

dynamics case are also treated here. Section 4 concerns a “projection” algorithm to handle state space constraints. Here, the limit satisfies a “projected” ODE. The asymptotics of $X^\varepsilon(t_\varepsilon + \cdot)$ are dealt with in § 5, where $t_\varepsilon \rightarrow \infty$ as $\varepsilon \rightarrow 0$. This yields the ultimately desired result concerning the location of the iterates for large n and small ε . Finally, the rate of convergence and comparison with a centralized processor is developed in §§ 6 and 7. A discussion of some of the probable advantages and uses of the algorithm appears in § 7. Section 8 contains a comment on the case where ε is replaced by $\varepsilon_n \rightarrow 0$.

The basic algorithm. We assume that there are q parallel processors, each with a state variable of dimension r . Let X_n^i denote the state of processor i at time n and define $X_n = (X_n^1, \dots, X_n^q)$. The symbol X generally denotes a qr -vector (not a random variable), which we partition as $X = (X^1, \dots, X^q)$, where each X^i is an r -vector. The “observation” of processor i at time n is $b^i(X_n^i, \xi_n^i)$, where ξ_n^i is the “noise”. Define the random variable $\xi_n = (\xi_n^1, \dots, \xi_n^q)$, and the generic dummy variable $\xi = (\xi^1, \dots, \xi^q)$, and write $B(X, \xi) = (b^1(X^1, \xi^1), \dots, b^q(X^q, \xi^q))$. Write $b^i(X^i, \xi^i) = (b_1^i(X^i, \xi^i), \dots, b_r^i(X^i, \xi^i))$, the $b_k^i(\cdot)$ being scalar valued. (All the above vectors are column vectors.) For vectors X^i in E^r , we often write simply x . The use of the “super vectors” X and X_n and their specializations below seems unavoidable, since all of the q processors are interconnected.

Let $\{A_n\}$ be a sequence of (possibly random) $qr \times qr$ matrices, where A_n can be written in the form

$$A_n = \begin{pmatrix} a_{11}(n) & \cdots & a_{q1}(n) \\ a_{1q}(n) & \cdots & a_{qq}(n) \end{pmatrix},$$

where each $a_{ij}(n)$ is a *diagonal* $r \times r$ matrix with nonnegative entries and $\sum_i a_{ij}(n) = I_r$, the identity matrix in E^r , Euclidean r -space (i.e. the “matrix valued” rows of A_n are “convexifying”). Note that the matrix indexing is not standard. The subscripts are of the form (source, destination) = (column, row). Suppose that there is a scalar $\alpha_0 > 0$ such that $a_{ii} \geq \alpha_0 I_r$ and, for $i \neq j$, either $a_{ij}(n) = 0$ or else $a_{ij}(n) \geq \alpha_0 I_r$.

The algorithm to be studied is

$$(1.1) \quad X_{n+1}^i = \sum_j a_{ji}(n) X_n^j + \varepsilon b^i(X_n^i, \xi_n^i), \quad X_{n+1} = A_n X_n + \varepsilon B(X_n, \xi_n).$$

At time n , processor i ($i = 1, \dots, q$) decides whether or not to communicate the current value of its state to any other processor and takes an observation $b^i(X_n^i, \xi_n^i)$. If there is no communication to processor i , then we set $a_{ii}(n) = I_r$ and $a_{ji}(n) = 0$ for $j \neq i$, and the iteration (for processor i at time n) is of the standard SA type: $X_{n+1}^i = X_n^i + \varepsilon b^i(X_n^i, \xi_n^i)$. If there are any communications to processor i from some processors $j \neq i$ at time n , then for such communicating processors j , $a_{ji}(n) \geq \alpha_0 I_r$, and the updated state X_{n+1}^i for processor i is a convex combination of X_n^i and of the states X_n^j communicated to it, added to its own SA increment $\varepsilon b^i(X_n^i, \xi_n^i)$. The requirement that either (for $j \neq i$) $a_{ji}(n) \geq \alpha_0 I_r$ or $a_{ji}(n) = 0$ simply means that if processor j communicates to processor i at time n , processor i can choose to ignore the communication, but if it incorporates the received X_n^j into its own state, it must do so in a “nontrivial” way. For notational simplicity, we omit the symbol for the ε dependence of X_n . The results (and methods) would not change if bounded communication delays were included in the model.

In [1], the algorithm was slightly more complex, since the dimensions of the X^i were not necessarily the same and a somewhat more complicated block structure of A_n was used. But, with no additional mathematical work (although with a more complex

notation), such extensions can readily be incorporated into our framework. It should be clear from the development that many related algorithms and conditions can be treated by essentially identical methods. Reference [14] discusses the asymptotic behavior of interacting stochastic approximations, in terms of the actual elapsed (random) computation (not iterate) times.

2. Some preparatory estimates. This section is devoted to obtaining the rate of convergence of the product $A_n \cdots A_k$ as $n \rightarrow \infty$. We use the following assumption.

(C2.1) Let F_n be an increasing sequence of σ -algebras such that F_n measures $\{X_i, i \leq n, \xi_i, A_i, i < n\}$. There are a scalar $p_0 > 0$ and integer m_0 such that

(2.1) P_{F_n} {processor i communicates to processor j on $[n, n + m_0) \cong p_0$ and j does not ignore the received message}

for all n and i, j , and $i \neq j$.

Remark. In [1], it was assumed that there is an m_0 such that $p_0 = 1$. Assumption (C2.1) covers the case where at each instant each processor flips a coin to decide whether to communicate or not. More generally, there often is a process $\{\tilde{A}_n\}$ such that $\{\tilde{A}_i, \xi_i, i < n, X_i, i \leq n\}$ is Markov, and A_n is a component of \tilde{A}_n . With this model, if F_n denotes the minimal σ -algebra that measures $\{\tilde{A}_i, \xi_i, i < n, X_i, i \leq n\}$, then (C2.1) covers many interesting cases where the intercommunication intervals are not bounded a priori, and might be "state" dependent. The condition seems to be unrestrictive.

For $n \geq k$, define $\Phi(n|k) = A_n \cdots A_k$ and set $\Phi(n|n+1) = I_{qr}$, the identity matrix in E^{qr} .

LEMMA 2.1. Assume (C2.1) and the conditions on $\{A_n\}$ in § 1. Then $\Phi_k \equiv \lim_n \Phi(n|k)$ exists w.p.1 and for each $i \leq r$, all the rows $i, i+r, \dots, i+qr-r$ of Φ_k are equal. Also

(2.1a) $E|\Phi(n|k) - \Phi_k| \rightarrow 0$ geometrically as $n - k \rightarrow \infty$,

(2.1b) $E_{F_k}|\Phi(n|k) - \Phi_k| \rightarrow 0$ geometrically as $n - k \rightarrow \infty$

uniformly in k and ω (w.p.1). Also $E_{F_n}\Phi(n|k)$ converges to Φ_k geometrically, uniformly in ω, k , as $n \rightarrow \infty$.

Remark. The fact that the limit Φ_k exists is almost obvious if we look at the $\{A_n\}$ as transition matrices for a Markov chain. The proof is in Appendix B.

Remark on other cases. One can readily work with the case where all of the processors do not necessarily communicate with each other. We comment on only one special case. Let processors $1, \dots, q_1$ communicate to each other but not to the other processors, and let processors $q_1 + 1, \dots, q$, communicate only to processors $1, \dots, q_1$ but not to each other. Then $\Phi(n|k)$ converges geometrically to a matrix Φ_k , which takes the form

$$\Phi_k = \begin{bmatrix} M_{q_1+1,1}^k & \cdots & M_{q_1}^k \\ \vdots & & \\ 0 & M_{q_1+1,q_1}^k & \cdots & M_{qq_1}^k \\ I_r & & & 0 \\ 0 & \ddots & & I_r \end{bmatrix}.$$

The $i, i+r, \dots$ rows of the upper right-hand block are not necessarily equal.

3. Convergence: The limit ODE. *Nonstate-dependent* $\{A_n, \xi_n\}$. We will work with several sets of assumptions. First, the basic convergence theorem will be proved when

the sequences $\{A_n\}$ and $\{\xi_n\}$ are nonstate-dependent and independent of each other, and then the restrictions will be weakened. Theorem 3.1 is the basic weak convergence theorem, from which most other results will follow. Let E_n denote expectation, conditioned on $\{X_i, i \leq n, A_i, \xi_i, i < n\}$. We will use subsets of the following assumptions. Recall that the $X = (X^1, \dots, X^q)$ and x are dummy variables.

(C3.1) $\{A_k\}$ and $\{\xi_n\}$ are independent of each other.

(C3.2) Let $\{\hat{\xi}_k\}$ be a sequence of bounded random variables and $\{\tilde{\xi}_k\}$ a sequence of random variables with zero mean and bounded 4th moment. Write $\xi = (\hat{\xi}, \tilde{\xi})$, $\xi_n = (\hat{\xi}_n, \tilde{\xi}_n)$, and let $B(X, \xi) = B_0(X, \hat{\xi}) + B_1(X) \tilde{\xi}$, where the $B_i(\cdot)$ are continuous ($B_0(\cdot, \hat{\xi})$ uniformly in $\hat{\xi}$).

(C3.3) There is a continuous function $\bar{B}(X) \equiv (\bar{b}^1(X^1), \dots, \bar{b}^q(X^q))$ such that

$$E_k B_0(X, \hat{\xi}_n) - \bar{B}(X) \rightarrow 0, \quad E_k \tilde{\xi}_n \rightarrow 0$$

in probability for each vector X , as $n - k \rightarrow \infty$.

(C3.4) There are a matrix $\bar{\Phi}$ and a sequence $m_\varepsilon \rightarrow \infty$ such that $\varepsilon m_\varepsilon \equiv \delta_\varepsilon \rightarrow 0$ and

$$E \left| \frac{1}{m_\varepsilon} \sum_{k=n}^{n+m_\varepsilon-1} E_n \Phi_k - \bar{\Phi} \right| \xrightarrow{\varepsilon} 0 \quad \text{uniformly in } n.$$

Remark and Definition. (C3) can be extended in various ways to time varying limits: e.g., let continuous $\Phi(t)$ replace Φ , when $n\varepsilon \rightarrow t$; we can even replace (C3.4) by a condition guaranteeing the ratios in (C3.4) are in some set of matrices M with a probability tending to unity as $\varepsilon \rightarrow 0$. But the statement of the limit result then becomes much more complicated.

Under the conditions of Lemma 2.1, $\bar{\Phi}$ must have the form

$$\bar{\Phi} = \begin{bmatrix} \bar{\phi}_1, \dots, \bar{\phi}_q \\ \bar{\phi}_1, \dots, \bar{\phi}_q \end{bmatrix} \equiv \begin{bmatrix} \hat{\Phi} \\ \hat{\Phi} \end{bmatrix}$$

where the $\bar{\phi}_i$ are diagonal $r \times r$ matrices with diagonal denoted by $(\bar{\phi}_{i1}, \dots, \bar{\phi}_{ir})$ and $\sum_j \bar{\phi}_{ji} = 1$. For any vector X we have the form $\bar{\Phi}X = (y, \dots, y)$ (column vector) and $\Phi_k X = (y_k, \dots, y_k)$ for some y and y_k in E^r . Let $\hat{\Phi}$ denote the row of $r \times r$ matrices $[\bar{\phi}_1, \dots, \bar{\phi}_q]$. Let $\bar{B}(x)$ denote $\bar{B}(x, x, \dots, x)$, and $B(x, \xi)$ denote $B(x, x, \dots, \xi)$.

(C3.5) The ODE (3.1) has a unique solution for each initial condition

$$\begin{aligned} \dot{x}_1 &= \bar{\phi}_{11} \bar{b}_1^1(x) + \dots + \bar{\phi}_{q1} \bar{b}_1^q(x) \\ &\vdots \\ \dot{x}_r &= \bar{\phi}_{1r} \bar{b}_r^1(x) + \dots + \bar{\phi}_{qr} \bar{b}_r^q(x) \end{aligned} = \hat{\Phi} \bar{B}(x). \quad (3.1)$$

(C3.3') There are a continuous $\bar{B}(\cdot)$ and $m_\varepsilon \rightarrow \infty$ such that $\varepsilon m_\varepsilon \equiv \delta_\varepsilon \rightarrow 0$ and

$$\frac{1}{m_\varepsilon} \sum_n^{n+m_\varepsilon-1} E_n B(X, \xi_k) \xrightarrow{\varepsilon} \bar{B}(X)$$

in probability for each vector X , uniformly in n .

(C3.4') There is a matrix $\bar{\Phi}$ such that, as $n - k \rightarrow \infty$,

$$E|E_k \Phi_n - \bar{\Phi}| \rightarrow 0.$$

Let n_ε be a sequence tending to ∞ and such that $\sqrt{\varepsilon} n_\varepsilon \rightarrow 0$, and, for $n \geq n_\varepsilon$,

$$\sup_k P\{|\Phi(k + n_\varepsilon | k) - \Phi_k| \geq \varepsilon^2\} \leq \varepsilon^2.$$

There is such a sequence, by Lemma 2.1. In fact, we can use $n_\varepsilon = O(\log 1/\varepsilon)$. Define

$$X_0^\varepsilon = \Phi(n_\varepsilon | 0) X_0 + \varepsilon \sum_{p=0}^{n_\varepsilon-1} \Phi_{k+1} B(X_k, \xi_k)$$

and for $t \geq 0$ define $X^\varepsilon(\cdot)$ by $X^\varepsilon(t) = X_n$ for $t \in [(n - n_\varepsilon)\varepsilon, (n - n_\varepsilon + 1)\varepsilon)$. Write $X^\varepsilon(\cdot) = (X^{\varepsilon,1}(\cdot), \dots, X^{\varepsilon,q}(\cdot))$. It will turn out that, for any initial conditions X_0^i , the vectors X_n^i , $i \leq q$, rapidly come close together (due to the communication and convexification). This leads to an (asymptotic in ε) jump in the process $X_{[t/\varepsilon]}$ at $t = 0$. For this reason, we start $X^\varepsilon(\cdot)$ slightly away (n_ε steps) from the origin of the $\{X_n\}$ process.

THEOREM 3.1. *Assume (C2.1), the conditions on $\{A_n\}$ in § 1, and (C3.1), (C3.2), (C3.5) and either (C3.3), (C3.4) or (C3.3'), (C3.4'). Define $X(0)$ and x_0 by $X(0) = \lim_\varepsilon X_0^\varepsilon \equiv (x_0, \dots, x_0)$. Then $X^\varepsilon(\cdot)$ is tight in $D[0, \infty)$ and converges weakly to a process $X(\cdot) = (x(\cdot), \dots, x(\cdot))$, where $x(\cdot)$ satisfies (3.1) with initial condition x_0 .*

Proof. Part 1. The proofs are essentially the same for the pairs (C3.3), (C3.4) and (C3.3'), (C3.4'), and we work only with the first pair. We often use Schwarz' inequality and the inequality (for $a \geq 0$), $E|\Phi(n|k) - \Phi_k|^{1+a} \leq \text{constant} \cdot E|\Phi(n|k) - \Phi_k|$, without specific mention. Iterating (1.1) and letting $n \geq n_\varepsilon$ yields

$$\begin{aligned} X_{n+1} &= \Phi(n|0)X_0 + \varepsilon \sum_0^{n_\varepsilon-1} \Phi(n|k+1)B(X_k, \xi_k) + \varepsilon \sum_{n_\varepsilon}^n \Phi(n|k+1)B(X_k, \xi_k) \\ (3.2) \quad &= X_0^\varepsilon + \varepsilon \sum_{n_\varepsilon}^n \Phi_{k+1} B(X_k, \xi_k) + \varepsilon \psi_n^\varepsilon + [\Phi(n|0) - \Phi(n_\varepsilon|0)]X_0 \end{aligned}$$

where

$$\psi_n^\varepsilon = \sum_0^n [\Phi(n|k+1) - \Phi_{k+1}]B(X_k, \xi_k).$$

For the purposes of the weak convergence proof, we can assume (w.l.o.g.) that $\{X_k\}$ is bounded by simply truncating the dynamical terms, i.e., changing $B(\cdot, \xi)$ so that it is zero for large $|X|$. If the theorem is true for *each such truncation*, then by the uniqueness assumption (C3.5), it is true as stated. Henceforth we assume this boundedness.

Part 2. Next, we show that $\sup_{\varepsilon, n} E|\psi_n^\varepsilon|^3 < \infty$. All norms are in the l_∞ sense. We have

$$\begin{aligned} E|\psi_n^\varepsilon|^3 &\leq \text{constant} \cdot \sum_{i,j,k} E|\Phi(n|i+1) - \Phi_{i+1}| |\Phi(n|j+1) - \Phi_{j+1}| |\Phi(n|k+1) - \Phi_{k+1}| \\ &\quad \cdot [1 + |\tilde{\xi}_i| |\tilde{\xi}_j| |\tilde{\xi}_k|]. \end{aligned}$$

By Holder's inequality, the summand is bounded above by

$$\begin{aligned} E^{1/12} [|\Phi(n|i+1) - \Phi_{i+1}|^{12} \cdot E^{1/12} |\Phi(n|j+1) - \Phi_{j+1}|^{12} \cdot E^{1/12} |\Phi(n|k+1) - \Phi_{k+1}|^{12} \\ \cdot [1 + E^{3/4} |\tilde{\xi}_i|^4 \cdot E^{3/4} |\tilde{\xi}_j|^4 E^{3/4} |\tilde{\xi}_k|^4]. \end{aligned}$$

By (C3.2) and the geometric convergence in Lemma 2.1 and the boundedness of $\Phi(n|i)$ and Φ_i , there is a $d \in [0, 1)$ such that this term is bounded above by (constant) $d^{n-i} d^{n-j} d^{n-k}$. Thus $\sup_{\varepsilon, n} E|\psi_n^\varepsilon|^3 < \infty$. From this and (3.2) (and the truncation of $B(\cdot, \cdot)$)

$$\sup_{\varepsilon, n \geq n_\varepsilon} E|X_{n+1} - X_n|^2 / \varepsilon^2 < \infty,$$

and $\{|X_{n+1} - X_n|/\varepsilon, n \geq n_\varepsilon, \varepsilon\}$ is uniformly integrable. Thus, $\{X^\varepsilon(\cdot)\}$ is tight in $D[0, \infty)$ and all limit paths are Lipschitz continuous (in t).

Part 3. We fix and work with a weakly convergent subsequence of $\{X^\varepsilon(\cdot)\}$, also indexed by ε , and with the limit denoted by $X(\cdot)$. Skorokhod imbedding (see the Appendix) will be used where useful, without specific mention. Thus, we can assume, where needed, that $X^\varepsilon(\cdot) \rightarrow X(\cdot)$ uniformly on bounded time intervals, w.p.1.

We will show, for each real valued function $f(\cdot)$ with compact support and continuous second derivatives, that the $M_f(\cdot)$ defined by

$$(3.3) \quad M_f(t) = f(X(t)) - f(X(0)) - \int_0^t f'_X(X(s)) \bar{\Phi} \bar{B}(X(s)) ds$$

is a (continuous) martingale. Since $M_f(\cdot)$ is a Lipschitz-continuous martingale (since $X(\cdot)$ is Lipschitz continuous), it is a constant. Thus, since $M_f(0) = 0$, we have $M_f(t) = 0$ or, equivalently, $\dot{X} = \bar{\Phi} \bar{B}(X)$. By the properties of Φ_k for each $i \leq r$, the $i, i+r, \dots, i+qr-r$ rows of $\bar{\Phi}$ are equal. Thus all r -vector components of the limit $X(\cdot)$ must be equal, i.e., $X(\cdot)$ is of the form $(x(\cdot), \dots, x(\cdot))$, for $x(t) \in E^r$. This and $\dot{X} = \bar{\Phi} \bar{B}(X)$ implies that $x(\cdot)$ satisfies (3.1).

We need only show the martingale property. To do this, we need only show that for any integer p and continuous bounded $h(\cdot)$ and $t_i \leq t, i \leq p, s > 0$,

$$(3.4) \quad Eh(X(t_i), i \leq p) \left[f(X(t+s)) - f(X(t)) - \int_t^{t+s} f'_X(X(u)) \bar{\Phi} \bar{B}(X(u)) du \right] = 0.$$

To simplify the notation (and w.l.o.g.), let t and s be integral multiples of $\varepsilon m_\varepsilon \equiv \delta_\varepsilon$ (see (C3.4) for the definition of m_ε) and define the index set $I_l^\varepsilon = \{n: lm_\varepsilon + n_\varepsilon \leq n < lm_\varepsilon + m_\varepsilon + n_\varepsilon\}$. By Taylor's Theorem and (3.2),

$$(3.5) \quad \begin{aligned} f(X^\varepsilon(t+s)) - f(X^\varepsilon(t)) &= \sum_{l \leq l\delta_\varepsilon < t+s} [f(X_{lm_\varepsilon+m_\varepsilon+n_\varepsilon}^\varepsilon) - f(X_{lm_\varepsilon+n_\varepsilon}^\varepsilon)] \\ &= \varepsilon \sum_{l \leq l\delta_\varepsilon < t+s} f'_X(X_{lm_\varepsilon+n_\varepsilon}^\varepsilon) \sum_{k \in I_l^\varepsilon} \Phi_{k+1} B(X_k, \xi_k) \\ &\quad + \text{error terms,} \end{aligned}$$

where the error term is of the order of the sum of (all norms are in the l_∞ sense)

$$\begin{aligned} &\varepsilon \sum_l |\psi_{lm_\varepsilon+n_\varepsilon}^\varepsilon|, \varepsilon^2 \sum_l |\psi_{lm_\varepsilon+n_\varepsilon}^\varepsilon|^2, \varepsilon, \\ &\sum_l |\Phi(lm_\varepsilon + m_\varepsilon + n_\varepsilon|0) - \Phi(n_\varepsilon|0)| \cdot |X_0|, \\ &\sum_k \varepsilon^2 (1 + |\tilde{\xi}_k|^2) \end{aligned}$$

where the sums are over all l such that $t \leq l\delta_\varepsilon < t+s$ and k is summed over $t \leq \varepsilon k - \varepsilon n_\varepsilon < t+s$. The mean values of the error terms go to zero as $\varepsilon \rightarrow 0$.

By (3.5),

$$(3.6) \quad \begin{aligned} &\lim_\varepsilon Eh(X^\varepsilon(t_i), i \leq p) [f(X^\varepsilon(t+s)) - f(X^\varepsilon(t))] \\ &= \lim_\varepsilon Eh(X^\varepsilon(t_i), i \leq p) \left[\varepsilon \sum_{l \leq l\delta_\varepsilon < t+s} f'_X(X_{lm_\varepsilon+n_\varepsilon}^\varepsilon) \sum_{k \in I_l^\varepsilon} \Phi_{k+1} B(X_k, \xi_k) \right]. \end{aligned}$$

We now rearrange the terms in a more convenient way. Define

$$\hat{B}_l^\varepsilon = \frac{1}{m_\varepsilon} \sum_{k \in I_l^\varepsilon} \Phi_{k+1} B(X_k, \xi_k)$$

and define the function $B^\varepsilon(\cdot)$ by

$$B^\varepsilon(t) = f'_X(X_{lm_\varepsilon+n_\varepsilon}^\varepsilon) E_{lm_\varepsilon+n_\varepsilon} \hat{B}_l^\varepsilon \quad \text{for } l\delta_\varepsilon \leq t < l\delta_\varepsilon + \delta_\varepsilon.$$

Since $X^\varepsilon(t_i)$, $i \leq p$, is measurable on the σ -algebra $F_{l_{m_\varepsilon+n_\varepsilon}}$, for $l\delta_\varepsilon \geq t$ (3.6) can be rewritten as

$$(3.7) \quad \begin{aligned} & \lim_{\varepsilon} Eh(X^\varepsilon(t_i), i \leq p) \left[\delta_\varepsilon \sum_{t \leq l\delta_\varepsilon < t+s} f'_X(X^\varepsilon_{l_{m_\varepsilon+n_\varepsilon}}) \hat{B}_l^\varepsilon \right] \\ &= \lim_{\varepsilon} Eh(X^\varepsilon(t_i), i \leq p) \int_t^{t+s} B^\varepsilon(u) du. \end{aligned}$$

If

$$(3.8) \quad B^\varepsilon(u) \xrightarrow{\varepsilon} f'_X(X(u)) \bar{\Phi} \bar{B}(X(u))$$

in probability for almost all u , then the second limit in (3.7) would be

$$Eh(X(t_i), i \leq p) \int_t^{t+s} f'_X(X(u)) \bar{\Phi} \bar{B}(X(u)) du.$$

Using this and taking limits in (3.6) yields the desired result (3.4), and we will be done. Thus, we need only show (3.8).

Fix u and for $\varepsilon > 0$, define l_ε by $u \in [l_\varepsilon \delta_\varepsilon, l_\varepsilon \delta_\varepsilon + \delta_\varepsilon)$. Then we need to show that

$$(3.9) \quad \frac{1}{m_\varepsilon} \sum_{k \in I_{l_\varepsilon}^\varepsilon} E_{l_\varepsilon m_\varepsilon + n_\varepsilon} f'_X(X_k) \Phi_{k+1} B(X_k, \xi_k) \xrightarrow{P} f'_X(X(u)) \bar{\Phi} \bar{B}(X(u)).$$

By (C3.2) (and the truncation), we can replace the X_k in (3.9) by $X_{l_\varepsilon m_\varepsilon + n_\varepsilon}$ without changing the limit. Using this and the independence assumption (C3.1), we can rewrite (3.9) as

$$(3.10) \quad \frac{1}{m_\varepsilon} \sum_{k \in I_{l_\varepsilon}^\varepsilon} f'_X(X_{l_\varepsilon m_\varepsilon + n_\varepsilon}) E_{l_\varepsilon m_\varepsilon + n_\varepsilon} \Phi_{k+1} E_{l_\varepsilon m_\varepsilon + n_\varepsilon} B(X_{l_\varepsilon m_\varepsilon + m_\varepsilon}, \xi_k) + \text{error term},$$

where the error term goes to zero in the mean. By the convergence of $X^\varepsilon(\cdot)$ to $X(\cdot)$ and using $l_\varepsilon \delta_\varepsilon \rightarrow u$, and (C3.3), (C3.4), we get that (3.10) converges in the mean to the right side of (3.9) as $\varepsilon \rightarrow 0$ and the proof is concluded. (The “intermediate” details in the last part of the proof are very similar to those in the “centralized” case. See [8, Chap. 5.2] or [9].) Q.E.D.

State dependent $\{A_n\}$ and $\{\xi_n\}$ and/or discontinuous dynamics. The state dependent “communication” and noise are most conveniently modeled by a “Markov” dependence. This will allow $\{A_n\}$ and $\{\xi_n\}$ to depend on the state in a variety of ways: A_n can depend (statistically) on recent events or changes in the X_n -sequence greater than a given magnitude over some time interval, or time elapsed since recent communications or on the “levels” of recent communications (i.e., the degree of “convexification” or incorporation of received data into one’s own estimate can depend on the nature or timing of recent receptions, transmissions, etc.). To be precise, we suppose that there is a bounded sequence of random variables $\{\tilde{A}_n\}$ such that A_n is a component of \tilde{A}_n and, for each $\varepsilon > 0$, $(X_n, \tilde{A}_{n-1}, \xi_{n-1})$ is a Markov process with a homogeneous transition function. The \tilde{A}_n can incorporate other data, e.g., time elapsed since last reception, transmission, etc. The case where some components of $B(\cdot, \cdot)$ are merely indicator functions (hence, not continuous functions) is of particular importance in applications. Such “Markovianizations” seem to be quite natural for many problems. It might be hard to explicitly evaluate the ODE’s here, but the character of the results is clear and precisely what is wanted.

Example. For one example of the appearance of state dependent noise, see the “routing” problem in [5]. In that example, inputs to a service or communication system occur at random, and the service times are random (correlated or not). The parameter

x (the state) determines the probability that incoming events are routed along particular channels. The effective noise is a consequence of the queue length or occupancy level of each channel; its statistics are dependent on the routing parameter. A Markov dependence model was appropriate there. The routing parameter at time n increased or decreased by ε , depending on whether or not certain events occurred at time n ; hence the dynamics were discontinuous. The model used in this section includes “decentralized” generalizations of such problems.

Assume that the marginal one-step transition function is of the product (conditionally independent) form. For some P_C and P_N

$$(3.11) \quad P\{\tilde{A}_1 \in B_1, \xi_1 \in B_2 | X_1, \tilde{A}_0, \xi_0\} = P_C\{\tilde{A}_1 \in B_1 | X_1, \tilde{A}_0\} P_N\{\xi_1 \in B_2 | X_1, \xi_0\}$$

(C denotes “communication”; N denotes “noise”). The P_C and P_N will not depend on ε . (We could allow some ε -dependence, but, in many applications, ε is merely a step size parameter and does not affect the distribution of the A_n , \tilde{A}_n or ξ_n other than via the values of the X_n (e.g., as in the above example).) The product from (3.11) is a natural generalization of (C3.1). Here the noise and intercommunication intervals are independent, conditional on the state. For each fixed X , the P_C and P_N in (3.11) can be considered to be one-step transition functions for “fixed X ” Markov chains which we denote by $\{\tilde{A}_n(X)\}$, $\{\xi_n(X)\}$. Let $P_C\{\tilde{A}_n, n, \cdot | X\}$ and $P_N\{\xi, n, \cdot | X\}$ denote the associated n -step transition functions. Then $P_C\{\tilde{A}, 1, \cdot | X\} = P_C\{\tilde{A}_1 \in \cdot | \tilde{A}_0 = A, X\}$, etc. The fixed- X processes are, of course, homogeneous. Let E_C^X and E_N^X denote the associated expectations.

Several assumptions will now be given, followed by some remarks concerning extensions. The assumptions are phrased so as to cover many potential applications. Again, recall that the X and x are vectors not random variables.

$$(C3.6) \quad E^X[A_n \cdots A_1 | \tilde{A}_0 = \tilde{A}] \equiv F_n(\tilde{A}, X) \text{ is continuous in } (\tilde{A}, X).$$

$$(C3.7) \quad \text{For each bounded and continuous functions } f_i(\cdot), \quad i=1,2, \int f_1(\xi_1) P_N(\xi, 1, d\xi_1 | X) \text{ and } \int f_2(\tilde{A}_1) P_C(\tilde{A}, 1, d\tilde{A}_1 | X) \text{ are continuous in } (\xi, X) \text{ and } (\tilde{A}, X) \text{ respectively.}$$

$$(C3.8) \quad \{\xi_n\} \text{ is bounded.}$$

$$(C3.9) \quad \text{For each vector } X \text{ of the form } X = (x, x, \cdots, x), \text{ let the pair of processes } \{\tilde{A}_n(X), \xi_n(X)\} \text{ associated with the } n\text{-step transition function } P_C\{\tilde{A}, n, \cdot | X\} P_N\{\xi, n, \cdot | X\} \text{ have a unique invariant measure and which is of the product form } P_C^x\{\cdot\} P_N^x\{\cdot\}.$$

$$(C3.10) \quad \int B(X, \xi_1) P(\xi, 1, d\xi_1 | X) \text{ is continuous in } (X, \xi).$$

Remark. Since the two fixed X -processes are independent, the product form in (C3.9) will hold if the processes are aperiodic. Under the conditions of Lemma 2.1, the $F_n(\tilde{A}, X)$ in (C3.6) converge geometrically (uniformly in \tilde{A}, X) to a function $\Phi(\tilde{A}, X)$, which must be continuous under (C3.6). By the discussion associated with Theorem 3.1, we see that $\Phi(\tilde{A}, X)$ has the form

$$\Phi(\tilde{A}, X) = \begin{bmatrix} \bar{\phi}_1(\tilde{A}, X) & \cdots & \bar{\phi}_q(\tilde{A}, X) \\ \vdots & & \vdots \\ \bar{\phi}_1(\tilde{A}, X) & \cdots & \bar{\phi}_q(\tilde{A}, X) \end{bmatrix} \equiv \begin{bmatrix} \hat{\Phi}(\tilde{A}, X) \\ \vdots \\ \hat{\Phi}(\tilde{A}, X) \end{bmatrix}$$

where $\phi_i(\tilde{A}, X)$ is a diagonal $(r \times r)$ matrix. Write $\phi_i(\tilde{A}, X) = \text{diag}[\phi_{i1}(\tilde{A}, X), \cdots, \phi_{ir}(\tilde{A}, X)]$.

If X takes the form $X = (x, x, \cdots, x)$ for $x \in E'$, we might simply write x for X .

$$(C3.11) \quad \text{The ODE (3.12) has a unique solution for each initial condition (analogous to (3.1)—the } \tilde{A} \text{ and } \tilde{\xi} \text{ are simply averaged out with respect to the invariant measure)}$$

$$(3.12) \quad \begin{aligned} \dot{x}_1 &= \sum_j \int \bar{\phi}_{j1}(\tilde{A}, x) P_C^x \{d\tilde{A}\} \int b_1^j(x, \xi^j) P_N^x(d\xi^j) \\ &\quad \vdots \\ \dot{x}_r &= \sum_j \int \bar{\phi}_{jr}(\tilde{A}, x) P_C^x \{d\tilde{A}\} \int b_r^j(x, \xi^j) P_N^x(d\xi^j) \end{aligned} = \hat{\Phi}(x) \bar{B}(x)$$

where

$$\hat{\Phi}(x) = \int \hat{\Phi}(\tilde{A}, x) P_C^x(d\tilde{A}), \quad \bar{B}(x) = \int B(x, \xi) P_N^x(d\xi).$$

Write

$$\bar{\Phi}(x) = \begin{bmatrix} \hat{\Phi}(X) \\ \vdots \\ \hat{\Phi}(x) \end{bmatrix}.$$

Under (C3.7), (C3.9) and (C3.10), the right side of (3.12) is continuous.

Remarks on the assumptions. In many applications, \tilde{A} takes only a finite number of values. Then the appropriate topology is the discrete topology and the \tilde{A} -continuity required in (C3.6) and (C3.7) always holds, since then all functions of \tilde{A} are continuous. The one-step smoothing assumption in (C3.10) can be replaced by a k -step smoothing assumption, and Theorem 3.2 will still hold. Since $\Phi(\tilde{A}, X)$ is continuous (see above remark), a Φ_k -analogue of (C3.10) is not needed. In (3.12) we are simply averaging the dynamics with respect to the invariant measures. If the invariant measure is not unique, then the right side of (3.12) is set valued and P_N^x and P_C^x range over all the invariant measures. We use (C3.8) here to avoid some details. Extensions to cover typical unbounded $\{\xi_n\}$ cases are possible via essentially the same method. To see how this can be done, see the proof for the “centralized” case in [8, Chap. 5.3] or in [9].

THEOREM 3.2. Assume (C2.1), the conditions on $\{A_n\}$ in § 1, (C3.2) (without the $\tilde{\xi}$ component) and (C3.6) to (C3.10). Then $\{X^\varepsilon(\cdot)\}$ is tight in $D[0, \infty)$ and converges weakly to $X(\cdot) = (x(\cdot), \dots, x(\cdot))$, where $x(\cdot)$ satisfies (3.12) and $X(0) = (x(0), \dots, x(0)) = \Phi_0 X_0$.

Proof. $\{X^\varepsilon(\cdot)\}$ is tight and all limits are Lipschitz continuous for the same reasons as in Theorem 3.1. Let ε index a weakly convergent subsequence with limit denoted by $X(\cdot)$. As in Theorem 3.1, $X(\cdot)$ has the form $X(\cdot) = (x(\cdot), \dots, x(\cdot))$. Owing to the Markov assumption, E_k denotes conditioning on $(X_k, \xi_{k-1}, \tilde{A}_{k-1})$. By the method of proof of Theorem 3.1, we need only show that the left side of (3.9) converges in probability to $f'_X(X(u))\bar{\Phi}(X(u))\bar{B}(X(u))$ for $X(u)$ of the form $X(u) = (x(u), \dots, x(u))$. The f_X term does not play an important role and we discard it henceforth.

We use the “truncation” method and notation discussed in Theorem 3.1. Thus, we can suppose that $B(\cdot, \cdot)$ and $\{X_n\}$ are bounded. For each ν , rewrite (3.9) as (using the conditional independence implied by (3.11))

$$(3.13) \quad \begin{aligned} H^\varepsilon &\equiv \frac{1}{m_\varepsilon} \sum_{k \in I_{I_\varepsilon}^\varepsilon} E_{m_\varepsilon I_\varepsilon + n_\varepsilon} E_k \Phi_{k+1} E_k B(X_k, \xi_k) \\ &= \frac{1}{m_\varepsilon} \sum_{k \in I_{I_\varepsilon}^\varepsilon} E_{m_\varepsilon I_\varepsilon + n_\varepsilon} E_k (A_{k+\nu} \cdots A_{k+1}) E_k B(X_k, \xi_k) + Q^\varepsilon, \end{aligned}$$

where $E|Q^\varepsilon| \rightarrow 0$ uniformly in ε , as $\nu \rightarrow \infty$, by Lemma 2.1.

We next estimate $E_{k+1} A_{k+\nu} \cdots A_{k+1}$. All norms are in the l_∞ sense. Since the $\{X_n\}$ and $\{\tilde{A}_n\}$ lie in a compact set, the function of δX defined by

$$\delta_i(|\delta X|) = \sup_{\tilde{A}, X} |F_i(\tilde{A}, X) - F_i(\tilde{A}, X + \delta X)|$$

can be supposed to go to zero as $|\delta X| \rightarrow 0$. We have $E_n A_n = F_1(\tilde{A}_{n-1}, X_n) = F_1(\tilde{A}_{n-1}, X_{n-1}) + \Delta_1(\tilde{A}_{n-1}, X_{n-1}, X_n)$, where $|\Delta_1(\tilde{A}_{n-1}, X_{n-1}, X_n)| \leq \delta_1(|X_n - X_{n-1}|)$. Next we can write $E_{n-1} A_n A_{n-1} = E_{n-1}(E_n A_n) A_{n-1} = E_{n-1} F_1(\tilde{A}_{n-1}, X_{n-1}) A_{n-1} + \Delta_1(\tilde{A}_{n-1}, X_{n-1}, X_n) A_{n-1}$. Note that

$$(3.14) \quad E_{n-1} F_1(\tilde{A}_{n-1}, X_{n-1}) A_{n-1} = F_2(\tilde{A}_{n-2}, X_{n-1}) = E_{C^{n-1}} A_n A_{n-1},$$

which is just the expectation for the two-step fixed X -process with X fixed at X_{n-1} . Using this and $|A_n| = 1$, we have

$$(3.15) \quad E_{n-1} A_n A_{n-1} = F_2(\tilde{A}_{n-2}, X_{n-2}) + \text{error terms},$$

where

$$\begin{aligned} |\text{error terms}| &= |(F_2(\tilde{A}_{n-2}, X_{n-1}) - F_2(\tilde{A}_{n-2}, X_{n-2})) + \Delta_1(\tilde{A}_{n-1}, X_{n-1}, X_n) A_{n-1}| \\ &\leq \delta_2(|X_{n-1} - X_{n-2}|) + \delta_1(|X_n - X_{n-1}|). \end{aligned}$$

Continuing in this way, we get

$$(3.16) \quad E_{k+1} A_{k+\nu} \cdots A_{k+1} = F_\nu(\tilde{A}_k, X_k) + T_k^{\varepsilon, \nu},$$

where

$$|T_k^{\varepsilon, \nu}| \leq \sum_{i=1}^{\nu} \delta_i(|X_{\nu+k-i+1} - X_{\nu+k-i}|)$$

and $E|T_k^{\varepsilon, \nu}| \xrightarrow{\varepsilon} 0$ for each ν , uniformly in k , owing to the convergence of the $X^\varepsilon(\cdot)$.

Putting the estimate (3.16) into (3.13) yields that

$$\begin{aligned} (3.17) \quad H^\varepsilon &= \frac{1}{m_\varepsilon} \sum_{k \in I_{I_\varepsilon}^\varepsilon} E_{m_\varepsilon l_\varepsilon + n_\varepsilon} F_\nu(\tilde{A}_k, X_k) B(X_k, \xi_k) + Q_\nu^\varepsilon \\ &\quad + \frac{1}{m_\varepsilon} \sum_{k \in I_{I_\varepsilon}^\varepsilon} E_{m_\varepsilon l_\varepsilon + n_\varepsilon} T_k^{\varepsilon, \nu} B(X_k, \xi_k). \end{aligned}$$

The last two right-hand terms in (3.17) go to zero in mean as $\varepsilon \rightarrow 0$ and then $\nu \rightarrow \infty$, and can be neglected. The sequence $F_\nu(\tilde{A}, X)$ converges uniformly to the continuous function $\Phi(\tilde{A}, X)$ as $\nu \rightarrow \infty$. Thus, the limit (as $\varepsilon \rightarrow 0$, $\nu \rightarrow \infty$) of the first term on the right side of (3.16) is the same if $\Phi(\tilde{A}, X)$ replaces $F_\nu(\tilde{A}, X)$.

Now we are in a position to use the result of [8, Chap. 5.3] or [9]. By the arguments (for the Markov model) in either of these references (which, when adapted to our current situation, require the continuity of $\Phi(\cdot, \cdot)$, and (C3.7), (C3.8) and (C3.10)) and the fact that $X^\varepsilon(\cdot) \rightarrow X(\cdot)$, we have

$$(3.18) \quad \frac{1}{m_\varepsilon} \sum_{k \in I_{I_\varepsilon}^\varepsilon} E_{m_\varepsilon l_\varepsilon + n_\varepsilon} \Phi(\tilde{A}_k, X_k) B(X_k, \xi_k) \rightarrow \int \Phi(\tilde{A}, X(u)) B(X(u), \xi) m^{x(u)}(d\tilde{A} d\xi),$$

where $m^X(\cdot)$ is an invariant measure for the process $\{\tilde{A}_n(X), \xi_n(X)\}$. Since $X(u) = (x(u), \cdots, x(u))$, the uniqueness and product form of the invariant measure in (C3.9) yields that $m^x(d\tilde{A} d\xi) = P_C^x(d\tilde{A}) \cdot P_N^x(d\xi)$. Thus the right side of (3.18) equals

$$\bar{\Phi}(x(u)) \bar{B}(x(u)) = \int \Phi(\tilde{A}, x(u)) P_C^{x(u)}(d\tilde{A}) \cdot \int B(x(u), \xi) P_N^{x(u)}(d\xi),$$

and the proof is concluded. Q.E.D.

4. State space constraints: a projection algorithm. In many applications, it is desirable to confine the iterates to a compact set L , and if they ever leave L , the algorithm will project them back onto L . Such algorithms are ubiquitous in applications, even if not explicitly defined or assumed; e.g., the ambiguous notion of “monitoring” in adaptive control which implicitly assumes some sort of projection. We treat two special, but useful, cases.

4.1. Assumptions and problem formulation.

(C4.1) Let $g_i(x)$, $i \leq \alpha$, be real valued continuously differentiable functions on E' and define $L = \{x: g_i(x) \leq 0, i \leq \alpha\}$. Let L be bounded, convex and the closure of its interior. Also (w.l.o.g.) assume that the gradient $g_{ix}(x)$ is not zero if $g_i(x) = 0$.

Let $\pi_L(y)$ denote the (unique) closest point on L to $y \in E'$. We use the projected form of algorithm (1.1):

$$(4.1) \quad \tilde{X}_{n+1} = A_n X_n + \varepsilon b(X_n, \xi_n), \quad X_{n+1}^i = \pi_L(\tilde{X}_{n+1}^i), \quad i \leq q.$$

Thus, each processor projects independently and the constraint set is the same for each. We now set the problem up so that previous results can be used.

Define $\rho_n = (\rho_n^1, \dots, \rho_n^q)$, where $\rho_n^i = [X_{n+1}^i - \tilde{X}_{n+1}^i]/\varepsilon$ and define $\tilde{\psi}_n^\varepsilon = \psi_n^\varepsilon + \sum_0^n [\Phi(n|k+1) - \Phi_{k+1}] \rho_k$. Then for $n \geq n_\varepsilon$ (n_ε was defined below (C3.4'))

$$(4.2) \quad \begin{aligned} X_{n+1} = X_0^\varepsilon + \varepsilon \sum_{n_\varepsilon}^n \Phi_{k+1} B(X_k, \xi_k) + \varepsilon \sum_{n_\varepsilon}^n \Phi_{k+1} \rho_k + \varepsilon \tilde{\psi}_n^\varepsilon \\ + [\Phi(n|0) - \Phi(n_\varepsilon|0)] X_0, \end{aligned}$$

where

$$X_0^\varepsilon = \Phi(n_\varepsilon|0) X_0 + \varepsilon \sum_0^{n_\varepsilon-1} \Phi_{k+1} [B(X_k, \xi_k) + \rho_k].$$

The two cases which we treat are covered by ((C4.1), (C4.2)) and (C4.3), respectively.

(C4.2) The matrices $a_{ij}(n)$ in A_n take the form $a_{ij}(n) = \alpha_{ij}(n) I_r$ where $\alpha_{ij}(n)$ is a scalar valued random variable and $\sum_i \alpha_{ij}(n) = 1$.

Under (C4.2) each of the scalar components “communicated” from a processor j to processor i are incorporated the same way into the updated estimates of processor i . In (C4.3), L takes the form of a hyper-rectangle.

(C4.3) There are bounded g_{1i} and g_{2i} such that $L = \{x: g_{1i} \leq x_i \leq g_{2i}, i \leq r\}$.

DEFINITIONS. For a vector field $h(\cdot)$ in E' , define the projection onto L by (for $x \in L$) $\pi(x, h(x)) = \lim_{\Delta \rightarrow 0} [\pi_L(x + \Delta h(x)) - x]/\Delta$. By the convexity of L , the limit is unique. Define the convex cone

$$C(x) = \left\{ y: y = \sum_{i \in A(x)} \lambda_i g_{ix}(x), \lambda_i \geq 0 \right\}$$

where $A(x)$ is the set of constraints $\{i: g_i(x) = 0\}$ (the active constraints at x). Note that $\rho_n^i \in -C(X_{n+1}^i)$. Write $A_n \rho_n = (Z_n^1, \dots, Z_n^q)$, where $Z_n^i \in E'$. Under (C4.2), each Z_n^i is a convex combination of vectors in the $-C(X_{n+1}^j)$, $j \leq q$. We will see below that the same property holds under (C4.2'). The same conditions apply when A_k or Φ_k replaces A_n .

The theorem is stated under the conditions of Theorem 3.1, but there is an analogous result under the conditions of Theorem 3.2.

THEOREM 4.1. Assume the conditions of Theorem 3.1, and either (C4.1), (C4.2) or

(C4.3). Let the solution to (4.3) (the projected form of (3.1)) be unique. Then $\{X^\varepsilon(\cdot)\}$ converges weakly to $X(\cdot)$, where $X(\cdot) = (x(\cdot), \dots, x(\cdot))$ and

$$(4.3) \quad \dot{x} = \pi(x, \hat{\Phi} \bar{B}(x)).$$

Equivalently,

$$(4.4) \quad \dot{x} = \hat{\Phi} \bar{B}(x) + \nu(x), \quad x(t) \in L$$

where $\nu(x(t)) \in -C(x(t))$. Also $X(0)$ takes the form $X(0) = \Phi_0 X_0 = (x(0), \dots, x(0))$, if $X_0^i \in L$.

Proof. Only (4.4) will be proved, since (4.4) implies (4.3). No truncation (see Theorem 3.1) is needed here since $X_n^i \in L$, a compact set. Define the process $\bar{R}^\varepsilon(\cdot)$ by

$$\bar{R}^\varepsilon(t) = \varepsilon \sum_{n_\varepsilon}^n \Phi_{k+1} \rho_k \quad \text{for } t \in [(n - n_\varepsilon)\varepsilon, (n - n_\varepsilon + 1)\varepsilon]$$

(analogous to the definition of $X^\varepsilon(\cdot)$ above Theorem 3.1). All norms below are in the l_∞ sense. For $X_n^i \in L$, the qr -vector components of $A_n X_n$ are all in L under either (C4.2) or (C4.3). Thus $|\rho_n^\varepsilon| \leq |B(X_n, \xi_n)|$. Hence, the proof of uniform integrability of $\{\psi_n^\varepsilon\}$ and $\{\rho_n^\varepsilon\}$ is the same as that for $\{\psi_n^\varepsilon\}$ given in Theorem 3.1. Thus $\{X^\varepsilon(\cdot), \bar{R}^\varepsilon(\cdot)\}$ is tight and all weak limits are Lipschitz continuous. Henceforth, we fix and work with a weakly convergent subsequence, also indexed by ε , and with a limit denoted by $(X(\cdot), \bar{R}(\cdot))$.

As in Theorem 3.1, for $i \leq q$, the $i, i+r, \dots, i+rq-r$ th rows of Φ_k are equal. Then so are the same components of $\Phi_{k+1} B(X_k, \xi_k)$ and of $\Phi_{k+1} \rho_k$. Thus (as in Theorem 3.1) $X(\cdot) = (x(\cdot), \dots, x(\cdot))$ and $\bar{R}(\cdot) = (R(\cdot), \dots, R(\cdot))$, where $x(t)$ and $R(t)$ are in E' , and

$$(4.5) \quad \dot{x} = \hat{\Phi} \bar{B}(x) + \dot{R}(t).$$

Obviously $x(t) \in L$. Thus, we need only show that $\dot{R}(t) \in -C(x(t))$ for almost all t .

Write $X^\varepsilon(\cdot) = (X^{\varepsilon,1}(\cdot), \dots, X^{\varepsilon,q}(\cdot))$. Let $x(t)$ be in the interior of L for $t \in [t_1, t_2]$ with $t_1 < t_2$. Then, by the weak convergence (i.e. convergence of all $X^{\varepsilon,i}(\cdot)$ to $x(\cdot)$) the $X^{\varepsilon,i}(t)$, $i \leq q$, are the strictly interior to L on $[t_1, t_2]$ with a probability which tends to unity as $\varepsilon \rightarrow 0$. Thus, for small ε , the cones $C(X^{\varepsilon,i}(t))$, $i \leq q$, $t_1 \leq t \leq t_2$, will be empty with a probability which tends to unity as $\varepsilon \rightarrow 0$. Thus $R(t) = 0$ for $t_1 \leq t \leq t_2$.

We will now consider the case where $x(t)$ is on the boundary of L for $t \in [t_1, t_2]$, $t_1 < t_2$. The general case follows by the same argument. Skorokhod imbedding will be used (see the Appendix), so that we can assume that the convergence is w.p.1 on each bounded time interval. Note that $C(x)$ is an upper semicontinuous function of x in the sense that if $x_n \rightarrow x$, then

$$(4.6) \quad C(x) \supset \bigcap_{n=k=n}^{\infty} C(x_k).$$

Let $(g_{i_1 x}(x(t)), \dots, g_{i_a x}(x(t))) \equiv (\nu_1, \dots, \nu_a)$ be the gradient vectors of the active constraints at $x = x(t)$, and let C_β denote the convex cone formed by the vectors in a β -neighborhood of (ν_1, \dots, ν_a) .

By the weak convergence (i.e., the convergence of all $X^{\varepsilon,i}(\cdot)$ to $x(\cdot)$) and (4.6), for each $\beta > 0$ and $\gamma > 0$, there are $\beta_1 > 0$ and $\varepsilon_1 > 0$ such that for $\varepsilon \leq \varepsilon_1$,

$$(4.7) \quad P\{\rho_k^i \in -C_\beta, i \leq q, \text{ all } k \text{ such that } |\varepsilon(k - n_\varepsilon) - t| \leq \beta_1\} \geq 1 - \gamma,$$

i.e., for $\varepsilon(k - n_\varepsilon)$ close enough to t , the ρ_k^i are in a "small neighborhood" of $-C(x(t))$ with probability close to unity.

Now, assume (C4.1), (C4.2). Then, each of the q r -vector components of $\Phi_{k+1} \rho_k$ is also in such a "small neighborhood" with a probability close to unity, for $\varepsilon(k - n_\varepsilon)$ close to t . This implies that $R(t) \in -C(x(t))$, for almost all t .

Write $x(t) = (x_1(t), \dots, x_r(t))$. Assume (C4.3), and let $\varepsilon(k - n_\varepsilon)$ be close to t . Then $C(x(t))$ is particularly simple. Write $\rho_k^j = (\rho_k^{j1}, \dots, \rho_k^{jr})$ where the ρ_k^{ji} are scalar valued. If $x_i(t) = g_{1i}$ (the lower limit) then (using the weak convergence) $X^\varepsilon(\cdot) \Rightarrow (x(\cdot), \dots, x(\cdot))$, the ρ_k^{ji} must be (asymptotically in ε) ≥ 0 for all j , with a probability arbitrarily close to unity. Similarly, if $x_i(t) = g_{2i}$ (the upper limit), then (asymptotically in ε) the ρ_k^{ji} must be ≤ 0 for all j . By the properties of Φ_{k+1} , the same property must hold for the respective components $(i, i+r, i+2r, \dots)$ of $\Phi_{k+1}\rho_k$.

The conclusion follows from this last remark, since if $x = (x_1, \dots, x_r)$, where $x_i = g_{1i}$, $i \leq r_1$, $x_i = g_{2i}$, $r_1 < i \leq r_2$ and $g_{1i} < x_i < g_{2i}$, $r_2 < i \leq r$, then we have that $-C(x)$ is the collection of vectors whose first r_1 components are nonnegative, the next $r_2 - r_1$ are nonpositive and the last $r - r_2$ are zero. Q.E.D.

5. The asymptotics of $X^\varepsilon(\cdot)$ for large t and small ε . Weak convergence in $D[0, \infty)$ or in $C[0, \infty)$ basically gives information on the locations and/or distribution of $X^\varepsilon(\cdot)$ for small ε , and for t confined to some large, but still bounded, interval. See, e.g., the discussion of the topology of these spaces in the Appendix. It is important to have a convergence result which is valid uniformly in (large) t for small ε , and such a result is readily available by appropriate modifications of the previous results. One usually requires that the ODE satisfied by the limit processes is stable; hence we assume the following:

(C5.1) Let (3.1) (or (3.12) for the state dependent $\{A_n, \xi_n\}$ case) have a unique stable (in the sense of Lyapunov) point θ which is globally attracting.

Let $t_\varepsilon \rightarrow \infty$ as $\varepsilon \rightarrow 0$. Frequently, if (C5.1) (and the conditions of Theorem 3.1 or 3.2) holds, then $X^\varepsilon(t_\varepsilon + \cdot)$ converges weakly to a constant process $\bar{X}(\cdot)$, where $\bar{X}(t) = (\theta, \dots, \theta)$. This is precisely the desired asymptotic result, since it says (roughly) that if the algorithm is "stable" then, after a fixed "transient period" (independent of ε), the $X^{t_\varepsilon}(\cdot)$ are arbitrarily close to θ in the sense of weak convergence. Condition (5.1) below will be dealt with later in this section.

We suppose now that the set

$$(5.1) \quad M = \{X^\varepsilon(t), t \geq 0, \varepsilon > 0\}$$

is bounded in probability (tight), i.e., for each $\eta > 0$ there is a $k_\eta < \infty$ such that $P\{|X^\varepsilon(t)| \geq k_\eta\} \leq \eta$ for all $\varepsilon > 0$, $t \geq 0$.

THEOREM 5.1. Assume tightness of (5.1), and the conditions of Theorem 3.1 or 3.2. Then $X^\varepsilon(t_\varepsilon + \cdot)$ converges weakly to the constant process (θ, \dots, θ) .

Discussion of the main idea of the development. Choose $T > 0$ and consider a convergent subsequence of the pair of processes $\{X^\varepsilon(t_\varepsilon + \cdot), X^\varepsilon(t_\varepsilon - T + \cdot)\}$, with limit denoted by $(X(\cdot), X_T(\cdot)) = (x(\cdot), \dots, x(\cdot); x_T(\cdot), \dots, x_T(\cdot))$ (recall that all the r -vector components of the limits are equal). We have $X(0) = X_T(T)$. The value of $X_T(0)$ is unknown, but all the possible such $X_T(0)$, over all T and convergent subsequences, belong to a tight set, with the same η and k_η as above. By this and the stability condition (A5.1) and Theorem 3.1 (or Theorem 3.2), for any $\delta > 0$ there is a $T_\delta < \infty$ such that for $T \geq T_\delta$, $X_T(T) = (x_T(T), \dots, x_T(T))$ will be in a δ -neighborhood of (θ, \dots, θ) with probability $\geq 1 - \delta$. This yields the desired conclusion, since it implies that $X(0) = (\theta, \dots, \theta)$ w.p.1. Thus, to get the asymptotic (in t and ε) result, only (5.1) must be shown.

Next, consider the projection algorithm of § 4 and assume (C5.1').

(C5.1') Let (4.4) have a unique stable (in the sense of Lyapunov) point θ which is attracting in L .

Under (C5.1'), (5.1) is automatically bounded, and if $t_\varepsilon \rightarrow \infty$ as $\varepsilon \rightarrow 0$ then under the additional conditions of § 4, $X^\varepsilon(t_\varepsilon + \cdot) \Rightarrow \bar{X}(\cdot)$, where $\bar{X}(t) = (\theta, \dots, \theta)$. Some

form of projection algorithm is usually used in practical algorithms, and so the tightness condition on (5.1) is not burdensome.

5.1. Sharper bounds on the asymptotic errors $(X_n^i - \theta)$, for large εn and small ε . Under additional "stability" conditions, one can get order of magnitude estimates for $(X^{i,\varepsilon}(t) - \theta)$ for large t and small ε . We do one case here in preparation for the rate of convergence work in § 6. We will need the following assumption.

(C5.2) There is a twice continuously differentiable Lyapunov function $0 \leq \bar{V}(x) \rightarrow \infty$ and $\bar{V}(x) > 0$ for $x \neq \theta$ such that for some $\lambda > 0$ and $K < \infty$, $\bar{V}'_x(x) \hat{\Phi} \bar{B}(x) \leq -\lambda \bar{V}(x)$, $|\bar{V}_x(x)|^2 \leq K(\bar{V}(x) + 1)$ and $\bar{V}_{xx}(\cdot)$ is bounded.

Define

$$(5.2) \quad V(X) = \sum_1^q \bar{V}(X^i) \quad \text{for } X = (X^1, \dots, X^q).$$

(C5.3) Case (C3.2), but where $B_0(X, \hat{\xi})$ and $B_1(X)$ are bounded and have bounded and continuous X -derivatives (uniformly in $\hat{\xi}$, for B_0).

(C5.4) There is a constant K such that

$$E \left| \sum_{\nu}^{\nu+m} E_{\nu}(\Phi_{k+1} B(X, \xi_k) - \bar{\Phi} \bar{B}(X)) \right|^2 \leq K[V(X) + 1],$$

for all positive m and ν . Similarly for the derivatives B_X and \bar{B}_X replacing B and \bar{B} , respectively.

Remark. Case (C5.4) essentially implies a "low" correlation between data in the remote past and in the distant future. There is an analogous result to Theorem 5.1 for the state dependent $\{A_n, \xi\}$ case, and for the constrained case.

THEOREM 5.2. Assume (C5.1) to (C5.4). There is an $N_{\varepsilon} < \infty$ for each small ε such that

$$(5.3) \quad EV(X_n) = O(\varepsilon), \quad n \geq N_{\varepsilon}.$$

Proof. We always assume $n \geq n_{\varepsilon}$, so that $E|\Phi(n|0) - \Phi_0|^a = O(\varepsilon^2)$ for any $a > 0$. Write

$$(5.4) \quad \begin{aligned} X_{n+1} - X_n &= [\Phi(n|0) - \Phi(n-1|0)]X_0 + \varepsilon(\psi_n^{\varepsilon} - \psi_{n-1}^{\varepsilon}) \\ &\quad + \varepsilon \bar{\Phi} \bar{B}(X_n) + \varepsilon[\Phi_{n+1} B(X_n, \xi_n) - \bar{\Phi} \bar{B}(X_n)] \end{aligned}$$

and

$$(5.5) \quad \begin{aligned} E_n V(X_{n+1}) - V(X_n) &= \varepsilon V'_X(X_n) E_n [\Phi(n|0) - \Phi(n-1|0)] X_0 \\ &\quad + \varepsilon V'_X(X_n) E_n (\psi_n^{\varepsilon} - \psi_{n-1}^{\varepsilon}) + \varepsilon V'_X(X_n) \bar{\Phi} \bar{B}(X_n) \\ &\quad + \varepsilon V'_X(X_n) E_n [\Phi_{n+1} B(X_n, \xi_n) - \bar{\Phi} \bar{B}(X_n)] + \text{error term}, \end{aligned}$$

where $E|\text{error term}| = O(\varepsilon^2)$. By (C5.2) and $n \geq n_{\varepsilon}$, the expectation of the first term on the right-hand side of (5.5) is $O(\varepsilon^2)(1 + EV(X_n))$. Write Φ_n in the form

$$\Phi_n = \begin{bmatrix} \hat{\Phi}_n \\ \vdots \\ \hat{\Phi}_n \end{bmatrix},$$

where $\hat{\Phi}_n$ is an $r \times qr$ matrix. For $n \geq n_{\varepsilon}$,

$$(5.6) \quad |X_n^i - X_n^j| = O_n(\varepsilon^2) + O(\varepsilon)|\psi_n^{\varepsilon}|,$$

where $E|O_n(\varepsilon^2)|^2 = O(\varepsilon^4)$, uniformly in $n \geq n_{\varepsilon}$.

Using (5.6), rewrite the last two terms on the right side of (5.5) as, respectively,

$$(5.7) \quad \begin{aligned} & \varepsilon \sum_i \bar{V}'_x(X_n^i) \hat{\Phi} \bar{B}(X_n^i) + \text{error term}, \\ & \varepsilon \sum_i \bar{V}'_x(X_n^i) E_n [\hat{\Phi}_{n+1} B(X_n^i, \xi_n) - \hat{\Phi} \bar{B}(X_n^i)] + \text{error term}, \end{aligned}$$

where by (C5.2), $E|\text{error term}| = O(\varepsilon^2)(1 + EV(X_n))$.

We now define the perturbations to the Lyapunov function.

Define

$$V_1^\varepsilon(n) \text{ by } V_1^\varepsilon(n) = -\varepsilon V'_x(X_n) \psi_{n-1}^\varepsilon.$$

We have

$$(5.8) \quad E|V_1^\varepsilon(n)| = O(\varepsilon)(1 + EV(X_n)),$$

$$(5.9) \quad EV_1^\varepsilon(n+1) - EV_1^\varepsilon(n) \leq -\varepsilon EV'_x(X_n)(\psi_n^\varepsilon - \psi_{n-1}^\varepsilon) + O(\varepsilon^2)E(1 + V(X_n)).$$

Define $V_2^{i\varepsilon}(n)$:

$$(5.10) \quad V_2^{i\varepsilon}(n) = \varepsilon \sum_{j=n}^{\infty} \bar{V}'_x(X_n^i) E_n [\hat{\Phi}_{j+1} B(X_n^i, \xi_j) - \hat{\Phi} \bar{B}(X_n^i)].$$

By (C5.2) and (C5.4),

$$(5.11) \quad E|V_2^{i\varepsilon}(n)| = O(\varepsilon)(1 + EV(X_n)).$$

Also,

$$(5.12) \quad E_n V_2^{i\varepsilon}(n+1) - V_2^{i\varepsilon}(n) = -\varepsilon \bar{V}'_x(X_n^i) E_n [\hat{\Phi}_{n+1} B(X_n^i, \xi_n) - \hat{\Phi} \bar{B}(X_n^i)] + \text{error term},$$

where by (C5.4), $E|\text{error term}| = O(\varepsilon^2)(1 + EV(X_n))$.

Now, define the perturbed Lyapunov function $V^\varepsilon(n) = V(X_n) + V_1^\varepsilon(n) + \sum_1^q V_2^{i\varepsilon}(n)$, and evaluate $E_n V^\varepsilon(n+1) - V^\varepsilon(n)$ and cancel the terms $\pm \varepsilon V'_x(X_n)(\psi_n^\varepsilon - \psi_{n-1}^\varepsilon)$ and $\pm \varepsilon \sum_j \bar{V}'_x(X_n^i) E_n [\hat{\Phi}_{n+1} B(X_n^i, \xi_j) - \hat{\Phi} \bar{B}(X_n^i)]$ to get

$$(5.13) \quad \begin{aligned} E_n V^\varepsilon(n+1) - V^\varepsilon(n) &= \varepsilon \sum_i \bar{V}'_x(X_n^i) \hat{\Phi} \bar{B}(X_n^i) + \text{error terms}, \\ E|\text{error term}| &= O(\varepsilon^2)(1 + EV(X_n)). \end{aligned}$$

Using (C5.2) and the bounds on $E|V_1^\varepsilon(n)|$ and on $E|V_2^{i\varepsilon}(n)|$, we get

$$(5.14) \quad \begin{aligned} EV^\varepsilon(n+1) - EV^\varepsilon(n) &\leq -\lambda \varepsilon \sum_i E \bar{V}(X_n^i) + O(\varepsilon^2)(1 + EV(X_n)) \\ &\leq -\lambda \varepsilon EV^\varepsilon(n) + O(\varepsilon^2)(1 + EV^\varepsilon(n)). \end{aligned}$$

Hence, for small $\varepsilon > 0$,

$$(5.15) \quad EV^\varepsilon(n) \leq \left(1 - \frac{\lambda \varepsilon}{2}\right)^{n-n_e} V^\varepsilon(n_e) + O(\varepsilon).$$

This, together with the bounds on $E|V_1^\varepsilon(n)|$ and on $E|V_2^{i\varepsilon}(n)|$, yields the theorem. Q.E.D.

6. Rate of convergence: qualitative asymptotic properties. The Lyapunov function in (C5.2) is often locally quadratic about θ in the sense that $\bar{V}(x) = (x - \theta)'Q(x - \theta) + O(|x - \theta|^3)$ for $Q > 0$. If this is true, then Theorem 5.1 implies that

$\{(X_n^i - \theta)/\varepsilon^{1/2}, i \leq q, n \geq N_\varepsilon, \varepsilon > 0\}$ is tight. In this section, we are interested in the asymptotic normalized errors, and we assume this tightness, whether or not the Lyapunov function is of the above form and the conditions of Theorem 5.1 hold. In particular, we will suppose that there are numbers $\tilde{N}_\varepsilon < \infty$ for each small $\varepsilon > 0$ so that

$$(6.1) \quad \left\{ \frac{X_n^i - \theta}{\sqrt{\varepsilon}}, i \leq q, n \geq \tilde{N}_\varepsilon, \varepsilon > 0 \right\} \text{ is tight, } \quad Eb^i(\theta, \xi_k^i) = 0.$$

Under (6.1), one can apply the methods of the “centralized” case to get a classical rate of convergence result.

Much information concerning the asymptotic behavior and comparison with other algorithms can be obtained from such a result. The method and results will be discussed in an informal way so that the main ideas are clear. Despite the informality, the conditions needed for the proof will generally be stated. The proofs follow standard lines in weak convergence theory and are not hard. Our aim is to exhibit the asymptotic behavior of the suitably normalized errors, then specialize them to simple cases where a comparison can be made with “centralized” forms of the algorithm, so that one can see the effects and value of the decentralization, and evaluate alternative forms of the communication and algorithms. The discussion is continued in § 7. Such insights are needed at this stage of development of the “decentralized” algorithms as a guide to future developments, and are perhaps more important than a rigorous development along the standard lines. We will use (6.1), the boundedness of $B(\cdot, \cdot)$ in each bounded X -set and that $B(\cdot, \xi)$ has a continuous (uniformly in ξ) derivative, and $EB(X, \xi) = 0$.

For any R^s valued function $p(\cdot) = (p^1(\cdot), \dots)$ of x (or X), let $(p(\theta))_x$ denote the (Jacobian) matrix whose i th row is the x (respectively, X) gradient of $p^i(\cdot)$. Recall the definitions of $\bar{\phi}_i$ and $\hat{\Phi}$ (above (C3.5)), and of $\hat{\Phi}_n$ (in Theorem 5.1). Define the matrix $M = (\hat{\Phi} \bar{B}(\theta))_x$ and suppose that it is stable. Let

$$\frac{1}{n} \sum_m^{n+m} (\hat{\Phi}_{k+1} B(\theta, \xi_k))_x \rightarrow (\hat{\Phi} \bar{B}(\theta))_x = M$$

in probability as $n \rightarrow \infty$ and $m \rightarrow \infty$.

Define $U_n^\varepsilon = (X_n - \bar{\theta})/\sqrt{\varepsilon}$, where $\bar{\theta} = (\theta, \theta, \dots, \theta)$. Recall the definition of n_ε given below (C3.4') and that n_ε can be chosen such that $\sqrt{\varepsilon} n_\varepsilon \rightarrow 0$. Given $N > 0$, we have, for $n \geq n_\varepsilon + N$,

$$(6.2) \quad \begin{aligned} U_{n+1}^\varepsilon &= \Phi(n|N) U_N^\varepsilon + \sqrt{\varepsilon} \sum_N^{N+n_\varepsilon} \Phi(n|k+1) B(X_k, \xi_k) \\ &\quad + \sqrt{\varepsilon} \sum_{N+n_\varepsilon+1}^n \Phi_{k+1} B(X_k, \xi_k) + \sqrt{\varepsilon} \hat{\psi}_n^\varepsilon, \end{aligned}$$

$$\hat{\psi}_n^\varepsilon = \sum_{N+n_\varepsilon}^n [\Phi(n|k+1) - \Phi_{k+1}] B(X_k, \xi_k).$$

Define (for $n \geq N + n_\varepsilon$)

$$W_n^\varepsilon = \sqrt{\varepsilon} \sum_{N+n_\varepsilon+1}^n \Phi_{k+1} B(\bar{\theta}, \xi_k).$$

Let $N \geq \tilde{N}_\varepsilon$. For $t \geq 0$, define the process $U^\varepsilon(\cdot)$ by $U^\varepsilon(t) = U_n^\varepsilon$ for $t \in [\varepsilon(n - N - n_\varepsilon), \varepsilon(n - N - n_\varepsilon + 1))$ and define $W^\varepsilon(\cdot)$ similarly from $\{W_n^\varepsilon\}$. By Taylor's Theorem and

the definition of n_ε ,

$$(6.3) \quad \begin{aligned} U_{n+1}^\varepsilon = & \Phi_N U_N^\varepsilon + O_n(\varepsilon^2) U_N^\varepsilon + \tilde{O}_n(n_\varepsilon \sqrt{\varepsilon}) + O(\sqrt{\varepsilon}) \hat{\psi}_n^\varepsilon \\ & + \varepsilon \sum_{N+n_\varepsilon+1}^n (\Phi_{k+1} B(\bar{\theta}, \xi_k))_X U_k^\varepsilon + W_n^\varepsilon + o(\varepsilon) \sum_{N+n_\varepsilon+1}^n O(|U_k^\varepsilon|), \end{aligned}$$

where $E|O_n(\varepsilon^2)|^2 = O(\varepsilon^4)$ since $n \geq n_\varepsilon$, and $E|\tilde{O}_n(n_\varepsilon \sqrt{\varepsilon})| = O(n_\varepsilon \sqrt{\varepsilon})$. Also $(\Phi_{k+1} B(\bar{\theta}, \xi_k))_X$ denotes the matrix whose rows are the X -gradients of the components of $\Phi_{k+1} B(X, \xi_k)$ evaluated at $X = (\theta, \theta, \dots) = \bar{\theta}$.

In order to study the weak convergence of $U^\varepsilon(\cdot)$, we can truncate the dynamics (as in Theorem 3.1) if $\{U_k^\varepsilon\}$ is not bounded: wherever U_k^ε appears in (6.3), we simply replace it by $U_k^\varepsilon q_m(U_k^\varepsilon)$, where $q_m(u) = 1$ for $|u| \leq m$, and is a smooth function with compact support. We get the weak convergence with use of q_m , and then let $m \rightarrow \infty$. The uniqueness of the solution to the limit (6.9) below guarantees that the procedure works. For notational simplicity we simply suppose that $\{U_k^\varepsilon\}$ is bounded. Suppose that $\{W^\varepsilon(\cdot)\}$ is tight and has continuous limits. Then this also holds for $\{U^\varepsilon(\cdot)\}$. Also, the second, third, fourth and last terms on the right side of (6.3) disappear in the limit. The limit of any convergent subsequence satisfies

$$(6.4) \quad U(t) = U(0) + \int_0^t (\bar{\Phi} \bar{B}(\bar{\theta}))_X U(s) ds + W(t)$$

where $W(\cdot)$ is the limit of $\{W^\varepsilon(\cdot)\}$.

6.1. The limits of $\{W^\varepsilon(\cdot)\}$. Under broad conditions $W(\cdot)$ is a Wiener process. The most convenient expression for the covariance matrix depends on the statistical properties of the $\{\Phi_{n+1}, \xi_n\}$. Under the appropriate mixing and stationarity (including the special case in § 7 below), we have

$$(6.5) \quad EW(t)W'(t) = t \sum_{-\infty}^{\infty} E\Phi_{k+1}B(\bar{\theta}, \xi_k)B'(\bar{\theta}, \xi_0)\Phi_1'.$$

We now give some conditions under which $W(\cdot)$ is the asserted Wiener process. Let

$$(6.6) \quad E \left| \sum_n^{n+m-1} \Phi_{k+1} B(\bar{\theta}, \xi_k) \right|^4 \leq \text{Constant} \cdot m^2,$$

then $\{W^\varepsilon(\cdot)\}$ is tight and all limits are continuous [6]. If

$$(6.7) \quad \sum_n^{m+n-1} \Phi_{k+1} B(\bar{\theta}, \xi_k) / \sqrt{m}$$

converges in distribution to a normal random variable (with mean zero) as $n \rightarrow \infty$ and $m \rightarrow \infty$, then $W(\cdot)$ is a Gaussian process. If, for $t_1 \leq t_2 \leq t_3 \leq t_4$,

$$(6.8) \quad E[W^\varepsilon(t_4) - W^\varepsilon(t_3)][W^\varepsilon(t_2) - W^\varepsilon(t_1)]' \xrightarrow{\varepsilon} 0,$$

then the increments of the limit $W(\cdot)$ are orthogonal and the limit is a (nonstandard) Wiener process. The proofs follow standard lines in weak convergence theory [6]. Properties (6.6) and (6.8) hold if the $\{A_k\}$ is independent of the $\{\xi_k\}$ and the dependence among the ξ_k decreases fast enough as the time difference increases. Henceforth we assume that $W(\cdot)$ is the zero mean Wiener process with covariance (6.5).

For the same reasons that the $X(\cdot)$ of § 3 took the form $X(\cdot) = (x(\cdot), \dots, x(\cdot))$ for $x(t) \in E^r$, we have $U(\cdot) = (u(\cdot), \dots, u(\cdot))$ and $W(\cdot) = (w(\cdot), \dots, w(\cdot))$. Then (6.4) reduces to

$$(6.9) \quad du = Mu \, dt + dw.$$

The $w(\cdot)$ is an r -dimensional Wiener process. In fact that $W(\cdot)$ reduces to $(w(\cdot), \dots, w(\cdot))$ is obvious via an examination of the form of W_n^ε , since all the q (r -vector) components of each component of the sum in W_n^ε must be equal. The covariance of $w(1)$ can be obtained from (6.5): if we write

$$\Phi_k = \begin{bmatrix} \phi_1(k), & \dots, & \phi_q(k) \\ \vdots & & \vdots \\ \phi_1(k), & \dots, & \phi_q(k) \end{bmatrix}$$

where the $\phi_i(k)$ are diagonal, (6.5) reduces to

$$(6.10) \quad \text{cov } w(1) = \bar{R} = \sum_{-\infty}^{\infty} E \left[\sum_1^q \phi_i(k+1) b^i(\theta, \xi_k^i) \right] \left[\sum_1^q \phi_i(k+1) b^i(\theta, \xi_k^i) \right]'$$

If $N \rightarrow \infty$ fast enough as $\varepsilon \rightarrow 0$, then the limit $u(\cdot)$ is the stationary solution to (6.9).

The stationary covariance

$$(6.11) \quad \int_0^\infty e^{M't} \bar{R} e^{M't'} dt$$

of (6.9) is a standard measure of the "rate of convergence" or asymptotic quality of the algorithm, and can be used as a basis of comparison among alternative algorithms.

6.2. A special case. We specialize to a simple case in order to get some insight into the asymptotic behavior. Let $\{\xi_k\}$ be independent of $\{A_k\}$ with $\{\xi_k^i, i \leq q, k = 1, 2, \dots\}$ mutually independent with $\text{cov } b^i(\theta, \xi_k^i) \equiv R_i$. Then

$$(6.12) \quad \text{cov } w(1) = \sum_1^q \lim_m \frac{1}{m} \sum_n^{n+m} \phi_i(k) R_i \phi_i(k).$$

6.3. A scalar system. Let $r = 1$. Then $\phi_i(k)$ and $\bar{\phi}_i$ are scalars and

$$\sum_1^q \bar{\phi}_i = 1 = \sum_1^q \phi_i(k).$$

Let $b^i(\cdot, \cdot) = b(\cdot, \cdot)$ and $R_i = R$ not depend on i . Then (6.9) becomes ($\bar{b}_x(\theta) < 0$),

$$(6.13) \quad du = \bar{b}_x(\theta) u \, dt + \sigma_D d\tilde{w}$$

where $\tilde{w}(\cdot)$ is a *standard* Wiener process and (where by the expectation E we mean the ergodic mean in (6.12))

$$\sigma_D^2 = R \sum_1^q E \phi_i^2(n).$$

The stationary variance of $u(\cdot)$ is $\sigma_D^2/2|\bar{b}_x(\theta)| \equiv \text{var}_D$.

6.4. Comparison with a "centralized" algorithm. Define the following centralized algorithm, under the scalar system assumptions of the above paragraph:

$$(6.14) \quad Z_{n+1} = Z_n + \varepsilon b(Z_n, \xi_n^1), \quad \{\xi_n^1, n = 1, 2, \dots\} \text{ i.i.d.}$$

Define $V_n = (Z_n - \theta)/\sqrt{\varepsilon}$ and define $v^\varepsilon(\cdot)$ by $v^\varepsilon(t) = V_n$ on $[n\varepsilon, n\varepsilon + \varepsilon)$. If $t_\varepsilon \rightarrow \infty$ fast enough as $\varepsilon \rightarrow 0$, then under appropriate conditions [8] $v^\varepsilon(t_\varepsilon + \cdot) \Rightarrow v(\cdot)$, where

$$(6.15) \quad dv = \bar{b}_x(\theta)v dt + \sqrt{\bar{R}} d\tilde{w}.$$

The stationary variance of (6.15) is $R/2|\bar{b}_x(\theta)| = \text{var}_C$. Since

$$\frac{\text{var}_D}{\text{var}_C} = E \sum_1^q \phi_i^2(n) < 1,$$

the decentralized algorithm yields an improvement. The infimum of the ratio occurs when the $E\phi_i^2(n)$, $i \leq q$, are all equal, an *unattainable* case (to which we can come close, see § 7). In this limit, $1/q = \text{var}_D/\text{var}_C$. This limit can be obtained if the communications are simultaneous, and the A_i are appropriately symmetric.

A fairer comparison accounts for the fact that the decentralized algorithm uses a total of q observations per iterate. Using the same number in the centralized algorithm (6.14), we rewrite it as

$$(6.16) \quad \bar{Z}_{n+1} = \bar{Z}_n + \frac{\varepsilon}{q} \sum_1^q b(\bar{Z}_n, \xi_n^i), \quad \{\xi_n^i, i \leq q, n = 1, 2, \dots\} \text{ i.i.d.}$$

Define \bar{V}_n^ε and $\bar{v}^\varepsilon(\cdot)$ as the V_n^ε and $v^\varepsilon(\cdot)$ were defined, but based on $\{\bar{Z}_n\}$. Under appropriate conditions $\bar{v}^\varepsilon(t_\varepsilon + \cdot) \Rightarrow \bar{v}(\cdot)$, where

$$(6.17) \quad d\bar{v} = \bar{b}_x(\theta)\bar{v} dt + \sqrt{R/q} d\tilde{w},$$

with stationary covariance $R/2q|b_x(\theta)| = \text{var}_{qC}$ and

$$(6.18) \quad \text{var}_D/\text{var}_{qC} = q \sum_1^q E\Phi_i^2(n) \geq 1.$$

The ratio (6.18) can be used to decide on the proper tradeoff between the asymptotic error and the communication policy. Reasons why the decentralized algorithm might be preferable are discussed in § 7. Analogous results are, of course, obtainable for the general vector case.

7. Asymptotic properties: discussion and comparison.

7.1. Independent $\{A_n\}$. We evaluate $\text{var}_D/\text{var}_C$ under the conditions of the last § 6.4 where $q = 2$ and the $\{A_n\}$ are i.i.d. In particular, let $c \in [0, 1)$, and let the processors act independently, with p = probability that i communicates to $j \neq i$ at time n . With no communication (probability $(1-p)^2$), $A_n = I$; if 2 communicates to 1, but not conversely (probability $p(1-p)$), then

$$A_n = \begin{bmatrix} 1-c & c \\ 0 & 1 \end{bmatrix} = A^{21}.$$

If 1 communicates to 2 (but not conversely), then

$$A_n = \begin{bmatrix} 1 & 0 \\ c & 1-c \end{bmatrix} = A^{12}.$$

If both communicate to each other, then

$$A_n = \begin{bmatrix} 1-c & c \\ c & 1-c \end{bmatrix} = A^0.$$

(See Table 1.) The *optimum* value of the ratio of the variances is unity, a value closely approximated by small c . Clearly a larger p is desirable. As $c \rightarrow 0$, the ratio improves,

TABLE 1
Values for $2 \text{ var}_D / \text{var}_C = \text{var}_D \text{ var}_{2C}$.

$p \backslash c$	0.05	0.25	0.5
0.1	1.036	1.13	1.312
0.3	1.016	1.104	1.26
0.7	1.008	1.048	1.13

but the size of the ψ_n^e would increase. This implies that one must wait longer for the stationary variance to be a good indicator of the actual performance (the effects of the communication are realized more slowly). Similarly, this is the case for small p . But, in all cases, the average performance is much better than that for the centralized algorithm (6.14). To compute (6.18) (and, hence, Table 1) one need only calculate $E\Phi\Phi'$, where $\Phi = \lim_n A_n \cdots A_1$ and the A_i are i.i.d. and distributed as above.

7.2. A deterministic communication scheme. We retain the assumptions of § 7.1, except for those on the communications. Let m and m_1 be integers with $m_1 \leq m/2$. Processor 2 communicates to 1 each m units of time, and 1 communicates to 2 m_1 units of time later. We use A^{12} , A^{21} (when $m_1 \neq 0$) and A^0 (when $m_1 = 0$). For $m_1 = 0$, $(2 \text{ var}_D / \text{var}_C) = 1$, for all $0 < 1 < c$. For $m_1 \neq 0$, we have Table 2. The values of m and m_1 appear only in the values of ψ_n^e , which increases as m and m_1 increase. The values of $\text{var}_D / \text{var}_{2C}$ are substantially worse when processor 1 communicates to processor 2 more often than the reverse communication rate, for deterministic communication times. This suggests that a relatively balanced communication strategy is better and that a processor should “respond” as soon as possible after it receives a “message” from another processor.

7.3. Discussion. It is clear that the decentralized algorithm takes advantage of the possibilities of parallel processing, since its variance is better than that of the classical algorithm (6.14), and can be nearly as good as that of the fully centralized algorithm (6.16). But there is another advantage, which can be considerable. Simulations with recursive algorithms such as (6.14) indicate that a key problem concerns the frequently slow recovery from the effects of large “bursts” of noise, i.e., from a large “random” jump in the state value. This effect would not show up in the asymptotic variance estimates, but is of considerable importance in practice, particularly when the algorithm is not in operation for a very long time. The nature of the “convexification” should often reduce the magnitude of this problem and “robustify” the process. In a sense, the decentralized algorithm would perform much better than the worst of q -identical (but not communicating) processors, and (in a tracking system, for example) would reduce the chances of any one processor losing track. In applications to optimization or systems evaluation by Monte Carlo simulation one can use “variance reduction”

TABLE 2
 $2 \text{ var}_D / \text{var}_C = \text{var}_D / \text{var}_{2C}$.

c	$2 \text{ var}_D / \text{var}_C$
0.1	1.0028
0.3	1.03
0.5	1.11

ideas in choosing appropriate correlations among the sets $\{\xi_n^i, n = 1, 2, \dots\}$, $i \leq q$. We hope that this, together with the above “robustifying” property, would yield good behavior.

7.4. An example. The following is an example which opens up many new possibilities. Consider two receivers—say, digital phase locked loops—each receiving a signal from the same source, but the two being physically separated. Each must estimate the phase or epoch of the signal pulse (and perhaps the phase of the carrier). Suppose that the source is much farther from the receivers than they are from each other, so that more reliable communication between the receivers is possible. It might be possible to improve each other’s estimates by occasional communications. This communication would transfer the estimates—as well as allow the receivers to improve the mutual synchronization of their clocks or oscillators—so that the transferred estimates could be meaningfully used.

7.5. Communication noise. In examples such as the preceding, one would normally have communication noise. This is readily incorporated into the analysis. Write (1.1) as

$$(7.1) \quad X_{n+1} = A_n(X_n + \tilde{\delta}_n) + \varepsilon B(X_n, \xi_n),$$

where $\tilde{\delta}_n$ represents the communication noise. For the algorithm to be useful at all, this noise should be of an order no larger than ε . Then write $\tilde{\delta}_n = \varepsilon \delta_n$, and proceed as before.

Even if $\tilde{\delta}_n = O(\sqrt{\varepsilon})$ and $E\tilde{\delta}_n = 0$, useful results can be obtained. If the interpolation of

$$\left\{ \sum_{k=0}^n \Phi(n|k) \tilde{\delta}_k \right\}$$

converges weakly to a Wiener process $\tilde{W}(\cdot)$, then we might have $X^\varepsilon(\cdot) \Rightarrow X(\cdot)$:

$$dX = \bar{\Phi} \bar{B}(X) dt + d\tilde{W}.$$

Again $X(\cdot)$ takes the form $(x(\cdot), \dots, x(\cdot))$, under appropriate conditions on $\{\tilde{\delta}_n\}$.

7.6. An alternative algorithm. To get additional insight into the behavior of decentralized algorithms, we formally compare (1.1) with a reasonable alternative. Suppose that the processors communicate and “convexify” only the changes in the states since the last communication. In particular, let $q = 2$ and let $\{\tau_n^i\}$, $i = 1, 2$, denote the communication times of the two processors, with $|\tau_{n+1}^i - \tau_n^i|$ bounded. Here processor 2 communicates to processor 1 at $\{\tau_n^1\}$, and similarly for the reverse communication. We proceed purely formally, and suppose that the dynamics are smooth and bounded. Define $\{\tilde{X}_n^i\}$ by $\tilde{X}_{\tau_k^i}^i = X_{\tau_k^i}^i$ and

$$(7.2) \quad \tilde{X}_{n+1}^i = \tilde{X}_n^i + \varepsilon b^i(\tilde{X}_n^i, \xi_n^i), \quad \tau_k^i \leq n < \tau_{k+1}^i - 1.$$

For $\alpha \varepsilon(0, 1/2]$, set

$$(7.3) \quad \begin{aligned} X_{\tau_{k+1}^1}^1 &= X_{\tau_k^1}^1 + (1 - \alpha) \varepsilon \sum_{\tau_k^1}^{\tau_{k+1}^1 - 1} b^1(\tilde{X}_n^1, \xi_n^1) + \alpha \varepsilon \sum_{\tau_k^1}^{\tau_{k+1}^1 - 1} b^2(\tilde{X}_n^2, \xi_n^2), \\ X_{\tau_{k+1}^2}^2 &= X_{\tau_k^2}^2 + \alpha \varepsilon \sum_{\tau_k^2}^{\tau_{k+1}^2 - 1} b^1(\tilde{X}_n^1, \xi_n^1) + (1 - \alpha) \varepsilon \sum_{\tau_k^2}^{\tau_{k+1}^2 - 1} b^2(\tilde{X}_n^2, \xi_n^2). \end{aligned}$$

Owing to the smoothness and boundedness assumptions, there are $O_n^i(\varepsilon^2) = O(\varepsilon^2)$ and a process $\hat{X}_n = (\hat{X}_n^1, \hat{X}_n^2)$ satisfying (7.4) and which equals (modulo $O(\varepsilon^2)$) $(\hat{X}_n^1, \hat{X}_n^2)$ and $(X_{\tau_k}^1, X_{\tau_k}^2)$ (at the communication times)

$$(7.4) \quad \begin{aligned} \hat{X}_{n+1}^1 &= \hat{X}_n^1 + (1-\alpha)\varepsilon b^1(\hat{X}_n^1, \xi_n^1) + \alpha\varepsilon b^2(\hat{X}_n^2, \xi_n^2) + O_n^1(\varepsilon^2), \\ \hat{X}_{n+1}^2 &= \hat{X}_n^2 + \alpha\varepsilon b^1(\hat{X}_n^1, \xi_n^1) + (1-\alpha)\varepsilon b^2(\hat{X}_n^2, \xi_n^2) + O_n^2(\varepsilon^2). \end{aligned}$$

The “size” of the O_n^i depend on the bound on $|\tau_{k+1}^i - \tau_k^i|$. From this point on, one can use standard theory for the centralized case to get both the ODE and the asymptotic normalized variance. Define $\hat{X}^\varepsilon(\cdot)$ as $X^\varepsilon(\cdot)$ was defined, and similarly for $\hat{U}^\varepsilon(\cdot)$. The limit ODE is

$$(7.5) \quad \dot{\hat{X}} = \frac{(1-\alpha)\bar{b}^1(\hat{X}^1) + \alpha\bar{b}^2(\hat{X}^2)}{\alpha\bar{b}^1(\hat{X}^1) + (1-\alpha)\bar{b}^2(\hat{X}^2)} = \hat{B}(\hat{X}^1, \hat{X}^2).$$

The limit $\hat{U}(\cdot)$ of $\{\hat{U}^\varepsilon(\cdot)\}$ satisfies

$$(7.6) \quad d\hat{U} = \dot{M}\hat{U}dt + d\hat{W},$$

where

$$\begin{aligned} \text{cov } \hat{W}(1) &= \sum_{-\infty}^{\infty} E\bar{\xi}_n\bar{\xi}_0', \\ \bar{\xi}_n &= \begin{bmatrix} (1-\alpha)b^1(\theta, \xi_n^1) + \alpha b^2(\theta, \xi_n^2) \\ \alpha b^1(\theta, \xi_n^1) + (1-\alpha)b^2(\theta, \xi_n^2) \end{bmatrix}, \\ \dot{M} &= (\hat{B}(\theta, \theta))_X, \end{aligned}$$

and we suppose that \dot{M} is a stable matrix.

7.7. Comparison of the alternative (7.2), (7.3) with (1.1). We use the special scalar case of § 7.1, where $\{\xi_n^i\}$ are i.i.d. and $b^1(\cdot, \cdot) = b^2(\cdot, \cdot) = b(\cdot, \cdot)$. Then (again $\bar{b}_x(\theta) < 0$)

$$\begin{aligned} \dot{\hat{X}} &= \begin{bmatrix} (1-\alpha)\bar{b}(\hat{X}^1) + \alpha\bar{b}(\hat{X}^2) \\ \alpha\bar{b}(\hat{X}^1) + (1-\alpha)\bar{b}(\hat{X}^2) \end{bmatrix}, \\ d\hat{U} &= \bar{b}_x(\theta) \begin{bmatrix} (1-\alpha) & \alpha \\ \alpha & (1-\alpha) \end{bmatrix} \hat{U}dt + d\hat{W} = M\hat{U}dt + d\hat{W}, \\ \text{cov } \hat{W}(1) &= E b^2(\theta, \xi_n^i) \begin{bmatrix} (1-\alpha)^2 + \alpha^2 & 2\alpha(1-\alpha) \\ 2\alpha(1-\alpha) & (1-\alpha)^2 + \alpha^2 \end{bmatrix}. \end{aligned}$$

Let var_{D2} denote the stationary variance of $\hat{U}(\cdot)$. As $\alpha \uparrow \frac{1}{2}$, this converges to the infimal value, equal to var_{2C} . But at $\alpha = \frac{1}{2}$, the matrix M is singular. Thus, again, there seems to be a trade-off between the “minimal asymptotic variance” and the length of time one must wait for the asymptotic estimates to be valid or, similarly, for the communication to be effective. At this point, the alternative algorithm does not seem to offer any clear advantages. It was investigated simply because of the idea that there might be an advantage in communicating only recent data.

8. Stochastic approximation with $\varepsilon_n \rightarrow 0$. The entire development can be repeated if ε is replaced by $0 < \varepsilon_n \rightarrow 0$, $\sum \varepsilon_n = \infty$. One then gets results of classical stochastic approximation type, and we only make a few formal comments. We use $X_{n+1} = A_n X_n + \varepsilon_n b(X_n, \xi_n)$. Define $t_n = \sum_{i=0}^{n-1} \varepsilon_i$ and (for $t \geq 0$) define $X^n(\cdot)$ by $X^n(t) = X_{n+i}$ for $t \in [t_{n+i} - t_n, t_{n+i+1} - t_n)$. Under the conditions of Theorem 5.1, $\overline{\lim}_n EV(X_n) < \infty$.

Given either this or the use of the projection algorithm of § 4, one can get the appropriate ODE which characterizes the limit paths. If this has the appropriate stability properties (as in § 5), we can show that $X^{\varepsilon,i}(\cdot) \Rightarrow x(\cdot) \equiv \theta$. The ODE is the same as that in the previous sections, for all the same cases.

If $\sum \varepsilon_n^2 < \infty$, then the idea in [10] can be adapted to get w.p.1 convergence results.

Appendix A. Some results in weak convergence. For some integer s , let $D[0, \infty)$ denote the space of E^s -valued functions on $[0, \infty)$ which are right continuous and have left-hand limits, with the Skorokhod topology [7, Chap. 2]. This topology is defined as follows. Let A be the set of strictly increasing Lipschitz continuous functions from $[0, \infty)$ onto $[0, \infty)$. Define the metric

$$d(x(\cdot), y(\cdot)) = \inf_{\lambda \in A} \max \left\{ \sup_{s > t \geq 0} \left| \log \left(\frac{\lambda(s) - \lambda(t)}{s - t} \right) \right|, \int_0^\infty e^{-\tau} d_\tau(x(\cdot), y(\cdot), \lambda) d\tau \right\}$$

where $d_\tau(x(\cdot), y(\cdot), \lambda) = \min(1, \sup_t |x(\lambda(t) \cap \tau) - y(t \cap \tau)|)$.

Define $\{Z_n^\varepsilon\}$ and $\{Z^\varepsilon(\cdot)\}$ by $Z_{n+1}^\varepsilon = Z_n^\varepsilon + \varepsilon F_n^\varepsilon$, $Z^\varepsilon(t) = Z_n^\varepsilon$ for $t \in [n\varepsilon, (n+1)\varepsilon)$. If $\{Z_0^\varepsilon\}$ is tight and the $\{F_n^\varepsilon\}$ are uniformly integrable, then $\{Z^\varepsilon(\cdot)\}$ is tight in $D[0, \infty)$ and all weak limits are absolutely continuous.

Let $Z^\varepsilon(\cdot) \Rightarrow Z(\cdot)$ in $D[0, \infty)$. By a suitable choice of the probability space, the weak convergence becomes convergence w.p.1 in the metric of $D[0, \infty)$ [13, Thm. 3.1.1]. That is, there is a probability space $(\tilde{\Omega}, \tilde{B}, \tilde{P})$ with processes $\{\tilde{Z}^\varepsilon(\cdot)\}$, $\tilde{Z}(\cdot)$ defined on it so that for each Borel set B in $D[0, \infty)$, $\tilde{P}\{\tilde{Z}^\varepsilon(\cdot) \in B\} = P\{Z^\varepsilon(\cdot) \in B\}$, $\tilde{P}\{\tilde{Z}(\cdot) \in B\} = P\{Z(\cdot) \in B\}$ and $\tilde{Z}^\varepsilon(\cdot) \rightarrow \tilde{Z}(\cdot)$ w.p.1 in the topology of $D[0, \infty)$. The use of this representation often facilitates the analysis and characterization of the limits.

Appendix B. Proof of Lemma 2.1. The proofs of (2.1a) and (2.1b) are essentially the same and only (2.1a) will be proved. We will evaluate $E|\Phi(n|k) - \Phi_k|$ by a slight variation of the proof of [1, Lemma 5.2.1]. Owing to the block diagonal structure of the A_n , when we calculate the product $\Phi(n|k)$, the r sets of rows $(i, i+r, \dots, i+qr-r)$, $i \leq r$, do not interact and we can (and will) let $r=1$ without loss of generality.

The geometric convergence of $\Phi(n|k)$ to Φ_k was proved in [1, Lemma 5.2.1] when $p_0=1$ (see (A2.4) below). By (C2.1), there are $\alpha_1 > 0$ and an increasing sequence of (finite w.p.1) random times $\{N_i\}$ such that the components of $\Phi(N_{2i+1}|N_{2i})$ are all $\geq \alpha_1$ w.p.1. This and the convergence result for $p_0=1$ imply that $\Phi(n|k)$ converges w.p.1 to some matrix Φ_k as $n \rightarrow \infty$. All the rows of such a limit must be equal, and the entries of each row must sum to unity. Let $\phi_1(k), \dots, \phi_q(k)$ denote the scalar elements of any row of Φ_k , and let the vectors v_1, \dots, v_q span E^q , and define $e = (1, 1, \dots, 1)$. Define $c(x) = \sum \phi_i(k)x_i$ where $x = (x_1, \dots, x_q)$. Both e and x are column vectors. Then $\Phi_k x = c(x)e$. All norms here and elsewhere are in the l_∞ sense.

For a matrix M ,

$$(A2.1) \quad |M| = \sup_{|x|=1} |Mx| \leq \sum |Mv_i|.$$

Thus, we need only show, for any vector x , that $E|\Phi(n|k)x - c(x)e| \xrightarrow{n} 0$ geometrically. Define $x(n|k) = \Phi(n|k)x$. Let $c(n|k)$ denote the minimum value of the components of $x(n|k)$. We can write $x(n|k) = y(n|k) + c(n|k)e$, where all the components of $y(n|k)$ are nonnegative and

$$(A2.2) \quad E|\Phi(n|k)x - c(x)e| \leq E|y(n|k)| + E|c(x) - c(n|k)|.$$

By the "convexification" properties of the A_n ,

$$(A2.3a) \quad |y(n+1|k)| \leq |y(n|k)|,$$

$$(A2.3b) \quad c(n|k) \leq c(n+1|k) \leq c(n|k) + |y(n|k)|.$$

By (C2.1), there is an $\alpha_0 > 0$ such that w.p. p_0 (conditioned on F_n) all the elements of $\Phi(n + m_0|n)$ are $\geq \alpha_0$. This, together with the "convexification" property of the A_n implies that

$$(A2.4) \quad E|y(n + m_0|k)| \leq (1 - \alpha_0 p_0) E|y(n|k)|.$$

(If $p_0 = 1$, then drop the E in (A2.4), and (A2.3), (A2.4) yield w.p.1 convergence.) The asserted geometric convergence is a consequence of (A2.3), (A2.4) and the w.p.1 convergence of $\Phi(n|k)$ (hence of $c(n|k)$ to $c(x)$). The last sentence of the lemma follows by a similar argument. Q.E.D.

REFERENCES

- [1] J. N. TSITSIKLIS, *Problems in decentralized decision making and computation*, Ph.D. thesis, Electrical Engineering Dept., Massachusetts Inst. of Technology, Cambridge, MA, 1984.
- [2] D. BERTSEKAS, J. N. TSITSIKLIS AND M. ATHANS, *Convergence theories of distributed iterative processes: A survey*, Technical Report for Information and Decision Systems, Massachusetts Inst. of Technology, Cambridge, MA, 1984.
- [3] A. P. KOROSTELEV, *Stochastic Recurrent Processes*, Nauka, Moscow, 1984.
- [4] P. DUPUIS AND H. J. KUSHNER, *Stochastic approximations via large deviations: Asymptotic properties*, this Journal, 23 (1985), pp. 675-696.
- [5] H. J. KUSHNER AND HAI HUANG, *Averaging methods for the asymptotic analysis of learning and adaptive systems with small adjustment rate*, this Journal, 19 (1981), pp. 635-650.
- [6] P. BILLINGSLEY, *Convergence on Probability Measures*, John Wiley, New York, 1968.
- [7] T. G. KURTZ, *Approximation of Population Processes*, in CBMS-NSF Regional Conference Series in Appl. Mathematics 36, Society for Industrial and Applied Mathematics, Philadelphia, PA, 1981.
- [8] H. J. KUSHNER, *Approximation and Weak Convergence Methods for Random Processes with Applications to Stochastic Systems Theory*, MIT Press, Cambridge, MA, 1984.
- [9] H. J. KUSHNER AND A. SHWARTZ, *An invariant measure approach to the convergence of stochastic approximations with state-dependent noise*, this Journal, 22 (1984), pp. 13-27.
- [10] H. J. KUSHNER, *An averaging method for stochastic approximations with discontinuous dynamics, constraints, and state-dependent noise*, in Recent Advances in Stochastics, M. H. Rizri, J. S. Rustagi and D. Siegmund, eds., Academic Press, New York, 1983.
- [11] G. BLANKENSHIP AND G. C. PAPANICOLAOU, *Stability and control of stochastic systems with wide band noise disturbances*, SIAM J. Appl. Math., 34(1978), pp. 437-476.
- [12] H. J. KUSHNER AND HAI HUANG, *Asymptotic properties of stochastic approximations with constant coefficients*, this Journal, 19 (1981), pp. 87-105.
- [13] A. V. SKOROKHOD, *Limit theorems for stochastic processes*, Theory Probab. Appl., 1 (1956), pp. 262-290.
- [14] H. J. KUSHNER AND G. YIN, *Stochastic approximation algorithms for parallel and distributed processing*, Lefschetz Center for Dynamical Systems Rept. 86-31, Brown Univ., Providence, RI, Stochastics, to appear.

THE RELATIONSHIP BETWEEN THE MAXIMUM PRINCIPLE AND DYNAMIC PROGRAMMING*

FRANK H. CLARKE† AND RICHARD B. VINTER‡

Abstract. Let $V(t, x)$ be the infimum cost of an optimal control problem, viewed as a function of the initial time and state (t, x) . Dynamic Programming is concerned with the properties of $V(\cdot, \cdot)$ and in particular with its characterization as a solution to the Hamilton–Jacobi–Bellman equation. Heuristic arguments have long been advanced relating the Maximum Principle to Dynamic Programming according to

$$p(t) = -V_x(t, x_0(t)).$$

Here $x_0(\cdot)$ is the minimizing state function under consideration and $p(\cdot)$ is the costate function of the Maximum Principle. In this paper we examine the validity of such claims and find that this relationship, interpreted as a differential inclusion involving the generalized gradient, is indeed true, almost everywhere and at the endpoints, for a very large class of nonsmooth optimal control problems.

Key words. optimal control, Maximum Principle, Dynamic Programming

AMS(MOS) subject classifications. 49B10, 49C05

1. Introduction. Our purpose in this paper is to relate the costate function which appears in the Maximum Principle and the value function associated with perturbations in the initial time and state, and thereby to clarify the relationship between the Maximum Principle and Dynamic Programming. For the most part our framework is that of the following free final state optimal control problem:

$$\begin{aligned} & \text{Minimize} && \int_0^1 L(t, x(t), u(t)) dt + h(x(1)) \\ (1.1) \quad & \text{subject to} && \dot{x}(t) = f(t, x(t), u(t)) \quad \text{a.e. } t \in [0, 1], \\ & && x(0) = x_0, \\ (1.2) \quad & && u(t) \in U_t \quad \text{a.e. } t \in [0, 1], \\ (1.3) \quad & && x(t) \in \Omega_t \quad \text{all } t \in [0, 1]. \end{aligned}$$

The data for the problem are the following: functions $f: [0, 1] \times \mathbb{R}^n \times \mathbb{R}^m \rightarrow \mathbb{R}^n$, $L: [0, 1] \times \mathbb{R}^n \times \mathbb{R}^m \rightarrow \mathbb{R}$, $h: \mathbb{R}^n \rightarrow \mathbb{R}$, a vector $x_0 \in \mathbb{R}^n$, and sets

$$U \in [0, 1] \times \mathbb{R}^m, \quad \Omega \in [0, 1] \times \mathbb{R}^n.$$

(For any set $A \subset [0, 1] \times \mathbb{R}^m$ and point $t \in [0, 1]$ we denote by A_t the set $\{x: (t, x) \in A\}$.)

Given a subinterval $I \subset [0, 1]$, a *control function* (on I) is a (Lebesgue) measurable function $u(\cdot): I \rightarrow \mathbb{R}^m$ which satisfies (1.2) (when I replaces $[0, 1]$).

A *process* (on I) is a pair of functions $(x(\cdot), u(\cdot))$ on I of which $u(\cdot)$ is a control function and $x(\cdot): I \rightarrow \mathbb{R}^n$ is an absolutely continuous function which satisfies the differential equation (1.1) and the constraint (1.3) (when I replaces the interval $[0, 1]$), and for which $s \mapsto L(s, x(s), u(s))$ is integrable. If $x(\cdot)$ is the first component of some process, $x(\cdot)$ is called a *state function*. A process $(x(\cdot), u(\cdot))$ on $[0, 1]$ is *admissible* if $x(0) = x_0$. The process is *minimizing* if it minimizes the value of the cost function $\int L dt + h$ over all admissible processes.

* Received by the editors January 6, 1986; accepted for publication (in revised form) November 5, 1986.

† Centre de Recherche de Mathématiques, Université de Montréal, C.P. 6128 Succursale A, Montréal, Quebec, Canada H3C 3J7. The work of this author was supported by the Natural Sciences and Engineering Research Council of Canada.

‡ Department of Electrical Engineering, Imperial College, London SW7 2BT, England.

A tube T is a subset of Ω of the form

$$\{(t, x): \|x - z(t)\| < \delta\}$$

for some $\delta > 0$ and some continuous function $z(\cdot)$. We also refer to this set as the δ -tube about $z(\cdot)$. We loosely refer to "state functions in a tube T about $z(\cdot)$ " when we have in mind state functions having graphs in a tube about $z(\cdot)$.

Let $(x_0(\cdot), u_0(\cdot))$ be a minimizing process contained in some tube.

Under certain conditions (the precise nature of which is unimportant at this juncture) the Maximum Principle asserts: there exists a costate function $p(\cdot): [0, 1] \rightarrow \mathbb{R}^n$ such that

$$(1.4) \quad -\dot{p}(t) = p(t) \cdot f_x(t, x_0(t), u_0(t)) - L_x(t, x_0(t), u_0(t)) \quad \text{a.e. } t \in [0, 1],$$

$$(1.5) \quad -p(1) = h_x(x_0(1))$$

and, a.e. $t \in [0, 1]$,

$$(1.6) \quad p(t) \cdot f(t, x_0(t), u_0(t)) - L(t, x_0(t), u_0(t)) = \max_{u \in U_t} \{p(t) \cdot f(t, x(t), u) - L(t, x_0(t), u)\}.$$

Now define the value function $V(\cdot, \cdot): [0, 1] \times \mathbb{R}^n \rightarrow \bar{\mathbb{R}}$ to be

$$(1.7) \quad V(t, x) = \inf \left\{ \int_t^1 L(s, x(s), u(s)) ds + h(x(1)) \right\}.$$

Here the infimum is taken over processes $(x(\cdot), u(\cdot))$ on $[t, 1]$ for which $x(t) = x$. When no such processes exist the value is $+\infty$.

If $V(\cdot, \cdot)$ happens to be continuously differentiable on the interior of some tube T about $x_0(\cdot)$ we find under mild assumptions that $V(\cdot, \cdot)$ satisfies the Hamilton-Jacobi-Bellman equation there:

$$(1.8) \quad V_t(t, x) + \min_{u \in U_t} \{V_x(t, x) \cdot f(t, x, u) + L(t, x, u)\} = 0, \quad (t, x) \in T.$$

Also

$$(1.9) \quad V(1, x) = h(x), \quad x \in \Omega_1,$$

and

$$(1.10) \quad (V_t + V_x \cdot f + L)(t, x_0(t), u_0(t)) = 0 \quad \text{a.e. } t \in [0, 1].$$

Solving (1.8) and (1.9) for $V(\cdot, \cdot)$ is the basis of the Dynamic Programming approach in Optimal Control Theory. In introductory treatments of optimal control theory the connection between the Maximum Principle and Dynamic Programming is usually described in the following terms (here we quote from [4, pp. 297ff.]): "the adjoint (i.e., costate) variables \cdots equal the negative of the rate of change of the performance index with respect to the corresponding state variables." This simply means

$$(1.11) \quad -\dot{p}(t) = V_x(t, x_0(t)).$$

See also [3] and [10].

The grounds for this assertion are as follows: (1.8) and (1.10) imply the following for each $t \in [0, 1]$ and x in a neighborhood of $x_0(t)$:

$$(1.12) \quad (V_t + V_x f + L)(t, x_0(t), u_0(t)) \leq (V_t + V_x f + L)(t, x, u_0(t)),$$

and for each $t \in [0, 1]$ and $u \in U_t$

$$(1.13) \quad (V_x f + L)(t, x_0(t), u_0(t)) \leq (V_x f + L)(t, x_0(t), u).$$

Now write

$$-q(t) = V_x(t, x_0(t)).$$

Then (1.13) amounts to the "maximization of the Hamiltonian condition" (1.6) (when $p(\cdot)$ replaces $q(\cdot)$). If for each t the right-hand side of (1.12) is a continuously differentiable function of x , we deduce from this inequality that

$$\frac{\partial}{\partial x} [V_t(t, x) + V_x(t, x)f(t, u_0(t), x) + L(t, x, u_0(t))] \Big|_{x=x_0(t)} = 0.$$

If we also assume $V(\cdot, \cdot)$ is twice continuously differentiable, and $f(\cdot, \cdot, \cdot)$ and $L(\cdot, \cdot, \cdot)$ are continuously differentiable in x , then we obtain

$$(V_{tx} + V_{xx}f + V_x f_x + L_x)(t, x_0(t), u_0(t)) = 0.$$

But

$$\frac{d}{dt} V_x(t, x_0(t)) = (V_{tx} + V_{xx}f)(t, x_0(t), u_0(t))$$

whence $q(t)$ ($= -V_x(t, x_0(t))$) satisfies the costate differential equation (1.4) for $p(\cdot)$. Finally (1.9) implies

$$-q(1) = V_x(1, x(1)) = h_x(x_0(1)),$$

which is the transversality condition (1.5).

The relationship between the Maximum Principle and Dynamic Programming would appear then to have been explained. But such impressions are illusory. Indeed it is difficult to justify our heuristic arguments in a setting of any generality. They are based on (among other things) the assumption that $V(\cdot, \cdot)$ is smooth and, to quote Pontryagin et al. [12, p. 7], "It must be noted that the assumption on the continuous differentiability of the functional does not hold in the simplest cases." The point is also emphasized by Berkovitz [2] and Fleming and Rischel [9].

This leads to our central preoccupation in this paper: the extent to which the relationship (1.11) bridging the Maximum Principle and the Dynamic Programming technique is valid, under more or less the weakest hypotheses evoked in proof of the Maximum Principle.

While it is not reasonable to assume continuous differentiability of $V(\cdot, \cdot)$, we can expect $V(t, \cdot)$ to be Lipschitz continuous for each $t \in [0, 1]$. In the circumstances it is natural to replace (1.11) by the inclusion

$$(1.14) \quad -p(t) \in \partial_x V(t, x_0(t))$$

where $\partial_x V$ denotes the partial generalized gradient.

Our main result is that, for a very large class of nonsmooth free endpoint problems, we can choose a costate function $p(\cdot)$ which satisfies the inclusion (1.14) for $t = 0$ and 1, and on a subset of $(0, 1)$ of full measure. We are able to show also that the inclusion (1.14) is still valid for some costate function when a terminal constraint of the form " $x(1) \in C_1$ " is present, provided some verification function is substituted for $V(\cdot, \cdot)$. Here we take advantage of recent advances in our understanding of verification functions [6] that permit us to replace the terminally constrained problem by an equivalent free endpoint problem involving a nonsmooth penalty term and thereby to reduce the problem to one we already know how to deal with.

We conclude this introduction by describing the basic idea behind our methods. For any measurable function $\alpha(\cdot): [0, 1] \rightarrow \{\xi: \|\xi\| < 1\}$ and control function $u(\cdot)$ on $[0, 1]$, let $x(\cdot)$ be a solution to

$$(1.15) \quad \dot{x}(t) = f(t, x(t), u(t)) + \alpha(t) \quad \text{a.e. } t \in [0, 1].$$

If (1.8) were valid we could write

$$(V_t + V_x \cdot (f + \alpha) - V_x \alpha)|_{(t, x(t), \alpha = \alpha(t))} + L(t, x(t), u(t)) \geq 0$$

for all $t \in [0, 1]$. Integrating over $[0, 1]$ we obtain

$$\int_0^1 \frac{d}{dt} V(t, x(t)) dt - \int_0^1 V_x(t, x(t)) \alpha(t) dt + \int_0^1 L(t, x(t), u(t)) dt \geq 0,$$

i.e.,

$$(1.16) \quad \int_0^1 L(t, x(t), u(t)) dt + h(x(1)) - V(0, x(0)) - \int_0^1 V_x(t, x(t)) \alpha(t) dt \geq 0.$$

We now treat the left-hand side of (1.16) as a cost function for an auxiliary optimal control problem with dynamics (1.15), where $\alpha(\cdot)$ is treated as an additional control variable and where the initial and terminal values of the state function are unconstrained. Since

$$\int_0^1 L(t, x_0(t), u_0(t)) dt + h(x_0(1)) - V(0, x_0(0)) = 0$$

by definition of the value function, it follows from (1.16) that $(x_0(\cdot), (u_0(\cdot), \alpha(\cdot) \equiv 0))$ is a solution to the auxiliary optimal control problem. Now apply the Maximum Principle to the auxiliary problem at $(x_0(\cdot), (u_0(\cdot), \alpha(\cdot) \equiv 0))$. This leads to the conclusion that there is an absolutely continuous function $p(\cdot): [0, 1] \rightarrow \mathbb{R}^n$ satisfying the usual conditions in the Maximum Principle for the original problem (i.e., $p(\cdot)$ is a costate function). But the presence of the terms involving $V(\cdot, \cdot)$ in the cost and the $\alpha(\cdot)$ control in the dynamics implies in addition

$$-p(t) = V_x(t, x_0(t)) \quad \text{a.e. } t \in [0, 1]$$

(we get this from the maximization of the Hamiltonian condition), and

$$-p(0) = V_x(0, x_0)$$

(from the transversality condition). These are the desired relations.

Of course we are not, in general, justified in using (1.8) because we cannot expect $V(\cdot, \cdot)$ to be continuously differentiable. However, a kind of nonsmooth version of the inequality (1.16) can be proved. This provides an auxiliary problem. Applying the Maximum Principle to that problem gives a costate function with the hoped-for properties. Derivation of the nonsmooth inequality involves consideration of state functions, in some extended sense, which are discontinuous, and some rather delicate approximations.

Earlier work is available on the connection between Dynamic Programming and the Maximum Principle, in the absence of restrictive hypotheses. This takes the form of necessary conditions expressed in terms of generalized gradients of the value function. It is established in [14] that the pseudo-Hamiltonian function evaluated along

some selector of $t \rightarrow \partial_x V(t, x_0(t))$ is maximized at the optimal control. See also [11] for a result of this nature. Showing that we can also arrange for the selector to satisfy the costate differential inclusion in a general nonsmooth setting, and thereby express the full Maximum Principle in terms of the value function, is a more formidable task. This is accomplished for the first time in the present paper. Expanded versions of some of the proofs involved are available in [7].

2. Hypotheses. Throughout the paper $(x_0(\cdot), u_0(\cdot))$ is taken to be a fixed minimizing process for the optimal control problem. Define \tilde{f} :

$$\tilde{f} = \begin{bmatrix} L \\ f \end{bmatrix}.$$

The following hypotheses remain always in force:

- (H1) $(t, u) \rightarrow \tilde{f}(t, x, u)$ is $\mathcal{L} \times \mathcal{B}^m$ measurable for each fixed $x \in \mathbb{R}^n$, where $\mathcal{L} \times \mathcal{B}^m$ denotes the σ -algebra, generated by product sets whose first element is a Lebesgue subset of $[0, 1]$ and whose second is a Borel set in \mathbb{R}^m .
- (H2) U is a Borel measurable set.
- (H3) For each $t \in [0, 1]$ and $u \in U_t$, $\tilde{f}(t, \cdot, u)$ is Lipschitz continuous on Ω_t , with rank at most $k(t)$ ($k(t)$ does not depend on u), and $k(\cdot) \in L^1(0, 1)$.
- (H4) There exists a function $c(\cdot) \in L^1(0, 1)$ such that

$$\|\tilde{f}(t, x_0(t), u)\| \leq c(t) \quad \text{for all } u \in U_t, t \in [0, 1].$$

- (H5) $h(\cdot)$ is a locally Lipschitz continuous function.

As before (see (1.7)), we take the value function $V(\cdot, \cdot): \Omega \rightarrow \bar{\mathbb{R}}$ to be

$$V(t, x) = \inf \left\{ \int_t^1 L(s, x(s), u(s)) ds + h(x(1)) \right\}$$

where the infimum is taken over processes on $[t, 1]$ which satisfy $x(t) = x$, and once again we set the value to $+\infty$ when no such processes exist.

Concerning $V(\cdot, \cdot)$ we impose the following.

- (H6) There exists a tube Ω' about $x_0(\cdot)$, $\Omega' \subset \Omega$, and a constant r such that for each $t \in [0, 1]$, $V(\cdot, t)$ is Lipschitz continuous on Ω_t with rank at most r .

Hypothesis (H6) on the value function is very mild. In fact it is almost superfluous, since one is most often interested in examples of the optimal control problem in which $\Omega = [0, 1] \times \mathbb{R}^n$; for such examples (H6) is implied by the other hypotheses. See [7], where this and some other properties of the value function are proved.

3. The main result. Define the pseudo-Hamiltonian function $H(\cdot, \cdot, \cdot, \cdot)$:

$$H(t, x, p, u) = p \cdot f(t, x, u) - L(t, x, u).$$

THEOREM 3.1. *There exists an absolutely continuous function $p(\cdot): [0, 1] \rightarrow \mathbb{R}^n$ and a subset $\Sigma \subset [0, 1]$ of full Lebesgue measure such that*

$$(3.1) \quad -\dot{p}(t) \in \partial_x H(t, x_0(t), p(t), u_0(t)) \quad \text{a.e. } t \in [0, 1],$$

$$(3.2) \quad H(t, x_0(t), p(t), u_0(t)) = \max_{u \in U_t} H(t, x_0(t), p(t), u) \quad \text{a.e. } t \in [0, 1]$$

and

$$(3.3) \quad -p(t) \in \partial_x V(t, x_0(t)) \quad \text{for all } t \in \{0\} \cup \{1\} \cup \Sigma.$$

Here $\partial_x(\cdot)$ denotes the partial generalized gradient in the x -variable [5, p. 63], i.e.,

$$\partial_x V(t, x) = \text{co} \{ \lim V_x(t, x_i) : x_i \rightarrow x, V(t, \cdot) \text{ is differentiable at } x_i, i = 1, 2, \dots \}.$$

Notice that the theorem incorporates the Maximum Principle since the new condition on the costate function (3.3) implies the transversality condition

$$-p(1) \in \partial_x h(x_0(1)),$$

the only component of the Maximum Principle not explicitly present. This follows from the fact that $V(1, \cdot) = h(\cdot)$.

4. An example. For nonsmooth problems it is possible that, associated with the minimizing process $(x_0(\cdot), u_0(\cdot))$, there are many possible costate functions. (The set of costate functions comprises all absolutely continuous functions $p(\cdot)$ which satisfy the costate equation (3.1), have the "maximization of the Hamiltonian" property, and for which $-p(1) \in \partial_x h(x_0(1))$.)

One might speculate that any costate function $p(\cdot)$ satisfies the inclusion

$$-p(t) \in \partial_x V(t, x_0(t))$$

(almost everywhere and at the endpoints). The purpose of the following example is to illustrate that this is not the case.

$$\begin{aligned} &\text{Minimize} && g(x(1)) \\ &\text{subject to} && \dot{x}(t) = x(t)u(t) \quad \text{a.e. } t \in [0, 1], \\ &&& x(0) = 0, \\ &&& u(t) \in [0, 1] \quad \text{a.e. } t \in [0, 1], \\ &&& x(t) \in \mathbb{R}, \quad \text{all } t \in [0, 1]. \end{aligned}$$

Here

$$g(x) = \begin{cases} -x & \text{if } x > 0, \\ -e^{1/2}x & \text{if } x \leq 0. \end{cases}$$

A minimizing process is $\{x_0(\cdot) \equiv 0, u_0(\cdot) \equiv 0\}$. Indeed $x(\cdot) \equiv 0$ is the only possible state function. In order to evaluate $V(t, x)$, $(t, x) \in [0, 1] \times \mathbb{R}^n$, we must look at the associated control problem with initial time t and initial state x . The terminal cost function is monotonic decreasing, so it is minimized by making $x(1)$ as large as possible. We achieve this by selecting $u(\cdot) \equiv 1$ if $x > 0$ and $u(\cdot) \equiv 0$ if $x \leq 0$. We can now calculate $V(t, x)$:

$$V(t, x) = \begin{cases} -e^{(1-t)}x & \text{if } x > 0, \\ -e^{1/2}x & \text{if } x \leq 0. \end{cases}$$

The information supplied by the Maximum Principle concerning the minimizing process $(x_0(\cdot), u_0(\cdot))$ is: there exists $p(\cdot) : [0, 1] \rightarrow \mathbb{R}$ satisfying

$$-\dot{p}(t) = u_0(t) \quad \text{a.e. } t \in [0, 1]$$

such that

$$p(t)x_0(t)u_0(t) = \max_{u \in [0, 1]} p(t)x_0(t)u \quad \text{a.e. } t \in [0, 1],$$

and because the generalized gradient of $g(\cdot)$ at $x_0(1)$ is

$$\partial_x g(x_0(1)) = [-e^{1/2}, -1],$$

it follows that

$$-p(1) \in [-e^{1/2}, -1].$$

But $x_0(\cdot) \equiv 0$, $u_0(\cdot) \equiv 0$. A costate function then is a function $p(\cdot)$ satisfying

$$p(\cdot) \equiv \lambda, \quad \lambda \in [-e^{1/2}, -1].$$

Notice that we can arrange that

$$-p(t) \in \partial_x V(t, x_0(t)), \quad \text{all } t \in [0, 1]$$

by choosing $\lambda = -e^{1/2}$. But this inclusion fails (on a set of positive measure) if any other value of λ is adopted.

In this example then there are uncountably many costate functions, but only one of them satisfies the inclusion.

5. Some special cases. The assertions of Theorem 3.1 do not exclude the possibility that $-p(t) \notin \partial_x V(t, x_0(t))$ for all t 's in some null set in $(0, 1)$. This is probably unavoidable under the hypotheses considered. However, the question arises whether this null set can be eliminated in special circumstances. We seek then additional hypotheses under which (3.3) can be strengthened to

$$(5.1) \quad -p(t) \in \partial_x V(t, x_0(t)) \quad \text{for all } t \in [0, 1].$$

One direction in which we can proceed is to introduce regularity hypotheses on the multifunction $t \rightarrow \partial_x V(t, x_0(t))$.

PROPOSITION 5.1. *Suppose that the multifunction $t \rightarrow \partial_x V(t, x_0(t))$ is upper semicontinuous on $[0, 1]$. Then condition (3.3) can be strengthened to (5.1).*

Proof. Take any $t \in [0, 1]$. Let $\{t_i\}$ be a sequence such that $t_i \rightarrow t$ as $i \rightarrow \infty$ and $p(t_i) \in \partial_x V(t_i, x_0(t_i))$ for $i = 1, 2, \dots$. Since $p(t_i) \rightarrow p(t)$ and $t_i \rightarrow t$, as $i \rightarrow \infty$, it follows from upper semicontinuity that $p(t) \in \partial_x V(t, x_0(t))$. \square

COROLLARY 5.2. *Suppose that for every $t \in [0, 1]$ $V(t, \cdot)$ is convex, condition (3.3) can be strengthened to (5.1).*

In the proof of this corollary, and subsequently in this paper, we adopt the notation: B^k is the open unit ball, centre the origin, in \mathbb{R}^k . When there is no need to emphasize the dimension of the vector space concerned, we write B^k briefly as B .

Proof. Under the hypotheses $V(\cdot, \cdot)$ is continuous on some tube T about $x_0(\cdot)$ (see [7]). Take any $t \in (0, 1)$, $p \in \mathbb{R}^n$ and sequences $\{t_i\}$, $\{p_i\}$ such that $t_i \rightarrow t$ and $p_i \rightarrow p$, and suppose that $p_i \in \partial_x V(t_i, x_0(t_i))$. In view of Proposition 5.1 it suffices to show that

$$(5.2) \quad p \in \partial_x V(t, x_0(t)).$$

Let $\delta > 0$ be such that $(t, x_0(t)) + \delta B^{n+1} \subset T$. Choose any $\xi \in \frac{1}{2} \delta B^n$. Since $V(t_i, \cdot)$ is convex,

$$V(t_i, \xi + x_0(t_i)) - V(t_i, x_0(t_i)) \geq p_i \cdot \xi.$$

For i sufficiently large, $(t_i, \xi + x_0(t_i))$ and $(t_i, x_0(t_i))$ both lie in T . By continuity we can pass to the limit:

$$V(t, \xi + x_0(t)) - V(t, x_0(t)) \geq p \cdot \xi.$$

This inequality holds for all $\xi \in \frac{1}{2} \delta B^n$. Since $V(t, \cdot)$ is convex the inequality extends however to all of \mathbb{R}^n . We have shown (5.2). \square

There is at least one important case when the hypotheses of Corollary 5.2 are satisfied, as we shall see later in the section.

Our next special case concerns control problems with smooth data. Our results here involve the concept of strict differentiability: a function $\psi(\cdot): \mathbb{R}^k \rightarrow \mathbb{R}^l$ admits a strict derivative at $x \in \mathbb{R}^k$, an $l \times k$ matrix denoted by $D_s\psi(x)$, if, for all $v \in \mathbb{R}^k$,

$$\lim_{\substack{x' \rightarrow x \\ \lambda \downarrow 0}} \frac{F(x' + \lambda v) - F(x')}{\lambda} = D_s\psi(x) \cdot v.$$

An important property of a function $\psi(\cdot)$ which is strictly differentiable at a point x is that the function is Lipschitz continuous in a neighborhood of x , and

$$\partial\psi(x) = \{D_s\psi(x)\}$$

where $\partial\psi(x)$ refers to the generalized Jacobian

$$\partial\psi(x) = \text{co} \{ \lim \psi_x(x_i): x_i \rightarrow x, \psi(\cdot) \text{ is differentiable at } x_i, i = 1, 2, \dots \}.$$

This follows from [5, Props. 2.2.1, 2.6.2(e)].

PROPOSITION 5.3. *Suppose that for almost every $t \in [0, 1]$ the functions*

$$x \rightarrow f(t, x, u_0(t)) \quad \text{and} \quad x \rightarrow L(t, x, u_0(t))$$

are strictly differentiable at $x_0(t)$. Suppose

$$(5.3) \quad x \rightarrow h(x) \text{ is strictly differentiable at } x_0(1).$$

Then condition (3.3) can be strengthened to (5.1).

Proof. Let $\delta > 0$ be such that $V(t, \cdot)$ is Lipschitz continuous on $x(t) + \delta B$ for all $t \in [0, 1]$. For any $t \in [0, 1]$ consider the auxiliary control problem (P_t) :

$$\text{Minimize } \int_t^1 L(t, x(t), u(t)) dt + h(x(1)) - V(t, x(t)) \text{ over processes}$$

$$(x(\cdot), u(\cdot)) \text{ on } [t, 1].$$

By definition of $V(\cdot, \cdot)$ and the principle of optimality, a minimizing process for (P_t) is $(x_0(\cdot), u_0(\cdot))$ restricted to $[t, 1]$. The maximum principle [5, Thm. 5.2.1] tells us: there exists $p'(\cdot): [t, 1] \rightarrow \mathbb{R}^n$ such that

$$-p'(s) = D_s H(t, x_0(s), p'(s), u_0(s)) \quad \text{a.e. } s \in [t, 1],$$

$$H(t, x_0(s), p'(s), u_0(s)) = \max_{u \in U_s} H(t, x_0(s), p'(s), u) \quad \text{a.e. } s \in [t, 1],$$

$$-p'(1) = D_s h(x(1)),$$

$$(5.4) \quad -p'(t) \in \partial_x V(t, x_0(t)).$$

($D_s(\cdot)$ denotes the strict derivative in the x -variable. Note that H is strictly differentiable in x since f and L are.) Now define $p(\cdot) := p^0(\cdot)$. Choose any $t \in [0, 1]$; then $p'(\cdot)$ and $p(\cdot)$ must coincide on $[t, 1]$ (and in particular at t) since they are both solutions there to the differential equation in $q(\cdot)$

$$-\dot{q}(s) = q(s) \cdot D_s f(s, x_0(s), u_0(s)) - D_s L(s, x_0(s), u_0(s))$$

with initial condition

$$-q(1) = D_s h(x(1))$$

and this differential equation has a unique solution. But now (5.1) follows from (5.4). \square

This covers the smooth case, since continuous differentiability implies strict differentiability. The proof of Proposition 5.3 is very much more straightforward than

that available for Theorem 3.1, and readers with interests only in smooth problems might be tempted to think that Proposition 5.3 and its proof were adequate for their purposes, although of course the nonsmoothness of V persists. This however, is not necessarily the case since, when we try to prove analogous results to (5.1) in the presence of a terminal constraint

$$(5.5) \quad x(1) \in C_1$$

(see § 6), we need to apply Theorem 3.1 to an auxiliary problem with a penalty term to accommodate the constraint (5.5). This penalty term must, in general, be non-differentiable and so, even if the data are smooth, the uniqueness argument above does not automatically provide an easy way to obtain the desired inclusions. We mention that the uniqueness argument in the proof of Proposition 5.3 has previously been used by Barbu [1, p. 209] to relate the costate function and the value function for certain optimal control problems involving distributed parameter systems.

We can slightly weaken hypothesis (5.3) in Proposition 5.3 to admit some degree of nonsmoothness.

PROPOSITION 5.4. *The conclusions of Proposition 5.3 remain valid if hypothesis (5.3) is replaced by either*

(a) *$h(\cdot)$ is expressible as a sum of functions*

$$h(\cdot) = h_1(\cdot) + h_2(\cdot),$$

in which $h_1(\cdot)$ is strictly differentiable at $x_0(1)$ and $h_2(\cdot)$ is concave at $x_0(1)$, or

(b) *$V(0, \cdot)$ is expressible as a sum of functions*

$$V(0, \cdot) = v_1(\cdot) + v_2(\cdot),$$

in which $v_1(\cdot)$ is strictly differentiable at x_0 and $v_2(\cdot)$ is convex at x_0 .

Proof. We prove just (b); (a) is proved in a similar way. Let $(x(\cdot), u(\cdot))$ be any admissible process on $[0, t]$. By Lemma 8.3 below

$$\int_0^t L(s, x(s), u(s)) ds + V(t, x(t)) - V(0, x(0)) \geq 0.$$

By assumption $V(0, \cdot) = v_1(\cdot) + v_2(\cdot)$, in which $v_1(\cdot)$ is strictly differentiable at x_0 and $v_2(\cdot)$ is convex. Select any $g \in \partial v_2(x_0)$. Then

$$\int_0^t L(s, x(s), u(s)) ds + V(t, x(t)) - v_1(x_0) - g \cdot (x(0) - x_0) - v_2(x_0) \geq 0.$$

However, by the principle of optimality,

$$\int_0^t L(s, x_0(s), u_0(s)) ds + V(t, x_0(t)) - v_1(x_0) - v_2(x_0) = 0.$$

It follows that, for any $t \in [0, 1]$, $(x_0(\cdot), u_0(\cdot))$ restricted to $[0, t]$ is minimizing for the control problem

$$\begin{aligned} &\text{Minimize } \int_0^t L ds + V(t, x(t)) - v_1(x(0)) - g \cdot (x(0) - x_0) \text{ over control processes} \\ &(x(\cdot), u(\cdot)) \text{ on } [0, t]. \end{aligned}$$

We now invoke the Maximum Principle for each $t \in [0, 1]$. There results a costate $p'(\cdot)$ with the usual properties, notably

$$(5.6) \quad -p'(t) \in \partial_x V(t, x_0(t)),$$

and

$$-p'(0)(= -D_s v_1(x_0) + g) \in \partial_x V(0, x_0).$$

Let $p(\cdot) := p^1(\cdot)$. Then since for each $t \in [0, 1]$, $p'(\cdot)$ and $p^1(\cdot)$ are both solutions to

$$-\dot{q}(s) = q(s) \cdot D_s f(s, x_0(s), u_0(s)) - D_s L(s, x_0(s), u_0(s)),$$

$$q(0) = -D_s v_1(x_0) - g,$$

and this equation has only one solution, (5.6) implies (5.1). \square

There is one case, of some importance, when additional hypothesis (b) of Proposition 5.4 is directly verifiable. This is the case of optimal control problems involving linear dynamics and a convex terminal cost function.

PROPOSITION 5.5. *Assume that $\Omega = [0, 1] \times \mathbb{R}^n$, $h(\cdot)$ is convex and that*

$$\begin{bmatrix} L \\ f \end{bmatrix}(t, x, u) = \begin{bmatrix} a(t)x + b(t)u \\ A(t)x + B(t)u \end{bmatrix}$$

for integrable vector and matrix valued functions $a(\cdot)$, $b(\cdot)$, $A(\cdot)$ and $B(\cdot)$. Suppose also that

$$U_t \text{ is compact a.e. } t \in [0, 1],$$

and

$$\operatorname{ess\,sup}_{t \in [0, 1]} \|U_t\| < \infty.$$

Then condition (3.3) can be strengthened to (5.1).

Proof. In view of Proposition 5.4 it suffices to show that $V(0, \cdot)$ is convex. In fact $V(t, \cdot)$ is convex for all $t \in [0, 1]$ as we shall see.

Define the $(n+1) \times (n+1)$ block matrix $\tilde{A}(t)$ and the $(n+1) \times m$ block matrix $\tilde{B}(t)$ by

$$\tilde{A}(t) = \begin{bmatrix} 0 & -a(t) \\ 0 & A(t) \end{bmatrix}, \quad \tilde{B}(t) = \begin{bmatrix} -b(t) \\ B \end{bmatrix}$$

and let $\Phi(\cdot, \cdot)$ be the transition matrix associated with the time-varying differential equation

$$\dot{y} = \tilde{A}(t)y.$$

Now for each $t \in [0, 1]$ define the mapping M_t from control functions into \mathbb{R}^{n+1} according to

$$M_t(u(\cdot)) = \int_t^1 \Phi(1, s) \tilde{B}(s) u(s) ds$$

and the linear mapping $\Lambda_t: \mathbb{R}^{n+1} \rightarrow \mathbb{R}^n$ by

$$\Lambda_t x = \Phi(1, t) \begin{bmatrix} 0 \\ x \end{bmatrix}.$$

Write

$$R_t = \operatorname{range} \{M_t\}.$$

By Aumann's Theorem (see, e.g., [5, p. 256]), R_t is a compact convex set.

Finally define $\tilde{h}: \mathbb{R}^{n+1} \rightarrow \mathbb{R}$ to be

$$\tilde{h}((y^0, y)) = y^0 + h(y).$$

Using the standard representation of the solution to inhomogeneous linear differential equations in terms of the transition matrix, we easily confirm that

$$(5.7) \quad V(t, x) = \min_{\xi \in R_t} \tilde{h}(\Lambda_t x + \xi).$$

Take arbitrary $x_1, x_2 \in \mathbb{R}^n$ and $\varepsilon \in [0, 1]$. Let ξ_1 (ξ_2) be a minimizing element in the expression on the right-hand side of (5.7) when x takes value x_1 (x_2). We now deduce from the convexity of the function \tilde{h} and the convexity of the set R_t that

$$\begin{aligned} \varepsilon V(t, x_1) + (1 - \varepsilon) V(t, x_2) &= \varepsilon \tilde{h}(\Lambda_t x_1 + \xi_1) + (1 - \varepsilon) \tilde{h}(\Lambda_t x_2 + \xi_2) \\ &\geq \tilde{h}(\Lambda_t(\varepsilon x_1 + (1 - \varepsilon)x_2) + \bar{\xi}) \end{aligned}$$

where $\bar{\xi} = \varepsilon \xi_1 + (1 - \varepsilon) \xi_2$

$$\geq V(t, \varepsilon x_1 + (1 - \varepsilon)x_2).$$

The convexity of $V(t, \cdot)$ is proved. \square

We note, incidentally, that the problem that Proposition 5.5 addresses is an instance for which the hypotheses of Corollary 5.2 are verifiable. Indeed, as we point out in the proof of Proposition 5.5, for the "linear-convex" problem $V(t, \cdot)$ is convex for all $t \in [0, 1]$. Thus Corollary 5.2 affords an independent proof of Proposition 5.5.

6. Problems with terminal constraints. In this section we aim to give new information about the costate function $p(\cdot)$, in situations where the state function $x(\cdot)$ is constrained to satisfy $x(1) \in C_1$. The properties to be described undoubtedly remain true within a broader framework than that which we adopt; we limit attention to a pure terminal cost problem and impose extra hypotheses though, in order to avail ourselves directly of results from [6] and thereby greatly to simplify our proofs. The problem considered is

$$\begin{aligned} &\text{Minimize} && h(x(1)) \\ &\text{subject to} && \dot{x}(t) = f(t, x(t), u(t)) \quad \text{a.e. } t \in [0, 1], \\ &&& x(0) = x_0, \\ &&& x(1) \in C_1, \\ &&& u(t) \in U_t \quad \text{a.e. } t \in [0, 1], \\ &&& x(t) \in \Omega_t, \quad \text{all } t \in [0, 1]. \end{aligned}$$

A process (on $[0, 1]$) will be called minimizing if it minimizes $h(x(1))$ over processes $(x(\cdot), u(\cdot))$ satisfying the problem's constraints (which now include " $x(1) \in C_1$ ").

Let $(x_0(\cdot), u_0(\cdot))$ be a minimizing process such that some tube about $x_0(\cdot)$ is contained in Ω .

Define the multifunction $F(\cdot, \cdot)$:

$$F(t, x) = f(t, x, U_t), \quad \text{all } (t, x) \in [0, 1] \times \mathbb{R}^n.$$

In addition to the hypotheses of § 2, we impose the following throughout this section:

(H7). $F(\cdot, \cdot)$ takes values compact, convex sets in \mathbb{R}^n and is continuous in the sense that

$$\text{dist}(F(t', x'), F(t, x)) \rightarrow 0 \quad \text{if } (t', x') \rightarrow (t, x) \quad \text{in } \Omega$$

($\text{dist}(\cdot, \cdot)$ is the Hausdorff distance function).

(H8) There is a constant \bar{k} such that

$$\text{dist}(F(t, x'), F(t, x)) \leq \bar{k}|x' - x| \quad \text{for all } t \in [0, 1], \quad x, x' \in \Omega_t.$$

(H9) $F(\cdot, \cdot)$ is uniformly bounded on Ω .

(H10) C_1 is closed.

The usual Maximum Principle now asserts existence of a real number λ ($\lambda = 0$ or 1) and an absolutely continuous function $p(\cdot): [0, 1] \rightarrow \mathbb{R}^n$ ($\lambda, p(\cdot)$ not both zero) such that

$$\begin{aligned} -\dot{p}(t) &\in \partial_x H(t, x_0(t), p(t), u_0(t)) \quad \text{a.e. } t \in [0, 1], \\ -p(1) &\in N_{C_1}(x_0(1)) + \lambda \partial_x g(x_0(1)), \end{aligned}$$

and

$$H(t, x_0(t), p(t), u_0(t)) = \max_{u \in U_t} H(t, x_0(t), p(t), u) \quad \text{a.e. } t \in [0, 1].$$

Here

$$H(t, x, p, u) = p \cdot f(t, x, u).$$

Let us recall in passing that a minimizing process is customarily called “normal” if the number λ in such conditions as above cannot be taken zero, i.e., if the conditions are nondegenerate in the sense that they make some reference to the cost function. The precise notion of normality required for later purposes is the following.

DEFINITION 6.1. The minimizing process $(x_0(\cdot), u_0(\cdot))$ will be said to be *normal* if the only absolutely continuous function $\tilde{p}(\cdot): [0, 1] \rightarrow \mathbb{R}^n$ which satisfies

$$(-\dot{\tilde{p}}(t), \dot{x}_0(t)) \in \partial \tilde{H}(t, x_0(t), \tilde{p}(t)) \quad \text{a.e. } t \in [0, 1]$$

and

$$-\tilde{p}(1) \in N_{C_1}(x_0(1))$$

is

$$\tilde{p}(\cdot) \equiv 0.$$

Here $\tilde{H}(t, x, p)$, the “true Hamiltonian,” is

$$\tilde{H}(t, x, p) = \sup_{u \in U_t} p \cdot f(t, x, u)$$

and $\partial \tilde{H}$ denotes the generalized gradient of $(x, p) \rightarrow \tilde{H}(t, x, p)$. Normality as defined here relates to nondegeneracy of first order optimality conditions expressed in terms of generalized gradients of the true Hamiltonian (see [5, p. 147]).

Even in the presence of a terminal constraint we may define the value function $V(\cdot, \cdot): [0, 1] \times \mathbb{R}^n \rightarrow \bar{\mathbb{R}}$:

$$V(t, x) = \inf \{h(x(1))\}$$

where the infimum is over the processes $\{x(\cdot), u(\cdot)\}$ on $[t, 1]$ which satisfy the constraints of the problem, modified to involve the time-interval $[t, 1]$ and the initial condition $x(t) = x$ ($V(\cdot, \cdot)$ takes values $+\infty$ if no such processes exist), and we might hope to show in certain circumstances that $p(\cdot)$ can be chosen to satisfy the inclusion

$$-p(t) \in \partial_x V(t, x_0(t)),$$

appropriately interpreted.

When we add the terminal constraint $x(1) \in C_1$ there is however the very real possibility that, for all $t \in [0, 1]$, $V(t, \cdot)$ will not be Lipschitz continuous on a neighborhood of $\{x_0(t)\}$ and, worse, might take finite values only on a set having empty interior. It is not clear how our methods could be modified to overcome these difficulties; establishing such a relationship between some costate function and $V(\cdot, \cdot)$ remains a challenging open problem. (Note that the inclusion, if it applies in situations having any degree of generality, must involve generalized gradients of possibly non-Lipschitz continuous, extended valued functions.)

However, the approach of earlier sections does provide an answer to a related question. To pose this we need to bring verification functions into the discussion.

DEFINITION 6.2. Let $(x(\cdot), u(\cdot))$ be some process on $[0, 1]$ which satisfies the constraints. Then a Lipschitz continuous function $W(\cdot, \cdot)$ defined on some δ -tube T^δ about $x(\cdot)$ is said to be a *verification function* (for $(x(\cdot), u(\cdot))$) if

$$(6.1) \quad \min_{(\alpha, \beta) \in \partial W(t, x)} \{\alpha - H(t, x, -\beta)\} = 0, \quad \text{for all } (t, x) \in \text{int } \{T^\delta\},$$

$$(6.2) \quad W(1, x) = h(x) \quad \text{for all } x \in \{\xi: \|\xi - z(1)\| < \delta\} \cap C_1,$$

and

$$(6.3) \quad W(0, x_0) = h(x(1)).$$

The significance of there being a verification function $W(\cdot, \cdot)$ on some tube $T \subset \Omega$ for an admissible process $(x(\cdot), u(\cdot))$ is that $(x(\cdot), u(\cdot))$ is a minimizing process among admissible processes whose state functions lie in some (possibly narrower) tube about $x(\cdot)$, i.e., $W(\cdot, \cdot)$ *verifies* the (local) optimality of $(x(\cdot), u(\cdot))$. This is shown in [6]. Of course there is an intimate relationship between verification functions for $(x_0(\cdot), u_0(\cdot))$ and the value function $V(\cdot, \cdot)$: $V(\cdot, \cdot)$ satisfies conditions (6.2) and (6.3) in the definition of a verification function $W(\cdot, \cdot)$ and we can expect $V(\cdot, \cdot)$ to satisfy (6.1), which will be recognized as a version of the Hamilton-Jacobi-Bellman equation, at least on neighborhoods where $V(\cdot, \cdot)$ is Lipschitz continuous. The difference is that there can exist a (Lipschitz continuous) verification function for $(x_0(\cdot), u_0(\cdot))$ even when $V(\cdot, \cdot)$ is very ill behaved (has effective domain with empty interior, say). Existence of verification functions has been established under quite mild conditions, as is illustrated by the following result, proved in [5] and [6].

PROPOSITION 6.3. Suppose the minimizing process $(x_0(\cdot), u(\cdot))$ is normal. Then there is a verification function for $(x_0(\cdot), u_0(\cdot))$.

After this detour we are ready to return to our modified question: when can we choose some costate function $p(\cdot)$ and some verification function $W(\cdot, \cdot)$ for $(x_0(\cdot), u_0(\cdot))$ such that

$$-p(t) \in \partial_x W(t, x_0(t))?$$

An answer is provided by the following theorem.

THEOREM 6.4. Suppose that the minimizing process $(x_0(\cdot), u_0(\cdot))$ is normal. Then there exists an absolutely continuous function $p(\cdot): [0, 1] \rightarrow \mathbb{R}^n$, a verification function $W(\cdot, \cdot)$ for $(x_0(\cdot), u_0(\cdot))$ and a subset $\Sigma \subset [0, 1]$ of full measure such that

$$-\dot{p}(t) \in \partial_x H(t, x_0(t), p(t), u_0(t)) \quad \text{a.e. } t \in [0, 1],$$

$$H(t, x_0(t), p(t), u_0(t)) = \max_{u \in U_t} H(t, x_0(t), p(t), u) \quad \text{a.e. } t \in [0, 1],$$

$$-p(1) \in N_{C_1}(x_0(1)) + \partial h(x_0(1)),$$

$$-p(t) \in \partial_x W(t, x_0(t)) \quad \text{for all } t \in \{0\} \cup \{1\} \cup \Sigma.$$

Proof. By Theorem 4.1 and Lemmas 6.3–6.6 of [6] there exist $K, \varepsilon > 0$ such that $(x_0(\cdot), u_0(\cdot))$ is a minimizing process for the *free* endpoint problem

$$\begin{aligned} \text{Minimize} \quad & K \int_0^1 \max \{ \|x(t) - z(t)\| - \varepsilon, 0 \} dt + K d_{C_1}(x(1)) + h(x(1)) \\ \text{subject to} \quad & \dot{x} = f(t, x, u), \quad x(0) = x_0, \\ & u(t) \in U_t \quad \text{a.e. } t \in [0, 1], \\ & x(t) \in X_t \quad \text{for all } t \in [0, 1]. \end{aligned}$$

Here X is the 2ε -tube about $x_0(\cdot)$. $d_{C_1}(\cdot)$ denotes the Euclidean distance function. It is shown further in [6] that the value function for this problem is Lipschitz continuous on some tube about $x_0(\cdot)$ and is a verification function (relative to the original terminally constrained problem) for $(x_0(\cdot), u_0(\cdot))$. Now apply Theorem 3.1 to this free endpoint problem. This is permissible since, as we easily check, the free endpoint problem satisfies the necessary hypotheses. The assertions of Theorem 6.4 follow. \square

7. Discrete approximation of integrals. An important step in the proof of Theorem 3.1 will be the construction of sequences of discrete approximations to certain integrals with desirable convergence properties. The necessary machinery is provided in this section.

Define the functions $\psi_N(\cdot): \mathbb{R} \rightarrow \mathbb{R}$, $N = 1, 2, \dots$:

$$\psi_N(s) = \sum_{j=-\infty}^{+\infty} \frac{j}{N} \chi_{[j/N, (j+1)/N)}(s)$$

in which $\chi_A(\cdot)$ denotes the function taking value 1 on A and zero elsewhere. We see that $\psi_N(\cdot)$ is a piecewise constant approximation to the straight line through the origin of unit slope.

Let $g(\cdot): [0, 1] \rightarrow \mathbb{R}^n$ be a function. In order to simplify certain formulae to follow we consider $g(\cdot)$ extended to the whole real line: off $[0, 1]$ we define its value to be zero.

Now for any $\tau \in [0, 1]$ and positive integer N , the function $\tilde{g}_{N,\tau}(\cdot)$:

$$(7.1) \quad \tilde{g}_{N,\tau}(s) = g(\tau + \psi_N(s - \tau)),$$

which can alternatively be written

$$\tilde{g}_{N,\tau}(s) = \sum_{j=-\infty}^{+\infty} g\left(\tau + \frac{j}{N}\right) \chi_{[j/N + \tau, (j+1)/N + \tau)}(s),$$

is a piecewise constant approximation to $g(\cdot)$.

The following proposition, drawn to our attention by P. Loewen, concerns the relationship between the indefinite integrals of $g(\cdot)$ and $\tilde{g}_{N,\tau}(\cdot)$. It is a special case of a result due to Doob [8, p. 440ff.], implicit in his construction of the stochastic integral.

PROPOSITION 7.1. *Let $g(\cdot)$ be an $\mathcal{L}^\infty(0, 1; \mathbb{R}^n)$ function, and let $\tilde{g}_{N,\tau}(\cdot)$ be the function defined by (7.1). Then there exists a subset $S \subset [0, 1]$, of full measure, and a subsequence $\{N_j\}$ (of the positive real numbers) such that*

$$\sup_{t \in [0, 1]} \left\| \int_0^t (\tilde{g}_{N_j,\tau}(s) - g(s)) ds \right\| \rightarrow 0$$

as $j \rightarrow \infty$ for all $\tau \in S$.

Proof. Extend $g(\cdot)$ to all of \mathbb{R} , as in the preceding discussion. We show first that

$$(7.2) \quad \lim_{h \rightarrow 0} \int_0^1 \|g(s+h) - g(s)\| ds = 0.$$

Let K be a uniform bound on the values of $g(\cdot)$. Take $\varepsilon > 0$. By Lusin's theorem (see [13, p. 53]), extended to the vector case, there exist a function \tilde{g} , continuous on $[0, 1]$, taking value 0 off $[0, 1]$ and having values bounded by K , and a measurable set Ω_ε , with $|\Omega_\varepsilon| \leq \varepsilon$, such that

$$\tilde{g}(s) = g(s) \quad \text{for all } s \in [0, 1] \setminus \Omega_\varepsilon.$$

For any h

$$\begin{aligned} \int_0^1 \|g(s+h) - g(s)\| \, ds &\leq \int_0^1 \|\tilde{g}(s+h) - \tilde{g}(s)\| \, ds \\ &\quad + \int_0^1 [\|\tilde{g}(s+h) - g(s+h)\| + \|\tilde{g}(s) - g(s)\|] \, ds \\ &\leq \int_0^1 \|\tilde{g}(s+h) - \tilde{g}(s)\| \, ds + 4K\varepsilon. \end{aligned}$$

But the integrals on the right-hand side have limit 0 as $h \rightarrow 0$, by the Dominated Convergence Theorem, since the integrals are majorized by the constant function $2K$ and since they converge pointwise to zero (except possibly at 0 or 1). It follows that

$$\limsup_{h \rightarrow 0} \int_0^1 \|g(s+h) - g(s)\| \, ds \leq 4K\varepsilon.$$

Since $\varepsilon > 0$ was arbitrary, (7.2) follows.

Now define $d_N(\tau)$:

$$d_N(\tau) = \int_0^1 \|g(\tau + \psi_N(s - \tau)) - g(s)\| \, ds.$$

Then

$$\begin{aligned} \int_0^1 d_N(\tau) \, d\tau &= \int_0^1 \int_0^1 \|g(\tau + \psi_N(s - \tau)) - g(s)\| \, ds \, d\tau \\ (7.3) \quad &= \int_{-1}^{+1} \left\{ \int_{0 \vee \sigma}^{1 \wedge (1+\sigma)} \|g(s - \sigma + \psi_N(\sigma)) - g(s)\| \, ds \right\} d\sigma. \end{aligned}$$

To obtain the last expression we have changed variables $\sigma = s - \tau$ ($a \wedge b, a \vee b$ denote $\min\{a, b\}, \max\{a, b\}$, respectively).

But for fixed σ ,

$$\int_{0 \vee \sigma}^{1 \wedge (1+\sigma)} \|\cdot \cdot \cdot\| \, ds \leq \int_0^1 \|g(s - \sigma + \psi_N(\sigma)) - g(s)\| \, ds \rightarrow 0$$

as $N \rightarrow \infty$, by (7.2). Another application of the Dominated Convergence Theorem yields

$$\int_0^1 d_N(\tau) \, d\tau \rightarrow 0$$

as $N \rightarrow \infty$, i.e., $d_N(\cdot) \rightarrow 0$ strongly in L^1 . But then, for a suitable subsequence $\{N_j\}$, $d_{N_j}(\cdot) \rightarrow 0$ almost everywhere. Since

$$\sup_{t \in [0, 1]} \left\| \int_0^t (g(\tau + \psi_{N_j}(s - \tau)) - g(s)) \, ds \right\| \leq d_{N_j}(\tau)$$

for all $\tau \in [0, 1]$, the proposition is proved. \square

Now we introduce an approximation $g_{N,\tau}$ to the measure associated with $g(\cdot)$, namely $A \rightarrow \int_A g(s) ds$:

$$g_{N,\tau} = \frac{1}{N} \sum_{j=-\infty}^{+\infty} g\left(\tau + \frac{j}{N}\right) \delta_{\{j/N + \tau\}}.$$

Here δ_t denotes the unit measure concentrated at t .

COROLLARY 7.2. *Let $g(\cdot)$ be an $\mathcal{L}^\infty(0, 1; \mathbb{R}^n)$ function. Then there exists a subsequence $\{N_j\}$ and a subset $S \subset [0, 1]$ of full measure, such that, for $\tau \in S$,*

$$\sup_{t \in [0, 1]} \left\| \int_0^t dg_{N_j,\tau}(s) - \int_0^t g(s) ds \right\| \rightarrow 0 \quad \text{as } j \rightarrow \infty.$$

Proof. Scrutiny of the formulae for $g_{N,\tau}(\cdot)$ and $\tilde{g}_{N,\tau}(\cdot)$ reveals that, for any t , $\tau \in [0, 1]$ and integer N ,

$$\left\| \int_0^t \tilde{g}_{N,\tau}(s) ds - \int_0^t dg_{N,\tau}(s) \right\| \leq \sup_{s \in [0, 1]} g(s) \cdot \frac{1}{N}.$$

The assertions of the corollary now follow from Proposition 7.1, since the values of $g(\cdot)$ are uniformly bounded. \square

8. Proof of Theorem 3.1. Fix $\delta > 0$ such that the δ -tube about $x_0(\cdot)$ is contained in the set Ω' of hypothesis (H6).

We shall say a triple $(x(\cdot), u(\cdot), \alpha(\cdot))$ is a *perturbed process* (on $[0, 1]$) if $\mu(\cdot)$ is a measurable \mathbb{R}^m -valued function such that $u(t) \in \Omega_t$, a.e., $\alpha(\cdot)$ is a measurable \mathbb{R}^n -valued function such that

$$\|\alpha(t)\| < 1 \quad \text{a.e. } t \in [0, 1],$$

and $x(\cdot)$ is an absolutely continuous \mathbb{R}^n -valued function such that

$$x(t) = f(t, x(t), u(t)) + \alpha(t) \quad \text{a.e. } t \in [0, 1].$$

Let $(x(\cdot), u(\cdot), \alpha(\cdot))$ be any perturbed process such that $x(\cdot)$ is contained in the δ -tube about $x_0(\cdot)$. Define

$$(8.1) \quad \sigma_\delta(t, \alpha) := \max \{p \cdot \alpha : p \in \partial_x V(t, x) : \|x - x_0(t)\| \leq \delta\}.$$

The following lemma is proved in [7].

LEMMA 8.1. *The function $t \rightarrow \sigma_\delta(t, \alpha(t))$ is measurable and essentially bounded.*

According to Corollary 7.2, we can choose an irrational number $\tau \in [0, 1]$ and a subsequence $\{N_i\}$ of the positive integers such that

$$(8.2) \quad \sup_{t \in [0, 1]} \left\| \int_{[0, t]} d\alpha_i(s) - \int_0^t \alpha(s) ds \right\| \rightarrow 0$$

and

$$(8.3) \quad \int_{[0, 1]} d\sigma_i(s) \rightarrow \int_0^1 \sigma_\delta(s, \alpha(s)) ds$$

as $i \rightarrow \infty$. Here the measures α_i and σ_i are defined by

$$\alpha_i = \frac{1}{N_i} \sum_{j=m_i}^{n_i} \alpha\left(\tau + \frac{j}{N_i}\right) \delta_{\{j/N_i + \tau\}}$$

and

$$\sigma_i = \frac{1}{N_i} \sum_{j=m_i}^{n_i} \sigma_\delta \left(\tau + \frac{j}{N_i}, \alpha \left(\tau + \frac{j}{N_i} \right) \right) \delta_{\{j/N_i + \tau\}}$$

where

$$(8.4) \quad m_i = \min \left\{ j: \tau + \frac{j}{N_i} > 0 \right\} \quad \text{and} \quad n_i = \max \left\{ j: \tau + \frac{j}{N_i} < 1 \right\}.$$

(τ is chosen irrational to ensure that neither 0 nor 1 can be a mesh point, and thereby to simplify certain formulae that follow.)

Consider now the integral equation

$$(8.5) \quad y(t) = x(0) + \int_0^t f(s, y(s), u(s)) ds + \int_{[0,t]} d\alpha_i(s) \quad \text{all } t \in [0, 1].$$

A fairly routine application of [5, Thm. 3.1.6] in which we use (8.2) (see [7] for details) yields the following lemma.

LEMMA 8.2. *There exists a positive integer i_0 such that for all $i > i_0$ the integral equation (8.5) has a solution, $x_i(\cdot)$, in the δ -tube about $x_0(\cdot)$, and*

$$\sup_{t \in [0,1]} \|x_i(t) - x(t)\| \rightarrow 0 \quad \text{as } i \rightarrow \infty.$$

We shall require also the following version of the principle of optimality.

LEMMA 8.3. *Let $A \subset \Omega$ be a tube about $x_0(\cdot)$ on which $V(\cdot, \cdot)$ is finite. Suppose that $(x(\cdot), u(\cdot))$ is a process on $[a, b]$ with graph in A . Then*

$$\int_a^b L(s, x(s), u(s)) ds + V(b, x(b)) - V(a, x(a)) \geq 0.$$

Proof. We readily deduce from hypotheses (H3) and (H4) that $s \rightarrow L(s, x(s), u(s))$ is an integrable function. Since $V(\cdot, \cdot)$ is finite at $(b, x(b))$ there exists a sequence of processes $(x_i(\cdot), u_i(\cdot))$ on $[b, 1]$ such that $x_i(b) = x(b)$ for each i and

$$(8.6) \quad V(b, x(b)) = \lim_i \int_b^1 L(s, x_i(s), u_i(s)) ds + h(x_i(1)).$$

We concatenate the processes $(x(\cdot), u(\cdot))$ on $[a, b]$ and $(x_i(\cdot), u_i(\cdot))$ on $[b, 1]$ to obtain the process $(\tilde{x}_i(\cdot), \tilde{u}_i(\cdot))$ on $[a, 1]$, for each i . Since $V(\cdot, \cdot)$ is the infimum cost

$$\int_a^1 L(s, \tilde{x}_i(s), \tilde{u}_i(s)) ds + h(\tilde{x}_i(1)) - V(a, x(a)) \geq 0$$

whence

$$\int_a^b L(s, x(s), u(s)) ds + \int_b^1 L(s, x_i(s), u_i(s)) ds + h(x_i(1)) - V(a, x(a)) \geq 0.$$

We now prove the lemma by noting (8.6) and passing to the limit. \square

LEMMA 8.4. *The generalized process satisfies the inequality*

$$(8.7) \quad \int_0^1 L(t, x(t), u(t)) dt + h(x(1)) + \int_0^1 \sigma_\delta(t, -\alpha(t)) dt - V(0, x(0)) \geq 0.$$

Proof. Take $i > i_0$. The proof hinges on the observation that the solution $x_i(\cdot)$ to the integral equation (8.5) evolves according to an ordinary differential equation

$$\dot{z}(t) = f(t, x(t), u(t))$$

on the subintervals $[j/N_i + \tau, (j+1)/N_i + \tau]$, $j = m_i, \dots, n_i - 1$, $[0, a_i]$ and $[b_i, 1]$. (Here $a_i = \tau + m_i/N_i$, $b_i = \tau + m_i/N_i$); the integers m_i, n_i were defined in (8.4).) The observation, coupled with Lemma 8.3, leads to the following conclusions:

$$\int_0^{a_i} L(t, x_i(t), u(t)) dt + V(a_i, x_i(a_i^-)) - V(0, x(0)) \geq 0,$$

$$\int_{j/N_i + \tau}^{(j+1)/N_i + \tau} L(t, x_i(t), u(t)) dt + V(t, x_i(t^-))|_{t=(j+1)/N_i + \tau} - V(t, x_i(t^+))|_{t=j/N_i + \tau} \geq 0$$

for $j = m_i, m_i + 1, \dots, n_i - 1$, and

$$\int_{b_i}^1 L(t, x_i(t), u(t)) dt + V(1, x_i(1)) - V(b_i, x(b_i^+)) \geq 0.$$

Adding these inequalities we have

$$(8.8) \quad \int_0^1 L(t, x_i(t), u(t)) dt + V(1, x_i(1)) - V(0, x(0)) + q_i \geq 0$$

where

$$q_i = \sum_{j=m_i}^{n_i} V(t, x_i(t^-)) - V(t, x_i(t^+))|_{t=(j/N_i + \tau)}.$$

Note that since $x_i(\cdot)$ satisfies the integral equation (8.5)

$$x_i(t^+) = x_i(t^-) + N_i^{-1} \alpha(t)$$

for $t = j/N_i + \tau$, $j = m_i, \dots, n_i$, and so q_i can alternatively be expressed as

$$(8.9) \quad q_i = \sum_{j=m_i}^{n_i} [V(t, x_i(t^-)) - V(t, x_i(t^-) + N_i^{-1} \alpha(t))]|_{t=(j/N_i + \tau)}.$$

Recall that, by assumption, $V(t, \cdot)$ is Lipschitz continuous on $\{x_0(t)\} + \delta B$, for each $t \in [0, 1]$. Since $x_i(\cdot)$ is in the δ -tube about $x_0(\cdot)$,

$$x_i(t^-), x_i(t^-) + \frac{1}{N_i} \alpha(t) \in \{x_0(t)\} + \delta B$$

for $t = j/N_i + \tau$, $j = m_i, \dots, n_i$. By the Mean Value Theorem for generalized gradients (see [5, Thm. 2.3.7])

$$V(t, x_i(t^-)) - V\left(t, x_i(t^-) + \frac{1}{N_i} \alpha(t)\right) \in \frac{1}{N_i} \partial_x V(t, \xi_i(t)) \cdot (-\alpha(t))$$

for some $\xi_i(t) \in \{x_0(t)\} + \delta B$, when t assumes values $j/N_i + \tau$, $j = m_i, \dots, n_i$. It follows from (8.9) that

$$q_i \leq \frac{1}{N_i} \sum_{j=m_i}^{n_i} \sigma_\delta(t, -\alpha(t))|_{t=(j/N_i + \tau)}.$$

(The function $\sigma_\delta(\cdot, \cdot)$ was defined by (8.1).) By (8.8)

$$\int_0^1 L(t, x_i(t), u(t)) dt + V(1, x_i(1)) - V(0, x(0)) + \frac{1}{N_i} \sum_{j=m_i}^{n_i} \sigma_\delta(t, -\alpha(t))|_{t=(j/N_i + \tau)} \geq 0.$$

We now pass to the limit $i \rightarrow \infty$. Since $x_i(\cdot)$ converges uniformly to $x(t)$, and in view of property (8.3), the result is (8.7). \square

We denote by P_δ the following optimal control problem:

$$\begin{aligned} \text{Minimize } & \int_0^1 L(t, x(t), u(t)) dt + \int_0^1 \sigma_\delta(t, -\alpha(t)) dt + h(x(1)) - V(0, x(0)) \\ \text{subject to } & \dot{x}(t) = f(t, x(t), u(t)) + \alpha(t) \quad \text{a.e. } t \in [0, 1], \\ & u(\cdot) \in U_t \quad \text{a.e. } t \in [0, 1], \\ & \alpha(t) \in B \quad \text{a.e. } t \in [0, 1], \\ & x(t) \in X_t^\delta \quad \text{all } t \in [0, 1]. \end{aligned}$$

Here X^δ is the δ -tube about $x_0(\cdot)$. As usual, B is the open unit ball (in \mathbb{R}^n). In P_δ , the control variable is the pair of vectors $(u(t), \alpha(t))$. Lemma 8.4 can now be re-expressed in the following terms.

LEMMA 8.5. *The extended process $(x_0(\cdot), u_0(\cdot), \alpha(\cdot) \equiv 0)$ solves P_δ .*

It is a straightforward task to check that the hypotheses are satisfied under which the Maximum Principle [5, Thm. 5.2.1] applies, at the minimizing process $(x_0(\cdot), u_0(\cdot), \alpha(\cdot) \equiv 0)$. We use in particular Lemma 8.1. Bearing in mind that P_δ is a free left and right endpoint problem, we conclude the following: let

$$H(t, x, p, u) = p \cdot f(t, x, u) - L(t, x, u)$$

as before. Then there exists an absolutely continuous function $p(\cdot): [0, 1] \rightarrow \mathbb{R}^n$ such that

$$(8.10) \quad \begin{cases} -\dot{p}(t) \in \partial_x H(t, x_0(t), p(t), u_0(t)) & \text{a.e. } t \in [0, 1], \\ -p(1) \in \partial h(x_0(1)), \quad -p(0) \in \partial_x V(0, x_0(0)), \\ H(t, x_0(t), p(t), u_0(t)) = \max_{u \in U_t} H(t, x_0(t), p(t), u) & \text{a.e. } t \in [0, 1], \end{cases}$$

and

$$(8.11) \quad \max_{\alpha \in B} \{p(t) \cdot \alpha - \sigma_\delta(t, -\alpha)\} = 0 \quad \text{a.e. } t \in [0, 1].$$

Take a point $t \in [0, 1]$ at which (8.11) is true. Then

$$(8.12) \quad -p(t) \in S_\delta(t)$$

where

$$S_\delta(t) = \overline{\text{co}} \cup \{\partial_x V(t, \xi): \|\xi - x_0(t)\| \leq \delta\},$$

for otherwise $-p(t)$ and the closed convex set $S_\delta(t)$ can be strictly separated, i.e., there exists an n -vector $\bar{\alpha} \in B$ such that

$$\begin{aligned} p(t) \cdot \bar{\alpha} &> \max \{-r \cdot \bar{\alpha}: r \in S_\delta(t)\} \\ &= \sigma_\delta(t, -\bar{\alpha}) \end{aligned}$$

in contradiction of (8.11).

Up to this point $\delta > 0$ has been fixed. Now let $\{\delta_i\}$, $\delta_i \downarrow 0$, be a sequence of numbers such that the δ_i -tube about $x_0(\cdot)$ is contained in Ω' for all i . Replace δ by δ_i , and write $p_i(\cdot)$ in place of the costate $p(\cdot)$ satisfying (8.10) et seq.

We deduce from the costate differential inclusion and the transversality conditions that

$$\frac{d}{dt} \|p_i(t)\| \leq k(t)[1 + \|p_i(t)\|]$$

where $k(\cdot)$ is the function of hypothesis (H3), and

$$p_i(1) \leq K$$

for all i , where K is the rank of $h(\cdot)$ on some suitable neighborhood of $x_0(1)$. Application of Gronwall's inequality tells us that the $p_i(\cdot)$'s are uniformly bounded and the $\dot{p}_i(\cdot)$'s are dominated by a common integrable function. The hypotheses are met then under which [5, Thm. 3.1.7] applies to the inclusions

$$\dot{p}_i(t) \in G(t, p_i(t)) \quad \text{a.e. } t \in [0, 1]$$

where $G(t, z) := \partial_x H(t, x_0(t), z, u_0(t))$. We conclude that, following extraction of a suitable subsequence,

$$\sup_{t \in [0, 1]} \|p_i(t) - p(t)\| \rightarrow 0$$

for some absolutely continuous function $p(\cdot): [0, 1] \rightarrow \mathbb{R}^n$ satisfying

$$-\dot{p}(t) \in \partial_x H(t, x_0(t), p(t), u_0(t)) \quad \text{a.e.}$$

We deduce

$$-p(1) \in \partial_x h(x_0(1)) \quad \text{and} \quad -p(0) \in \partial_x V(0, x_0(0))$$

from the transversality conditions and the upper semicontinuity of the generalized gradient of a locally Lipschitz function. A simple contradiction argument, in which we employ hypotheses (H3) and (H4), leads to the following conclusion:

$$(8.13) \quad H(t, x_0(t), p(t), u_0(t)) = \max_{u \in U_t} H(t, x_0(t), p(t), u)$$

for all $t \in \cap_i M_i$, where M_i is the set on which the Hamiltonian is maximized for problem P_{δ_i} . So (8.13) is true almost everywhere.

Finally we examine the implications of " $-p_i(t) \in S_{\delta_i}(t)$ a.e." (see (8.12)).

Evidently,

$$(8.14) \quad -p(t) \in \bigcap_{\varepsilon > 0} \overline{\text{co}} \{ \partial_x V(t, \xi) : \|\xi - x_0(t)\| \leq \varepsilon \}$$

for all t belonging to some subset $S \subset [0, 1]$ of full measure. We claim that, for any $t \in S$,

$$(8.15) \quad -p(t) \in \partial_x V(t, x_0(t)).$$

Otherwise we can strictly separate the point $-p(t)$ and the closed convex set $\partial_x V(t, x_0(t))$, i.e., there exists $q \in \mathbb{R}^n$ and $\gamma > 0$ such that

$$-p(t) \cdot q - \gamma > \max_{s \in \partial_x V(t, x_0(t))} s \cdot q = D_x^0 V(t, x_0(t); q).$$

Here D_x^0 denotes the generalized directional derivative with respect to the x variable. (We have used the fact that $D_x^0 V$ is the polar function of the set $\partial_x V$.) Since the generalized directional derivative is upper semicontinuous in its arguments [5, Prop. 2.1.1]

$$-p(t) \cdot q - \frac{\gamma}{2} > D_x^0 V(t, \xi; q)$$

whenever $\xi \in \{x_0(t)\} + \varepsilon_1 B$, for some $\varepsilon_1 > 0$. Then

$$\begin{aligned} -p(t) \cdot q - \frac{\gamma}{2} &> \sup \{ s \cdot q : s \in \bigcup \{ \partial_x V(t, \xi) : \|\xi - x_0(t)\| \leq \varepsilon_1 \} \} \\ &= \max \{ s \cdot q : s \in \overline{\text{co}} \bigcup \{ \partial_x V(t, \xi) : \|\xi - x_0(t)\| \leq \varepsilon_1 \} \}. \end{aligned}$$

But this means that

$$-p(t) \notin \overline{\text{co}} \{ \partial_x V(t, \xi) : \|\xi - x_0(t)\| \leq \varepsilon_1 \}$$

in contradiction of (8.14). (8.15) is then true on a subset of full measure.

This completes the proof of Theorem 3.1. \square

REFERENCES

- [1] V. BARBU, *Optimal Control of Variational Inequalities*, Research Notes in Mathematics, 100, Pitman, London, 1984.
- [2] L. D. BERKOVITZ, *Optimal Control Theory*, Springer, New York, 1974.
- [3] A. E. BRYSON AND Y. C. HO, *Applied Optimal Control*, Blaisdell, Waltham, MA, 1969.
- [4] S. J. CITRON, *Elements of Optimal Control*, Holt, Rinehart and Winston, New York, 1969.
- [5] F. H. CLARKE, *Optimization and Nonsmooth Analysis*, John Wiley, New York, 1983.
- [6] F. H. CLARKE AND R. B. VINTER, *Local optimality conditions and Lipschitzian solutions to the Hamilton-Jacobi equation*, this Journal, 21 (1983), pp. 856-870.
- [7] ———, *The Maximum Principle and the Dynamic Programming technique: how are they related?* Technical Report CRM-1300, Univ. of Montréal, Montreal, Quebec, Canada, 1985.
- [8] J. L. DOOB, *Stochastic Processes*, John Wiley, New York, 1953.
- [9] W. H. FLEMING AND R. W. RISCHER, *Deterministic and Stochastic Optimal Control*, Springer, New York, 1975.
- [10] O. L. R. JACOBS, *Introduction to Control Theory*, Clarendon, Oxford, 1974.
- [11] F. MIGNANEGO AND G. PIERI, *On a generalized Bellman equation for the optimal-time problem*, Systems Control Lett., 3 (1983), pp. 235-241.
- [12] L. S. PONTRYAGIN, V. G. BOLTJANSKII, R. V. GRAMKRELIDZE AND E. F. MISHCHENKO, *The Mathematical Theory of Optimal Processes*, Interscience, New York, 1962.
- [13] W. RUDIN, *Real and Complex Analysis*, McGraw-Hill, New York, 1966.
- [14] T. ZOLEZZI, *On generalised dynamic programming*, Istituto di Matematica, Univ. di Genova, Genova, Italy, preprint, 1985.

NONSMOOTH CALCULUS IN FINITE DIMENSIONS*

D. E. WARD† AND J. M. BORWEIN‡

Abstract. The notion of subgradient, originally defined for convex functions, has in recent years been extended, via the “upper subderivative,” to cover functions that are not necessarily convex or even continuous. A number of calculus rules have been proven for these generalized subgradients. This paper develops the finite-dimensional generalized subdifferential calculus for (strictly) lower semicontinuous functions under considerably weaker hypotheses than those previously used. The most general finite-dimensional convex subdifferential calculus results are recovered as corollaries. Other corollaries given include new necessary conditions for optimality in a nonsmooth mathematical program. Various chain rule formulations are considered. Equality in the subdifferential calculus formulae is proven under hypotheses weaker than the usual “subdifferential regularity” assumptions.

Key words. Clarke tangent cone, upper subderivative, subgradient, contingent cone, subdifferential regularity

AMS(MOS) subject classifications. Primary 46G05; secondary 58C20, 90C30.

1. Introduction. Let E be a real locally convex Hausdorff topological vector space (l.c.t.v.s.) with dual space E^* . For $x \in E$, denote by $\mathcal{N}(x)$ the set of all neighborhoods of x .

DEFINITION 1.1. Let $C \subset E$ and $x_0 \in \text{cl } C$. The *Clarke tangent cone* to C at x_0 is the set

$$(1.1) \quad T_C(x_0) := \{y \in E \mid \forall Y \in \mathcal{N}(y), \exists X \in \mathcal{N}(x_0), \text{ and } \exists \lambda > 0 \text{ such that } \forall x' \in X \cap C \text{ and } \forall t \in (0, \lambda), \exists y' \in Y \text{ with } x' + ty' \in C\}.$$

The *normal cone* to C at x_0 is the set

$$(1.2) \quad N_C(x_0) := T_C(x_0)^0 := \{z \in E^* \mid \langle y, z \rangle \leq 0 \ \forall y \in T_C(x_0)\}.$$

$T_C(x_0)$ is always closed and contains the origin. More important, it is always a convex set. (See [17], [20, Chap. 1].)

DEFINITION 1.2. Let $f: E \rightarrow \bar{\mathbf{R}}$ be an extended-real-valued function on E . The *domain* of f is the set

$$(1.3) \quad \text{dom } f := \{x \in E \mid f(x) < +\infty\}.$$

The *epigraph* of f is the set

$$(1.4) \quad \text{epi } f := \{(x, r) \in E \times \mathbf{R} \mid f(x) \leq r\}.$$

Let $x_0 \in E$ be a point at which f is finite, and let $y \in E$. The upper subderivative of f at x_0 in the direction y is defined by

$$(1.5) \quad f^\uparrow(x_0; y) := \inf \{r \mid (y, r) \in T_{\text{epi } f}(x_0, f(x_0))\}.$$

Notice that $f^\uparrow(x_0; y)$ is defined precisely so that

$$(1.6) \quad \text{epi } f^\uparrow(x_0; y) = T_{\text{epi } f}(x_0, f(x_0)).$$

By (1.6), $f^\uparrow(x_0; \cdot)$ has several important properties. Since $T_{\text{epi } f}(x_0, f(x_0))$ is closed, $f^\uparrow(x_0; \cdot)$ is lower semicontinuous (l.s.c.), and since $T_{\text{epi } f}(x_0, f(x_0))$ is a convex cone, $f^\uparrow(x_0; \cdot)$ is convex and positively homogeneous. As a result, either

* Received by the editors April 3, 1984; accepted for publication (in revised form) February 6, 1986.

† Department of Mathematics and Statistics, Miami University, Oxford, Ohio 45056.

‡ Department of Mathematics, Statistics, and Computing Science, Dalhousie University, Halifax, Nova Scotia, Canada B3H 3J5.

(a) $f^\uparrow(x_0; 0) = 0$, in which case $f^\uparrow(x_0; \cdot)$ is *proper*, i.e., is somewhere finite and never $= -\infty$, or

(b) $f^\uparrow(x_0; 0) = -\infty$, in which case $f^\uparrow(x_0; \cdot)$ is equal to $-\infty$ throughout its domain. For further discussion of the idea of associating directional derivatives with approximating cones to epigraphs of functions, in particular the Clarke tangent cone, see [12], [7] and [20, Chap. 3].

DEFINITION 1.3. Let $f: E \rightarrow \bar{\mathbf{R}}$ and $x_0 \in E$ be such that $f(x_0)$ is finite. The *subgradient* of f at x_0 is the set

$$(1.7) \quad \partial f(x_0) := \{x^* \in E^* \mid \langle y, x^* \rangle \leq f^\uparrow(x_0; y) \forall y \in E\}.$$

The notation $\partial f(x_0)$ is the same as that for the subgradient of a convex function or the Clarke subgradient of a locally Lipschitzian function, and justifiably so. If f is locally Lipschitzian, $\partial f(x_0)$ coincides with the Clarke subgradient, which in turn coincides with the ordinary subgradient for convex functions if f is convex. (See [18], [19], [20].)

In [19], Rockafellar introduced the following direct characterization of $f^\uparrow(x_0; \cdot)$:

$$(1.8) \quad f^\uparrow(x_0; y) = \sup_{Y \in \mathcal{N}(y)} \inf_{\substack{N \in \mathcal{N}(x_0, f(x_0)) \\ \lambda > 0}} \sup_{\substack{(x, \alpha) \in \text{epi } f \cap N \\ t \in (0, \lambda)}} \inf_{y' \in Y} \frac{f(x + ty) - \alpha}{t}.$$

If f is l.s.c., this definition reduces to

$$(1.9) \quad \begin{aligned} f^\uparrow(x_0; y) &= \limsup_{\substack{x \rightarrow x_0 \\ f \\ t \downarrow 0}} \inf_{y' \rightarrow y} \frac{f(x + ty') - f(x)}{t} \\ &:= \sup_{\substack{y \in \mathcal{N}(y) \\ \lambda > 0}} \inf_{\substack{X \in \mathcal{N}(x_0) \\ \lambda > 0}} \sup_{\substack{x \in X \\ t \in (0, \lambda) \\ f(x) \equiv f(x_0) + \lambda}} \inf_{y' \in Y} \frac{f(x + ty') - f(x)}{t}. \end{aligned}$$

If f is continuous, $f^\uparrow(x_0; y)$ may be written more simply as

$$(1.10) \quad \begin{aligned} f^\uparrow(x_0; y) &= \limsup_{\substack{x \rightarrow x_0 \\ t \downarrow 0}} \inf_{y' \rightarrow y} \frac{f(x + ty') - f(x)}{t} \\ &:= \sup_{Y \in \mathcal{N}(y)} \inf_{\substack{x \in \mathcal{N}(x_0) \\ \lambda > 0}} \sup_{\substack{x \in X \\ t \in (0, \lambda)}} \inf_{y' \in Y} \frac{f(x + ty') - f(x)}{t}. \end{aligned}$$

Rockafellar used characterization (1.8) in [18] to extend the subdifferential calculus to the subgradient of (1.7). (See also [6, § 2.9].)

In this paper, we show that if E is finite-dimensional and $f: E \rightarrow \bar{\mathbf{R}}$ is strictly l.s.c., the hypotheses in the subdifferential calculus theorems of [18] can be considerably weakened, and the hypotheses required in proving multiplier rules for nonsmooth mathematical programs can be correspondingly weakened. This is in analogy with the convex case, where in finite dimensions, interiority assumptions are replaced by assumptions about relative interiors. In fact, the convex subgradient calculus of § 23 of [16] can be entirely recaptured as a special case of the results presented here.

Our method of proof differs from that of [18] in that instead of working directly with $f^\uparrow(x_0; \cdot)$ as defined in (1.8), we use an “inversion theorem” (see [4]), the relationship in (1.6), and underlying properties of the Clarke tangent cone.

Here is an outline of the remainder of the paper. In § 2, we list the definitions and collect the preliminary results that we will use in proving subdifferential calculus

formulae. In § 3, we prove our two main subdifferential calculus formulae and give a number of corollaries. In § 4, we apply the results of § 3 to the study of necessary optimality conditions in nonsmooth mathematical programming. In particular, we prove a finite-dimensional version of Theorem 6 of [18] under weaker hypotheses and apply a “Dubovitskii–Milyutin” approach to prove a “Fritz John type” result as in [26]. We investigate in § 5 the possibility of extending a chain rule of Hiriart-Urruty [11, Chap. 8] to functions which are not necessarily locally Lipschitzian. In § 6 we prove inequalities involving the contingent [1] and Ursescu [9] directional derivatives. We then combine these results with those of § 3 to give conditions for equality in the subdifferential calculus theorems of § 3 that somewhat relax the usual “subdifferential regularity” conditions. We also give a generalization of one of our results of § 4 through the use of “upper convex approximates” to these directional derivatives.

An excellent background reference for this paper is Chapter 7 of Aubin and Ekeland’s recent book [3]. In particular, Corollary 3.4 and Propositions 3.10 and 3.14 are derived in [3] by an approach in many ways similar to that employed here.

Two entirely different approaches to this subject can be found in the significant papers [13] and [22]. In [13], Ioffe essentially proves Theorem 3.2 and its corollaries as special cases of corresponding formulae in the calculus of “approximate subdifferentials.” Rockafellar in [22] derives subgradient inclusions (3.3), (3.6), (3.30) and (4.2) by an approach that is “dual” to that taken here and in [18]. The methods used in [22] center around the concepts of the normal cone and proximal normals. Both [13] and [22] go further than this present work in some directions. On the other hand, the chain rule formulation in Theorem 3.17 and the material in § 6 are not discussed in either [13] or [22]. In fact, it is not clear how one might derive Theorem 3.17 by the methods of [13], and the methods of [22] do not readily yield conditions for equality in subdifferential calculus formulae. References [3], [6], [11], [13], [18], [22] and this work combine to form quite an extensive body of information about the generalized subdifferential calculus.

2. Preliminaries.

DEFINITION 2.1. Let E, E_1 be l.c.t.v.s. A function $G: E \rightarrow E_1$ is said to be *strictly differentiable* at $x_0 \in E$ if there exists a linear mapping $\nabla G(x_0): E \rightarrow E_1$ such that

$$(2.1) \quad \lim_{\substack{x \rightarrow x_0 \\ y' \rightarrow y \\ t \downarrow 0}} \frac{G(x + ty') - G(x)}{t} = \nabla G(x_0)y$$

for all $y \in E$.

DEFINITION 2.2. Let E be an l.c.t.v.s.

(a) The set $C \subset E$ is said to be closed near $x_0 \in C$ if there exists $X \in \mathcal{N}(x_0)$ such that $X \cap C$ is closed.

(b) The function $f: E \rightarrow \bar{\mathbf{R}}$ is *strictly l.s.c.* at $x_0 \in E$ if for some $\alpha > f(x_0)$, the function $\min\{f, \alpha\}$ is l.s.c.

It is observed in [21] that if f is strictly l.s.c. at x_0 , then the set $\text{epi } f$ is closed near $(x_0, f(x_0))$.

DEFINITION 2.3. $F: \mathbf{R}^n \rightarrow \bar{\mathbf{R}}$ is *isotone* on $D \subset \mathbf{R}^n$ if $F(x) \leq F(y)$ whenever $x, y \in D$ and $x \leq y$ (with respect to the coordinate ordering). F is *strictly isotone* in the i th coordinate at $x := (x_1, \dots, x_n) \in \mathbf{R}^n$ if $F(x) < F(y)$ whenever $x \leq y$ and $x_i < y_i$.

The result that allows us to weaken the “constraint qualifications” of [18] is the following special case of Theorem 4.1 of [4], which itself follows from Ekeland’s variational principle [8] (see also [3, § 7.6]).

THEOREM 2.4. *Let $G: \mathbf{R}^p \rightarrow \mathbf{R}^q$ be strictly differentiable at $x_0 \in D \cap G^{-1}(0)$, where $D \subset \mathbf{R}^p$ is closed near x_0 . Assume that*

$$(2.2) \quad \nabla G(x_0) T_D(x_0) = \mathbf{R}^q.$$

Then

$$(2.3) \quad T_D(x_0) \cap \nabla G(x_0)^{-1}(0) \subset T_{D \cap G^{-1}(0)}(x_0).$$

We will use the following two properties of the Clarke tangent cone in our proofs in § 3. The first follows easily from (1.1).

PROPOSITION 2.5. *Let E, E_1 be l.c.t.v.s., and let $x_0 \in C \subset E$ and $y_0 \in D \subset E_1$. Then*

$$(2.4) \quad T_{C \times D}(x_0, y_0) = T_C(x_0) \times T_D(y_0).$$

PROPOSITION 2.6. *Let E, E_1 be l.c.t.v.s. and $A: E \rightarrow E_1$ be linear and continuous. Let $z_0 \in C \subset E$. Suppose that A is relatively open on C at z_0 ; i.e.,*

$$(2.5) \quad \text{For each } X \in \mathcal{N}(z_0), \text{ there exists } Z \in \mathcal{N}(Az_0) \text{ such that } Z \cap A(C) \subset A(X \cap C).$$

Then

$$(2.6) \quad A(T_C(z_0)) \subset T_{A(C)}(Az_0).$$

Proof. Let $y \in T_C(z_0)$ and $Y \in \mathcal{N}(A(y))$. Then $Y' := A^{-1}(Y)$ is in $\mathcal{N}(y)$. There exists $\lambda > 0$, $X \in \mathcal{N}(z_0)$ such that for all $t \in (0, \lambda)$ and for all $x' \in X \cap C$,

$$(x' + tY') \cap C \neq \emptyset.$$

By (2.5), there exists $Z \in \mathcal{N}(Az_0)$ such that

$$Z \cap A(C) \subset A(X \cap C).$$

Then for each $z \in Z \cap A(C)$, there exists $x' \in X \cap C$ such that $A(x') = z$. For such an x' , $(x' + tY') \cap C \neq \emptyset$ for all $t \in (0, \lambda)$. Hence $Ax' + tY \cap A(C) \neq \emptyset$ for all $t \in (0, \lambda)$, and so $A(y) \in T_{A(C)}(Az_0)$. Thus $A(T_C(z_0)) \subset T_{A(C)}(Az_0)$. \square

Remark 2.7. (a) Condition (2.5) holds in particular whenever A is open and one-to-one on C .

(b) Proposition 2.6 is a special case of Corollary 4.2 of [4] (see also [14]).

In § 3, we will establish calculus rules involving functions of two forms:

(a) $h := f_1 + f_2 \circ F$, where $f_1: \mathbf{R}^n \rightarrow \bar{\mathbf{R}}$ is strictly l.s.c. at x_0 , $f_2: \mathbf{R}^m \rightarrow \bar{\mathbf{R}}$ is strictly l.s.c. at $F(x_0)$, and $F: \mathbf{R}^n \rightarrow \mathbf{R}^m$ is strictly differentiable at $x_0 \in \text{dom } f_1 \cap F^{-1}(\text{dom } f_2)$.

(b) $h := F \circ f$, where $f = (f_1, \dots, f_n)$, each $f_i: \mathbf{R}^m \rightarrow \bar{\mathbf{R}}$ is strictly l.s.c. at x_0 , and $F: \mathbf{R}^n \rightarrow \bar{\mathbf{R}}$ is isotone.

In (b), we define $F(f(x))$ to be $+\infty$ whenever $f_i(x) = +\infty$ for some i . We also adopt the convention that if $f_i(x) = -\infty$ for some i ,

$$h(x) = \inf \{F(y) \mid f_i(x) < y_i, i = 1, \dots, n, y = (y_1, \dots, y_n)\}.$$

Under this convention, we have

$$(2.7) \quad \text{epi } h = \{(x, z) \mid \exists y_i \in \mathbf{R} \text{ with } f_i(x) \leq y_i, i = 1, \dots, n, F(y_1, \dots, y_n) \leq z\}$$

because F is isotone. We will use this fact in § 3.

The proofs of our calculus rules will consist of two stages:

Stage 1: Establish an inequality involving upper subderivatives with the help of (1.6), Proposition 2.6, Theorem 2.4 and Proposition 2.5.

Stage 2: Use that inequality and a convex subdifferential calculus formula to establish a corresponding inclusion for the subgradients of (1.7).

As part of stage 1, we will make use of the following lemma, which shows that upper subderivatives preserve isotonicity.

LEMMA 2.8. *Let $F: \mathbf{R}^n \rightarrow \bar{\mathbf{R}}$ be finite at $x_0 \in \mathbf{R}^n$ and isotone on a neighborhood of x_0 . Then $F^\dagger(x_0; \cdot)$ is isotone on \mathbf{R}^n .*

Proof. Let $y_1, y_2 \in \mathbf{R}^n$ with $y_1 \leq y_2$ and $F^\dagger(x_0; y_2) \leq d$. It suffices to show that $F^\dagger(x_0; y_1) \leq d$. To this end, let $\varepsilon > 0$ and $U \in \mathcal{N}(0)$ be given. Suppose F is isotone on $X_1 \in \mathcal{N}(x_0)$. There exists $X_2 \in \mathcal{N}(x_0)$ and $\lambda_0 > 0$ such that

$$X_2 + (0, \lambda_0)(y_i + U) \subset X_1 \quad \text{for } i = 1, 2.$$

Since $(y_2, d) \in T_{\text{epi } F}(x_0, F(x_0))$, there exist $X \subset X_2$, $X \in \mathcal{N}(x_0)$, $\delta > 0$, and $\lambda \in (0, \lambda_0)$ such that for all $x \in X$, $r \in (F(x_0) - \delta, F(x_0) + \delta)$ with $F(x) \leq r$, and for all $t \in (0, \lambda)$, there exists $h \in U$ such that

$$\frac{F(x + t(y_2 + h)) - r}{t} \leq d + \varepsilon.$$

Now $x + t(y_1 + h) \leq x + t(y_2 + h)$ and both $x + t(y_1 + h)$ and $x + t(y_2 + h)$ are in X_1 . By isotonicity of F on X_1 ,

$$\frac{F(x + t(y_1 + h)) - r}{t} \leq \frac{F(x + t(y_2 + h)) - r}{t},$$

and hence $(y_1, d) \in T_{\text{epi } F}(x_0, F(x_0))$, as required. \square

In order to carry out Stage 2, we will require two convex subdifferential calculus formulae. The first is simply a combination of Theorems 23.8 and 23.9 of [16]. We provide a proof of the second result.

THEOREM 2.9 (cf. [16, Thms. 23.8, 23.9]). *Let $f_1: \mathbf{R}^m \rightarrow \bar{\mathbf{R}}$, $f_2: \mathbf{R}^n \rightarrow \bar{\mathbf{R}}$ be proper convex functions, and let $A: \mathbf{R}^m \rightarrow \mathbf{R}^n$ be linear. Let $x_0 \in \text{dom } f_1 \cap A^{-1}(\text{dom } f_2)$. Assume that*

$$(2.8) \quad A(\text{ri dom } f_1) \cap \text{ri dom } f_2 \neq \emptyset.$$

Then

$$(2.9) \quad \partial(f_1 + f_2 \circ A)(x_0) = \partial f_1(x_0) + A^T \partial f_2(Ax_0).$$

THEOREM 2.10. *Let $f_i: \mathbf{R}^{m_i} \rightarrow \bar{\mathbf{R}}$, $i = 1, \dots, n$, be proper convex functions and let $F: \mathbf{R}^n \rightarrow \bar{\mathbf{R}}$ be convex, proper and isotone. Call $f := (f_1, \dots, f_n)$. Assume $x_0 \in \bigcap_{i=1}^n \text{dom } f_i$ and $f(x_0) \in \text{dom } F$. Suppose that*

$$(2.10) \quad \text{int}(\text{dom } F) \cap \text{range } f \neq \emptyset.$$

Then

$$(2.11) \quad \partial(F \circ f)(x_0) = \{\partial(\lambda \cdot f)(x_0) \mid \lambda \in \partial F(f(x_0))\}.$$

If, in addition,

$$(2.12) \quad \bigcap_{i=1}^n \text{ri dom } f_i \neq \emptyset,$$

then

$$(2.13) \quad \partial(F \circ f)(x_0) = \{\lambda \cdot (\partial f_1(x_0), \dots, \partial f_n(x_0)) \mid \lambda \in \partial F(f(x_0))\}.$$

Proof. Let $\lambda \in \partial F(f(x_0))$ and $x^* \in \partial(\lambda \cdot f)(x_0)$. Then $\lambda \geq 0$ and

$$\begin{aligned} x^*(x - x_0) &\leq (\lambda \cdot f)(x) - (\lambda \cdot f)(x_0) \\ &= \lambda \cdot (f(x) - f(x_0)) \\ &\leq F(f(x)) - F(f(x_0)) \\ &= (F \circ f)(x) - (F \circ f)(x_0). \end{aligned}$$

Thus $x^* \in \partial(F \circ f)(x_0)$, so

$$\{\lambda \cdot (\partial f_1(x_0), \dots, \partial f_n(x_0)) \mid \lambda \in \partial F(f(x_0))\} \subset \{\partial(\lambda \cdot f)(x_0) \mid \lambda \in \partial F(f(x_0))\} \subset \partial(F \circ f)(x_0).$$

Conversely, suppose $x_0^* \in \partial(F \circ f)(x_0)$.

Call $h := F \circ f$. Since F is isotone, $h(x) = \min \{F(y) \mid f(x) \leq y\}$ for all $x \in \bigcap_{i=1}^n \text{dom } f_i$. Then $h^*(x_0) := \sup_x x_0^* \cdot x - h(x) = \sup_{x,y} \{x_0^* \cdot x - F(y) \mid y - f(x) \geq 0\}$. Assumption (2.10) guarantees the existence of a Slater point for the above concave program, so we may apply the Lagrange multiplier theorem of [16, § 28]: There exists $\lambda \geq 0$ such that

$$\begin{aligned} h^*(x_0^*) &= \sup_{x,y} x_0^* \cdot x - F(y) + \lambda \cdot (y - f(x)) \\ &= F^*(\lambda) + (\lambda \cdot f)^*(x_0^*). \end{aligned}$$

Then

$$\begin{aligned} x_0^* \cdot x &= h(x_0) + h^*(x_0^*) \\ &= F(f(x_0)) + (\lambda \cdot f)^*(x_0^*) + F^*(\lambda), \end{aligned}$$

and so

$$x_0^* \cdot x_0 + \lambda \cdot f(x_0) = \lambda \cdot f(x_0) + (\lambda \cdot f)^*(x_0^*) + F(f(x_0)) + F^*(\lambda).$$

We now have

$$x_0^* \cdot x_0 = \lambda \cdot f(x_0) + (\lambda \cdot f)^*(x_0^*) \quad \text{and} \quad \lambda \cdot f(x_0) = F(f(x_0)) + F^*(\lambda),$$

implying that

$$\lambda \in \partial F(f(x_0)) \quad \text{and} \quad x_0^* \in \partial(\lambda \cdot f)(x_0).$$

Thus $\partial(F \circ f)(x_0) \subset \{\partial(\lambda \cdot f)(x_0) \mid \lambda \in \partial F(f(x_0))\}$. If in addition (2.12) holds, Theorem 23.8 of [16] gives $x_0^* \in \partial(\lambda \cdot f)(x_0) = \lambda \cdot (\partial f_1(x_0), \dots, \partial f_n(x_0))$, so that (2.13) holds. \square

Remark 2.11. Since $\text{int}(\text{dom } F) \neq \emptyset$ (because F is proper and isotone), (2.10) is equivalent to the seemingly weaker condition

$$\text{ri}(\text{dom } F) \cap \text{ri}(\text{range } f) \neq \emptyset.$$

Remark 2.12. If $\lambda = 0$ in Theorem 2.10, λf_i should be interpreted as the indicator function of $\text{dom } f_i$ (Definition 3.3). This is in keeping with the convention that $(F \circ f)(x) = +\infty$ whenever some $f_i(x) = +\infty$. As a result, $\partial f_i(x_0)$ should be interpreted as $\partial^\infty f_i(x_0)$ in (2.13) and (3.27) (see Definition 3.13).

3. The main theorems and their corollaries. In Stage 1 of the proof of our first calculus formula, we will need a technical lemma verifying that condition (2.5) is satisfied for the appropriate A , C , and z_0 . In the proof of this lemma, for $x = (x_1, \dots, x_p) \in \mathbb{R}^p$ and $\varepsilon > 0$, we use the notation

$$B_\varepsilon(x) := \{y = (y_1, \dots, y_p) \in \mathbb{R}^p : |y_i - x_i| \leq \varepsilon, i = 1, \dots, p\}.$$

LEMMA 3.1. *Let $f_1: \mathbb{R}^n \rightarrow \bar{\mathbb{R}}$, $f_2: \mathbb{R}^m \rightarrow \bar{\mathbb{R}}$ be l.s.c., and let $F: \mathbb{R}^n \times \mathbb{R}^m \rightarrow \mathbb{R}^m$ be continuous at $x_0 \in \mathbb{R}^n$. Assume $f_1(x_0)$ and $f_2(F(x_0))$ are finite. Define $A: \mathbb{R}^n \times \mathbb{R} \times \mathbb{R}^m \times \mathbb{R} \rightarrow \mathbb{R}^n \times \mathbb{R}$ by $A(x, y, z, r) := (x, y + r)$ and $G: \mathbb{R}^n \times \mathbb{R} \times \mathbb{R}^m \times \mathbb{R} \rightarrow \mathbb{R}^m$ by $G(x, y, z, r) := F(x) - z$.*

Then (2.5) is satisfied with A as above, $C := (\text{epi } f_1 \times \text{epi } f_2) \cap G^{-1}(0)$, and $z_0 := (x_0, f_1(x_0), F(x_0), f_2(F(x_0)))$.

Proof. Let $\varepsilon > 0$ be given, and let $X := B_\varepsilon(z_0)$. Since f_1, f_2 are l.s.c. and F is continuous, there exists $\delta \in (0, \varepsilon/3)$ such that for all $x \in B_\delta(x_0)$, we have

$$\begin{aligned} f_1(x) &\geq f_1(x_0) - \varepsilon/3, \\ F(x) &\in B_{\varepsilon/3}(F(x_0)), \\ (f_2 \circ F)(x) &\geq (f_2 \circ F)(x_0) - \varepsilon/3. \end{aligned}$$

Define $N := B_\delta(z_0)$. Since A is surjective, $Z := A(N) \in \mathcal{N}(A z_0)$. We will now verify that $Z \cap A(C) \subset A(X \cap C)$. To do so, let $(\bar{x}, \bar{r}) \in Z \cap A(C)$. Since $(\bar{x}, \bar{r}) \in Z$, there exists $(x, y, z, r) \in N$ with $x = \bar{x}$, $y + r = \bar{r}$. Since $(\bar{x}, \bar{r}) \in A(C)$, there exists $(x', y', z', r') \in C$ with $x' = \bar{x}$, $y' + r' = \bar{r}$, and $f_1(x') \leq y'$, $z' = F(x')$, $f_2(z') \leq r'$. Now $x' \in B_\delta(x_0)$, so $z' \in B_{\varepsilon/3}(F(x_0))$. Also $\bar{r} \in B_{\varepsilon/3}(f_1(x_0)) + B_{\varepsilon/3}(f_2(F(x_0)))$, so $\bar{r} \in B_{2\varepsilon/3}(f_1(x_0) + f_2(F(x_0)))$. Finally, since $y' + r' = \bar{r}$ and $y' \geq f_1(x_0) - \varepsilon/3$, we conclude that $y' \in B_\varepsilon(f_1(x_0))$ and $r' \in B_\varepsilon(f_2(F(x_0)))$. Thus $(x', y', z', r') \in X$, and so $(\bar{x}, \bar{r}) \in A(X \cap C)$. We conclude that $Z \cap A(C) \subset A(X \cap C)$. \square

We now proceed to the first of our two main theorems.

THEOREM 3.2. *Let $F: \mathbf{R}^n \rightarrow \mathbf{R}^m$ be strictly differentiable at x_0 , $f_1: \mathbf{R}^n \rightarrow \bar{\mathbf{R}}$ finite and strictly l.s.c. at x_0 , and $f_2: \mathbf{R}^m \rightarrow \bar{\mathbf{R}}$ finite and strictly l.s.c. at $F(x_0)$. Assume that*

$$(3.1) \quad \nabla F(x_0) \text{ dom } f_1^\dagger(x_0; \cdot) - \text{dom } f_2^\dagger(x_0; \cdot) = \mathbf{R}^m.$$

Then for all $y \in \mathbf{R}^n$,

$$(3.2) \quad (f_1 + f_2 \circ F)^\dagger(x_0; y) \leq f_1^\dagger(x_0; y) + f_2^\dagger(F(x_0); \nabla F(x_0)y).$$

Moreover,

$$(3.3) \quad \partial(f_1 + f_2 \circ F)(x_0) \subset \partial f_1(x_0) + (\nabla F(x_0))^T \partial f_2(F(x_0)).$$

Proof. Call $f := f_1 + f_2 \circ F$. Then

$$\begin{aligned} \text{epi } f &= \{(x_1, r) \in \mathbf{R}^n \times \mathbf{R} \mid f_1(x_1) \leq r_1, f_2(x_2) \leq r_2, r = r_1 + r_2, \\ &\quad F(x_1) - x_2 = 0, \text{ for some } x_2 \in \mathbf{R}^m, r_1, r_2 \in \mathbf{R}\}. \end{aligned}$$

Define A , G , and C , and z_0 as in Lemma 3.1, and define $D := \text{epi } f_1 \times \text{epi } f_2$. Then $A(D \cap G^{-1}(0)) = \text{epi } f$.

Now

$$\begin{aligned} \text{epi } f^\dagger(x_0; \cdot) &= T_{A(D \cap G^{-1}(0))}(x_0, f(x_0)) \quad \text{by (1.6)} \\ &\supset A(T_{D \cap G^{-1}(0)}(z_0)) \end{aligned}$$

by Lemma 3.1 and Proposition 2.6. Next observe that (3.1) and Proposition 2.5 ensure that $\nabla G(z_0) T_D(z_0) = \mathbf{R}^m$. We can therefore apply Theorem 2.4 to obtain

$$T_{D \cap G^{-1}(0)}(z_0) \supset T_D(z_0) \cap \nabla G(z_0)^{-1}(0).$$

Thus

$$\begin{aligned} A(T_{D \cap G^{-1}(0)}(z_0)) &\supset A(T_D(z_0) \cap \nabla G(z_0)^{-1}(0)) \\ &= A((\text{epi } f_1^\dagger(x_0; \cdot) \times \text{epi } f_2^\dagger(F(x_0); \cdot)) \\ &\quad \cap \nabla G(z_0)^{-1}(0)) \quad (\text{by Proposition 2.5}) \\ &= A(\{(h_1, r_1, h_2, r_2) \in \mathbf{R}^n \times \mathbf{R} \times \mathbf{R}^m \times \mathbf{R} \mid f_1^\dagger(x_0; h_1) \\ &\quad \leq r_1, f_2^\dagger(F(x_0); h_2) \leq r_2, \nabla F(x_0)h_1 = h_2\}) \\ &= \{(h, r_1 + r_2) \in \mathbf{R}^n \times \mathbf{R} \mid f_1^\dagger(x_0; h) \leq r_1, f_2^\dagger(F(x_0); \nabla F(x_0)h) \leq r_2\} \\ &= \text{epi } [f_1^\dagger(x_0; \cdot) + f_2^\dagger(F(x_0); \nabla F(x_0)(\cdot))]. \end{aligned}$$

Therefore $\text{epi } f^\uparrow(x_0; \cdot) \supset \text{epi } [f_1^\uparrow(x_0; \cdot) + f_2^\uparrow(F(x_0)(\cdot))]$, and so (3.2) holds. The rest of the proof is much as in [18]. Set $p_1 := f_1^\uparrow(x_0; \cdot)$ and $p_2 := f_2^\uparrow(F(x_0); \cdot)$. If either $p_1(0)$ or $p_2(0)$ is $-\infty$, equality holds in (3.3), since (3.2) shows that both sides of the inclusion are then empty. Assume, then, that $p_1(0) = p_2(0) = 0$.

Then

$$\begin{aligned} \partial f(x_0) &= \{z \in \mathbf{R}^n \mid (f_1 + f_2 \circ F)^\uparrow(x_0; y) \geq \langle y, z \rangle \forall y \in \mathbf{R}^n\} \\ &\subset \{z \mid p_1(y) + p_2(\nabla F(x_0)y) \geq \langle y, z \rangle \forall y \in \mathbf{R}^n\} \\ &= \partial(p_1 + p_2 \circ \nabla F(x_0))(0). \end{aligned}$$

Since p_1 and p_2 are convex and proper, and since (3.1) holds, we may apply Theorem 2.9 to obtain

$$\begin{aligned} \partial(p_1 + p_2 \circ \nabla F(x_0))(0) &= \partial p_1(0) + \nabla F(x_0)^T \partial p_2(0) \\ &= \partial f_1(x_0) + \nabla F(x_0)^T \partial f_2(F(x_0)) \end{aligned}$$

by definition, and so (3.3) holds. \square

We can now obtain improved versions, in the finite-dimensional case, of results in [17] and [18].

DEFINITION 3.3. Let $C \subset E$. The *indicator function* of C , denoted i_C , is defined by

$$i_C(x) = \begin{cases} 0 & \text{if } x \in C, \\ +\infty & \text{otherwise.} \end{cases}$$

Observe that for $x_0 \in C$, $i_C^\uparrow(x_0; \cdot) = i_{T_C(x_0)}(\cdot)$ and $\partial i_C(x_0) = N_C(x_0)$. The following is a strengthening of Theorem 5 of [17], which has also been proved by Aubin [2], [3].

COROLLARY 3.4. Let $F_2: \mathbf{R}^m \rightarrow \mathbf{R}^m$ be strictly differentiable at x_0 , and let $C_1 \subset \mathbf{R}^n$ be closed near x_0 and $C_2 \subset \mathbf{R}^m$ closed near $F(x_0)$. Suppose that

$$(3.4) \quad \nabla F(x_0) T_{C_1}(x_0) - T_{C_2}(F(x_0)) = \mathbf{R}^m.$$

Then

$$(3.5) \quad T_{C_1 \cap F^{-1}(C_2)}(x_0) \supset T_{C_1}(x_0) \cap \nabla F(x_0)^{-1} T_{C_2}(F(x_0))$$

and

$$(3.6) \quad N_{C_1 \cap F^{-1}(C_2)}(x_0) \subset N_{C_1}(x_0) + \nabla F(x_0)^T N_{C_2}(F(x_0)).$$

Proof. Let $f_1 := i_{C_1}$, $f_2 := i_{C_2}$ in Theorem 3.2. Then (3.1) becomes (3.4), and (3.5) and (3.6) follow from (3.2) and (3.3), respectively. \square

It is interesting to note that Theorem 2.4, which is used in proving Corollary 3.4, is itself a special case of Corollary 3.4.

Remark 3.5. If $F = A$ is linear, (3.4) can be weakened to

$$(3.7) \quad A(T_{C_1}(x_0)) - T_{C_2}(Ax_0) = \text{span}(AC_1 - C_2).$$

To see this, suppose $C \subset \mathbf{R}^p$, $x_0 \in C$. For a given affine set S with $\text{aff } C \subset S \subset \mathbf{R}^p$, denote by $T_C^S(x_0)$ the Clarke tangent cone of C at x_0 where C is considered as a subset of S rather than as a subset of \mathbf{R}^p . It is easy to see that $T_C^S(x_0) = T_C(x_0)$. Without loss of generality, we may assume $x_0 = 0$, since $T_{C-x_0}(0) = T_C(x_0)$ and $x_0 \in C_1 \cap A^{-1}(C_2)$ if and only if $0 \in (C_1 - x_0) \cap A^{-1}(C_2 - Ax_0)$. Call $V := \text{span } C_1$ and $W := \text{span}(AC_1 - C_2)$. If (3.7) holds, then $A(T_{C_1}(x_0)) - T_{C_2}^W(Ax_0) = W$ and A may be considered as $A: V \rightarrow W$. By Corollary 3.4, $T_{C_1 \cap A^{-1}(C_2)}^V(x_0) \supset T_{C_1}^V(x_0) \cap A^{-1} T_{C_2}^W(Ax_0)$, which is the same as (3.5).

COROLLARY 3.6. Let $f_1: \mathbf{R}^n \rightarrow \bar{\mathbf{R}}, f_2: \mathbf{R}^n \rightarrow \bar{\mathbf{R}}$ be finite and strictly l.s.c. at x_0 . Assume that

$$(3.8) \quad \text{dom } f_1^\dagger(x_0; \cdot) - \text{dom } f_2^\dagger(x_0; \cdot) = \mathbf{R}^n.$$

Then for all $y \in \mathbf{R}^n$,

$$(3.9) \quad (f_1 + f_2)^\dagger(x_0; y) \leq f_1^\dagger(x_0; y) + f_2^\dagger(x_0; y).$$

In addition,

$$(3.10) \quad \partial(f_1 + f_2)(x_0) \subset \partial f_1(x_0) + \partial f_2(x_0).$$

Proof. Set $m = n$ and $F := I$ in Theorem 3.2. \square

Remark 3.7. (a) Corollary 3.6 is more general than the specialization of Theorem 2 of [18] to strictly l.s.c. functions with finite-dimensional domains, since Rockafellar's assumption

$$(3.11) \quad \text{dom } f_1^\dagger(x_0; \cdot) \cap \text{int dom } f_2^\dagger(x_0; \cdot) \neq \emptyset$$

implies, but is not implied by (3.8). Here is an example which satisfies (3.8) but not (3.11): Define $f_1: \mathbf{R}^2 \rightarrow \mathbf{R}, f_2: \mathbf{R}^2 \rightarrow \mathbf{R}$ by $f_1(x, y) := |x|^{1/2}, f_2(x, y) := |y|^{1/2}$. Then $\text{dom } f_1^\dagger(0; \cdot) = 0 \times \mathbf{R}$ and $\text{dom } f_2^\dagger(0; \cdot) = \mathbf{R} \times 0$, both of which have empty interior.

(b) Condition (3.8) can actually be weakened to

$$(3.12) \quad \text{dom } f_1^\dagger(x_0; \cdot) - \text{dom } f_2^\dagger(x_0; \cdot) = \text{span}(\text{dom } f_1 - \text{dom } f_2).$$

To see this, simply observe that we can again assume that $x_0 = 0$, and we can then replace \mathbf{R}^n by $S := \text{span}(\text{dom } f_1 - \text{dom } f_2)$, considering f_1 and f_2 as $f_1: S \rightarrow \bar{\mathbf{R}}$ and $f_2: S \rightarrow \bar{\mathbf{R}}$.

It is not possible, however, to weaken the hypothesis still further to

$$(3.13) \quad \text{ri dom } f_1^\dagger(x_0; \cdot) \cap \text{ri dom } f_2^\dagger(x_0; \cdot) \neq \emptyset.$$

For example, let $C_1 := \{(x, y) \mid y = x^2\}, C_2 := \{(x, y) \mid y = -x^2\}$, and $f_1 := i_{C_1}, f_2 := i_{C_2}$. Let $x_0 = (0, 0)$. Then $f_1^\dagger(x_0; \cdot) = f_2^\dagger(x_0; \cdot) = i_{\mathbf{R} \times 0}(\cdot)$, while

$$(f_1 + f_2)^\dagger(x_0; \cdot) = i_{T_{C_1 \cap C_2}(x_0)}(\cdot) = i_{\{(0,0)\}}(\cdot).$$

Hence (3.9) does not hold for $y = (x, 0)$ with $x \neq 0$, although (3.13) is satisfied.

(c) If f_1 and f_2 are not strictly l.s.c., then (3.8) may hold without either (3.9) or (3.10) being satisfied. For example, let $C_1 := \mathbf{Q}$, the set of rational numbers, and let $C_2 = (\mathbf{R}/\mathbf{Q}) \cup \{0\}$. Define $f_1 := i_{C_1}, f_2 := i_{C_2}$, and let $x_0 = 0$. Then $T_{C_1}(0) = T_{C_2}(0) = \mathbf{R}$, so (3.8) holds, and $f_1^\dagger(x_0; \cdot) = f_2^\dagger(x_0; \cdot) = i_{\mathbf{R}}(\cdot)$. However, $C_1 \cap C_2 = \{0\}$, so we have $(f_1 + f_2)^\dagger(x_0; \cdot) = i_{\{0\}}(\cdot)$, and (3.9) does not hold for $y \neq 0$.

DEFINITION 3.8. Let $C \subset \mathbf{R}^p$. Define

$$\Delta^n C := \{(x_1, \dots, x_n) \mid x_i \in C, x_1 = x_2 = \dots = x_n\}.$$

DEFINITION 3.9. Let $C_i \subset \mathbf{R}^m, i = 1, \dots, n$ be convex sets. The sets $C_i, i = 1, \dots, n$, are said to be in *strong general position* [27] if

$$(3.14) \quad 0 \in \text{int} \left[\Delta^{n-1} C_1 - \prod_{j=2}^n C_j \right].$$

If the sets are cones, (3.14) is equivalent to

$$(3.15) \quad \Delta^{n-1} C_1 - \prod_{j=2}^n C_j = \mathbf{R}^{(n-1)m}.$$

Another equivalent way to write (3.14) is

$$(3.16) \quad 0 \in \text{int} \left[\Delta^n \mathbf{R}^m - \prod_{j=1}^n C_j \right].$$

(see [27] for a thorough discussion of this concept.) By (3.16), the order in which the sets in (3.14) are listed does not matter.

The following is an important special case of Corollary 3.4:

PROPOSITION 3.10 ([26], [3]). *Let $D_i \subset \mathbf{R}^m$, $i = 1, \dots, n$ be closed near $y_0 \in \bigcap_{i=1}^n D_i$. Assume $T_{D_i}(y_0)$, $i = 1, \dots, n$ are in strong general position. Then*

$$(3.17) \quad T_{D_1 \cap \dots \cap D_n}(y_0) \supset \bigcap_{i=1}^n T_{D_i}(y_0)$$

and

$$(3.18) \quad N_{D_1 \cap \dots \cap D_n}(y_0) \subset \sum_{i=1}^n N_{D_i}(y_0).$$

Proof. Call $C_1 := D_1 \times \dots \times D_n$, and let C_2 be the origin in $\mathbf{R}^{(n-1)m}$. Define $F: \mathbf{R}^{nm} \rightarrow \mathbf{R}^{(n-1)m}$ by $F(x_1, \dots, x_n) := (x_1 - x_2, \dots, x_1 - x_n)$. Apply Corollary 3.4 with $x_0 := (y_0, \dots, y_0)$. By Proposition 2.5, (3.4) reduces to $T_{D_i}(y_0)$, $i = 1, \dots, n$, being in strong general position. By (3.5),

$$T_{\Delta^n(D_1 \cap \dots \cap D_n)}(y_0, \dots, y_0) \supset (T_{D_1}(y_0) \times \dots \times T_{D_n}(y_0)) \cap \Delta^n \mathbf{R}^m, \quad \text{or}$$

$$\Delta^n T_{D_1 \cap \dots \cap D_n}(y_0) \supset \Delta^n \bigcap_{i=1}^n T_{D_i}(y_0).$$

Thus (3.17) holds. Finally, (3.18) follows from (3.6). \square

Remark 3.11. By (3.7), the strong general position assumption can be weakened to

$$(3.19) \quad \Delta^{n-1} T_{D_1}(y_0) - \prod_{i=2}^n T_{D_i}(y_0) = \prod_{j=2}^n \text{span}(D_1 - D_j).$$

We can derive from Proposition 3.10 a formula for the subgradient of $f(x) := \max_{1 \leq i \leq n} f_i(x)$, where $f_i: \mathbf{R}^m \rightarrow \bar{\mathbf{R}}$ are strictly l.s.c. We start with a lemma which we will use to show that the condition “ $\text{dom } f_i^\uparrow(x_0; \cdot)$, $i = 1, \dots, n$, are in strong general position” is sufficient to guarantee this subgradient formula.

LEMMA 3.12. *Suppose $f_i: \mathbf{R}^m \rightarrow \bar{\mathbf{R}}$, $i = 1, \dots, n$ are such that*

$$\Delta^{n-1} \text{dom } f_1 - \prod_{i=2}^n \text{dom } f_i = \mathbf{R}^{m(n-1)}.$$

Then

$$\Delta^{n-1} \text{epi } f_1 - \prod_{i=2}^n \text{epi } f_i = (\mathbf{R}^m \times \mathbf{R})^{n-1}.$$

Proof. We must show $(\mathbf{R}^m \times \mathbf{R})^{n-1} \subset \Delta^{n-1} \text{epi } f_1 - \prod_{i=2}^n \text{epi } f_i$. Let $(y_1, s_1, \dots, y_{n-1}, s_{n-1}) \in (\mathbf{R}^m \times \mathbf{R})^{n-1}$. Then there exist $x_i \in \text{dom } f_i$, $i = 1, \dots, n$ such that $x_1 - x_i = y_{i-1}$, $i = 2, \dots, n$. Choose r_i , $i = 1, \dots, n$, such that $r_i \geq f_i(x_i)$. If $r_1 - r_2 < s_1$, replace r_1 by $s_1 + r_2$. If $r_1 - r_2 > s_1$, replace r_2 by $r_1 - s_1$. Now if $r_1 - r_3 > s_2$, replace r_3 by $r_1 - s_2$. If $r_1 - r_3 < s_2$, replace r_1 by $s_2 + r_3$ and r_2 by $s_2 + r_3 - s_1$. Proceed in this manner. After $k-1$ steps, we have $r_1 - r_i = s_{i-1}$, $i = 2, \dots, k$. If $r_1 - r_{k+1} > s_k$, replace r_{k+1} by $r_1 - s_k$. If $r_1 - r_{k+1} < s_k$, replace r_1 by $s_k + r_{k+1}$ and replace r_j , $j = 2, \dots, k$ by $s_k + r_{k+1} - s_{j-1}$.

After $n-1$ steps, we obtain $r_1 - r_i = s_{i-1}$, $i = 2, \dots, n$, and $(x_i, r_i) \in \text{epi } f_1$. Thus $(y_i, s_j) \in \Delta^{n-1} \text{epi } f_1 - \prod_{i=2}^n \text{epi } f_i$ and so $(\mathbf{R}^m \times \mathbf{R})^{n-1} \subset \Delta^{n-1} \text{epi } f_1 - \prod_{i=2}^n \text{epi } f_i$. \square

DEFINITION 3.13. Let $f: E \rightarrow \bar{\mathbf{R}}$ be finite at $x \in E$. The *asymptotic generalized gradient* of f at x is the set

$$\partial^\infty f(x) = \{z \in E^* \mid (z, 0) \in N_{\text{epi } f}(x, f(x))\}.$$

In our next result, we will use the fact that if $\partial f(x) \neq \emptyset$, then

$$(3.20) \quad N_{\text{epi } f}(x, f(x)) = \bigcup_{\lambda > 0} \lambda(\partial f(x), -1) \cup (\partial^\infty f(x), 0)$$

([18], [6]).

PROPOSITION 3.14. Let $f_i: \mathbf{R}^n \rightarrow \bar{\mathbf{R}}$, $i = 1, \dots, n$ be strictly l.s.c. and finite at x_0 . Define $f(x) := \max_{1 \leq i \leq n} f_i(x)$ and $I(x) := \{i \in \{1, \dots, n\} \mid f_i(x) = f(x)\}$. Suppose that f_i is continuous at x_0 for each $i \notin I(x_0)$, and suppose that $\text{dom } f_i^\uparrow(x_0; \cdot)$, $i = 1, \dots, n$, are in strong general position. Then for all $y \in \mathbf{R}^n$,

$$(3.21) \quad f^\uparrow(x_0; y) \leq \max_{I(x_0)} f_i^\uparrow(x_0; y).$$

If also $\partial f_i(x_0)$, $i \in I(x_0)$, are nonempty, then

$$(3.22) \quad \partial f(x_0) \subset \sum_{I(x_0)} (\lambda_i \partial f_i(x_0) \cup \partial^\infty f_i(x_0))$$

for some $\lambda_i \geq 0$ with $\sum_{I(x_0)} \lambda_i = 1$.

Proof. Call $D_i := \text{epi } f_i$, $i = 1, \dots, n$. Since f_j , $j \in I(x_0)$, are strictly l.s.c. and f_j , $j \notin I(x_0)$, are continuous at x_0 , $T_{D_j}(x_0, f(x_0)) = \mathbf{R}^{m+1}$ for all $j \notin I(x_0)$. Thus we only need to consider D_j with $j \in I(x_0)$. By Lemma 3.12, our assumption that $\text{dom } f_i^\uparrow(x_0; \cdot)$, $i = 1, \dots, n$, are in strong general position implies that $T_{D_i}(x_0, f(x_0))$, $i \in I(x_0)$, are also in strong general position. We may then apply Proposition 3.10. By (3.17),

$$\text{epi } f^\uparrow(x_0; \cdot) \supset \text{epi } \max_{i \in I(x_0)} f_i^\uparrow(x_0; \cdot),$$

which gives (3.21). If also $\partial f_i(x_0)$, $i \in I(x_0)$, are nonempty, we have by (3.18) and (3.20) that

$$\bigcup_{\lambda > 0} \lambda(\partial f(x_0), -1) \subset \sum_{i \in I(x_0)} \left(\bigcup_{\lambda_i > 0} \lambda_i(\partial f_i(x_0), -1) \cup (\partial^\infty f_i(x_0), 0) \right).$$

To obtain (3.22), set $\lambda = 1$ on the left-hand side of the above inclusion. \square

In the proof of our second main result, we will use another technical lemma to ensure that we may apply Proposition 2.6.

LEMMA 3.15. Let $f_i: \mathbf{R}^m \rightarrow \bar{\mathbf{R}}$, $i = 1, \dots, n$ be finite and strictly l.s.c. at x_0 , and let $F: \mathbf{R}^n \rightarrow \bar{\mathbf{R}}$ be l.s.c. Call $f := (f_1, \dots, f_n)$. Assume $F(f(x_0))$ is finite and F is isotone on $B_{t_0}(f(x_0)) + \mathbf{R}_+^n$ for some $t_0 > 0$. Define

$$A: \mathbf{R}^{nm+2n+1} \rightarrow \mathbf{R}^{m+1} \quad \text{by}$$

$$A(x_1, y_1, \dots, x_n, y_n, z_1, \dots, z_n, r) := (x_1, r).$$

(Here $x_i \in \mathbf{R}^m$, $y_i, z_i \in \mathbf{R}$.) Define

$$G: \mathbf{R}^{nm+2n+1} \rightarrow \mathbf{R}^{m(n-1)+n} \quad \text{by}$$

$$G(x_1, y_1, \dots, x_n, y_n, z_1, \dots, z_n, r) = (x_1 - x_2, \dots, x_1 - x_n, y_1 - z_1, \dots, y_n - z_n).$$

Assume for each $j \in \{1, \dots, n\}$ that either

- (1) f_j is continuous at x_0 , or
- (2) F is strictly isotone in the j th coordinate at $f(x_0)$.

Then (2.5) is satisfied with A as above,

$$C := (\text{epi } f_1 \times \dots \times \text{epi } f_n \times \text{epi } F) \cap G^{-1}(0), \text{ and}$$

$$z_0 := (x_0, f_1(x_0), \dots, x_0, f_n(x_0), f_1(x_0), \dots, f_n(x_0), F(f(x_0))).$$

Proof. For a given $\varepsilon > 0$, let $X := B_\varepsilon(z_0)$. Let $I \subset \{1, \dots, n\}$ be the set of coordinates in which F is strictly isotone at $f(x_0)$. For $t > 0$, define $U_t := \{(y_1, \dots, y_n) \in \mathbb{R}^n \mid y_i \geq f_i(x_0) - t, i = 1, \dots, n\}$. Suppose $i \in I$. We claim that there exists $\mu_i \in (0, t_0)$ such that

$$F(f_1(x_0) - \mu_i, \dots, f_i(x_0) + \varepsilon, \dots, f_n(x_0) - \mu_i) > F(f(x_0)).$$

If not, then for all $t \in (0, t_0)$,

$$F(f_1(x_0) - t, \dots, f_i(x_0) + \varepsilon, \dots, f_n(x_0) - t) \leq F(f(x_0)).$$

But since F is l.s.c., it follows that

$$F(f_1(x_0), \dots, f_i(x_0) + \varepsilon, \dots, f_n(x_0)) \leq F(f(x_0)).$$

This contradicts our assumption that $i \in I$. So for each $i \in I$, we can choose $\mu_i \in (0, t_0)$ with

$$\delta_i := F(f_1(x_0) - \mu_i, \dots, f_i(x_0) + \varepsilon, \dots, f_n(x_0) - \mu_i) - F(f(x_0)) > 0.$$

Let $\mu = \min_I \mu_i$, and let $\delta = \frac{1}{2} \min_I \delta_i$. Then if $r \in B_\delta(F(f(x_0))) \cap F(U_\mu)$, it follows from the isotonicity of F that $r = F(y)$ for some $y = (y_1, \dots, y_n)$ with $y_i \leq f_i(x_0) + \varepsilon$ for all $i \in I$.

Now by assumptions (1) and (2), there exists $\delta_0 \in (0, \min(\mu, \delta))$ such that $f_j(x) \in B_\mu(f_j(x_0))$ whenever $x \in B_{\delta_0}(x_0)$, $j \notin I$ and $f_j(x) \geq f_j(x_0) - \mu$ whenever $x \in B_{\delta_0}(x_0)$, $j \in I$. Let $N = B_{\delta_0}(z_0)$. Again $Z := A(N) \in \mathcal{N}(Az_0)$. We will now show that $Z \cap A(C) \subset A(X \cap C)$. Suppose $(\bar{x}, \bar{r}) \in Z \cap A(C)$. Since $(\bar{x}, \bar{r}) \in A(N)$, \bar{x} must be in $B_{\delta_0}(x_0)$, and \bar{r} in $B_{\delta_0}(F(f(x_0)))$. Since $(\bar{x}, \bar{r}) \in A(C)$, there exists $y \in \mathbb{R}^n$ with $f(\bar{x}) \leq y$ and $F(y) \leq \bar{r}$. Thus $\bar{y} = f(\bar{x})$ satisfies $F(\bar{y}) \leq \bar{r}$, $\bar{y}_i \in B_\mu(f_j(x_0))$ for all $j \notin I$, and $f_j(x_0) - \mu \leq \bar{y}_j \leq f_j(x_0) + \varepsilon$ for all $j \in I$. We conclude that $(\bar{x}, \bar{y}_1, \dots, \bar{x}, \bar{y}_n, \bar{y}_1, \dots, \bar{y}_n, \bar{r}) \in X \cap C$, and so $(\bar{x}, \bar{r}) \in A(X \cap C)$. \square

Remark 3.16. The hypotheses of Lemma 3.15 hold in the following important cases:

- (a) $F(y_1, \dots, y_n) := \sum_{i=1}^n y_i$ and each f_n strictly l.s.c. at x_0 .
- (b) $F(y_1, \dots, y_n) := \prod_{i=1}^n y_i$ and each f_j positive and strictly l.s.c. at x_0 .
- (c) $F(y_1, \dots, y_n) := \max_{1 \leq j \leq n} y_j$ and each f_j strictly l.s.c. at x_0 , with f_j continuous at x_0 for all $j \notin I(x_0)$, where $I(x_0) := \{i \in \{1, \dots, n\} \mid f_i(x_0) = \max_{1 \leq j \leq n} f_j(x_0)\}$.

THEOREM 3.17. Let $f_i: \mathbb{R}^m \rightarrow \bar{\mathbb{R}}$, $i = 1, \dots, n$ be finite and strictly l.s.c. at x_0 , and define $f := (f_1, \dots, f_n)$. Let $F: \mathbb{R}^n \rightarrow \bar{\mathbb{R}}$ be finite at $f(x_0)$, isotone on the union of some neighborhood of $f(x_0)$ and $\text{Range } f + \mathbb{R}_+^n$, and l.s.c. Assume that for each $j \in \{1, \dots, n\}$, either

$$(3.23) \quad f_j \text{ is continuous at } x_0, \text{ or}$$

$$(3.24) \quad F \text{ is strictly isotone in the } j\text{th coordinate at } f(x_0).$$

Assume also that

$$(3.25) \quad (\Delta^n \mathbb{R}^n \times \text{dom } F^\uparrow(f(x_0); \cdot)) - S = \mathbb{R}^{nm+n}$$

where

$$S := \{(y_1, \dots, y_n, r_1, \dots, r_n) \mid (y_i, r_i) \in \text{epi } f_i^\uparrow(x_0; \cdot), i = 1, \dots, n\}.$$

Then for all $y \in \mathbf{R}^m$,

$$(3.26) \quad (F \circ f)^\uparrow(x_0; y) \leq F^\uparrow(f(x_0); f_1^\uparrow(x_0; y), \dots, f_n^\uparrow(x_0; y)).$$

If in addition each $f_i^\uparrow(x_0; \cdot)$ is proper, then

$$(3.27) \quad \partial(F \circ f)(x_0) \subset \{\lambda \cdot (\partial f_1(x_0), \dots, \partial f_n(x_0)) \mid \lambda \in \partial F(f(x_0))\}.$$

Proof. Call $h := F \circ f$. Since F is isotone on $\text{Range } f + \mathbf{R}_+^n$,

$$\text{epi } h = \{(x, r) \in \mathbf{R}^m \times \mathbf{R}, \exists (y_1, \dots, y_n) \in \mathbf{R}^n \text{ with } F(y_1, \dots, y_n) \leq r, f_i(x) \leq y_i, 1 \leq i \leq n\}.$$

Define the functions G and A as in Lemma 3.15, and define

$$D := \text{epi } f_1 \times \dots \times \text{epi } f_n \times \text{epi } F,$$

$$C := D \cap G^{-1}(0).$$

Then $\text{epi } h = A(C)$. Assumptions (3.23) and (3.24) allow us to apply Lemma 3.15 and Proposition 2.6. Thus

$$\begin{aligned} \text{epi } h^\uparrow(x_0; \cdot) &= T_{A(C)}(x_0, h(x_0)) \quad (\text{by (1.6)}) \\ &\supset A(T_C(z_0)) \quad \text{where } z_0 \\ &:= (x_0, f_1(x_0), \dots, x_0, f_n(x_0), f_1(x_0), \dots, f_n(x_0), h(x_0)) \\ &\quad (\text{by Lemma 3.15 and Proposition 2.6}) \\ &\supset A(T_D(z_0) \cap \nabla G(z_0)^{-1}(0)) \end{aligned}$$

by Theorem 2.4, since (3.25) says exactly that

$$\begin{aligned} \nabla G(z_0) T_D(z_0) &= \mathbf{R}^{(n-1)m+n}, \\ &= \{(x, r) \in \mathbf{R}^m \times \mathbf{R} \mid \exists y \in \mathbf{R}^n \text{ with } f_i^\uparrow(x_0; x) \leq y_i, F^\uparrow(f(x_0); y) \leq r\} \\ &= \{(x, r) \mid F^\uparrow(f(x_0); f_1^\uparrow(x_0; x), \dots, f_n^\uparrow(x_0; x)) \leq r\} \end{aligned}$$

since $F^\uparrow(f(x_0); \cdot)$ is itself isotone (Lemma 2.8). Thus $\text{epi } h^\uparrow(x_0; \cdot) \supset \text{epi } F^\uparrow(f(x_0); f_1^\uparrow(x_0; \cdot), \dots, f_n^\uparrow(x_0; \cdot))$, and so (3.26) holds.

Now if $F^\uparrow(f(x_0); 0) = -\infty$, (3.26) shows that both sides of (3.27) are empty. Assume, then, that $F^\uparrow(f(x_0); \cdot)$ is proper. Since each $f_i^\uparrow(x_0; \cdot)$ is convex and proper and (3.25) implies that (2.10) and (2.12) hold for $F^\uparrow(f(x_0); \cdot)$ and $(f_1^\uparrow(x_0; \cdot), \dots, f_n^\uparrow(x_0; \cdot))$, we may apply Theorem 2.10. We have

$$\begin{aligned} \partial(F \circ f)(x_0) &= \{z \in \mathbf{R}^m \mid (F \circ f)^\uparrow(x_0; y) \geq \langle y, z \rangle \forall y \in \mathbf{R}^m\} \\ &\subset \{z \in \mathbf{R}^m \mid F^\uparrow(f(x_0); f_1^\uparrow(x_0; y), \dots, f_n^\uparrow(x_0; y)) \geq \langle y, z \rangle \forall y \in \mathbf{R}^m\} \\ &= \partial((F^\uparrow(f(x_0); \cdot)) \circ (f_1^\uparrow(x_0; \cdot), \dots, f_n^\uparrow(x_0; \cdot)))(0) \\ &= \{\lambda \cdot (\partial f_1^\uparrow(x_0; \cdot)(0), \dots, \partial f_n^\uparrow(x_0; \cdot)(0)) \mid \lambda \in \partial F^\uparrow(f(x_0); \cdot)(0)\} \\ &= \{\lambda \cdot (\partial f_1(x_0), \dots, \partial f_n(x_0)) \mid \lambda \in \partial F(f(x_0))\}. \quad \square \end{aligned}$$

Remark 3.18. (a) Assumption (3.25) reduces to a simple, familiar-looking form in important special cases. For example, if $n = 1$, (3.25) becomes

$$(3.28) \quad \text{Range } f_1^\uparrow(x_0; \cdot) - \text{dom } F^\uparrow(f_1(x_0); \cdot) = \mathbf{R}.$$

If F is as in Remark 3.16(a) or (b), (3.25) reduces to the assumption that $\text{dom } f_i^\uparrow(x_0; \cdot)$, $i = 1, \dots, n$ are in strong general position.

(b) The assumption that $f_i^\uparrow(x_0; 0) = 0$, $i = 1, \dots, n$ is not needed in important special cases. For example, if F is as in Remark 3.16(a) or (b) and $f_i^\uparrow(x_0; 0) = -\infty$ for some i , then both sides of (3.27) will equal \emptyset .

(c) Corollaries of Theorem 3.17 include Corollary 3.6, Proposition 3.10, and Proposition 3.14 (without our having to use Lemma 3.12).

(d) Any isotone function $F: \mathbf{R}^n \rightarrow \bar{\mathbf{R}}$ is directionally Lipschitzian for all $y \leq 0$; i.e.,

$$\inf_{\substack{Y \in \mathcal{N}(y) \\ N \in \mathcal{N}(x_0, F(x_0)) \\ \lambda > 0}} \sup_{\substack{y' \in Y \\ (x, \alpha) \in N \cap \text{epi } F \\ t \in (0, \lambda)}} \frac{F(x + ty') - \alpha}{t} < +\infty$$

for all $y \leq 0$ (see [19, Prop. 4]). This property plays a crucial role in the subdifferential calculus results of [18]. However, there is no analogue of Theorem 3.17 for F merely directionally Lipschitzian, as we will see in § 5.

From Theorem 3.17 we can derive an extension of Corollary 3.6 to n functions, and a product rule for (locally) nonnegative functions.

COROLLARY 3.19. *Let $f_i: \mathbf{R}^m \rightarrow \bar{\mathbf{R}}$, $i = 1, \dots, n$ be strictly l.s.c. and finite at x_0 , and suppose that $\text{dom } f_i^\uparrow(x_0; \cdot)$, $i = 1, \dots, n$ are in strong general position. Then for all $y \in \mathbf{R}^m$,*

$$(3.29) \quad (f_1 + \dots + f_n)^\uparrow(x_0; y) \leq \sum_{i=1}^n f_i^\uparrow(x_0; y).$$

Moreover,

$$(3.30) \quad \partial(f_1 + \dots + f_n)(x_0) \subset \sum_{i=1}^n \partial f_i(x_0).$$

Proof. Let $F(x_1, \dots, x_n) = \sum_{i=1}^n x_i$ in Theorem 3.17. F is continuous and strictly isotone in each coordinate. Assumption (3.25) in this case reduces to $\text{dom } f_i^\uparrow(x_0; \cdot)$, $i = 1, \dots, n$ being in strong general position. As explained in Remark 3.18(b), the assumption that each $f_i^\uparrow(x_0; 0) = 0$ is not needed. Then (3.29) follows from (3.26) and (3.30) from (3.27). \square

COROLLARY 3.20. *Let $f_i: \mathbf{R}^m \rightarrow \bar{\mathbf{R}}$, $i = 1, \dots, n$ be nonnegative on \mathbf{R}^m and strictly l.s.c. and positive at $x_0 \in \bigcap_{i=1}^n \text{dom } f_i$. Suppose that $\text{dom } f_i^\uparrow(x_0; \cdot)$, $i = 1, \dots, n$, are in strong general position. Then for all $y \in \mathbf{R}^m$,*

$$(3.31) \quad \left(\prod_{i=1}^n f_i \right)^\uparrow(x_0; y) \leq \sum_{i=1}^n \left(\prod_{j \neq i} f_j(x_0) \right) f_i^\uparrow(x_0; y).$$

Moreover,

$$(3.32) \quad \partial \left(\prod_{i=1}^n f_i \right)(x_0) \subset \sum_{i=1}^n \left(\prod_{j \neq i} f_j(x_0) \right) \partial f_i(x_0).$$

Proof. Let $F(x_1, \dots, x_n) = \prod_{i=1}^n x_i$ in Theorem 3.17. F is continuous and strictly isotone in each coordinate since f_i is nonnegative. Condition (3.25) again reduces to $\text{dom } f_i^\uparrow(x_0; \cdot)$, $i = 1, \dots, n$, being in strong general position. As explained in Remark 3.18(b), the assumption that each $f_i^\uparrow(x_0; \cdot) = 0$ is not needed. Then (3.31) follows from (3.26) and (3.32) from (3.27). \square

One special case of Corollary 3.20 is a quotient rule where the denominator is continuous and positive.

LEMMA 3.21. Suppose $g : E \rightarrow \bar{\mathbf{R}}$ is continuous and positive at $x_0 \in \text{dom } g$. Then

$$(3.33) \quad \left(\frac{1}{g}\right)^\uparrow(x_0; y) = \frac{(-g)^\uparrow(x_0; y)}{(g(x_0))^2} \quad \text{for all } y \in E,$$

and

$$(3.34) \quad \partial\left(\frac{1}{g}\right)(x_0) = \frac{1}{(g(x_0))^2} \partial(-g)(x_0).$$

Proof. First observe that (3.34) follows immediately from (3.33) because $1/(g(x_0))^2 > 0$. To prove (3.33), use (1.10) to write

$$\begin{aligned} \left(\frac{1}{g}\right)^\uparrow(x_0; y) &= \limsup_{\substack{x \rightarrow x_0 \\ t \downarrow 0}} \inf_{y' \rightarrow y} \frac{(1/g(x + ty')) - (1/g(x))}{t} \\ &= \limsup_{\substack{x \rightarrow x_0 \\ t \downarrow 0}} \inf_{y' \rightarrow y} \frac{(-g)(x + ty') - (-g)(x)}{t} \cdot \frac{1}{g(x + ty')(g(x))} \\ &= \limsup_{\substack{x \rightarrow x_0 \\ t \downarrow 0}} \inf_{y' \rightarrow y} \frac{(-g)(x + ty') - (-g)(x)}{t} \lim_{\substack{x \rightarrow x_0 \\ t \downarrow 0 \\ y' \rightarrow y}} \frac{1}{g(x + ty')g(x)} \end{aligned}$$

(by Lemma 5.2, which we prove in § 5).

$$= \frac{1}{(g(x_0))^2} (-g)^\uparrow(x_0; y). \quad \square$$

PROPOSITION 3.22. Let $f : \mathbf{R}^m \rightarrow \bar{\mathbf{R}}$ be nonnegative on $\text{dom } f$ and strictly l.s.c. and positive at x_0 , and let $g : \mathbf{R}^m \rightarrow \mathbf{R}$ be positive and continuous at $x_0 \in \text{dom } g \cap \text{dom } f$. Assume that

$$(3.35) \quad \text{dom } f^\uparrow(x_0; \cdot) - \text{dom } (-g)^\uparrow(x_0; \cdot) = \mathbf{R}^m.$$

Then

$$(3.36) \quad \left(\frac{f}{g}\right)^\uparrow(x_0; y) \leq \frac{f(x_0)(-g)^\uparrow(x_0; y) + g(x_0)f^\uparrow(x_0; y)}{(g(x_0))^2} \quad \text{for all } y \in \mathbf{R}^m$$

and

$$(3.37) \quad \partial\left(\frac{f}{g}\right)(x_0) \subset \frac{f(x_0)\partial(-g)(x_0) + g(x_0)\partial f(x_0)}{(g(x_0))^2}.$$

Proof. By Lemma 3.21, $\text{dom } (1/g)^\uparrow(x_0; \cdot) = \text{dom } (-g)^\uparrow(x_0; \cdot)$, so (3.35) ensures that we can apply Corollary 3.20 with $n = 2$, $f_1 := f$, and $f_2 := 1/g$. By (3.31) and (3.33),

$$\begin{aligned} \left(\frac{f}{g}\right)^\uparrow(x_0) &\leq f(x_0)\left(\frac{1}{g}\right)^\uparrow(x_0; y) + \frac{1}{g(x_0)}f^\uparrow(x_0; y) \\ &= \frac{f(x_0)}{(g(x_0))^2}(-g)^\uparrow(x_0; y) + \frac{1}{g(x_0)}f^\uparrow(x_0; y) \\ &= \frac{f(x_0)(-g)^\uparrow(x_0; y) + g(x_0)f^\uparrow(x_0; y)}{(g(x_0))^2}. \end{aligned}$$

Similarly, by (3.32) and (3.34),

$$\begin{aligned} \partial\left(\frac{f}{g}\right)(x_0) &\subset f(x_0)\partial\left(\frac{1}{g}\right)(x_0) + \frac{1}{g(x_0)}\partial f(x_0) \\ &= \frac{f(x_0)}{(g(x_0))^2}\partial(-g)(x_0) + \frac{1}{g(x_0)}\partial f(x_0) \\ &= \frac{f(x_0)\partial(-g)(x_0) + g(x_0)\partial f(x_0)}{(g(x_0))^2}. \end{aligned} \quad \square$$

Remark 3.23. If g is locally Lipschitzian near x_0 , then $(-g)^\uparrow(x_0; y) = -g^\uparrow(x_0; y)$, $\partial(-g)(x_0) = -\partial g(x_0)$, and (3.35) is automatically satisfied. If in addition f is locally Lipschitzian near x_0 , Proposition 3.22 almost reduces to the quotient rule of [11, Chap. 8] and [6, Chap. 2]. (The hypotheses of Proposition 3.22 include the assumption that $f(x_0) \neq 0$, which is not needed in [6] or [11].)

4. Application to constrained optimization problems.

PROPOSITION 4.1 ([18], [20]). *Suppose $f: E \rightarrow \bar{\mathbf{R}}$ has a local minimum at $x_0 \in E$. Then $0 \in \partial f(x_0)$.*

Combining Proposition 4.1 with subdifferential calculus results of the preceding section, we can give a strengthened version, in finite dimensions, of [18, Thm. 6].

PROPOSITION 4.2. *Suppose the problem*

$$\min \{f(x) \mid x \in C\}$$

has a local minimum at $x_0 \in C$, where $f: \mathbf{R}^n \rightarrow \bar{\mathbf{R}}$ is strictly l.s.c. at x_0 and $C \subset \mathbf{R}^n$ is closed near x_0 . Assume

$$(4.1) \quad \text{dom } f^\uparrow(x_0; \cdot) - T_C(x_0) = \mathbf{R}^n.$$

Then

$$(4.2) \quad 0 \in \partial f(x_0) + N_C(x_0).$$

Proof. Let $f_1 := f$, $f_2 := i_C$ in Corollary 3.6. The function $f + i_C$ is minimized at x_0 , so by Proposition 4.1, $0 \in \partial(f + i_C)(x_0)$. Assumption (4.1) is (3.8) in this case, so $\partial(f + i_C)(x_0) \subset \partial f(x_0) + \partial i_C(x_0) = \partial f(x_0) + N_C(x_0)$, and $0 \in \partial f(x_0) + N_C(x_0)$. \square

We will next consider a particular constraint set, the set $C := \{x \mid g(x) \leq 0\}$ for a strictly l.s.c. function $g: \mathbf{R}^m \rightarrow \bar{\mathbf{R}}$.

PROPOSITION 4.3 (cf. [18, Thm. 5]). *Suppose $g: \mathbf{R}^m \rightarrow \bar{\mathbf{R}}$ is strictly l.s.c. at $x_0 \in g^{-1}(0)$, and let $C := \{x \mid g(x) \leq 0\}$. Assume $0 \notin \partial g(x_0)$. Then*

$$(4.3) \quad T_C(x_0) \subset \{y \mid g^\uparrow(x_0; y) \leq 0\}.$$

If in addition $\partial g(x_0)$ is nonempty, then

$$(4.4) \quad N_C(x_0) \subset \left(\bigcup_{\lambda > 0} \lambda \partial g(x_0) \right) \cup \partial^\infty g(x_0).$$

Proof. In Proposition 3.10, let $n = 2$, $D_1 := \{(z, u) \in \mathbf{R}^m \times \mathbf{R} \mid u = 0\}$, and $D_2 := \text{epi } g$. To apply Proposition 3.10, we must verify that $T_{D_1}(x_0, 0) - T_{D_2}(x_0, 0) = \mathbf{R}^{m+1}$; i.e., that $\{(z, u) \mid u = 0\} - \text{epi } g^\uparrow(x_0; \cdot) = \mathbf{R}^{m+1}$. To do so, suppose that $(x', r') \in \mathbf{R}^m \times \mathbf{R}$. Since $0 \notin \partial g(x_0)$, there exists $y \in \mathbf{R}^m$ with $g^\uparrow(x_0; y) < 0$. Now $g^\uparrow(x_0; \cdot)$ is positively homogeneous, so for any $\bar{r} \in \mathbf{R}$, there exists $\bar{y} \in \mathbf{R}^m$ with $(\bar{y}, \bar{r}) \in \text{epi } g^\uparrow(x_0; \cdot)$. Let y' be such that $g^\uparrow(x_0; y') \leq -r'$. Then $(x', r') = (x' + y', 0) - (y', -r')$. Thus $\{(z, u) \mid u = 0\} - \text{epi } g^\uparrow(x_0; \cdot) = \mathbf{R}^{m+1}$. Now by (3.17), $T_{D_1 \cap D_2}(x_0, 0) \supset T_{D_1}(x_0, 0) \cap T_{D_2}(x_0, 0)$; i.e.,

$T_{C \times 0}(x_0, 0) \supset (\mathbf{R}^m \times 0) \cap \text{epi } g^\uparrow(x_0; \cdot)$. Hence $T_C(x_0) \supset \{y \mid g^\uparrow(x_0; y) \leq 0\}$. By (3.18), $N_{D_1 \cap D_2}(x_0, 0) \subset N_{D_1}(x_0, 0) + N_{D_2}(x_0, 0)$; i.e., $N_{C \times 0}(x_0, 0) \subset 0 \times \mathbf{R} + N_{D_2}(x_0, 0)$. If $\partial g(x_0) \neq \emptyset$, we have by (3.20) that

$$N_{D_2}(x_0) = \bigcup_{\lambda > 0} \lambda(\partial g(x_0), -1) \cup (\partial^\infty g(x_0), 0).$$

Hence

$$N_{C \times 0}(x_0, 0) \subset (0 \times \mathbf{R}) + \left(\bigcup_{\lambda > 0} \lambda(\partial g(x_0), -1) \cup (\partial^\infty g(x_0), 0) \right)$$

and (4.4) holds.

By combining Propositions 4.2 and 4.3 and applying Proposition 3.10, we can obtain the following Kuhn–Tucker Theorem:

THEOREM 4.4. *Let $f: \mathbf{R}^m \rightarrow \bar{\mathbf{R}}$ and $g_i: \mathbf{R}^m \rightarrow \bar{\mathbf{R}}$, $i = 1, \dots, n$, be strictly l.s.c. at x_0 , $D \subset \mathbf{R}^m$ closed near x_0 , and $G: \mathbf{R}^m \rightarrow \mathbf{R}^p$ strictly differentiable at x_0 , a local minimizer of*

$$(P) \quad \min \{f(x) \mid g_i(x) \leq 0, i = 1, \dots, n, G(x) = 0, x \in D\}.$$

Call $I(x) := \{i \in \{1, \dots, n\} \mid g_i(x) = 0\}$. Assume that $0 \notin \partial g_i(x_0)$ and $\partial g_i(x_0) \neq \emptyset$ for each $i \in I(x_0)$, that g_i is continuous at x_0 for each $i \notin I(x_0)$, and that $\nabla G(x_0)\mathbf{R}^m = \mathbf{R}^p$. In addition, assume that $\text{dom } f^\uparrow(x_0; \cdot)$, $\nabla G(x_0)^{-1}(0)$, $T_D(x_0)$, and $\{y \mid g^\uparrow(x_0; y) \leq 0\}$, $i \in I(x_0)$, are in strong general position. Then

$$(4.5) \quad 0 \in \partial f(x_0) + \sum_{i \in I(x_0)} (\lambda_i \partial g_i(x_0) \cup \partial g_i^\infty(x_0)) + \lambda \nabla G(x_0) + N_D(x_0)$$

for some $\lambda_i \geq 0$ and $\lambda \in \mathbf{R}$.

Proof. Denote by C the constraint set in Program (P). Since $\nabla G(x_0)\mathbf{R}^m = \mathbf{R}^p$, $\nabla G(x_0)^{-1}(0) \subset T_{G^{-1}(0)}(x_0)$ by Theorem 2.4. (In fact, these two sets are equal). By our strong general position assumption, we have

$$\{y \mid g^\uparrow(x_0; y) \leq 0, i \in I(x_0), y \in T_D(x_0), \nabla G(x_0)y = 0\} \subset T_C(x_0)$$

by Propositions 4.3 and 3.10. Thus the strong general position assumption guarantees that (4.1) holds. By Proposition 4.2,

$$0 \in \partial f(x_0) + N_C(x_0),$$

and (4.5) then follows immediately from (3.18) of Proposition 3.10 and Proposition 4.3. \square

Remark 4.5. In the case in which f and g_i , $i \in I(x_0)$ are strictly differentiable, our general position assumption and $\nabla G(x_0)\mathbf{R}^m = \mathbf{R}^p$ comprise the familiar Mangasarian–Fromovitz constraint qualification.

It is also possible to derive necessary optimality conditions for (P) via the following “Dubovitskii–Milyutin” result, due to Watkins [26]:

THEOREM 4.6. *Let $C_i \subset \mathbf{R}^m$, $i = 1, \dots, p$ be closed near x_0 with $\bigcap_{i=1}^p C_i = \{x_0\}$, and suppose that at least one of the sets $T_{C_i}(x_0)$, $i = 1, \dots, p$ is not a subspace. Then there exist $a_i \in N_{C_i}(x_0)$, $i = 1, \dots, p$, not all equal to zero, such that $\sum_{i=1}^p a_i = 0$.*

In [26], a “Fritz John type” theorem is derived from Theorem 4.6 for a constrained optimization problem involving locally Lipschitzian functions. Using Proposition 4.3, we can generalize this result beyond the locally Lipschitzian case.

THEOREM 4.7. *Posit the assumptions of Theorem 4.4, with the exception of the strong general position hypothesis. Assume, in addition, that f is directionally Lipschitzian at x_0 , that $\partial f(x_0)$ is nonempty, and that $0 \notin \partial f(x_0)$. Then*

$$(4.6) \quad 0 \in (\lambda_0 \partial f(x_0) \cup \partial^\infty f(x_0)) + \sum_{i \in I(x_0)} (\lambda_i \partial g_i(x_0) \cup \partial^\infty g_i(x_0)) + \lambda \nabla G(x_0) + N_D(x_0)$$

for some $\lambda_0 \geq 0$, $\lambda_i \geq 0$, and $\lambda \in \mathbf{R}$. Moreover, λ_0 , λ_i , $i \in I(x_0)$, λ , and the elements of $\partial^\infty f(x_0)$, $\partial^\infty g_i(x_0)$, and $N_D(x_0)$ in (4.6) are not all equal to zero.

Proof. Again call C the feasible set in (P). There exists $\varepsilon > 0$ such that $f(x) \geq f(x_0)$ for all $x \in B_\varepsilon(x_0)$. Let

$$C_0 = \{x \in B_\varepsilon(x_0) \mid f(x) \leq f(x_0) - \|x - x_0\|^2\},$$

$$C_i := \{x \in \mathbf{R}^n \mid g_i(x) \leq 0\}, i = 1, \dots, n,$$

$$C_{n+1} := \{x \in \mathbf{R}^n \mid G(x) = 0\} \quad \text{and} \quad C_{n+2} = D.$$

Since x_0 is a local minimizer for (P), $\bigcap_{i=0}^{n+2} C_i = \{x_0\}$. Consider first the case in which $T_{C_0}(x_0)$ is not a subspace. Then by Theorem 4.6, $0 \in \sum_{i=0}^{n+2} N_{C_i}(x_0)$. Call $\tilde{f}(x) := f(x) - f(x_0) + \|x - x_0\|^2$. By Corollary 3.6, $\partial \tilde{f}(x_0) \subset \partial f(x_0)$ and $\partial^\infty \tilde{f}(x_0) = \partial^\infty f(x_0)$ (see [22, Prop. 2.4]). Thus

$$\begin{aligned} N_{C_0}(x_0) &\subset \bigcup_{\lambda > 0} \lambda \partial \tilde{f}(x_0) \cup \partial^\infty \tilde{f}(x_0) \\ &\subset \bigcup_{\lambda > 0} \lambda \partial f(x_0) \cup \partial^\infty f(x_0). \end{aligned}$$

Again $\nabla G(x_0)^{-1}(0) = T_{C_{n+1}}(x_0)$. For $i \in \{1, \dots, n\} \setminus I(x_0)$, $N_{C_i}(x_0) = \{0\}$. Combining these facts and (4.4), we obtain (4.6).

Now suppose $T_{C_0}(x_0)$ is a subspace. Since f is directionally Lipschitzian at x_0 and $0 \notin \partial f(x_0)$, $T_{C_0}(x_0)$ has nonempty interior [18], which means that $T_{C_0}(x_0) = \mathbf{R}^n$. Hence $x_0 \in \text{int } C_0$ [17], contradicting the fact that x_0 is a local minimizer for (P). \square

The additional assumption that $\nabla G(x_0)^{-1}(0)$, $T_D(x_0)$, and $\{y \mid g_i^+(x_0; y) \leq 0\}$, $i \in I(x_0)$ are in strong general position is enough to guarantee that the element of $\lambda_0 \partial f(x_0) \cup \partial^\infty f(x_0)$ in (4.6) is nonzero (see [26, Remarks 1, 2]). This produces another Kuhn-Tucker type result.

5. Limitations of the generalized subdifferential calculus. In § 3, we obtain “chain rules” for $(F \circ f)^\dagger(x_0; \cdot)$ and $\partial(F \circ f)(x_0)$ with F and f of these types:

(a) $F: \mathbf{R}^m \rightarrow \mathbf{R}$ l.s.c., $f: \mathbf{R}^m \rightarrow \mathbf{R}^n$ strictly differentiable at x_0 .

(b) $F: \mathbf{R}^n \rightarrow \bar{\mathbf{R}}$ l.s.c. and isotone, $f = (f_1, \dots, f_n)$ with each $f_i: \mathbf{R}^m \rightarrow \bar{\mathbf{R}}$ l.s.c., and (3.23) or (3.24) satisfied.

(In fact, (a) can be encompassed by (b)—e.g. [18, Thm. 3]).

Are other general chain rules possible? In particular, is there a chain rule for $F: \mathbf{R}^n \rightarrow \mathbf{R}$ directionally Lipschitzian (or strictly differentiable) and $f_i: \mathbf{R}^m \rightarrow \bar{\mathbf{R}}$ l.s.c.? We answer this question by considering possible extensions of the following chain rule for Clarke subgradients. (See [5, Prop. 10] and [11, Chap. 8].)

THEOREM 5.1. *Let E be a normed space and $f: E \rightarrow \bar{\mathbf{R}}$ be locally Lipschitzian near x_0 , and let $F: \mathbf{R} \rightarrow \mathbf{R}$ be continuously differentiable near x_0 . Then*

$$(5.1) \quad (F \circ f)^0(x_0; y) = F'(f(x_0))f^0(x_0; y) \quad \text{for all } y \in E,$$

and

$$(5.2) \quad \partial(F \circ f)(x_0) = F'(f(x_0))\partial f(x_0).$$

In attempting to extend Theorem 5.1, it does not seem possible to apply the techniques used in the proofs of Theorems 3.2 and 3.17. We will proceed instead by examining limits of the form (1.10). We start by deriving a technical lemma involving these limits which we have already applied in the proof of Lemma 3.21.

LEMMA 5.2. Let E, E_1 be l.c.t.v.s., and let $g: E \times E_1 \rightarrow \bar{\mathbf{R}}$ and $h: E \times E_1 \rightarrow \bar{\mathbf{R}}$ be extended-real-valued functions. Suppose

$$\lim_{\substack{x \rightarrow x_0 \\ y \rightarrow y_0}} h(x, y) = A \quad \text{and} \quad \limsup_{x \rightarrow x_0} \inf_{y \rightarrow y_0} g(x, y) = B.$$

Then

- (i) $\limsup_{x \rightarrow x_0} \inf_{y \rightarrow y_0} h(x, y)g(x, y) = AB$ if $0 < A < +\infty$.
(ii) The same equation holds for $-\infty < A \leq 0$ if g is bounded in a neighborhood of (x_0, y_0) .

Proof of (i). For any given $\lambda \in (0, A)$, there exist $X_1 \in \mathcal{N}(x_0)$, $Y_1 \in \mathcal{N}(y_0)$ such that $A - \lambda \leq h(x, y) \leq A + \lambda$ for all $x \in X_1$, $y \in Y_1$. We first establish $\limsup_{x \rightarrow x_0} \inf_{y \rightarrow y_0} h(x, y)g(x, y) \leq AB$.

Case 1. $0 \leq B < +\infty$. Let $Y \in \mathcal{N}(y)$ and $\varepsilon > 0$ be given. Choose $\lambda \in (0, A)$ small enough so that $(A + B)\lambda + \lambda^2 < \varepsilon$ and X_1, Y_1 as above. Then there exists $X_2 \in \mathcal{N}(x_0)$ such that for all $x \in X_2$, there exists $y \in Y \cap Y_1$ with $g(x, y) \leq B + \lambda$. Thus for all $x \in X_1 \cap X_2$, there exists $y \in Y \cap Y_1$ with $h(x, y)g(x, y) \leq (A + \lambda)(B + \lambda) = AB + (A + B)\lambda + \lambda^2 \leq AB + \varepsilon$. Since ε was arbitrary, $\limsup_{x \rightarrow x_0} \inf_{y \rightarrow y_0} h(x, y)g(x, y) \leq AB$.

Case 2. $-\infty < B < 0$. Let $Y \in \mathcal{N}(y)$ and $\varepsilon > 0$ be given. Let $\lambda \in (0, \min(A, -B))$ be such that $(A - B)\lambda \leq \varepsilon$, and choose X_1, Y_1 as above. There exists $X_2 \in \mathcal{N}(x_0)$ such that for each $x \in X_2$, there exists $y \in Y \cap Y_1$ with $g(x, y) \leq B + \lambda$. So for each $x \in X_1 \cap X_2$, there exists $y \in Y \cap Y_1$ with $h(x, y)g(x, y) \leq (A - \lambda)(B + \lambda) = AB + (A - B)\lambda - \lambda^2 \leq AB + \varepsilon$. Since ε was arbitrary, $\limsup_{x \rightarrow x_0} \inf_{y \rightarrow y_0} h(x, y)g(x, y) \leq AB$.

Case 3. $B = -\infty$. Let $m < 0$ and $Y \in \mathcal{N}(y)$ be given, and choose $\lambda \in (0, A)$ and X_1, Y_1 as above. There exists $X_2 \in \mathcal{N}(x_0)$ such that for all $x \in X_2$, there exists $y \in Y \cap Y_1$ with $g(x, y) \leq m/(A - \lambda)$. Then for all $x \in X_1 \cap X_2$, there exists $y \in Y \cap Y_1$ such that $h(x, y)g(x, y) \leq (A - \lambda)(m/(A - \lambda)) = m$. Since m was arbitrary, $\limsup_{x \rightarrow x_0} \inf_{y \rightarrow y_0} h(x, y)g(x, y) = -\infty = AB$.

Case 4. $B = +\infty$. Choose $\lambda \in (0, A)$, and let $X \in \mathcal{N}(x_0)$ and $M > 0$ be given. Choose X_1, Y_1 as above. There exists $Y \in \mathcal{N}(y_0)$ such that for some $x \in X \cap X_1$, $g(x, y) \geq M$ for all $y \in Y$. So there exists $x \in X \cap X_1$ such that for all $y \in Y \cap Y_1$, $h(x, y)g(x, y) \geq M(A - \lambda)$. Since M and X were arbitrary, $\lim_{x \rightarrow x_0} \sup \inf_{y \rightarrow y_0} g(x, y)h(x, y) = +\infty = AB$. We have now established that $\limsup_{x \rightarrow x_0} \inf_{y \rightarrow y_0} g(x, y)h(x, y) \leq AB$. It then follows that

$$\begin{aligned} B &= \limsup_{x \rightarrow x_0} \inf_{y \rightarrow y_0} g(x, y) \\ &\leq \lim_{\substack{x \rightarrow x_0 \\ y \rightarrow y_0}} \frac{1}{h(x, y)} \limsup_{x \rightarrow x_0} \inf_{y \rightarrow y_0} g(x, y)h(x, y) \\ &= \frac{1}{A} \limsup_{x \rightarrow x_0} \inf_{y \rightarrow y_0} g(x, y)h(x, y). \end{aligned}$$

Hence $\limsup_{x \rightarrow x_0} \inf_{y \rightarrow y_0} g(x, y)h(x, y) \geq AB$, and (i) holds.

We leave the proof of (ii), which is similar, to the reader. We make the important observation that if $A \leq 0$, local boundedness of g is required in both the proof of $\limsup_{x \rightarrow x_0} \inf_{y \rightarrow y_0} h(x, y)g(x, y) \leq AB$ and that of $\limsup_{x \rightarrow x_0} \inf_{y \rightarrow y_0} h(x, y)g(x, y) \geq AB$. \square

We can now prove what seems to be the best possible extension of Theorem 5.1.

THEOREM 5.3. *Let E be an l.c.t.v.s. and $f: E \rightarrow \bar{\mathbf{R}}$ continuous at $x_0 \in E$, and let $F: \mathbf{R} \rightarrow \mathbf{R}$ be continuously differentiable near $f(x_0)$. Assume $F'(f(x_0)) > 0$. Then for all $y \in E$,*

$$(5.3) \quad (F \circ f)^\uparrow(x_0; y) = F'(f(x_0))f^\uparrow(x_0; y).$$

Moreover,

$$(5.4) \quad \partial(F \circ f)(x_0) = F'(f(x_0))\partial f(x_0).$$

Proof.

$$\begin{aligned} (F \circ f)^\uparrow(x_0; y) &= \limsup_{\substack{x \rightarrow x_0 \\ t \downarrow 0}} \inf_{y' \rightarrow y} \frac{(F \circ f)(x + ty') - (F \circ f)(x)}{t} \\ &= \limsup_{\substack{x \rightarrow x_0 \\ t \downarrow 0}} \inf_{y' \rightarrow y} F'(z(t, x, y)) \frac{f(x + ty') - f(x)}{t} \end{aligned}$$

(for some $z(t, x, y)$ between $f(x)$ and $f(x + ty')$ by the classical mean value theorem)

$$= \lim_{\substack{x \rightarrow x_0 \\ t \downarrow 0 \\ y' \rightarrow y}} F'(z(t, x, y)) \limsup_{\substack{x \rightarrow x_0 \\ t \downarrow 0}} \inf_{y' \rightarrow y} \frac{f(x + ty') - f(x)}{t}$$

(by Lemma 5.2(i))

$$= F'(f(x_0))f^\uparrow(x_0; y)$$

since f is continuous at x_0 and F is continuously differentiable near $f(x_0)$.

Now if $f^\uparrow(x_0; 0) = -\infty$, $(F \circ f)^\uparrow(x_0; 0)$ will by (5.3) also equal $-\infty$, and both sides of (5.4) will be empty. Otherwise,

$$\begin{aligned} \partial(F \circ f)(x_0) &= \partial(F \circ f)^\uparrow(x_0; \cdot)(0) \\ &= \partial(F'(f(x_0))f^\uparrow(x_0; \cdot))(0) \\ &= F'(f(x_0))\partial f^\uparrow(x_0; \cdot)(0) \end{aligned}$$

since $\partial f^\uparrow(x_0; \cdot)(0)$ is nonempty

$$= F'(f(x_0))\partial f(x_0). \quad \square$$

Remark 5.4. (a) If $F'(f(x_0)) < 0$, neither inclusion of (5.4) will hold in general. For example, define $F: \mathbf{R} \rightarrow \mathbf{R}$ by $F(x) := -x$ and $f: \mathbf{R} \rightarrow \mathbf{R}$ by $f(x) := -|x|^{1/2}$. Then $\partial(F \circ f)(0) = \mathbf{R}$, but $\partial f(0) = \emptyset$. On the other hand, if $F(x) := -x$ but $f(x) := |x|^{1/2}$, $\partial(F \circ f)(x_0) = \emptyset$ while $F'(f(x_0))\partial f(x_0) = \mathbf{R}$.

(b) There are also counterexamples for $F'(f(x_0)) = 0$. Define $F: \mathbf{R} \rightarrow \mathbf{R}$ by $F(x) := x^2$ and $f: \mathbf{R} \rightarrow \mathbf{R}$ as in (a). Then $\partial(F \circ f)(0) = [-1, 1]$, but $F'(f(x_0)) = 0$ and $\partial f(0) = \emptyset$. On the other hand, if $F(x) := -x^{4/3}$ and $f(x) := |x|^{1/2}$, then $\partial(F \circ f)(0) = \emptyset$ while $F'(f(0)) = 0$ and $\partial f(0) = \mathbf{R}$.

We now have at least a partial answer to our original questions. Even if F is continuously differentiable and f is continuous, the requirement that F be locally isotone seems to be needed in order to prove a chain rule for upper subderivatives.

The difficulty seems to lie in the “nonradial” nature of ∂f for general f . If f is locally Lipschitzian, $f^0(x_0; -y) = (-f)^0(x_0; y)$ and $\partial(-f)(x_0) = -\partial f(x_0)$. These relationships would allow us to extend Theorem 5.3 to $F'(f(x_0)) < 0$ if they held in general, but they do not—e.g. consider $f: \mathbf{R} \rightarrow \mathbf{R}$ defined by $f(x) := |x|^{1/2}$ and $x_0 = 0$.

A more thorough discussion of the subject of this section is given in [24, § 3.5].

6. Contingent derivatives, Ursescu derivatives and subdifferential regularity. In this section, we consider the tangent cones

$$(6.1) \quad K_C(x_0) := \{y \in E \mid \forall Y \in \mathcal{N}(y), \forall \lambda > 0, \exists t \in (0, \lambda), \\ \exists y' \in Y \text{ such that } x_0 + ty' \in C\}$$

and

$$(6.2) \quad k_C(x_0) := \{y \in E \mid \forall Y \in \mathcal{N}(y), \exists \lambda > 0 \text{ such that } \forall t \in (0, \lambda), \exists y' \in Y \text{ with } x_0 + ty' \in C\}.$$

$K_C(x_0)$ is the well-known *contingent cone*. For more information on these cones, see [23], [7], [25], [24].

From (6.1) and (6.2), we see that

$$T_C(x_0) \subset k_C(x_0) \subset K_C(x_0)$$

for any $C \subset E$ and $x_0 \in \text{cl } C$. If C is a convex set or a differentiable manifold, all three cones coincide ([20, Chap. 2] and Theorem 2.4).

Let $f: E \rightarrow \bar{\mathbf{R}}$ be an extended real-valued function. Paralleling the development in § 1, define

$$(6.3) \quad f_+(x_0; y) := \inf \{r \mid (y, r) \in K_{\text{epi } f}(x_0, f(x_0))\}$$

and

$$(6.4) \quad f_{\square}(x_0; y) := \inf \{r \mid (y, r) \in k_{\text{epi } f}(x_0, f(x_0))\}.$$

The directional derivative $f_+(x_0; y)$ (respectively, $f_{\square}(x_0; y)$) has been called the *contingent derivative* (*Ursescu derivative*) of f at x_0 in the direction y [1], [9]. Since $T_{\text{epi } f}(x_0, f(x_0)) \subset k_{\text{epi } f}(x_0, f(x_0)) \subset K_{\text{epi } f}(x_0, f(x_0))$, $f_+(x_0; \cdot) \leq f_{\square}(x_0; \cdot) \leq f^{\uparrow}(x_0; \cdot)$. Definitions (6.3) and (6.4) are made exactly so that

$$(6.5) \quad \text{epi } f_+(x_0; \cdot) := K_{\text{epi } f}(x_0, f(x_0))$$

and

$$(6.6) \quad \text{epi } f_{\square}(x_0; \cdot) := k_{\text{epi } f}(x_0, f(x_0)).$$

There are also direct characterizations of $f_+(x_0; y)$ and $f_{\square}(x_0; y)$. It is well known that

$$(6.7) \quad f_+(x_0; y) = \liminf_{\substack{t \downarrow 0 \\ y' \rightarrow y}} \frac{f(x_0 + ty') - f(x_0)}{2}$$

(e.g., [20, Chap. 3]), and it is not difficult to show [7], [25], [9] that

$$(6.8) \quad f_{\square}(x_0; y) = \limsup_{t \downarrow 0} \inf_{y' \rightarrow y} \frac{f(x_0 + ty') - f(x_0)}{t}.$$

$K_C(x_0)$ and $k_C(x_0)$ are closed cones, so $f_+(x_0; \cdot)$ and $f_{\square}(x_0; \cdot)$ are l.s.c. and positively homogeneous. $K_C(x_0)$ and $k_C(x_0)$ are not in general convex, so it is not possible to directly associate convex subgradient sets with $f_+(x_0; \cdot)$ and $f_{\square}(x_0; \cdot)$ as we did with $f^{\uparrow}(x_0; \cdot)$ in Definition 1.3. It is possible, however, to prove calculus formulas for $f_+(x_0; \cdot)$ and $f_{\square}(x_0; \cdot)$, and in particular to derive the subdifferential calculus for convex functions by combining these formulae with those in § 3. We now proceed to do so, beginning with a counterpart to Theorem 3.17. Notice that it holds quite generally, with no need for the assumption of finite dimensionality or conditions (3.23), (3.24) and (3.25).

PROPOSITION 6.1. Let E be an l.c.t.v.s., let $f_i: E \rightarrow \bar{\mathbf{R}}$, $i = 1, \dots, n$ be finite at x_0 , and call $f := (f_1, \dots, f_n)$. Let $F: \mathbf{R}^n \rightarrow \bar{\mathbf{R}}$ be finite at $f(x_0)$ and isotone on the union of $\text{Range } f + \mathbf{R}_+^n$ and $B_\varepsilon(f(x_0))$ for some $\varepsilon > 0$. Then if $y \in E$ is such that $(f_i)_+(x_0; y)$, $i = 1, \dots, n$ are finite,

$$(6.9) \quad (F \circ f)_+(x_0; y) \geq F_+(f(x_0); (f_1)_+(x_0; y), \dots, (f_n)_+(x_0; y)).$$

Alternatively, if $(f_1)_\square(x_0; y)$, $(f_i)_+(x_0; y)$, $i = 2, \dots, n$ are finite, then

$$(6.10) \quad (F \circ f)_\square(x_0; y) \geq F_+(f(x_0); (f_1)_\square(x_0; y), (f_2)_+(x_0; y), \dots, (f_n)_+(x_0; y)).$$

Proof. Let $(y, r) \in K_{\text{epi}(F \circ f)}(x_0, (F \circ f)(x_0))$. To prove (6.9), it suffices to show that

$$((f_1)_+(x_0; y), \dots, (f_n)_+(x_0; y), r) \in K_{\text{epi } F}(f(x_0), F(f(x_0))).$$

To this end, let $\lambda > 0$ and $\delta > 0$ be given. Call $z_i = (f_i)_+(x_0; y) - \delta$, and choose $\lambda_0 \in (0, \lambda)$ such that for all i , $f_i(x_0) + [0, \lambda_0]z_i \geq f_i(x_0) - \varepsilon$. Then there exist $Y_i \in \mathcal{N}(y)$, $\lambda_i \in (0, \lambda_0)$ such that for all $t \in (0, \lambda_i)$ and all $y' \in Y_i$,

$$\frac{f_i(x_0 + ty') - f_i(x_0)}{t} \geq z_i, \quad i = 1, \dots, n.$$

In addition, there exist $\bar{y} \in \bigcap_{i=1}^n Y_i$, $\bar{t} \in (0, \min_i \lambda_i)$ such that

$$\frac{(F \circ f)(x_0 + \bar{t}\bar{y}) - (F \circ f)(x_0)}{\bar{t}} \leq r + \delta.$$

Now since $f_i(x_0) + \bar{t}z_i \leq f_i(x_0 + \bar{t}\bar{y})$ and F is isotone,

$$\frac{F(f(x_0) + \bar{t}(z_1, \dots, z_n)) - F(f(x_0))}{\bar{t}} \leq r + \delta.$$

Thus $((f_1)_+(x_0; y), \dots, (f_n)_+(x_0; y), r) \in K_{\text{epi } F}(f(x_0), F(f(x_0)))$, and so (6.9) holds.

Now let $(y, r) \in k_{\text{epi}(F \circ f)}(x_0, (F \circ f)(x_0))$. To prove (6.10), it suffices to show that $((f_1)_\square(x_0; y), (f_2)_+(x_0; y), \dots, (f_n)_+(x_0; y), r) \in K_{\text{epi } F}(f(x_0), F(f(x_0)))$. To this end, let $\delta > 0$ and $\lambda > 0$ be given. Call $z_1 = (f_1)_\square(x_0; y) - \delta$, $z_i = (f_i)_+(x_0; y) - \delta$, $i = 2, \dots, n$. Again choose $\lambda_0 \in (0, \lambda)$ such that $f_i(x_0) + [0, \lambda_0]z_i \geq f_i(x_0) - \varepsilon$. There exist $Y_i \in \mathcal{N}(y)$, $\lambda_i \in (0, \lambda_0)$ such that for all $t \in (0, \lambda_i)$ and all $y' \in Y_i$, $((f_i(x_0 + ty') - f_i(x_0))/t) \geq z_i$, $i = 2, \dots, n$, and there exists $Y_1 \in \mathcal{N}(y)$ such that for all $t > 0$, there exists $t' \in (0, t)$ with $((f_1(x_0 + t'y) - f_1(x_0))/t') \geq z_1$ for all $y' \in Y_1$. In addition, there exists $\bar{y} \in \bigcap_{i=1}^n Y_i$ and $\lambda_1 \in (0, \lambda_0)$ such that for all $t \in (0, \lambda_1)$, $((F \circ f)(x_0 + t\bar{y}) - (F \circ f)(x_0))/t \leq r + \delta$. Thus for some $\bar{t} \in (0, \min_i \lambda_i)$, $f_i(x_0) + \bar{t}z_i \leq f_i(x_0 + \bar{t}\bar{y})$. By the isotonicity of F , $((F(f(x_0) + \bar{t}(z_1, \dots, z_n)) - F(f(x_0)))/\bar{t}) \leq r + \delta$. Hence $((f_1)_\square(x_0; y), (f_2)_+(x_0; y), \dots, (f_n)_+(x_0; y), r) \in K_{\text{epi } F}(f(x_0), F(f(x_0)))$, and so (6.10) holds. \square

Inequality (6.9) does not hold with “+” replaced by “ \square ”. For example, let $n = 2$ and $F(x_1, x_2) := x_1 + x_2$; and consider $f_1: \mathbf{R} \rightarrow \mathbf{R}$ and $f_2: \mathbf{R} \rightarrow \mathbf{R}$ defined by

$$f_1(x) := \begin{cases} 0 & \text{if } x \leq 0, \\ x & \text{if } \frac{1}{2^{n+1}} < x \leq \frac{1}{2^n} \text{ and } n \text{ is odd,} \\ 2x & \text{if } \frac{1}{2^{n+1}} < x \leq \frac{1}{2^n} \text{ and } n \text{ is even,} \end{cases}$$

$$f_2(x) := \begin{cases} 0 & \text{if } x \leq 0, \\ x & \text{if } \frac{1}{2^{n+1}} < x \leq \frac{1}{2^n} \text{ and } n \text{ is even,} \\ 2x & \text{if } \frac{1}{2^{n+1}} < x \leq \frac{1}{2^n} \text{ and } n \text{ is odd.} \end{cases}$$

Then if $y \geq 0$,

$$(f_1)_{\square}(0; y) = (f_2)_{\square}(0; y) = 2y$$

while

$$(f_1 + f_2)(x) = \begin{cases} 0 & \text{if } x \leq 0, \\ 3x & \text{if } x > 0, \end{cases}$$

so that $(f_1 + f_2)_{\square}(0; y) = 3y$. Thus for $y > 0$,

$$(f_1 + f_2)_{\square}(0; y) = 3y < 4y = (f_1)_{\square}(0; y) + (f_2)_{\square}(0; y).$$

PROPOSITION 6.2. *Let E, E^1 be l.c.t.v.s., let $F: E \rightarrow E^1$ be strictly differentiable at $x_0 \in E$, and let $f: E^1 \rightarrow \bar{\mathbf{R}}$ be finite at $F(x_0)$. Then for all $y \in E^1$,*

$$(6.11) \quad (f \circ F)_+(x_0; y) \geq f_+(F(x_0); \nabla F(x_0)y)$$

and

$$(6.12) \quad (f \circ F)_{\square}(x_0; y) \geq f_{\square}(F(x_0); \nabla F(x_0)y).$$

Proof. Let $(y, r) \in K_{\text{epi}(f \circ F)}(x_0, (f \circ F)(x_0))$. To prove (6.11), it suffices to show that $(\nabla F(x_0)y, r) \in K_{\text{epi } f}(F(x_0), f(F(x_0)))$. Let $Z \in \mathcal{N}(\nabla F(x_0)y)$, $\lambda > 0$, and $\delta > 0$ be given. There exist $Y \in \mathcal{N}(y)$ and $\lambda_1 \in (0, \lambda)$ such that $((F(x_0 + ty') - F(x_0))/t) \in Z$ for all $y' \in Y$ and $t \in (0, \lambda_1)$. In addition, there exist $\bar{y} \in Y$ and $\bar{t} \in (0, \lambda_1)$ such that $((f \circ F)(x_0 + \bar{t}\bar{y}) - (f \circ F)(x_0))/\bar{t} \leq r + \delta$. Call $\bar{z} := ((F(x_0 + \bar{t}\bar{y}) - F(x_0))/\bar{t})$. Then $\bar{z} \in Z$ and $((f(F(x_0) + \bar{t}\bar{z}) - f(F(x_0)))/\bar{t}) \leq r + \delta$. Hence $(\nabla F(x_0)y, r) \in K_{\text{epi } f}(F(x_0), f(F(x_0)))$, and (6.11) holds. The proof of (6.12) is very similar to the proof of (6.11), and we leave it to the reader. \square

We can now give conditions under which equality holds in Theorems 3.2 and 3.17.

DEFINITION 6.3. The function $f: E \rightarrow \bar{\mathbf{R}}$ is said to be *subdifferentially regular* at $x_0 \in \text{dom } f$ if $f_+(x_0; y) = f^{\dagger}(x_0; y)$ for all $y \in E$. It is said to be *subdifferentially weakly regular* at x_0 if $f_{\square}(x_0; y) = f^{\dagger}(x_0; y)$ for all $y \in E$.

PROPOSITION 6.4. *Suppose the hypotheses of Theorem 3.17 are satisfied, including the assumption that each $f_i^{\dagger}(x_0; \cdot)$ is proper. Assume in addition that F is subdifferentially regular at $f(x_0)$, f_2, \dots, f_n are subdifferentially regular at x_0 , and f_1 is subdifferentially weakly regular at x_0 . Then equality holds in (3.26) and (3.27).*

Proof. By (6.10) and our regularity hypothesis, for all $y \in \mathbf{R}^n$,

$$\begin{aligned} (F \circ f)^{\dagger}(x_0; y) &\geq (F \circ f)_{\square}(x_0; y) \\ &\geq F_+(f(x_0); (f_1)_{\square}(x_0; y), (f_2)_+(x_0; y), \dots, (f_n)_+(x_0; y)) \\ &= F^{\dagger}(f(x_0); f_1^{\dagger}(x_0; y), \dots, f_n^{\dagger}(x_0; y)), \end{aligned}$$

so equality holds in (3.26). Equality in (3.27) follows from equality in (3.26). \square

PROPOSITION 6.5. *Suppose the hypotheses of Theorem 3.2 are satisfied, and that either*

$$(6.13) \quad f_1 \text{ is subdifferentially regular at } x_0 \text{ and } f_2 \text{ is subdifferentially weakly regular at } F(x_0), \text{ or}$$

$$(6.14) \quad f_1 \text{ is subdifferentially weakly regular at } x_0, \text{ and } f_2 \text{ is subdifferentially regular at } F(x_0).$$

Then equality holds in (3.3). If in addition $f_1^\uparrow(x_0; 0) = f_2^\uparrow(F(x_0); 0) = 0$, equality holds in (3.2) also.

Proof. If either $f_1^\uparrow(x_0; 0)$ or $f_2^\uparrow(F(x_0); 0)$ is $-\infty$, (3.2) implies that $(f_1 + f_2 \circ F)^\uparrow(x_0; 0) = -\infty$, and both sides of (3.3) are empty. Otherwise apply Proposition 6.1 with $n = 2$ and $F: \mathbf{R}^2 \rightarrow \mathbf{R}$ defined by $F(x_1, x_2) = x_1 + x_2$. If (6.13) holds, then for all $y \in \mathbf{R}^n$,

$$\begin{aligned} (f_1 + f_2 \circ F)^\uparrow(x_0; y) &\geq (f_1 + f_2 \circ F)_\square(x_0; y) \\ &\geq (f_1)_+(x_0; y) + (f_2 \circ F)_\square(x_0; y) \quad (\text{by (6.10)}) \\ &\geq (f_1)_+(x_0; y) + (f_2)_\square(F(x_0); \nabla F(x_0)y) \quad (\text{by (6.12)}) \\ &= (f_1)^\uparrow(x_0; y) + f_2^\uparrow(F(x_0); \nabla F(x_0)y). \end{aligned}$$

If (6.14) holds, then similarly, for all $y \in \mathbf{R}^n$,

$$\begin{aligned} (f_1 + f_1 \circ F)^\uparrow(x_0; y) &\geq (f_1 + f_2 \circ F)_\square(x_0; y) \\ &\geq (f_1)_\square(x_0; y) + (f_2 \circ F)_+(x_0; y) \quad (\text{by (6.10)}) \\ &\geq (f_1)_\square(x_0; y) + (f_2)_+(F(x_0); \nabla F(x_0)y) \quad (\text{by (6.11)}) \\ &= (f_1)^\uparrow(x_0; y) + f_2^\uparrow(F(x_0); \nabla F(x_0)y). \end{aligned}$$

In either case, equality holds in (3.2), and equality in (3.3) follows. \square

Remark 6.6. The conditions in Propositions 6.4 and 6.5 guaranteeing equality in the subdifferential calculus formulae of § 3 are less stringent than those given in [18]. For example, let $C_1 := \bigcup_{n=0}^\infty [2^{2n}, 2^{2n+1}]$ and $C := C_1 \cup (-C_1) \cup \{0\}$. Then $T_C(0) = k_C(0) = \{0\}$, but $K_C(0) = \mathbf{R}$. In Proposition 6.5, let $F := I$ and $x_0 := 0$. Define $f_1: \mathbf{R} \rightarrow \mathbf{R}$ by $f_1(x) := 0$, and let $f_2: \mathbf{R} \rightarrow \bar{\mathbf{R}}$ be i_C . Since (3.8) and (6.13) are satisfied, equality holds in (3.9) and (3.10), even though the condition for equality given in Theorem 2 of [18]—that both f_1 and f_2 be subdifferentially regular at x_0 —is not satisfied. Observe also that the regularity conditions in Propositions 6.4 and 6.5 are still valid in the infinite-dimensional setting of [18].

Since convex functions are subdifferentially regular, Propositions 6.4 and 6.5 enable us to rederive convex subdifferential calculus results, in particular Theorems 2.9 and 2.10. Thus Theorems 3.2 and 3.17, which are proved by means of Theorems 2.9 and 2.10, combine with Propositions 6.4 and 6.5 to yield Theorems 2.9 and 2.10 as special cases. We list below one further convex subdifferential calculus result.

PROPOSITION 6.7. *Let $F: \mathbf{R}^n \rightarrow \mathbf{R}^m$ be strictly differentiable at x_0 , let $f_1: \mathbf{R}^n \rightarrow \bar{\mathbf{R}}$ be convex, proper, and strictly l.s.c. at x_0 , and let $f_2: \mathbf{R}^m \rightarrow \bar{\mathbf{R}}$ be convex, proper, and strictly l.s.c. at $F(x_0)$, where $x_0 \in \text{dom } f_1 \cap F^{-1}(\text{dom } f_2)$. Assume that*

$$(6.15) \quad \nabla F(x_0)x_0 - F(x_0) \in \text{int}(\nabla F(x_0) \text{dom } f_1 - \text{dom } f_2).$$

Then

$$(6.16) \quad \partial(f_1 + f_2 \circ F)(x_0) = \partial f_1(x_0) + \nabla F(x_0)^T \partial f_2(x_0).$$

Proof. By Theorem 3.2 and Proposition 6.5, equality will hold in (6.16) as long as $\nabla F(x_0) \text{dom } f_1^\uparrow(x_0; \cdot) - \text{dom } f_2^\uparrow(F(x_0); \cdot) = \mathbf{R}^m$, which in this case is equivalent to

$$\nabla F(x_0) \text{cone}(\text{dom } f_1 - x_0) - \text{cone}(\text{dom } f_2 - F(x_0)) = \mathbf{R}^m$$

(see [16], § 23). This reduces to

$$\text{cone}(\nabla F(x_0) \text{dom } f_1 - \text{dom } f_2 - \nabla F(x_0)x_0 + F(x_0)) = \mathbf{R}^m,$$

which is equivalent to (6.15). \square

COROLLARY 6.8 (cf. [17], Thm. 23.8). *Let $f_1: \mathbf{R}^m \rightarrow \bar{\mathbf{R}}$ and $f_2: \mathbf{R}^m \rightarrow \bar{\mathbf{R}}$ be convex, proper and l.s.c., and let $x_0 \in \text{dom } f_1 \cap \text{dom } f_2$. Assume that*

$$(6.17) \quad \text{ri dom } f_1 \cap \text{ri dom } f_2 \neq \emptyset.$$

Then

$$(6.18) \quad \partial(f_1 + f_2)(x_0) = \partial f_1(x_0) + \partial f_2(x_0).$$

Proof. By Proposition 6.7 with $m = n$ and $F := I$, (6.18) will hold if $\text{cone}(\text{dom } f_1 - \text{dom } f_2) = \mathbf{R}^m$. By (3.12) of Remark 3.7 (b), this assumption can be weakened to $\text{cone}(\text{dom } f_1 - \text{dom } f_2) = \text{span}(\text{dom } f_1 - \text{dom } f_2)$, which is equivalent to (6.17). \square

REMARK 6.9. We saw in Remark 3.7 (b) that hypothesis (3.8) of Corollary 3.6 cannot be weakened to (3.13). The reason is that $\text{aff dom } f^\uparrow(x_0; \cdot)$ can be of smaller dimension than $\text{aff dom } f$, as in the example given there. If f is convex and proper, however, $\text{aff dom } f^\uparrow(x_0; \cdot) = \text{aff dom } f$, and assumption (6.17) is sufficient to give (6.18).

It is possible by the techniques of § 3 to prove analogues of (3.2) and (3.26) involving contingent and Ursescu directional derivatives. To do so, we need counterparts of Theorem 2.4 and Propositions 2.5 and 2.6.

THEOREM 6.10. *Let $G: \mathbf{R}^p \rightarrow \mathbf{R}^q$ be strictly differentiable at $x_0 \in G^{-1}(0)$, and let $D \subset \mathbf{R}^p$ be closed near x_0 . Assume (2.2) holds. Then*

$$(6.19) \quad K_D(x_0) \cap \nabla G(x_0)^{-1}(0) \subset K_{D \cap G^{-1}(0)}(x_0)$$

and

$$(6.20) \quad k_D(x_0) \cap \nabla G(x_0)^{-1}(0) \subset k_{D \cap G^{-1}(0)}(x_0).$$

Proof. The inclusion (6.19) is a special case of [4, Thm. 4.1(a)]. The proof of (6.20) is entirely analogous to that of [4, Thm. 4.1]. \square

PROPOSITION 6.11. *Let E, E_1 be l.c.t.v.s., and let $x_0 \in C \subset E$ and $y_0 \in D \in E_1$. Then*

$$(6.21) \quad k_{C \times D}(x_0, y_0) = k_C(x_0) \times k_D(y_0),$$

$$(6.22) \quad K_{C \times D}(x_0, y_0) \subset K_C(x_0) \times K_D(y_0),$$

$$(6.23) \quad K_{C \times D}(x_0, y_0) \supset K_C(x_0) \times k_D(y_0).$$

The proof of Proposition 6.11 is a straightforward consequence of (6.1) and (6.2). That equality does not in general hold in (6.22) is demonstrated by an example in § 2 of [4].

PROPOSITION 6.12. *Let E, E_1 be l.c.t.v.s. and $A: E \rightarrow E_1$ be linear and continuous. Let $z_0 \in C \subset E$. Then*

$$(6.24) \quad A(K_C(z_0)) \subset K_{A(C)}(Az_0)$$

and

$$(6.25) \quad A(k_C(z_0)) \subset k_{A(C)}(Az_0).$$

Inclusion (6.24) is well known (see, for example, [1], [3]). The proof of (6.25) parallels that of (6.24).

THEOREM 6.13. *Let $f_i: \mathbf{R}^m \rightarrow \bar{\mathbf{R}}$, $i = 1, \dots$, be finite and strictly l.s.c. at x_0 , and call $f := (f_1, \dots, f_n)$. Let $F: \mathbf{R}^n \rightarrow \bar{\mathbf{R}}$ be finite and strictly l.s.c. at $f(x_0)$ and isotone on the union of $\text{Range } f + \mathbf{R}_+^n$ and $B_\varepsilon(f(x_0))$ for some $\varepsilon > 0$. Suppose (3.25) holds. Then for all $y \in \mathbf{R}^n$,*

$$(6.26) \quad (F \circ f)_\square(x_0; y) \leq F_\square(f(x_0); (f_1)_\square(x_0; y), \dots, (f_n)_\square(x_0; y))$$

and

$$(6.27) \quad (F \circ f)_+(x_0; y) \leq F_{\square}(f(x_0); (f_1)_+(x_0; y), (f_2)_{\square}(x_0; y), \dots, (f_n)_{\square}(x_0; y)).$$

Equality holds in (6.26) if $(f_i)_{\square}(x_0; y) = (f_i)_+(x_0; y)$, $i = 2, \dots, n$ for all $y \in \mathbf{R}^m$, $F_{\square}(f(x_0); z) = F_+(f(x_0); z)$ for all $z \in \mathbf{R}^n$, and $(f_i)_+(x_0; y)$, $i = 1, \dots, n$ are never $-\infty$.

Proof of (6.26). Let $h := F \circ f$, and define A and G as in Lemma 3.15 and Theorem 3.17. Define

$$D := \text{epi } f_1 \times \dots \times \text{epi } f_n \times \text{epi } F \text{ and } C := D \cap G^{-1}(0).$$

As in Theorem 3.17, $\text{epi } h = A(C)$. Then

$$\begin{aligned} \text{epi } h_{\square}(x_0; \cdot) &= k_{A(C)}(x_0, h(x_0)) \\ &\supset A(k_C(z_0)) \end{aligned}$$

where $z_0 := (x_0, f_1(x_0), \dots, x_0, f_n(x_0), f_1(x_0), \dots, f_n(x_0), h(x_0))$, by (6.25). Now assumption (3.25) guarantees that $\nabla G(x_0)T_D(z_0) = \mathbf{R}^{(n-1)m+n}$, so we may apply Theorem 6.10. By (6.20) and (6.21),

$$\begin{aligned} A(k_C(z_0)) &\supset A(k_D(z_0) \cap \nabla G(z_0)^{-1}(0)) \\ &= \{(x, r) \in \mathbf{R}^m \times \mathbf{R} \mid \exists y \in \mathbf{R}^n \text{ with} \\ &\quad (f_i)_{\square}(x_0; x) \leq y_i, F_{\square}(f(x_0); y) \leq r\} \\ &= \text{epi } F_{\square}(f(x_0); (f_1)_{\square}(x_0; \cdot), \dots, (f_n)_{\square}(x_0; \cdot)) \end{aligned}$$

since $F_{\square}(f(x_0); \cdot)$ is itself isotone (as in Lemma 2.8). Hence $\text{epi } h_{\square}(x_0; \cdot) \supset \text{epi } F_{\square}(f(x_0); (f_1)_{\square}(x_0; \cdot), \dots, (f_n)_{\square}(x_0; \cdot))$, and (6.26) holds. The proof of (6.27) is similar and uses (6.24), (6.19), and (6.23). That equality holds in (6.26) and (6.27) under the given assumptions is a consequence of Proposition 6.1. \square

THEOREM 6.14. Let $F: \mathbf{R}^n \rightarrow \mathbf{R}^m$ be strictly differentiable at x_0 , let $f_1: \mathbf{R}^m \rightarrow \bar{\mathbf{R}}$ be finite and strictly l.s.c. at x_0 , and let $f_2: \mathbf{R}^m \rightarrow \bar{\mathbf{R}}$ be finite and strictly l.s.c. at $F(x_0)$. Assume (3.1) holds. Then for all $y \in \mathbf{R}^n$,

$$(6.28) \quad (f_1 + f_2 \circ F)_{\square}(x_0; y) \leq (f_1)_{\square}(x_0; y) + (f_2)_{\square}(F(x_0); \nabla F(x_0)y)$$

and

$$(6.29) \quad (f_1 + f_2 \circ F)_+(x_0; y) \leq (f_1)_{\square}(x_0; y) + (f_2)_+(F(x_0); \nabla F(x_0)y),$$

$$(6.30) \quad (f_1 + f_2 \circ F)_+(x_0; y) \leq (f_1)_+(x_0; y) + (f_2)_{\square}(F(x_0); \nabla F(x_0)y).$$

If $(f_1)_+(x_0; \cdot)$ and $(f_2)_+(F(x_0); \nabla F(x_0)(\cdot))$ are never $-\infty$, equality holds in (6.28) and (6.29) if $(f_1)_{\square}(x_0; \cdot) = (f_1)_+(x_0; \cdot)$ and in (6.30) if $(f_2)_{\square}(F(x_0); \nabla F(x_0)(\cdot)) = (f_2)_+(F(x_0); \nabla F(x_0)(\cdot))$.

Proof. The proofs of (6.28), (6.29) and (6.30) are analogous to that of Theorem 3.2, using Theorem 6.10 and Propositions 6.11 and 6.12. That equality holds under the given assumptions is a consequence of Propositions 6.1 and 6.2. \square

We give here two consequences of Theorem 6.14.

COROLLARY 6.15. Let $C_1 \subset \mathbf{R}^n$ be closed near x_0 , let $F: \mathbf{R}^n \rightarrow \mathbf{R}^m$ be strictly differentiable at x_0 , and let $C_2 \subset \mathbf{R}^m$ be closed near $F(x_0)$. Assume that (3.4) holds. Then

$$(6.31) \quad k_{C_1 \cap F^{-1}(C_2)}(x_0) = k_{C_1}(x_0) \cap \nabla F(x_0)^{-1}k_{C_2}(F(x_0)),$$

$$\begin{aligned} (6.32) \quad K_{C_1}(x_0) \cap \nabla F(x_0)^{-1}K_{C_2}(F(x_0)) &\supset K_{C_1 \cap F^{-1}(C_2)}(x_0) \\ &\supset k_{C_1}(x_0) \cap \nabla F(x_0)^{-1}K_{C_2}(F(x_0)), \end{aligned}$$

and

$$(6.33) \quad K_{C_1 \cap F^{-1}(C_2)}(x_0) \supset K_{C_1}(x_0) \cap \nabla F(x_0)^{-1} k_{C_2}(F(x_0)).$$

Proof. In (6.31), $k_{C_1 \cap F^{-1}(C_2)}(x_0) \subset k_{C_1}(x_0) \cap \nabla F(x_0)^{-1} k_{C_2}(F(x_0))$ always holds. The opposite inclusion follows from (6.28) with $f_1 := i_{C_1}$ and $f_2 := i_{C_2}$. The first inclusion in (6.32) always holds, while the second is (6.29) with $f_1 := i_{C_1}$ and $f_2 := i_{C_2}$. Similarly, (6.33) follows from (6.30). \square

COROLLARY 6.16. *Let $f_1: \mathbf{R}^n \rightarrow \bar{\mathbf{R}}$, $f_2: \mathbf{R}^n \rightarrow \bar{\mathbf{R}}$ be finite and strictly l.s.c. at $x_0 \in \mathbf{R}^n$. Assume (3.8) holds. Then for all $y \in \mathbf{R}^n$,*

$$(6.34) \quad (f_1 + f_2)_{\square}(x_0; y) \leq (f_1)_{\square}(x_0; y) + (f_2)_{\square}(x_0; y)$$

and

$$(6.35) \quad (f_1 + f_2)_{+}(x_0; y) \leq (f_1)_{+}(x_0; y) + (f_2)_{\square}(x_0; y).$$

Equality holds in (6.34) and (6.35) if $(f_2)_{\square}(x_0; \cdot) = (f_2)_{+}(x_0; y)$ and if $(f_1)_{+}(x_0; \cdot)$ and $(f_2)_{+}(x_0; \cdot)$ are never equal to $-\infty$.

Proof. Take $m = n$ and $F := I$ in Theorem 6.14. \square

In the special case in which $f_{\square}(x_0; \cdot)$ is convex, we can define $\partial^k f(x_0) := \partial(f_{\square}(x_0; \cdot))(0)$ and can obtain results like the following:

PROPOSITION 6.17. *Let $f_1: \mathbf{R}^n \rightarrow \bar{\mathbf{R}}$ and $f_2: \mathbf{R}^n \rightarrow \bar{\mathbf{R}}$ be finite and strictly l.s.c. at x_0 . Assume that $(f_1)_{\square}(x_0; \cdot)$ and $(f_2)_{\square}(x_0; \cdot)$ are convex, and that (3.8) holds. Then*

$$(6.36) \quad \partial^k(f_1 + f_2)(x_0) \subset \partial^k f_1(x_0) + \partial^k f_2(x_0).$$

Equality holds in (6.36) if $(f_1)_{\square}(x_0; \cdot) = (f_1)_{+}(x_0; \cdot)$ or $(f_2)_{\square}(x_0; \cdot) = (f_2)_{+}(x_0; \cdot)$ —in particular, if f_1 or f_2 is convex.

Proof. Since (3.8) holds, so does (6.34), and (6.36) follows from (6.34). Equality under the given assumptions follows from Corollary 6.16 and the fact that if either $(f_1)_{\square}(x_0; \cdot)$ or $(f_2)_{\square}(x_0; \cdot)$ is not proper, both sides of (6.36) are empty. \square

It is interesting to note that assumptions involving $f^{\uparrow}(x_0; \cdot)$ are required in these results on $f_{+}(x_0; \cdot)$ and $f_{\square}(x_0; \cdot)$. The same is true of Ioffe's results on approximate subdifferentials [13].

We now present an example which satisfies the hypothesis of Proposition 6.17 even though f_1 and f_2 are not convex.

Example 6.18. Define $f_1: \mathbf{R} \rightarrow \mathbf{R}$ by

$$f_1(x) := \begin{cases} 0 & \text{if } x \leq 0, \\ \frac{1}{2^{n+1}} & \text{if } \frac{1}{2^{n+1}} < x \leq \frac{1}{2^n}, \quad n = 0, \pm 1, \pm 2, \dots, \end{cases}$$

and define $f_2: \mathbf{R} \rightarrow \mathbf{R}$ by

$$f_2(x) := \begin{cases} 0 & \text{if } x \leq 0, \\ \frac{1}{4^{(n+1)}} & \text{if } \frac{1}{4^{(n+1)}} < x \leq \frac{1}{4^n}, \quad n = 0, \pm 1, \pm 2, \dots \end{cases}$$

It is not hard to see that

$$f_1^{\uparrow}(0; y) = f_2^{\uparrow}(0; y) = \begin{cases} 0 & \text{if } y \leq 0, \\ +\infty & \text{if } y > 0, \end{cases}$$

$$(f_1)_\square(0; y) = (f_2)_\square(0; y) = \begin{cases} 0 & \text{if } y \leq 0, \\ y & \text{if } y > 0, \end{cases}$$

$$(f_1 + f_2)_\square(0; y) = \begin{cases} 0 & \text{if } y \leq 0, \\ 2y & \text{if } y > 0, \end{cases}$$

$$(f_1)_+(0; y) = \begin{cases} 0 & \text{if } y \leq 0, \\ y/2 & \text{if } y > 0, \end{cases}$$

$$(f_2)_+(0; y) = \begin{cases} 0 & \text{if } y \leq 0, \\ y/4 & \text{if } y > 0, \end{cases}$$

$$(f_1 + f_2)_+(0; y) = \begin{cases} 0 & \text{if } y \leq 0, \\ 3y/4 & \text{if } y > 0. \end{cases}$$

Then $\text{dom } f_1^\uparrow(0; \cdot) = \text{dom } f_2^\uparrow(0; \cdot) = \{y \mid y \leq 0\}$, so (3.8) is satisfied and (6.36) holds. In fact $\partial_\square(f_1 + f_2)(0) = [0, 2]$ while $\partial_\square f_1(0) = \partial_\square f_2(0) = [0, 1]$, so equality holds in (6.36) even though $(f_i)_+(x_0; \cdot) \neq (f_i)_\square(x_0; \cdot)$, $i = 1, 2$.

If $f_\square(x_0; \cdot)$ is not convex, it is still possible to define a notion of subgradient for f by means of upper convex approximates.

DEFINITION 6.19 (cf. [15]). The function $h: E \rightarrow \bar{\mathbf{R}}$ is an *upper convex approximate* for $f: E \rightarrow \bar{\mathbf{R}}$ at $x_0 \in \text{dom } f$ if

(a) $h(y) \leq f_\square(x_0; y)$ for all $y \in E$, and

(b) $h(\cdot)$ is convex and positively homogeneous. The set $\partial_h^k(x_0) := \{x' \in E^* \mid h(y) \leq \langle y, x' \rangle \forall y \in E\}$ is called an *h -subgradient* of f at x_0 .

Specific examples of upper convex approximates include $f^\uparrow(x_0; \cdot)$, $f^0(x_0; \cdot)$, the asymptotic epiderivatives of [9], [10], and the upper convex approximations of [15]. As one might expect, not as much can be said about general upper convex approximates as can be determined about specific ones like $f^0(x_0; \cdot)$ or $f^\uparrow(x_0; \cdot)$; however, necessary conditions for minimality can be expressed in terms of upper convex approximates [24], [25]. We will not develop this concept in detail here, except to mention the following generalizations of Propositions 4.1 and 4.2:

PROPOSITION 6.20. Suppose $f: E \rightarrow \bar{\mathbf{R}}$ has a local minimum at $x_0 \in E$ and $h(x_0; \cdot): E \rightarrow \bar{\mathbf{R}}$ is an upper convex approximate for f at x_0 . Then

$$0 \leq f_\square(x_0; y) \leq h(x_0; y) \quad \text{for all } y \in E$$

and

$$0 \in \partial_h^k f(x_0).$$

PROPOSITION 6.21. Let $f: \mathbf{R}^n \rightarrow \bar{\mathbf{R}}$ be strictly l.s.c. at x_0 , and let $C \subset \mathbf{R}^n$ be closed near x_0 . Suppose $x_0 \in X$ is a local minimizer of the problem $\min \{f(x) \mid x \in C\}$. Assume (4.1) holds. Let $h(x_0; \cdot): \mathbf{R}^n \rightarrow \bar{\mathbf{R}}$ be an upper convex approximate for f at x_0 such that $h(x_0; y) \leq f^\uparrow(x_0; y)$ for all $y \in \mathbf{R}^n$, and let D be a convex subcone of $k_C(x_0)$ satisfying $T_C(x_0) \subset D \subset k_C(x_0)$. Then

$$(6.37) \quad 0 \in \partial_h^k f(x_0) + D^0.$$

Proof. $f + i_C$ is locally minimized at x_0 , so for all $y \in \mathbf{R}^n$,

$$\begin{aligned} 0 &\leq (f + i_C)_\square(x_0; y) \\ &\leq f_\square(x_0; y) + (i_C)_\square(x_0; y) \end{aligned}$$

(by Corollary 6.16, since (4.1) holds)

$$\begin{aligned} &= f_\square(x_0; y) + i_{k_C}(x_0)^{(y)} \\ &\leq h(x_0; y) + i_D(y). \end{aligned}$$

Now $\text{dom } h(x_0; \cdot) \supset \text{dom } f^\dagger(x_0; \cdot)$ and $D \subset T_C(x_0)$ by hypothesis, so (4.1) implies that $\text{dom } h(x_0; \cdot) - D = \mathbf{R}^n$. Hence

$$0 \in \partial(h(x_0; \cdot) + i_D(\cdot))(0) = \partial_h f(x_0) + D^0$$

by Theorem 23.8 of [16] and so (6.37) holds. \square

Acknowledgments. We are grateful to G. G. Watkins for his help with the proof of Theorem 4.7, and to the referee for suggestions which improved the presentation of this work.

REFERENCES

- [1] J.-P. AUBIN, *Contingent derivatives of set-valued maps and existence of solutions to nonlinear inclusions and differential inclusions*, in *Advances in Mathematics Supplementary Studies 7A*, L. Nachbin, ed., Academic Press, New York, 1981, pp. 159–229.
- [2] ———, *Lipschitz behavior of solutions to convex minimization problems*, *Math. Oper. Res.*, 9 (1984), pp. 87–111.
- [3] J.-P. AUBIN AND I. EKELAND, *Applied Nonlinear Analysis*, John Wiley, New York, 1984.
- [4] J. M. BORWEIN, *Stability and regular points of inequality systems*, *J. Optim. Theory Appl.*, 48 (1986), pp. 9–52.
- [5] F. H. CLARKE, *A new approach to Lagrange multipliers*, *Math. Oper. Res.*, 1 (1976), pp. 165–174.
- [6] ———, *Optimization and Nonsmooth Analysis*, John Wiley, New York, 1983.
- [7] S. DOLECKI, *Tangency and differentiation: some applications of convergence theory*, *Ann. Mat. Pura Appl.*, CXXX (1982), pp. 223–255.
- [8] I. EKELAND, *On the variational principle*, *J. Math. Anal. Appl.*, 47 (1974), pp. 324–353.
- [9] H. FRANKOWSA, *Inclusions adjointes associées aux trajectoires minimales d'inclusions différentielles*, *C. R. Acad. Sci. Paris Sér. A-B*, 296 (1983), pp. 721–724.
- [10] ———, *Necessary conditions for the Bolza problem*, *Math. Oper. Res.*, 10 (1985), pp. 361–366.
- [11] J.-B. HIRIART-URRUTY, *Contributions à la programmation mathématique: déterministe et stochastique*, Thèse, Université de Clermont-Ferrand II, 1977.
- [12] ———, *Tangent cones, generalized gradients and mathematical programming in Banach spaces*, *Math. Oper. Res.*, 4 (1979), pp. 79–97.
- [13] A. D. IOFFE, *Approximate subdifferentials and applications I: the finite-dimensional theory*, *Trans. Amer. Math. Soc.*, 281 (1984), pp. 389–416.
- [14] A. G. KUSRAEV, *On a general method of subdifferentiation*, *Soviet Math. Dokl.*, 23 (1981), pp. 367–371.
- [15] B. N. PSHENICHNYI AND R. A. KHACHATRYAN, *Constraints of equality type in nonsmooth optimization problems*, *Soviet Math. Dokl.*, 26 (1982), pp. 659–662.
- [16] R. T. ROCKAFELLAR, *Convex Analysis*, Princeton University Press, Princeton, NJ, 1970.
- [17] ———, *Clarke's tangent cone and the boundaries of closed sets in \mathbf{R}^n* , *Nonlinear Anal. Theory Methods Appl.*, 3 (1979), pp. 145–154.
- [18] ———, *Directionally Lipschitzian functions and subdifferential calculus*, *Proc. London Math. Soc.*, 39 (1979), pp. 331–355.
- [19] ———, *Generalized directional derivatives and subgradients of nonconvex functions*, *Canad. J. Math.*, 32 (1980), pp. 257–280.
- [20] ———, *The Theory of Subgradients and its Applications to Problems of Optimization of Convex and Nonconvex Functions*, Heldermann Verlag, Berlin, 1981.
- [21] ———, *Lagrange multipliers and subderivatives of optimal value functions in nonlinear programming*, *Math. Programming Stud.*, 17 (1982), pp. 28–66.
- [22] ———, *Extensions of subgradient calculus with applications to optimization*, *Nonlinear Anal. Theory Methods Appl.*, 9 (1985), pp. 665–698.
- [23] C. URSESCU, *Tangent sets' calculus and necessary conditions for extremality*, *this Journal*, 20 (1982), pp. 563–574.
- [24] D. E. WARD, *Tangent cones, generalized subdifferential calculus, and optimization*, Ph.D. thesis, Dalhousie University, Halifax, Nova Scotia, Canada, 1984.
- [25] ———, *Isotone tangent cones and nonsmooth optimization*, preprint, 1985.
- [26] G. G. WATKINS, *Nonsmooth Milyutin–Dubovitskii theory and Clarke's tangent cone*, *Math. Oper. Res.*, 11 (1986), pp. 70–80.
- [27] C. ZALINESCU, *On convex sets in general position*, *Linear Algebra Appl.*, 64 (1985), pp. 191–198.

FEEDBACK CONTROL OF ANALYTIC NONLINEAR SYSTEMS BY EXTENDED LINEARIZATION*

WILLIAM T. BAUMANN† AND WILSON J. RUGH‡

Abstract. For multi-input, multi-output, analytic, nonlinear systems, a design method based on the family of linearizations of the system, parameterized by the family of constant operating points, is considered. Using the Cauchy-Kowalewski Theorem it is shown that nonlinear state feedback control laws and output/observer feedback control laws exist such that the eigenvalues of the family of closed-loop linearizations have specified, analytically-scheduled values with respect to the family of closed-loop operating points.

Key words. nonlinear control, feedback control, eigenvalue placement, nonlinear observers

AMS(MOS) subject classification. 93C10

1. Introduction. In many practical situations, control of a nonlinear system is approached by linearizing the system about a nominal constant operating point, and then applying linear feedback control methods. When control over a wider range of operation is required, the process is repeated at distinct constant operating points, and an overall controller is pieced together from the resulting linear control designs. This approach is often referred to as gain scheduling. Recently the authors have been working on a systematic reformulation and extension of this practical approach, called design by extended linearization.

The goal of the extended-linearization method is to design a nonlinear feedback controller for the nonlinear system such that each member of the family of linearizations of the closed-loop system achieves some specified (linear) control objective. In this paper, which is an extension of [1] to the multi-input case, we show that a nonlinear state feedback control law can be constructed so as to place the eigenvalues of the family of closed-loop linearizations at specified locations, which may be a function of the closed-loop operating point. A similar approach is used to design a nonlinear observer such that the eigenvalues of the family of linearized error equations have specified locations. Feedback of the observed state yields a closed-loop system whose family of linearizations exhibits the eigenvalue separation property, and has the specified eigenvalue locations. Although this approach typically is local in nature, it applies to a broad class of systems, and yields output feedback control laws.

The initial step in the design approach involves determining parameterized families of linearized feedback gains for the family of system linearizations. Although this is related to work on parameterized linear systems, the local nature of our problem makes this first step essentially trivial. The second, and major, step is to synthesize a nonlinear feedback gain that does not explicitly depend on the parameters in the linearized feedback gains. In the multi-input case, this requires the solution of total differential equations that involve nontrivial integrability conditions (the classical Frobenius theorem).

* Received by the editors August 28, 1985; accepted for publication (in revised form) August 11, 1986. This research was sponsored by the Air Force Office of Scientific Research, Air Force Systems Command, United States Air Force, under grant AFOSR-83-0079.

† Department of Electrical Engineering, Virginia Polytechnic Institute and State University, Blacksburg, Virginia 24061.

‡ Department of Electrical Engineering, The Johns Hopkins University, Baltimore, Maryland 21218.

Recent work on so-called pseudo-linearization, by variable change and state feedback, is related to our approach in that the closed-loop linearization family is required to be constant with respect to the closed-loop family of operating points, though the issues of state observation and output feedback were not addressed [3], [4]. For single-input systems with state feedback, comparison of [1] and [3] shows that similar hypotheses are required for the somewhat different objectives. However, for multi-input systems with state feedback, comparison of the results here with those in [4] shows that the objective of eigenvalue placement requires weaker hypotheses.

The mathematical developments in the sequel involve real-analytic functions. Results typically apply in a sufficiently small neighborhood of an appropriate point, although this local restriction often will be left understood. The following notation for differentiation will be used. If $f(x): \mathbf{R} \rightarrow \mathbf{R}^p$, then Df is a column vector $[df_1/dx \cdots df_p/dx]^T$, and if $f(x): \mathbf{R}^n \rightarrow \mathbf{R}^p$, then Df denotes the $p \times n$ Jacobian matrix of partial derivatives. If $f(x, y): \mathbf{R}^m \times \mathbf{R}^n \rightarrow \mathbf{R}^p$, then D_1f represents the partial derivative with respect to x , and D_2f the partial derivative with respect to y . Evaluation of a derivative is indicated in the customary fashion, e.g., $D_1f(x_0, y_0)$.

2. State feedback control. We assume that the nonlinear system to be controlled can be described by a constant-parameter, analytic differential equation, and that the nominal constant operating (equilibrium) point corresponds to zero input and zero state. That is, the system can be described by (suppressing t -arguments)

$$(2.1) \quad \dot{x} = f(x, w)$$

where $f(\cdot, \cdot): \mathbf{R}^n \times \mathbf{R}^m \rightarrow \mathbf{R}^n$ is analytic in a neighborhood of the origin, with $f(0, 0) = 0$. The family of constant operating points corresponding to nonzero constant inputs, $w(t) = \varepsilon$, also will be of interest. We will assume throughout that the linearization of (2.1) about zero satisfies the following hypotheses:

- (H1) (1) $[D_1f(0, 0), D_2f(0, 0)]$ is a controllable pair,
 (2) $D_1f(0, 0)$ is invertible,
 (3) $D_2f(0, 0)$ is full rank.

Then, by the implicit function theorem,

$$(2.2) \quad f(x, \varepsilon) = 0$$

has a unique analytic solution $x(\varepsilon)$ in an open neighborhood of $\varepsilon = 0$. Typically the argument of x will be suppressed.

The state feedback control laws to be considered have the form

$$(2.3) \quad w = u - k(x)$$

where $k(\cdot): \mathbf{R}^n \rightarrow \mathbf{R}^m$ is analytic, $k(0) = 0$, and $u \in \mathbf{R}^m$ is an external input. This control law yields a closed-loop system described by the analytic state equation

$$(2.4) \quad \dot{x} = f(x, u - k(x))$$

that retains a constant operating point at $u = 0$, $x = 0$. Again, the family of constant operating points corresponding to nonzero constant inputs, $u(t) = \beta$, will be of interest, so it will be assumed that the specified eigenvalues of the linearized closed-loop system are nonzero (in addition to self-conjugate). Then $D_1f(0, 0) - D_2f(0, 0)Dk(0)$ is invertible, and, as a consequence,

$$(2.5) \quad f(x_c, \beta - k(x_c)) = 0$$

has a unique analytic solution $\mathbf{x}_c(\beta)$ in an open neighborhood of $\beta = 0$. Let

$$(2.6) \quad \varepsilon(\beta) = \beta - k(\mathbf{x}_c(\beta));$$

then $\varepsilon(\beta)$ is analytic, with $\varepsilon(0) = 0$, and from (2.2),

$$(2.7) \quad \mathbf{x}(\varepsilon(\beta)) = \mathbf{x}_c(\beta).$$

In the remainder of the paper, the families of constant operating points for both the open- and closed-loop systems will be parameterized by ε , and $\mathbf{x}(\varepsilon)$ (or simply \mathbf{x}) will represent the equilibrium state. For the closed-loop system, the parameter ε is related to the constant input $u(t) = \beta$ via (2.6), and it is easy to show that this relation is invertible.

Linearization of the closed-loop system about a closed-loop operating point yields the linear state equation

$$(2.8) \quad \dot{x} = [D_1 f(\mathbf{x}, \varepsilon) - D_2 f(\mathbf{x}, \varepsilon) Dk(\mathbf{x})]x + D_2 f(\mathbf{x}, \varepsilon)u$$

where x and u now indicate deviations from \mathbf{x} and β , respectively. The goal of this section is to prove the following theorem:

THEOREM 2.1. *Suppose the analytic system (2.1) satisfies (H1). Then there exists an analytic feedback gain $k(\cdot): \mathbf{R}^n \rightarrow \mathbf{R}^m$ with $k(0) = 0$ such that the eigenvalues of the closed-loop linearization (2.8) are specified, analytic functions of ε in an open neighborhood of $\varepsilon = 0$.*

Since the linearization of (2.1) about the nominal operating point at zero is assumed to be controllable, it is not difficult to find an analytic $Q(\cdot): \mathbf{R}^m \rightarrow \mathbf{R}^{m \times n}$ such that the matrix $D_1 f(\mathbf{x}, \varepsilon) - D_2 f(\mathbf{x}, \varepsilon)Q(\varepsilon)$ has the specified eigenvalues. The problem lies in synthesizing a state feedback gain $k(\cdot)$ that does not depend explicitly on ε , and that satisfies

$$(2.9) \quad Dk(\mathbf{x}(\varepsilon)) = Q(\varepsilon).$$

However, unless $Q(\varepsilon)$ satisfies certain integrability conditions there can be no solution to (2.9). Thus our approach to the problem consists of several steps. After determining the integrability conditions, a general form for $Q(\cdot)$ that gives the specified eigenvalue locations and contains $(n-1)m$ undetermined functions is given in Lemma 2.2. It is then shown in Lemma 2.4 that these undetermined functions can be chosen so that $Q(\cdot)$ satisfies the integrability conditions. Finally, the proof of Theorem 2.1 is completed by showing how to construct a function $k(\cdot)$ satisfying (2.9).

To address the integrability conditions, consider the function $\hat{k}(\varepsilon) = k(\mathbf{x}(\varepsilon))$. If $k(\cdot): \mathbf{R}^n \rightarrow \mathbf{R}^m$ is an analytic solution of (2.9) then $\hat{k}(\cdot): \mathbf{R}^m \rightarrow \mathbf{R}^m$ is an analytic solution of

$$(2.10) \quad D\hat{k}(\varepsilon) = Q(\varepsilon)D\mathbf{x}(\varepsilon).$$

Thus existence of a solution to (2.10) is necessary for existence of a solution to (2.9). Let $\hat{k}_i(\varepsilon)$ denote the i th entry of $\hat{k}(\varepsilon)$ and $q_i(\varepsilon)$ the i th row of $Q(\varepsilon)$, then (2.10) can be written row-wise as

$$(2.11) \quad D\hat{k}_i = q_i D\mathbf{x}.$$

By the Frobenius theorem, a necessary and sufficient condition that (2.11) have a solution is that mixed second partial derivatives on the right side commute [6]. That is,

$$\frac{\partial}{\partial \varepsilon_j} [q_i D_k \mathbf{x}] = \frac{\partial}{\partial \varepsilon_k} [q_i D_j \mathbf{x}], \quad j, k = 1, \dots, m, \quad k > j.$$

Expanding this, and using the fact that mixed second partial derivatives of \mathbf{x} commute, we get

$$D_j q_i D_k \mathbf{x} = D_k q_i D_j \mathbf{x}, \quad j, k = 1, \dots, m, \quad k > j.$$

Therefore, transposing the left side of this expression and considering each value of j separately, the integrability conditions for (2.11) can be written in matrix form as the set of conditions

$$(2.12) \quad H_j D_j q_i^T = \begin{bmatrix} D_{j+1} q_i \\ \vdots \\ D_m q_i \end{bmatrix} D_j \mathbf{x}, \quad i = 1, \dots, m, \quad j = 1, \dots, m-1$$

where

$$(2.13) \quad H_j = [D_{j+1} \mathbf{x} | \dots | D_m \mathbf{x}]^T.$$

To show that an appropriate $Q(\varepsilon)$ exists, two technical lemmas will be given, the proofs of which are simplified if a convenient transformation is made at the outset. From the hypotheses on (2.1), it can be assumed without loss of generality that a linear coordinate change has been applied to (2.1) so that the pair $[D_1 f(0, 0), D_2 f(0, 0)]$ is in Brunovsky canonical form (without feedback) [5, p. 501]. It is also assumed that an input transformation has been performed so that $D_2 f(0, 0)$ has zero entries, except for ones in positions $(1, 1), (k_1 + 1, 2), \dots, (k_1 + \dots + k_{m-1} + 1, m)$, where the Kronecker indices k_i are ordered as $k_1 \leq k_2 \leq \dots \leq k_m$. Once the closed-loop eigenvalues are specified, this form allows the straightforward choice of a matrix $Q^0 \in \mathbb{R}^{m \times n}$ such that $D_1 f(0, 0) - D_2 f(0, 0) Q^0$ is in top companion form, has the eigenvalues specified at $\beta = 0$, and is controllable from the first component of the input. We let $q_{i+} = [q_{i1}, \dots, q_{i, n-k_i}]^T$ and $q_{i-} = [q_{i, n-k_i+1}, \dots, q_{in}]^T$ and define q_{i+}^0 and q_{i-}^0 in the obvious, consistent manner.

LEMMA 2.2. *Suppose the analytic system (2.1) satisfies (H1). Then given m analytic functions $q_{i+}(\varepsilon): \mathbb{R}^m \rightarrow \mathbb{R}^{n-k_i}$ with $q_{i+}(0) = q_{i+}^0$, there exist analytic functions $q_{i-}(\varepsilon, q_{1+}, \dots, q_{m+})$ with $q_{i-}(0, q_{1+}^0, \dots, q_{m+}^0) = q_{i-}^0$ such that the linearized closed-loop system (2.8), (2.9) has the specified eigenvalues that are analytic functions of ε in an open neighborhood of $\varepsilon = 0$.*

Proof. For convenience, let

$$\begin{aligned} P &= sI - D_1 f(\mathbf{x}, \varepsilon) + D_2 f(\mathbf{x}, \varepsilon) Q, \\ \det P &= s^n + p_1(\varepsilon, Q) s^{n-1} + \dots + p_n(\varepsilon, Q), \\ p(\varepsilon, Q) &= [p_1(\varepsilon, Q) \dots p_n(\varepsilon, Q)]^T, \\ d &= [q_{m-}^T | \dots | q_{1-}^T]^T. \end{aligned}$$

Then, for the given analytic $q_{i+}(\varepsilon)$, we want to choose analytic $q_{i-}(\varepsilon, q_{1+}, \dots, q_{m+})$ such that for all ε in an open neighborhood of $\varepsilon = 0$, $p(\varepsilon, Q)$ equals the specified vector of characteristic polynomial coefficients (which are functions of ε). This will be accomplished by showing that $\partial p(\varepsilon, Q) / \partial d|_{\varepsilon=0, Q=Q^0}$ is invertible, so that the implicit function theorem can be applied to complete the proof. The analyticity of the q_{i-} follows from the fact that p is analytic [6, p. 272].

We first find an expression for $\partial p(\varepsilon, Q) / \partial q_i|_{\varepsilon=0, Q=Q^0}$. Using Laplace's expansion about the j th column of P gives

$$\det P = \sum_{k=1}^n [P]_{kj} [P]^{kj}$$

where $[M]_{ij}$ is the (i, j) element of M and $[M]^{ij}$ is the cofactor of $[M]_{ij}$. Since q_{ij} , the j th entry of q_i , appears only in the j th column of P ,

$$\begin{aligned} \frac{\partial}{\partial q_{ij}} \det P|_{\varepsilon=0, Q=Q^0} &= \sum_{k=1}^n [P]^{kj}|_{\varepsilon=0, Q=Q^0} \frac{\partial}{\partial q_{ij}} [P]_{kj}|_{\varepsilon=0, Q=Q^0} \\ &= \sum_{k=1}^n [\text{Adj}(sI - D_1 f(0, 0) + D_2 f(0, 0) Q^0)]_{jk} [D_2 f(0, 0)]_{ki} \end{aligned}$$

where $\text{Adj}(M)$ is the adjugate matrix of M . Letting $A = D_1 f(0, 0)$, $B = D_2 f(0, 0)$, and b_i be the i th column of B , we can write

$$\frac{\partial}{\partial q_i} \det P|_{\varepsilon=0, Q=Q^0} = b_i^T \text{Adj}(sI - A + BQ^0)^T.$$

Using a standard resolvent identity we obtain

$$\begin{aligned} \frac{\partial}{\partial q_i} \det P|_{\varepsilon=0, Q=Q^0} &= \{b_i s^{n-1} + [(A - BQ^0)b_i + p_1(0, Q^0)b_i]s^{n-2} + \cdots \\ &\quad + [(A - BQ^0)^{n-1}b_i + \cdots + p_{n-1}(0, Q^0)b_i]\}^T. \end{aligned}$$

Finally,

$$(2.14) \quad \frac{\partial}{\partial q_i} p(\varepsilon, Q)|_{\varepsilon=0, Q=Q^0} = R^T [b_i | (A - BQ^0)b_i | \cdots | (A - BQ^0)^{n-1}b_i]^T$$

where R is an upper triangular Toeplitz matrix with first row $[1 \ p_1(0, Q^0) \cdots p_{n-1}(0, Q^0)]$. Now we can write

$$(2.15) \quad \frac{\partial}{\partial d} p(\varepsilon, Q)|_{\varepsilon=0, Q=Q^0} = R^T C^T$$

where C consists of particular rows chosen from $[b_i | (A - BQ^0)b_i | \cdots | (A - BQ^0)^{n-1}b_i]$. Because of the special forms chosen for A , B and Q^0 , it is not hard to verify that C is an upper triangular matrix with ones on the main diagonal. Thus, (2.15) is invertible. \square

The following corollary, obtained by considering (2.14) for $i = 1$, will be needed in § 3.

COROLLARY 2.3. *Suppose the analytic system (2.1) satisfies (H1). Then given $m - 1$ analytic functions $q_2(\varepsilon), \cdots, q_m(\varepsilon): \mathbf{R}^m \rightarrow \mathbf{R}^n$, with $q_2(0) = q_2^0, \cdots, q_m(0) = q_m^0$, there exists an analytic function $q_1(\varepsilon, q_2, \cdots, q_m)$ with $q_1(0, q_2^0, \cdots, q_m^0) = q_1^0$ such that the linearized closed-loop system (2.8), (2.9) has the specified eigenvalues for all ε in an open neighborhood of $\varepsilon = 0$.*

Now it must be shown that a matrix $Q(\varepsilon)$ of the form in Lemma 2.2 can be chosen to satisfy the integrability conditions (2.12), in order to establish the existence of a solution $\hat{k}(\varepsilon)$ to (2.10).

LEMMA 2.4. *Suppose the analytic system (2.1) satisfies (H1). Then there exists an analytic $Q(\varepsilon)$ such that, in an open neighborhood of $\varepsilon = 0$, (2.8), (2.9) has eigenvalues that are specified, analytic functions of ε , and (2.12) is satisfied.*

Proof. The notation

$$(2.16) \quad q_{i-} = q_{i-}(\varepsilon_1, \cdots, \varepsilon_m, q_{1+}, \cdots, q_{m+}), \quad q_{i+} = q_{i+}(\varepsilon_1, \cdots, \varepsilon_m)$$

$i = 1, \cdots, m$, will be used, where the q_{i+} are as yet unspecified, and the q_{i-} are as in Lemma 2.2 for the specified eigenvalues. Using the chain rule, with careful attention to the differentiation notation, the derivative of q_{i-} with respect to ε_j can be written as

$$(2.17) \quad D_{m+1}q_{i-} - D_j q_{1+} + \cdots + D_{2m}q_{i-} - D_j q_{m+} + D_j q_{i-}.$$

Then the set of conditions (2.12) can be expressed solely in terms of the q_{i+} as

$$(2.18) \quad G_j \begin{bmatrix} D_j q_{1+} \\ \vdots \\ D_j q_{m+} \end{bmatrix} = F_j(\varepsilon, q_{1+}, \dots, q_{m+}, D_{j+1} q_{1+}, \dots, D_m q_{m+}), \quad j = 1, \dots, m-1$$

where each F_j is a known, analytic function,

$$(2.19) \quad G_j = \begin{bmatrix} H_j^{1+} & & 0 \\ & \ddots & \\ 0 & & H_j^{m+} \end{bmatrix} + \begin{bmatrix} H_j^{1-} D_{m+1} q_{1-} & \cdots & H_j^{1-} D_{2m} q_{1-} \\ \vdots & & \vdots \\ H_j^{m-} D_{m+1} q_{m-} & \cdots & H_j^{m-} D_{2m} q_{m-} \end{bmatrix},$$

$j = 1, \dots, m-1$

is an $m(m-j) \times n(m-1)$ matrix partitioned so that the diagonal blocks are of dimension $(m-j) \times (n-k_i)$, $i = 1, \dots, m$, and where

$$[H_j^{i+} \quad H_j^{i-}] = H_j$$

is a partitioning of H_j into $(m-j) \times (n-k_i)$ and $(m-j) \times k_i$ blocks.

Using the fact that $D\mathbf{x}(0) = -[D_1 f(0, 0)]^{-1} D_2 f(0, 0)$, and the special forms of $D_1 f(0, 0)$ and $D_2 f(0, 0)$, we can see that $H_1|_{\varepsilon=0}$ has zero entries, except in columns $k_1, k_1 + k_2, \dots, n$. By using a preliminary constant state feedback the elements in these columns can be specified arbitrarily within the constraint that $H_1|_{\varepsilon=0}$ be full rank. Thus, we will assume that a preliminary state feedback has been performed (prior to choosing Q^0), such that $H_1|_{\varepsilon=0}$ is full rank and its n th column is zero. Keeping in mind the ordering of the Kronecker indices, this implies that $H_j^{i-}|_{\varepsilon=0} = 0$ and $H_j^{i+}|_{\varepsilon=0}$ is full rank for $i = 1, \dots, m$; $j = 1, \dots, m-1$. Hence $G_j|_{\varepsilon=0}$ is full rank for $j = 1, \dots, m-1$.

Now, independent equations can be appended to each equation in (2.18) in such a way that each augmented G_j, \tilde{G}_j , is invertible in an open neighborhood of $\varepsilon = 0$. Multiplying the augmented equations by \tilde{G}_j^{-1} yields the set of equations

$$(2.20) \quad \begin{bmatrix} D_j q_{1+} \\ \vdots \\ D_j q_{m+} \end{bmatrix} = \tilde{F}_j(\varepsilon, q_{1+}, \dots, q_{m+}, D_{j+1} q_{1+}, \dots, D_m q_{m+}), \quad j = 1, \dots, m-1.$$

For a given j and appropriate initial conditions, there exists a solution to (2.20), and hence a solution to (2.12) for that value of j , by the Cauchy-Kowalewski Theorem. What must be shown, however, is that there is a solution satisfying (2.12) for $j = 1, \dots, m-1$ simultaneously. To this end, we will use (2.20) to construct a trial solution, \hat{q}_i , to (2.12), and then verify that this is indeed a solution to (2.12). For $j = m-1$, set $\varepsilon_1 = \dots = \varepsilon_{m-2} = 0$ in (2.20) and consider the resulting equation in the independent variables $\varepsilon_{m-1}, \varepsilon_m$. Using the initial conditions $q_{i+}(0, \dots, 0, \varepsilon_m) = q_{i+}^0$, $i = 1, \dots, m$, an analytic solution exists in an open neighborhood of $\varepsilon_{m-1} = \varepsilon_m = 0$ by the Cauchy-Kowalewski Theorem. Addressing (2.20) for $j = m-2$ in the same fashion, with $\varepsilon_1 = \dots = \varepsilon_{m-3} = 0$, there exists an analytic solution using as the initial condition the previously determined solution of the $j = m-1$ case. Continuing in this manner, there exist analytic solutions $\hat{q}_{i+}(\varepsilon_1, \dots, \varepsilon_m)$, $i = 1, \dots, m$, for (2.20) with $j = 1$ in an open neighborhood of $\varepsilon = 0$.

Finally, it will be shown that

$$\hat{q}_i(\varepsilon_1, \dots, \varepsilon_m) = [\hat{q}_{i+}^T q_{i-}^T(\varepsilon_1, \dots, \varepsilon_m, \hat{q}_{1+}, \dots, \hat{q}_{m+})], \quad i = 1, \dots, m$$

satisfy the conditions (2.12). Since the following argument is independent of i , the index i will be dropped and \hat{q}_k will be written for \hat{q}_{ik} . Assuming that (2.12) is satisfied for $j = 1, \dots, J-1$, it will be shown that \hat{q} satisfies (2.12) for $j = J$. In scalar terms, (2.12) with $j = J$ is

$$(2.21) \quad \sum_{k=1}^n D_J q_k D_l \mathbf{x}_k = \sum_{k=1}^n D_l q_k D_J \mathbf{x}_k, \quad l = J+1, \dots, m.$$

Now, \hat{q} satisfies (2.21) if and only if

$$(2.22) \quad D_m^{d_m} \dots D_1^{d_1} \left[\sum_{k=1}^n D_J \hat{q}_k D_l \mathbf{x}_k \right] \Big|_0 = D_m^{d_m} \dots D_1^{d_1} \left[\sum_{k=1}^n D_l \hat{q}_k D_J \mathbf{x}_k \right] \Big|_0$$

for $l = J+1, \dots, m$ and $d_1, \dots, d_m = 0, 1, \dots$. By the inductive hypothesis

$$(2.23) \quad \sum_{k=1}^n D_j \hat{q}_k D_l \mathbf{x}_k = \sum_{k=1}^n D_l \hat{q}_k D_j \mathbf{x}_k, \quad j = 1, \dots, J-1, \quad l = j+1, \dots, m.$$

Applying D_J to (2.23) we get

$$(2.24) \quad \sum_{k=1}^n D_J D_j \hat{q}_k D_l \mathbf{x}_k + \sum_{k=1}^n D_j \hat{q}_k D_J D_l \mathbf{x}_k = \sum_{k=1}^n D_J D_l \hat{q}_k D_j \mathbf{x}_k + \sum_{k=1}^n D_l \hat{q}_k D_J D_j \mathbf{x}_k, \\ j = 1, \dots, J-1, \quad l = j+1, \dots, m.$$

Specializing (2.23) to $l = J$ and then applying D_l we obtain

$$(2.25) \quad \sum_{k=1}^n D_l D_j \hat{q}_k D_J \mathbf{x}_k + \sum_{k=1}^n D_j \hat{q}_k D_l D_J \mathbf{x}_k = \sum_{k=1}^n D_l D_J \hat{q}_k D_j \mathbf{x}_k + \sum_{k=1}^n D_J \hat{q}_k D_l D_j \mathbf{x}_k \\ j = 1, \dots, J-1, \quad l = 1, \dots, m.$$

Subtraction of (2.25) from (2.24) results in

$$\sum_{k=1}^n D_J D_j \hat{q}_k D_l \mathbf{x}_k + \sum_{k=1}^n D_j \hat{q}_k D_l D_J \mathbf{x}_k = \sum_{k=1}^n D_l D_j \hat{q}_k D_J \mathbf{x}_k + \sum_{k=1}^n D_l \hat{q}_k D_J D_j \mathbf{x}_k.$$

That is,

$$(2.26) \quad D_j \left[\sum_{k=1}^n D_J \hat{q}_k D_l \mathbf{x}_k \right] = D_j \left[\sum_{k=1}^n D_l \hat{q}_k D_J \mathbf{x}_k \right], \quad j = 1, \dots, J-1, \quad l = j+1, \dots, m,$$

is satisfied in an open neighborhood of $\varepsilon = 0$, since (2.23) is satisfied in an open neighborhood of $\varepsilon = 0$. Therefore, (2.22) is satisfied for all values of d_1, \dots, d_m , with the possible exception of $d_1 = \dots = d_{J-1} = 0$. But this case follows from the fact that $\hat{q}(0, \dots, 0, \varepsilon_J, \dots, \varepsilon_m)$ satisfies (2.21) with $\varepsilon_1 = \dots = \varepsilon_{J-1} = 0$ due to the iterative construction.

Since $Q(\varepsilon)$ can be chosen to satisfy the integrability conditions (2.12), there exists a solution $\hat{k}(\varepsilon)$ to (2.10). To complete the proof of Theorem 2.1, we need only show that existence of $\hat{k}(\varepsilon)$ implies existence of a solution $k(\cdot)$ to (2.9).

Proof of Theorem 2.1. Since (H1) is satisfied, $D\mathbf{x}(0) = -[D_1 f(0, 0)]^{-1} D_2 f(0, 0)$ is full rank and there are m components, say $\mathbf{x}_1, \dots, \mathbf{x}_m$ for convenience, such that the

function $\mathbf{g}(\varepsilon) = [\mathbf{x}_1, \dots, \mathbf{x}_m]^T$ has a local inverse by the inverse function theorem. Using $k_i(\varepsilon)$ and $Q(\varepsilon)$ from Lemma 2.4, we complete the proof by verifying that

$$(2.27) \quad k_i(x) = \hat{k}_i(\mathbf{g}^{-1}(x_1, \dots, x_m)) + \sum_{j=m+1}^n q_{ij}(\mathbf{g}^{-1}(x_1, \dots, x_m)) \\ \times [x_j - \mathbf{x}_j(\mathbf{g}^{-1}(x_1, \dots, x_m))]$$

satisfies (2.9). Differentiating (2.27) with respect to x_j , $j = m+1, \dots, n$, and evaluating at \mathbf{x} we obtain

$$\frac{\partial k_i}{\partial x_j}(\mathbf{x}) = q_{ij}(\mathbf{g}^{-1}(\mathbf{x}_1, \dots, \mathbf{x}_m)) = q_{ij}(\varepsilon), \quad i = 1, \dots, n, \quad j = m+1, \dots, n.$$

Proceeding similarly for x_j , $j = 1, \dots, m$ we get

$$\begin{aligned} \frac{\partial k_i}{\partial x_j}(\mathbf{x}) &= \sum_{k=1}^n \sum_{l=1}^m q_{ik}(\varepsilon) D_l \mathbf{x}_k D_j \mathbf{g}_l^{-1}(\mathbf{x}_1, \dots, \mathbf{x}_m) \\ &\quad - \sum_{k=m+1}^n \sum_{l=1}^m q_{ik}(\varepsilon) D_l \mathbf{x}_k D_j \mathbf{g}_l^{-1}(\mathbf{x}_1, \dots, \mathbf{x}_m) \\ &= \sum_{k=1}^m \sum_{l=1}^m q_{ik}(\varepsilon) D_l \mathbf{x}_k D_j \mathbf{g}_l^{-1}(\mathbf{x}_1, \dots, \mathbf{x}_m) \\ &= q_{ij}(\varepsilon), \quad i = 1, \dots, n, \quad j = 1, \dots, m. \end{aligned}$$

Remark 2.5. It is tempting to attack the proof of Theorem 2.1 by first reducing the controllable multi-input nonlinear system to a controllable single-input nonlinear system, and then applying the single-input results. However, the single-input system has a one-dimensional operating point manifold, and hence the result would hold only on a one-dimensional subset of the m -dimensional operating point manifold.

3. State observation. The approach in § 2 also can be used to address the problem of observing the state of (2.1) given an output signal

$$(3.1) \quad y = h(x)$$

where $h(\cdot): \mathbf{R}^n \rightarrow \mathbf{R}^p$, $m \leq p < n$, is analytic with $h(0) = 0$. Corresponding to the constant operating points for (2.1), the constant-operating-point output is defined by

$$(3.2) \quad \mathbf{y}(\varepsilon) = h(\mathbf{x}(\varepsilon)).$$

Also, it will be assumed that (2.1), (3.1) satisfy the following hypotheses:

- (H2) (1) $[D_1 f(0, 0), Dh(0)]$ is an observable pair,
 (2) $D_1 f(0, 0)$ is invertible,
 (3) $D\mathbf{y}(0), Dh(0)$ are full rank.

The proposed observer for (2.1), (3.1) is described by

$$(3.3) \quad \dot{\hat{\mathbf{x}}} = f(\hat{\mathbf{x}}, w) + g(y) - g(\hat{y}), \quad \hat{y} = h(\hat{\mathbf{x}})$$

where $g(\cdot): \mathbf{R}^p \rightarrow \mathbf{R}^n$ is analytic with $g(0) = 0$. Denoting the difference (error) between the actual and observed state by $\tilde{\mathbf{x}} = \mathbf{x} - \hat{\mathbf{x}}$, $\tilde{\mathbf{x}}$ satisfies

$$(3.4) \quad \dot{\tilde{\mathbf{x}}} = f(\mathbf{x}, w) - f(\hat{\mathbf{x}}, w) - g(y) + g(\hat{y}).$$

This error equation has constant operating points for $w = \varepsilon$, $x = \hat{x} = \mathbf{x}$, and $y = \hat{y} = \mathbf{y}$. Linearizing (3.4) about such an operating point yields the linearized error equation,

$$(3.5) \quad \dot{\tilde{x}} = [D_1 f(\mathbf{x}, \varepsilon) - Dg(\mathbf{y}) Dh(\mathbf{x})] \tilde{x}.$$

The main result is as follows.

THEOREM 3.1. *Suppose the analytic system (2.1), (3.1) satisfies (H2). Then there exists an analytic observer gain $g(\cdot): \mathbf{R}^p \rightarrow \mathbf{R}^n$ with $g(0) = 0$ such that the eigenvalues of the linearized error equation (3.5) are specified, analytic functions of ε in an open neighborhood of $\varepsilon = 0$.*

Remark 3.2. With the choice of $g(\cdot)$ described above, (3.3) is an observer for the original system in the following sense. When the system state, x , is close to the constant operating point \mathbf{x} , and the estimate of x , \hat{x} , is close to \mathbf{x} , then \hat{x} will approach x and the error will be described approximately by (3.5).

Proof. It will be convenient to perform a linear change of output coordinates, and write (3.1), (3.2) as

$$(3.6) \quad y = Gh(x), \quad \mathbf{y}(\varepsilon) = Gh(\mathbf{x}(\varepsilon))$$

where G is a real, invertible, $p \times p$ matrix that is chosen as follows. Differentiating (3.6), we obtain

$$D\mathbf{y}(0) = -G \quad Dh(0)[D_1 f(0, 0)]^{-1} D_2 f(0, 0).$$

Deletion of the first column from both sides yields a $p \times (m-1)$ matrix equation, and since $\text{rank } D\mathbf{y}(0) = m$, with $m \leq p$, it follows that an invertible G can be chosen such that the first row is $[D_2 \mathbf{y}_1(0) \cdots D_m \mathbf{y}_1(0)] = 0$, and the $m-1$ columns are linearly independent.

From (H2) it follows that $[D_1 f(0, 0), GDh(0)]$ is an observable pair, and it can be assumed, as in the proof of Theorem 2.1, that the pair $[D_1 f(0, 0)^T, Dh(0)^T G^T]$ is in Brunovsky canonical form (without feedback) [5, p. 501]. Then by Corollary 2.3 there exists an $n \times p$ matrix $Q(\varepsilon)$ such that the eigenvalues of

$$(3.7) \quad D_1 f(\mathbf{x}, \varepsilon)^T - Dh^T(\mathbf{x}) G^T Q^T(\varepsilon)$$

have specified values with respect to ε . Furthermore, all columns of Q except the first can be left unspecified. The proof will be completed by showing that these unspecified columns can be chosen so that there exists an analytic $g(\cdot)$ with $g(0) = 0$ and

$$(3.8) \quad Dg(\mathbf{y}(\varepsilon)) = Q(\varepsilon).$$

When $\hat{g}(\varepsilon) = g(\mathbf{y}(\varepsilon))$, (3.8) implies

$$(3.9) \quad D\hat{g}(\varepsilon) = Q(\varepsilon) D\mathbf{y}(\varepsilon)$$

and existence of an analytic solution to (3.9) is necessary for existence of an analytic solution to (3.8). Following the development leading to (2.12), the integrability conditions for (3.9) can be written in the form

$$(3.10) \quad H_j D_j q_i^T = \begin{bmatrix} D_{j+1} q_i \\ \vdots \\ D_m q_i \end{bmatrix} D_j \mathbf{y}, \quad i = 1, \dots, n, \quad j = 1, \dots, m-1$$

where

$$(3.11) \quad H_j = [D_{j+1} \mathbf{y} | \cdots | D_m \mathbf{y}]^T.$$

The i th entry of the first column of Q will be written as $q_{i1} = q_{i1}(\varepsilon_1, \dots, \varepsilon_m, q_{1*}, \dots, q_{n*})$, where $q_{k*} = [q_{k2} \dots q_{kp}]$, and then the derivative of q_{i1} with respect to ε_j is given by

$$D_j q_{1*} D_{m+1} q_{i1}^T + \dots + D_j q_{n*} D_{m+n} q_{i1}^T + D_j q_{i1}.$$

Using this expression we can write the set of conditions in (3.10) in the form

$$(3.12) \quad W_j \begin{bmatrix} D_j q_{1*}^T \\ \vdots \\ D_j q_{n*}^T \end{bmatrix} = F_j(\varepsilon, q_{1*}, \dots, q_{n*}, D_{j+1} q_{1*}, \dots, D_m q_{n*}), \quad j = 1, \dots, m-1$$

where W_j is the $n(m-j) \times n(p-1)$ matrix partitioned into $(m-j) \times (p-1)$ blocks given by

$$(3.13) \quad W_j = \begin{bmatrix} V_j & & 0 \\ & \ddots & \\ 0 & & V_j \end{bmatrix} + \begin{bmatrix} z_j D_{m+1} q_{11} & \dots & z_j D_{m+n} q_{11} \\ & \ddots & \\ z_j D_{m+1} q_{n1} & \dots & z_j D_{m+n} q_{n1} \end{bmatrix},$$

$$V_j = \begin{bmatrix} D_{j+1} y_2 & \dots & D_{j+1} y_p \\ \vdots & & \vdots \\ D_m y_2 & \dots & D_m y_p \end{bmatrix}, \quad z_j = \begin{bmatrix} D_{j+1} y_1 \\ \vdots \\ D_m y_1 \end{bmatrix}.$$

Now, by virtue of the choice of G , $z_1(0), \dots, z_{m-1}(0)$ are zero and $V_1(0), \dots, V_{m-1}(0)$ have full rank. Thus the coefficient matrices W_1, \dots, W_{m-1} have full rank in an open neighborhood of $\varepsilon = 0$. The remainder of the proof follows from the proofs of Lemma 2.4 and Theorem 2.1, with $g(\cdot)$ replacing $k(\cdot)$, and y replacing x .

Remark 3.3. An observer of dimension $n-p$ can be constructed as shown in [1] and a result similar to Theorem 3.1 can be proved. Also it should be noted that the observer and state feedback results can be combined to give an output feedback control law in the usual way [1].

4. Example. In the preceding sections, special forms for linear systems and particular matrix partitionings were used to prove the invertibility of certain matrices. However, for specific examples, there are easier ways to proceed. The following simple example is meant to clarify the basic steps of the design methodology. Only the state feedback computations will be illustrated. Less simple examples probably would not yield closed-form solutions and would have to be approached via (truncated) series expansion methods.

Consider the bilinear system

$$\dot{x} = \begin{bmatrix} 1 & 0 \\ 0 & 2 \end{bmatrix} x + \begin{bmatrix} 1 & 0 \\ 1 & 0 \end{bmatrix} x u_1 + \begin{bmatrix} -1 \\ -1 \end{bmatrix} u_1 + \begin{bmatrix} 0 \\ -1 \end{bmatrix} u_2.$$

An easy calculation gives

$$x = \frac{1}{2(\varepsilon_1 + 1)} \begin{bmatrix} 2\varepsilon_1 \\ \varepsilon_1 + \varepsilon_2 + \varepsilon_1 \varepsilon_2 \end{bmatrix}, \quad D_x = \begin{bmatrix} \frac{1}{(1+\varepsilon_1)^2} & 0 \\ \frac{1}{2(1+\varepsilon_1)^2} & \frac{1}{2} \end{bmatrix},$$

$$D_1 f(x, \varepsilon) = \begin{bmatrix} 1 + \varepsilon_1 & 0 \\ \varepsilon_1 & 2 \end{bmatrix}, \quad D_2 f(x, \varepsilon) = \frac{1}{\varepsilon_1 + 1} \begin{bmatrix} -1 & 0 \\ -1 & -1 - \varepsilon_1 \end{bmatrix}.$$

From (2.10), dropping the arguments of each $q_{ij}(\varepsilon_1, \varepsilon_2)$,

$$D\hat{k}(\varepsilon) = \begin{bmatrix} \frac{2q_{11} + q_{12}}{2(1 + \varepsilon_1)^2} & \frac{q_{12}}{2} \\ \frac{2q_{21} + q_{22}}{2(1 + \varepsilon_1)^2} & \frac{q_{22}}{2} \end{bmatrix}$$

and the integrability conditions in (2.12) are

$$(4.1) \quad D_1 q_{12} = \frac{2D_2 q_{11} + D_2 q_{12}}{(1 + \varepsilon_1)^2}, \quad D_1 q_{22} = \frac{2D_2 q_{21} + D_2 q_{22}}{(1 + \varepsilon_1)^2}.$$

If the closed-loop poles are to be placed at -2 , independent of the operating point, $Q(\varepsilon)$ must satisfy

$$\det[sI - D_1 f(\mathbf{x}, \varepsilon) + D_2 f(\mathbf{x}, \varepsilon)Q(\varepsilon)] = s^2 + 4s + 4.$$

Ignoring the q_{i+} , q_{i-} partitioning as unnecessary in this case, and solving this equation for q_{11} and q_{12} in terms of q_{21} and q_{22} yields

$$q_{11} = -\frac{q_{22}(q_{21}(\varepsilon_1 + 1) + \varepsilon_1 + \varepsilon_1^2) + q_{21}(7 + 8\varepsilon_1 + \varepsilon_1^2) - 9 - 8\varepsilon_1 + \varepsilon_1^2}{q_{22} + q_{21} + 1},$$

$$q_{12} = -\frac{(\varepsilon_1 + 1)q_{22}^2 + (8\varepsilon_1 + 8)q_{22} + 16\varepsilon_1 + 16}{q_{22} + q_{21} + 1}.$$

Substituting $q_{11}(\varepsilon_1, \varepsilon_2, q_{21}, q_{22})$ and $q_{12}(\varepsilon_1, \varepsilon_2, q_{21}, q_{22})$ for $q_{11}(\varepsilon_1, \varepsilon_2)$ and $q_{12}(\varepsilon_1, \varepsilon_2)$, respectively, in (4.1), we obtain (2.18) for this case as

$$(4.2) \quad \begin{bmatrix} D_3 q_{12} & D_4 q_{12} \\ 0 & 1 \end{bmatrix} \begin{bmatrix} D_1 q_{21} \\ D_1 q_{22} \end{bmatrix} = \begin{bmatrix} -D_1 q_{12} \\ 0 \end{bmatrix}.$$

Note that since ε_2 does not appear explicitly, the solution will not depend upon ε_2 if the initial condition does not, and therefore all terms involving derivatives with respect to ε_2 have been set to zero. Choosing the initial conditions

$$q_{21}(0, \varepsilon_2) = 0, \quad q_{22}(0, \varepsilon_2) = 0,$$

we can solve (4.2) to yield

$$q_{21}(\varepsilon_1, \varepsilon_2) = \varepsilon_1, \quad q_{22}(\varepsilon_1, \varepsilon_2) = 0.$$

Thus (2.10) has the form

$$D\hat{k}(\varepsilon) = \begin{bmatrix} \frac{-\varepsilon_1^2 - 8\varepsilon_1 + 1}{(1 + \varepsilon_1)^2} & -8 \\ \frac{\varepsilon_1}{(1 + \varepsilon_1)^2} & 0 \end{bmatrix}$$

which can be integrated to yield

$$\hat{k}(\varepsilon) = \begin{bmatrix} -\varepsilon_1 - \frac{8}{1 + \varepsilon_1} - 6 \ln |1 + \varepsilon_1| - 8\varepsilon_2 + 8 \\ \frac{1}{1 + \varepsilon_1} + \ln |1 + \varepsilon_1| - 1 \end{bmatrix}.$$

Since $\varepsilon_1 = \mathbf{x}_1/(1 - \mathbf{x}_1)$ and $\varepsilon_2 = 2\mathbf{x}_2 - \mathbf{x}_1$, (2.27) gives

$$k(\mathbf{x}) = \begin{bmatrix} \frac{-x_1}{1 - x_1} + 6 \ln |1 - x_1| + 16x_1 - 16x_2 \\ -x_1 - \ln |1 - x_1| \end{bmatrix}.$$

REFERENCES

- [1] W. T. BAUMANN AND W. J. RUGH, *Feedback control of nonlinear systems by extended linearization*, IEEE Trans. Automat. Control, AC-31 (1986), pp. 40-46.
- [2] W. J. RUGH, *Design of nonlinear compensators for nonlinear systems by an extended linearization technique*, Proc. IEEE Conference on Decision and Control, Las Vegas, NV, 1984, pp. 69-73.
- [3] C. REBOULET AND C. CHAMPETIER, *A new method for linearizing nonlinear systems: the Pseudolinearization*, Internat. J. Control, 40 (1984), pp. 631-638.
- [4] C. CHAMPETIER, P. MOUYON AND C. REBOULET, *Pseudolinearization of multi-input nonlinear systems*, Proc. IEEE Conference on Decision and Control, Las Vegas, NV, 1984, pp. 96-97.
- [5] T. KAILATH, *Linear Systems*, Prentice-Hall, Inc., Englewood Cliffs, NJ, 1980.
- [6] J. DIEUDONNE, *Foundations of Modern Analysis*, Academic Press, New York, 1969.

NECESSARY OPTIMALITY CONDITIONS FOR THE CONTROL OF SEMILINEAR HYPERBOLIC BOUNDARY VALUE PROBLEMS*

MARTIN BROKATE†

Abstract. This paper develops the differential version of Pontryagin's principle for optimal (interior and boundary) control of a first-order semilinear hyperbolic system in one space dimension with nonlinear boundary conditions. The proof uses first order necessary conditions for constrained optimization in Banach space as well as the method of characteristics.

Key words. hyperbolic system, optimal control, necessary conditions, characteristics

AMS(MOS) subject classifications. 49B22, 35L50

1. Introduction. We consider optimal control of semilinear hyperbolic systems in one space dimension of type

$$z_t + A(t, x)z_x = f(t, x, u, z)$$

in a domain Ω whose boundary consists of spacelike, timelike and characteristic portions. We prove necessary optimality conditions along the following line of reasoning: In one space dimension, the method of characteristics enables one to rewrite the IBVP as a Volterra system of integral equations with regular kernel. This provides sup-norm estimates for the state z . On the other hand, a differentiable source f and differentiable (nonlinear) boundary conditions lead to substitution (Nemytskii) operators which are differentiable in sup-norm function spaces. It is the strength of the method of characteristics that this interplay works for quite a general form of the domain Ω , of the boundary conditions and without additional assumptions on the source f . On the other hand, these arguments usually cannot be extended to hyperbolic systems in more than one space dimension, which is a severe limitation.

Anyway, if one proceeds in the manner described above, one obtains a constrained optimization problem of differentiable type in Banach spaces. We show that the "abstract" maximum principle from [30] can be applied and derive the first order conditions.

In the context of optimal control, the constraint qualification (which guarantees nondegenerate first-order necessary conditions) usually is equivalent to well-posedness of the IVP, respectively, IBVP plus some controllability property. Since we do not want to study the latter, we restrict ourselves to a control problem consisting of a well-posed IBVP and pure control constraints only.

Pontryagin-type and variational maximum principles for control of hyperbolic systems have been developed at least since the early 1960's (see Cesari [1], Butkovsky et al. [2], [3]). Many results for linear-quadratic problems in n space are presented in Lions [4]. Recently, convex problems in one space dimension have been treated in [5]–[7]. If more general nonlinearities enter the picture, the simplest situation is the control problem consisting of the characteristic IVP with nonlinear source f ,

$$(1.1) \quad \begin{aligned} z_{xt} &= f(t, x, u, z, z_x, z_t) \quad \text{in } (0, T) \times (0, 1), \\ z(0, x) &= f_0(x), \quad z(t, 0) = f_1(t), \end{aligned}$$

* Received by the editors February 18, 1985; accepted for publication (in revised form) May 1, 1986.

† Naturwissenschaftliche Fakultät, Universität Augsburg, Memminger Straße 6, 8900 Augsburg, West Germany.

with distributed control $u = u(t, x)$, plus some cost functional and control restriction. Consequently, this problem (or slight generalizations) has been tackled by various authors with various techniques (see [8]–[12] and [13, pp. 233 ff.]), and a Pontryagin-type maximum principle (adjoint BVP, Hamiltonian is maximized by optimal control) has been established. Also, optimal singular controls have been investigated [14]–[16].

Nonlinear control of second-order IBVP has been treated with controls in the coefficients [17] and with piecewise continuous interior and boundary controls [18]–[20], even in n space [21], [22], and necessary conditions are given there.

The method adopted in this paper is a standard one for control of ODE's [23]–[25] and also has been applied to control of FDE's [26], but it has not been used for hyperbolic equations except in the linear-quadratic case [27]. Because it leads directly only to the differential version of the maximum condition, the results presented here are slightly weaker if applied to the problems in the paper cited above. On the other hand, this method permits a unified treatment of various IBVP's and is quite easily adapted if one modifies the way the control variable enters the problem (e.g. for the situation in [28]).

It would be interesting to know whether the techniques of [23]–[25] can be generalized in order to yield the assertion that the Hamiltonian is maximized along the optimal trajectory.

In this paper, we do not want to discuss the question of existence of optimal controls seriously. Due to the general geometry and boundary conditions there is no existence result, which could be cited, for the control problem considered here. However, if the controls are constrained pointwise by a priori bounds, then from Lions [4] and Ahmed and Teo [13] and the references cited there, one gets the impression that the overall picture is similar to the situation in optimal control of ODE's: a bounded measurable optimal control exists, if the control problem has convex structure with respect to the control (e.g., if the set of admissible velocities at a given point is convex, or if some Cesari property holds). Otherwise, convexity must be obtained by admitting generalized controls (see e.g. [37], [38]).

The paper is organized as follows: in § 2, notation and formulas needed later are developed to transform the hyperbolic control problem (consisting of (2.2)–(2.4), (2.12), (2.13)) into the Banach space setting of § 2.6. The main theorem is stated in § 3 and proved in § 4. It is illustrated with a control problem for the semilinear wave equation in § 5.

2. The control problem.

2.1. The domain Ω and its boundary. Let Ω be an open bounded set in \mathbb{R}^2 . We assume $\partial\Omega$ to be of the form

$$\begin{aligned} \partial\Omega &= \Gamma_0 \cup \Gamma_1 \cup \Gamma_2 \cup \Gamma_T, \\ (2.1) \quad \Gamma_i &= \{(t, b_i(t)) \mid 0 \leq t \leq T\}, \quad i = 1, 2, \\ \Gamma_0 &= \{0\} \times [b_1(0), b_2(0)], \quad \Gamma_T = \{T\} \times [b_1(T), b_2(T)] \end{aligned}$$

where $b_1, b_2: [0, T] \rightarrow \mathbb{R}$ are piecewise C^1 and $b_1(t) < b_2(t)$ for $t \in (0, T)$. We allow $b_1(0) = b_2(0)$ and $b_1(T) = b_2(T)$. (See Fig. 1.)

2.2. Cost functional and control restrictions. We consider a state $z: \bar{\Omega} \rightarrow \mathbb{R}^n$, an interior control $u: \Omega \rightarrow \mathbb{R}^m$ and a boundary control $u^B: [0, T] \rightarrow \mathbb{R}^p$. We want to minimize

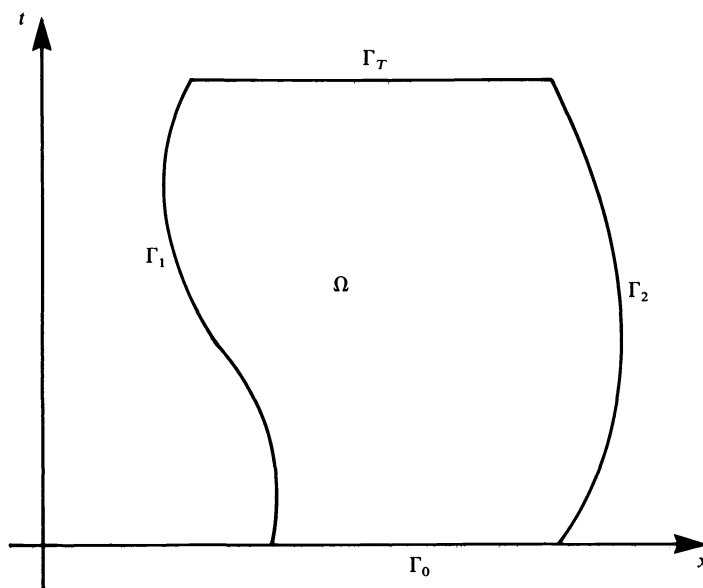


FIG. 1

the functional

$$(2.2) \quad \begin{aligned} \tilde{J}(z, u, u^B) = & \int \int_{\Omega} L(t, x, u, z) \, dx \, dt + \int_{b_1(T)}^{b_2(T)} L^T(x, z(T, x)) \, dx \\ & + \int_0^T L^B(t, u^B(t), z(t, b_1(t)), z(t, b_2(t))) \, dt \end{aligned}$$

subject to the pointwise control restrictions

$$(2.3) \quad u(t, x) \in U, \quad u^B(t) \in U^B \quad \text{a.e.}$$

where U and U^B are closed convex sets in \mathbb{R}^m , respectively \mathbb{R}^p .

2.3. The differential equation in Ω . We consider the semilinear first order system

$$(2.4) \quad z_t + A(t, x)z_x = f(t, x, u, z) \quad \text{for } (t, x) \in \Omega$$

where $A(t, x)$ is a diagonal $n \times n$ -matrix. If A is not diagonal, but diagonalizable, a transformation of z yields a system of form (2.4) with A diagonal [29, p. 44 ff.]. The classical method of characteristics consists of that transformation and the observation that one obtains a solution of (2.4) if one integrates f along the characteristics $(t, x_i(t))$ given by $x'_2 = a_i(t, x_i)$. Or, equivalently, if the coordinate transformation $\varphi_i = \varphi_i(t, s)$ satisfies $D_s \varphi_i(t, s) = (1, a_i(\varphi_i(t, s)))$, and if we set

$$y_i(t, s) = z_i(\varphi_i(t, s)),$$

then (2.4) becomes

$$y_i(t, s) = y_i(t, s_0) + \int_{s_0}^s f(\varphi_i(t, \tau), u, y) \, d\tau.$$

2.4. The characteristic transformation φ_i . We need a formal description of φ_i corresponding to a diagonal entry a_i of A . For momentary convenience, we drop

subscripts i, j and write $a(P), b'(P)$ instead of $a_i(t, b_j(t)), b'_j(t)$, if $P = (t, b_j(t)) \in \Gamma_1 \cup \Gamma_2$. We need characteristics which reach the boundary to do so transversally.

DEFINITION 2.1. We say that $a = a(t, x)$ is a regular left characteristic for Ω , if

- (i) a is C^1 on an open set containing $\bar{\Omega}$;
- (ii) There exists an $\varepsilon > 0$ such that for all $P \in \Gamma_1 \cup \Gamma_2$ either $a(P) = b'(P)$ or $a(P) - b'(P) \geq \varepsilon$;
- (iii) The initial and terminal portions $\Gamma_{\text{in}}, \Gamma_{\text{ter}}$ of $\partial\Omega$ are connected, where we define

$$\Gamma_{\text{char}} := \{P \mid P \in \Gamma_1 \cup \Gamma_2 \text{ and } a(P) = b'(P)\},$$

$$\Gamma_{\text{in}} := \text{cl}(\Gamma_0 \cup (\Gamma_1 \setminus \Gamma_{\text{char}})),$$

$$\Gamma_{\text{ter}} := \text{cl}(\Gamma_T \cup (\Gamma_2 \setminus \Gamma_{\text{char}})).$$

A regular right characteristic is defined by (i)–(iii) with the roles of Γ_1 and Γ_2 interchanged and the inequality sign in (ii) reversed. The matrix A is a regular characteristic matrix for Ω if each entry is a regular left or right characteristic.

As an example, two standard situations for the 1D wave equation written as system $z_{1t} + z_{1x} = f_1, z_{2t} - z_{2x} = f_2$ are depicted in Fig. 2. The second corresponds to (1.1) rotated by $\pi/4$. If $a(t, x)$ is a regular left characteristic, we define the corresponding φ by

$$(2.5) \quad \varphi(t, s) = (t + s, x(t, s))$$

where $x(t, s)$ is the solution of the parametrized IVP

$$(2.6) \quad \begin{aligned} \frac{\partial}{\partial s} x(t, s) &= a(t + s, x(t, s)), \\ x(t, -t) &= b_1(0) - t, \quad t \leq 0, \\ x(0, t) &= b_1(t), \quad t \geq 0. \end{aligned}$$

For right characteristics, take $x(t, -t) = b_2(0) + t$, respectively, $x(t, 0) = b_2(t)$, as initial conditions. Standard theory of ODE's implies that $\varphi(t, s)$ is well defined, continuous

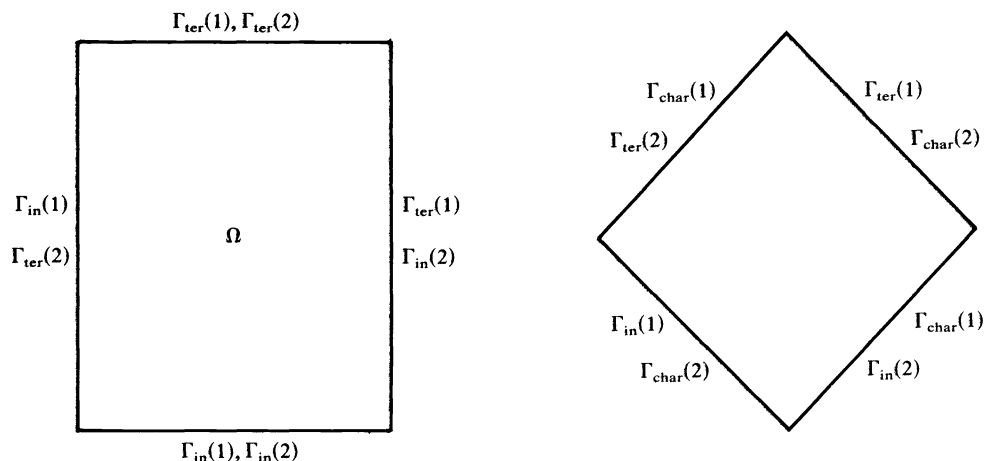


FIG. 2

in t , C^1 in s , and that the same holds for $D\varphi(t, s)$ for $t < 0$, and $t > 0$ if b_1 is C^1 around t . For such t we have

$$(2.7) \quad D\varphi(t, s) = \begin{pmatrix} 1 & 1 \\ \frac{\partial}{\partial t}x(t, s) & a(\varphi(t, s)) \end{pmatrix}$$

and if we define

$$(2.8) \quad \Psi(t, s) = \det D\varphi(t, s),$$

the chain rule and (2.6) imply

$$(2.9) \quad \begin{aligned} \frac{\partial}{\partial s}\Psi(t, s) &= \frac{\partial a}{\partial x}(\varphi(t, s)) \cdot \Psi(t, s), \\ \Psi(t, 0) &= a(P) - b'(P), \quad P = \varphi(t, 0), \quad t > 0, \\ \Psi(t, -t) &= 1 \quad (-1 \text{ for right characteristic}), \quad t < 0. \end{aligned}$$

To define the transformed region $G (= \varphi^{-1}(\Omega))$ we set

$$(2.10) \quad \begin{aligned} \gamma^0(t) &= \begin{cases} 0, & t \geq 0, \\ -t, & t \leq 0, \end{cases} \\ \gamma^s(t) &= \sup \{s \mid \varphi(t, s) \in \bar{\Omega}\}, \\ T^- &= b_1(0) - b_2(0), \\ G &= \{(t, s) \mid T^- \leq t \leq T, \gamma^0(t) \leq s \leq \gamma^s(t)\}. \end{aligned}$$

Hence γ^0 , γ^s correspond to Γ_{in} , Γ_{ter} . Γ_{char} corresponds to boundary portions of G with constant t value. It helps to visualize (2.10) for the examples in Fig. 2. Because of Definition 2.1, (2.8), (2.9) we know that

$$(2.11) \quad \Psi(t, s) \text{ is bounded away from zero in } G.$$

Now we have the following.

LEMMA 2.2. *Let φ_i denote the transformation corresponding to a regular characteristic $a_i = a_i(t, x)$. Then:*

- (i) $\varphi_i: G_i \rightarrow \Omega$ is well defined by (2.5), (2.6), (2.10);
- (ii) φ_i is continuous and bijective;
- (iii) *There are finitely many intervals I_k covering \mathbb{R} such that $\varphi_i|_{G_i \cap (I_k \times \mathbb{R})}$ can be extended to a C^1 -diffeomorphism defined on an open subset of \mathbb{R}^2 .*

Furthermore let $\varphi_{ij}: G_i \rightarrow G_j$ be defined by $\varphi_{ij} = \varphi_j^{-1} \circ \varphi_i$ and denote

$$\varphi_{ij}(t, s) = (t_{ij}(t, s), s_{ij}(t, s)).$$

Then

$$(iv) \quad t_{ij}(t, s) + s_{ij}(t, s) = t + s.$$

Proof. Equation (2.11) and the inverse function theorem yield (i)–(iii). Equation (2.5) implies (iv).

2.5. The boundary conditions on $\partial\Omega$. According to standard theory, we expect the IBVP for (2.4) with fixed control u to be well posed if, along each portion of $\partial\Omega$, we prescribe the values of the z_i 's whose characteristics originate at that portion as a function of the z_j 's whose characteristics end there. Thus, if Γ_0 does not degenerate to a point, we have the initial conditions for $P = (0, x) \in \Gamma_0$:

$$(2.12) \quad z_i(P) = z_{i0}(x), \quad 1 \leq i \leq n.$$

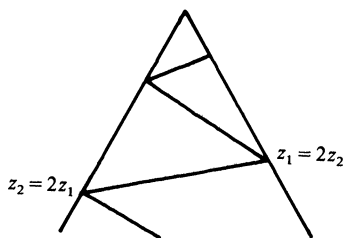


FIG. 3

For $P = (t, x) \in \Gamma_{\text{in}}(i)$, $t > 0$, we take, slightly generalizing,

$$(2.13) \quad z_i(P) = g_i(t, u^B(t), z^P(P), z^Q(Q))$$

where z^P is the subvector of z consisting of components z_j with $P \in \Gamma_{\text{ter}}(j)$, and $Q = (t, \bar{x})$ lies on the opposite boundary.

In order to write (2.13) for the transformed domains G_i , we introduce auxiliary functions r_j, \bar{r}_j . If $P = (t, x) \in \Gamma_{\text{ter}}(j) \setminus \Gamma_T$, we define $r_j(t)$ by

$$(2.14) \quad \varphi_j(r_j(t), t - r_j(t)) = (t, x)$$

(this is possible because of 2.1(iii) and (2.5)). The implicit function theorem and (2.7), (2.8), (2.11) imply that r_j is piecewise C^1 and

$$(2.15) \quad r'_j(t) \Psi_j(r_j(t), t - r_j(t)) = a_j(P) - b'(P).$$

If $P = (T, x) \in \Gamma_T$, we define $\bar{r}_j(x)$ by

$$(2.16) \quad \varphi_j(\bar{r}_j(x), T - \bar{r}_j(x)) = (T, x).$$

Analogous considerations yield a well-defined \bar{r}_j which is piecewise C^1 and satisfies

$$(2.17) \quad \bar{r}'_j(x) \Psi_j(\bar{r}_j(x), T - \bar{r}_j(x)) = -1.$$

These considerations prove the following.

LEMMA 2.3. *If $a_j = a_j(t, x)$ is a regular characteristic for Ω , then r_j, \bar{r}_j are well defined on intervals determined by $\Gamma_{\text{ter}}(j)$ and piecewise C^1 . Furthermore, $r'_j, \bar{r}'_j, (\bar{r}_j^{-1})'$ are bounded.*

If Γ_T or Γ_0 degenerates to a single point, we have to exclude the infinite zigzagging of Fig. 3. This is done formally by the following.

Assumption 2.4. We say that (j, i) is a degenerate pair at T if Γ_T consists of a single point and g_i depends explicitly on z_j .

Consider the directed graph Z_T with integers i as nodes and all degenerate pairs (j, i) as vertices. Then Z_T must not contain a cycle.

The same is assumed at 0.

2.6. The hyperbolic control problem (C) in characteristic form. We state the control problem (C) in a compact notation. See § 2.7. Assume that A is a regular characteristic matrix. With T^- from (2.10), set $S = T - T^-$. For

$$(2.18) \quad \begin{aligned} u &\in L_\infty(\Omega; \mathbb{R}^m), & u^B &\in L_\infty(0, T; \mathbb{R}^P), \\ y &\in C(0, S; L_\infty(T^-, T))^n \\ &(\text{i.e. } y_i(\cdot, s) \in L_\infty(T^-, T) \text{ for } s \in [0, S]) \end{aligned}$$

minimize

$$\begin{aligned}
 J(y, u, u^B) = & \int \int_{\Omega} L(t, x, u(t, x), y(\varphi^{-1}(t, x))) \, dx \, dt \\
 (2.19) \quad & + \int_0^T L^B(t, u^B(t), y(r(t), S)) \, dt \\
 & + \int_{b_1(T)}^{b_2(T)} L^T(x, y(\bar{r}(x), S)) \, dx
 \end{aligned}$$

subject to the state equations

$$(2.20) \quad y(t, s) = y(t, 0) + \int_0^s (F(u, y))(t, \tau) \, d\tau,$$

$$(2.21) \quad y(t, 0) = g(t, u^B(t), y(r(t), S))$$

and the control restrictions

$$(2.22) \quad u(t, x) \in U \quad \text{a.e.}, \quad u^B(t) \in U^B \quad \text{a.e.}$$

where U, U^B are closed convex sets and $F = (F_1, \dots, F_n)$ is defined by

$$(2.23) \quad (F_i(u, y))(t, \tau) = f_i(\varphi_i(t, \tau), u(\varphi_i(t, \tau)), y(\varphi^{-1}(\varphi_i(t, \tau))))$$

if $(t, \tau) \in G_i$ and $(F_i(u, y))(t, \tau) = 0$ otherwise.

2.7. Comments and explanations concerning problem (C).

(1) We have used the abbreviation

$$y(r(t), S) \equiv (y_1(r_1(t), S), \dots, y_n(r_n(t), S))$$

and likewise $y(\bar{r}(x), S), y(\varphi^{-1}(t, x))$.

(2) The convention $F_i(u, y) = 0$ outside $G_i = \varphi_i^{-1}(\Omega)$ enables one to pick up the initial values at $y(t, 0)$ and the terminal values at $y(t, S)$.

(3) The initial conditions (2.12) are included in (2.21) by a suitable definition $g = g(t)$ for $t < 0$.

(4) In accordance with § 2.5, for $t > 0$, the function $g_i(t, \cdot)$ depends on y_j only if the corresponding $P = (t, x) \in \Gamma_{\text{ter}}(j)$. Therefore, $y_j(r_j(t), S)$ is well defined.

(5) The foregoing remarks show that problem (C) is indeed a weak formulation of (2.2)–(2.4), (2.12), (2.13).

(6) Concerning the states y_i , the space $C([T^-, T] \times [0, S])$ is too small if we include discontinuous controls, and the space $L_\infty([T^-, T] \times [0, S])$ only causes additional complications concerning the trace $y(\cdot, S)$.

3. The maximum principle. We consider problem (C) as a special case of

$$\begin{aligned}
 (P) \quad & \text{Minimize} \quad J(y, v) \\
 & \text{subject to} \quad K(y, v) = 0, \quad v \in C
 \end{aligned}$$

where $J: Y \times V \rightarrow \mathbb{R}$, $K: Y \times V \rightarrow Y$ are continuously Fréchet-differentiable mappings defined on Banach spaces Y and V (whose duals are denoted by Y^*, V^*) and C is a closed convex subset of V .

THEOREM 3.1 (Lagrange Multiplier Theorem). *Let (y_*, v_*) be a solution of problem (P) with assumptions as stated, and let $D_y K(y_*, v_*): Y \rightarrow Y$ be surjective. Then there is a unique $q^* \in Y^*$ which solves the adjoint equation*

$$(3.1) \quad D_y J(y_*, v_*) - q^* \circ D_y K(y_*, v_*) = 0.$$

Furthermore, the maximum condition

$$(3.2) \quad \langle D_v J(y_*, v_*) - q^* \circ D_v K(y_*, v_*), v - v_* \rangle \geq 0 \quad \forall v \in C$$

holds.

Proof. This is a special case of Theorem 4.1(a) in [30]. Uniqueness trivially follows from (3.1) since $D_y K$ is surjective. \square

If $\text{int}(C) \neq \emptyset$, Theorem 3.1 is already part of the theory in [23] (see [24]). The work by Zowe and Kurcyusz [30] relies mainly on the stability theory of Robinson (see e.g. [31], [32]). A systematic exposition of the latter with emphasis on necessary optimality conditions may be found in [33].

As can be seen from [30]–[33], the existence part of Theorem 3.1 still holds if the assumption of surjectivity of $D_y K(y_*, v_*)$ is relaxed to the condition of Slater type

$$0 \in \text{int}(D_y K(y_*, v_*)Y + D_v K(y_*, v_*)(C - v_*)).$$

However, having also guaranteed uniqueness of q^* by (3.1) will be very convenient later on, since a general element of Y^* will be quite irregular, but we will be able to show that there exists a solution to (3.1) which has more regularity.

The differentiability needed in Theorem 3.1 will be guaranteed by the following.

Assumption 3.2. The functions L, L^B, L^T, g, f in problem (C) are measurable in (t, x) , differentiable in (y, u, u^B) . They are, together with their first derivatives, bounded on bounded sets in (t, x, y, u, u^B) -space and continuous in (y, u, u^B) uniformly with respect to (t, x) . For the derivatives at the optimal point we use an abbreviated notation; for example, $D_{yi} F(t, \tau)^T$ denotes the transpose of the vector $D_{yi} F(y_*, u_*)(t, \tau)$, the j th component therefore being

$$D_{yi} f(\varphi_j(t, \tau), u_*(\varphi_j(t, \tau)), y_{*1}(\varphi_1^{-1} \varphi_i(t, \tau)), \dots, y_{*n}(\varphi_n^{-1} \varphi_j(t, \tau))).$$

We apply Theorem 3.1 to problem (C) and get the following.

THEOREM 3.3. *Let Assumptions 2.4 and 3.2 hold. Let (y_*, u_*, u_*^B) be a solution of problem (C). Then there exists a unique $q \in C(0, S; L_\infty(T^-, T))^n$ which solves the adjoint equation*

$$(3.3) \quad \begin{aligned} q_i(t, s) = & q_i(t, S) + \int_s^S D_{yi} L(t, \tau) + \frac{\partial a_i}{\partial x}(\varphi_i(t, \tau)) q_i(t, \tau) \\ & + D_{yi} F(t, \tau)^T q(\varphi^{-1} \varphi_i(t, \tau)) d\tau \end{aligned}$$

with the boundary conditions

$$(3.4) \quad \begin{aligned} q_i(\bar{r}_i(x), S) &= D_{yi} L^T(x) \quad \text{if } x \in \text{dom } \bar{r}_i, \\ q_i(r_i(t), S) &= \alpha_i(t)^{-1} \left[D_{yi} L^B(t) + \sum_{j=1}^n D_{yi} g_j(t) |\Psi_j(t, 0)| q_j(t, 0) \right] \quad \text{if } t \in \text{dom } r_i, \\ q_i(t, S) &= 0 \quad \text{otherwise} \end{aligned}$$

where

$$\begin{aligned} \alpha_i(t) &= |a_i(P) - b'(P)| \quad \text{if } P = (t, x) \in \Gamma_{\text{ter}}(i), \\ \Psi_j(t, 0) &= a_j(P) - b'(P) \quad \text{if } P = (t, x) \in \Gamma_{\text{in}}(j). \end{aligned}$$

Moreover, the following maximum conditions hold

$$(3.5) \quad \langle D_u f(t, x)^T q(\varphi^{-1}(t, x)) + D_u L(t, x), u - u_*(t, x) \rangle \geq 0 \quad \text{a.e. in } \Omega,$$

$$(3.6) \quad \left\langle \sum_{j=1}^n D_u g_j(t) |\Psi_j(t, 0)| q_j(t, 0) + D_u L^B(t), u - u_*^B(t) \right\rangle \geq 0 \quad \text{a.e. in } t$$

for all $u \in U$, respectively, $u \in U^B$.

Proof. The proof is given in § 4. \square

Despite the somewhat technical character of this theorem we can observe that the adjoint q has the same regularity as the state y (namely, absolutely continuous along characteristics, L_∞ across them). Furthermore, portions of $\partial\Omega$ which are terminal for y are initial for q and vice versa. Characteristic portions of $\partial\Omega$ are identical for y and q .

For a rectangular domain Ω we write down the maximum principle in original (t, x) -coordinates. In this case, regularity of $A(t, x)$ in Definition 2.1 means that A is nonsingular along the vertical boundaries.

THEOREM 3.4. *Let $\Omega = (0, T) \times (0, 1)$, let Assumption 3.2 hold and let (y_*, u_*, u_*^B) be a solution of problem (C). Then there exists a unique $p \in L_\infty(\Omega)^n$ whose components p_i are absolutely continuous along the i th characteristic family and are L_∞ along the boundaries, which solves the adjoint IBVP in reverse time:*

$$(3.7) \quad \frac{\partial p}{\partial t} + A(t, x) \frac{\partial p}{\partial x} = -D_y L(t, x) - \left(D_y f(t, x)^T + \frac{\partial A}{\partial x}(t, x) \right) p,$$

$$(3.8) \quad p(T, x) = D_y L^T(x), \quad x \in (0, 1),$$

$$(3.9) \quad |a_i(t, X)| p_i(t, X) = D_{y_i} g(t)^T |A(t, X)| p(t, X) + D_{y_i} L^B(t).$$

Moreover, the maximum conditions

$$(3.10) \quad \langle D_u f(t, x)^T p(t, x) + D_u L(t, x), u - u_*(t, x) \rangle \geq 0 \quad \text{a.e. in } \Omega,$$

$$(3.11) \quad \langle D_u g(t)^T |A(t, X)| p(t, X) + D_u L^B(t), u - u_*^B(t) \rangle \geq 0 \quad \text{a.e. in } t$$

hold for all $u \in U$, respectively, $u \in U^B$. In (3.9), the left side is evaluated at the boundary $X = 0$ or $X = 1$, which is terminal for the i th characteristic family, whereas the j th component of the vector $|A(t, X)| p(t, x)$ in (3.9), (3.11) is evaluated at the boundary which is initial for the j th characteristic family ($|A|$ denotes $(|a_{ij}|)_{i,j}$).

Proof. Take q from Theorem 3.3, set

$$p_i(t, x) = q_i(\varphi_i^{-1}(t, x)).$$

With (2.5), (2.6), (3.3) evaluate the identity

$$\frac{\partial}{\partial s} p_i(\varphi_i(t, s)) = \frac{\partial}{\partial s} q_i(t, s)$$

to obtain (3.7). The boundary conditions are a direct transcription from (3.4). \square

For the linear-quadratic problem with interior control and space-dependent coefficients, Theorem 3.4 is given in [13, p. 211].

From (2.4) and (3.7) one sees that singularities in u_* , u_*^B spread forward respectively backwards in time along the characteristics. Jump conditions may be developed to describe that situation (compare [18], [20], [21]).

4. Proof of Theorem 3.3. We denote

$$(4.1) \quad D := (T^-, T) \times (0, S).$$

4.1. The operators J and K . Following § 2.6, we set

$$Y = C(0, S; L_\infty(T^-, T))^n, \quad V = L_\infty(\Omega; \mathbb{R}^m) \times L_\infty(0, T; \mathbb{R}^p),$$

equipped with the sup-norm. The closed convex set $C \subset V$ is defined by

$$C = \{(u, u^B) \mid u(t, x) \in U \text{ a.e. and } u^B(t) \in U^B \text{ a.e.}\}.$$

We want to define $J: Y \times V \rightarrow \mathbb{R}$ by (2.19) and $K: Y \times V \rightarrow Y$ by

$$(Ky)(t, s) = y(t, s) - g(t, u^B(t), y(r(t), S)) - \int_0^s (F(y, u))(t, \tau) d\tau$$

with F from (2.23). We have the canonical embedding $Y \rightarrow L_\infty(D)$ (it is easily seen to be well defined on the set of continuous, piecewise linear maps $[0, S] \rightarrow L_\infty(T^-, T)$ which is dense in Y) and Assumption 3.2 implies that

$$F: L_\infty(D; \mathbb{R}^n) \times L_\infty(\Omega; \mathbb{R}^m) \rightarrow L_\infty(D; \mathbb{R}^n)$$

is well defined and continuously Fréchet-differentiable.

Fubini's theorem then implies that

$$y \rightarrow \int_0^s y(t, \tau) d\tau$$

defines a linear continuous map from $L_\infty(D)$ to Y .

The boundary term is composed of an evaluation at $s = S$, a continuous Fréchet-differentiable operator from $L_\infty(0, T; \mathbb{R}^p) \times L_\infty(T^-, T; \mathbb{R}^n)$ to $L_\infty(T^-, T; \mathbb{R}^n)$ and the canonical map $L_\infty(T^-, T; \mathbb{R}^n) \rightarrow Y$. Therefore, K (and similarly J) is well defined and continuously Fréchet-differentiable. The derivatives of J and K are given by

$$(4.2) \quad \begin{aligned} DJ(y_*, u_*, u_*^B)(y, u, u^B) &= \iint_\Omega D_y L(t, x) y(\varphi^{-1}(t, x)) + D_u L(t, x) u(t, x) dx dt \\ &+ \int_0^T D_y L^B(t) y(r(t), S) + D_u L^B(t) u^B(t) dt \\ &+ \int_{b_1(T)}^{b_2(T)} D_y L^T(x) y(\bar{r}(x), S) dx, \end{aligned}$$

$$(4.3) \quad \begin{aligned} (DK(y_*, u_*, u_*^B)(y, u, u^B))_i(t, s) &= y_i(t, s) - D_y g_i(t) y(r(t), S) \\ &- D_u g_i(t) u^B(t) - \int_0^s D_y F_i(t, \tau) y(\varphi^{-1} \varphi_i(t, \tau)) + D_u F_i(t, \tau) u(\varphi_i(t, \tau)) d\tau. \end{aligned}$$

4.2. The linearized integral equation. Theorem 3.1 requires $D_y K: Y \rightarrow Y$ to be surjective. From (4.3) we see that this is a direct consequence of the following lemma.

LEMMA 4.1. Consider the integral equation for y_i , $1 \leq i \leq n$,

$$(4.4) \quad y_i(t, s) = h_i(t, s) + \sum_{j=1}^n g_{ij}(t) y_j(r_j(t), S) + \int_0^s \sum_{j=1}^n f_{ij}(t, \tau) y_j(\varphi_{ij}(t, \tau)) d\tau$$

with given functions $h_i \in C(0, S; L_\infty(T^-, T))$, $g_{ij} \in L_\infty(T^-, T)$, $f_{ij} \in L_\infty(D)$.

Assume $f_{ij}(t, \tau) = 0$ outside G_i and if Γ_T consists of a single point, then for each (i, j) either $g_{ij}(t) = 0$ in a neighbourhood of T or (i, j) is a degenerate pair at T (and the same for 0).

Then (4.4) has a unique solution $y \in Y$.

Proof. It is well known that unique solvability of hyperbolic problems in dimension 1 can be proved by successive approximation (see [34] for IVP's and [35, pp. 471 ff.] for IBVP's). To account for the present situation, we define an equivalent norm on Y such that the linear part of the right-hand side of (4.4) has operator norm less than 1. For $y = (y_1, \dots, y_n) \in Y$ we define

$$\begin{aligned} \|y\| &= \max \{\|y_i\|_i : 1 \leq i \leq n\}, \\ (4.5) \quad \|y_i\|_i &= \sup \{n_i(t, s) | y_i(t, s) | : (t, s) \in D\}, \\ n_i(t, s) &= c_i(t) \exp [-M(t + \gamma_i^0(t)) - L(\min \{s, \gamma_i^s(t)\} - \gamma_i^0(t))]. \end{aligned}$$

We set $c_i \equiv 1$ if no degeneracy in the sense of Assumption 2.4 occurs; otherwise, if e.g. Γ_0 is degenerate, we change c_i on some small interval $[0, \varepsilon_0]$ to

$$(4.6) \quad c_i(t) = \min \{1, c_j(t) [4n \|g_{ij}\|_\infty]^{-1} : (j, i) \in Z_0\}.$$

Because of Assumption 2.4 this procedure is well defined, and there exists a global bound

$$c_i(t)^{-1} \leq C_0, \quad 1 \leq i \leq n, \quad T^- \leq t \leq T.$$

Since Y can be identified with a subspace of $L_\infty(D)$, (4.5) defines an equivalent norm on Y . We set

$$\delta_i(t) = \gamma_i^s(t) - \gamma_i^0(t)$$

and choose $C_1 > 0$ such that

$$(4.7) \quad \frac{c_i(t)}{c_j(t)} |g_{ij}(t)| \exp (-C_1 \delta_j(t)) \leq \frac{1}{2n}$$

for all $1 \leq i, j \leq n$, $T^- \leq t \leq T$. This is possible because of (4.6) and the hypothesis of the lemma. In (4.5) we now choose

$$M = L + C_1$$

and with the notation

$$\sigma = \min \{s, \gamma_i^s(t)\},$$

we estimate the integral term in (4.4) for fixed (t, s) :

$$\begin{aligned} n_i(t, s) \left| \int_0^s \sum_{j=1}^n f_{ij}(t, \tau) y_j(\varphi_{ij}(t, \tau)) d\tau \right| \\ \leq n_i(t, s) \int_0^\sigma \|f\|_\infty \sum_{j=1}^n n_j(\varphi_{ij}(t, \tau))^{-1} \|y_j\|_j d\tau \\ \leq n_i(t, s) C_0 \|f\|_\infty \int_0^\sigma \sum_{j=1}^n \|y_j\|_j \exp [(M - L)(t_{ij}(t, \tau) + \gamma_j^0(t_{ij}(t, \tau)))] \\ \cdot \exp [L(t + \tau)] d\tau \quad \text{from Lemma 2.2 (iv)} \\ \leq c_i(t) C_0 \|f\|_\infty \cdot \frac{1}{L} \sum_{j=1}^n \|y_j\|_j \cdot \exp [C_1(T + S)]. \end{aligned}$$

We therefore get the norm of the integral term as small as we want if we choose L large enough. The boundary term is estimated as follows:

$$\begin{aligned}
 n_i(t, s) \left| \sum_{j=1}^n g_{ij}(t) y_j(r_j(t), S) \right| &\leq n_i(t, s) \sum_{j=1}^n |g_{ij}(t)| n_j(r_j(t), \gamma_j^s(r_j(t)))^{-1} \|y_j\|_j \\
 &= \sum_{j=1}^n \frac{c_i(t)}{c_j(t)} |g_{ij}(t)| \|y_j\|_j \exp[(L-M)(\gamma_j^s(r_j(t)) - \gamma_j^0(r_j(t)))] \\
 &\quad \cdot \exp[(L-M)\gamma_j^0(t)] \cdot \exp[-L \min\{s, \gamma_i^s(t)\}] \\
 &\quad \text{(where we have used } r_j(t) + \gamma_j^s(r_j(t)) = t) \\
 &\leq \sum_{j=1}^n \frac{c_i(t)}{c_j(t)} |g_{ij}(t)| \exp(-C_1 \delta_j(t)) \leq \frac{1}{2} \quad \text{by (4.7).}
 \end{aligned}$$

The lemma is proved. \square

4.3. The adjoint equation. From Theorem 3.1, we obtain a unique $q^* \in Y^*$ which satisfies (3.1), (3.2). Now, Y^* is a subspace of $L_\infty(D)^*$ which is the space of all finitely additive measures on D [36]. We show that actually (3.1) has a solution q^* of the form

$$(4.8) \quad q_i^*(y_i) = \int_D \rho_i(t, s) y_i(t, s) ds dt + \int_{T^-}^T \rho_i^B(t) y_i(t, S) dt$$

where $\rho_i \in L_\infty(D)$, $\rho_i^B \in L_\infty(T^-, T)$. To prove this, we apply the left side of (3.1) to an arbitrary $y_i \in C(0, S; L_\infty(T^-, T))$, using (4.2), (4.3) and assuming (4.8):

$$\begin{aligned}
 B_i(y_i) &:= \left[D_{yi} J - \sum_{j=1}^n (q_j^* \circ D_{yi} K_j) \right] (y_i) \\
 &= \iint_{\Omega} D_{yi} L(t, x) y_i(\varphi_i^{-1}(t, x)) dx dt + \int_0^T D_{yi} L^B(t) y_i(r_i(t), S) dt \\
 &\quad + \int_{b_1(T)}^{b_2(T)} D_{yi} L^T(x) y_i(\bar{r}_i(x), S) dx - \iint_D \rho_i(t, s) y_i(t, s) ds dt \\
 &\quad - \int_{T^-}^T \rho_i^B(t) y_i(t, S) dt + \sum_{j=1}^n \int_{T^-}^T D_{yi} g_j(t) y_i(r_i(t), S) \\
 &\quad \cdot \left[\rho_j^B(t) + \int_0^S \rho_j(t, s) ds \right] dt \\
 &\quad + \sum_{j=1}^n \int_{T^-}^T \int_0^S \rho_j(t, s) \int_0^S D_{yi} f_i(\varphi_j(t, \tau)) y_i(\varphi_{ji}(t, \tau)) d\tau ds dt \\
 &\quad + \sum_{j=1}^n \int_{T^-}^T \rho_j^B(t) \int_0^S D_{yi} f_j(\varphi_j(t, \tau)) y_i(\varphi_{ji}(t, \tau)) d\tau dt.
 \end{aligned}$$

We define

$$(4.9) \quad \tilde{\rho}_i(t, s) = \rho_i^B(t) + \int_s^S \rho_i(t, \tau) d\tau$$

and obtain by partial integration and substitution (which is valid because of Lemmas 2.2, 2.3 and Fubini's theorem)

$$\begin{aligned}
 B_i(y_i) = & \int \int_D y_i(t, s) \left[D_{yi} L(\varphi_i(t, s)) |\Psi_i(t, s)| - \rho_i(t, s) \right. \\
 & \left. + \sum_j D_{yi} f_j(\varphi_i(t, s)) \tilde{\rho}_j(\varphi_{ij}(t, s)) |\Psi_i(t, s)| |\Psi_j(t, s)|^{-1} \right] ds dt \\
 & + \int_{T^-}^T y_i(t, S) \left[-\tilde{\rho}_i(t, S) + D_{yi} L^T(\bar{r}_i(t)) |D\bar{r}_i^{-1}(t)| + D_{yi} L^B(r_i^{-1}(t)) |Dr_i^{-1}(t)| \right. \\
 & \left. + \sum_j D_{yi} g_j(r_j^{-1}(t)) \tilde{\rho}_j(r_j^{-1}(t), 0) |Dr_j^{-1}(t)| \right] dt.
 \end{aligned}$$

Therefore, a q^* of form (4.8) satisfies (3.1) if and only if the bracketed terms vanish identically. The first term vanishes if and only if

$$\begin{aligned}
 (4.10) \quad \tilde{\rho}_i(t, s) = & \tilde{\rho}_i(t, S) + \int_s^S |\Psi_i(t, s)| \\
 & \cdot \left[D_{yi} L(\varphi_i(t, s)) + \sum_j D_{yi} f_j(\varphi_i(t, s)) \tilde{\rho}_j(\varphi_{ij}(t, s)) \cdot |\Psi_j(t, s)|^{-1} \right] ds.
 \end{aligned}$$

But (4.10) with the boundary condition inserted from the second bracket constitutes a system of integral equations for $\tilde{\rho}_i$ of form (4.4) in reverse coordinates $(T-t, S-s)$. For this reverse problem, the considerations of §§ 2.4, 2.5 as well as the hypotheses of Lemma 4.1 also hold. Therefore, a unique solution $\tilde{\rho} \in Y$ exists and ρ, ρ^B can be obtained via (4.9). We now set

$$q_i(t, s) = \tilde{\rho}_i(t, s) |\Psi_i(t, s)|^{-1}$$

and since we have from (2.9) that

$$\frac{\partial}{\partial s} q_i = |\Psi_i|^{-1} \frac{\partial}{\partial s} \tilde{\rho}_i - \tilde{\rho}_i |\Psi_i|^{-2} \frac{\partial}{\partial s} a_i,$$

we see that (4.10) implies the adjoint equation (3.3) and the boundary conditions for $\tilde{\rho}_i(t, S)$ yield (3.4) with the aid of (2.15), (2.17). Since $\tilde{\rho}$ is unique, q is unique.

4.4. The maximum condition. Since q^* is unique, the adjoint computed in § 4.3 also satisfies the maximum condition (3.2). From (4.2), (4.3) and (4.8), we evaluate (3.2) for arbitrary $u \in L_\infty(\Omega; \mathbb{R}^n)$:

$$\begin{aligned}
 (D_u J - q^* \circ D_u K)(u) &= \int \int_D D_u L(t, x) u(t, x) dx dt \\
 &+ \sum_{i=1}^n \int \int_D \rho_i(t, s) \int_0^s D_u f_i(\varphi_i(t, \tau)) u(\varphi_i(t, \tau)) d\tau ds dt \\
 &+ \sum_{i=1}^n \int_{T^-}^T \rho_i^B(t) \int_0^S D_u f_i(\varphi_i(t, s)) u(\varphi_i(t, s)) ds dt \\
 &= \int \int_\Omega u(t, x) \left[D_u L(t, x) + \sum_{i=1}^n \tilde{\rho}_i(\varphi_i^{-1}(t, x)) D_u f_i(t, x) |\Psi_i^{-1}(t, x)| \right] dx dt
 \end{aligned}$$

where again we use partial integration and substitution. Therefore, (3.2) implies

$$\iint_{\Omega} \langle D_u L(t, x) + D_u f(t, x)^T q(\varphi^{-1}(t, x)), u(t, x) - u_*(t, x) \rangle dx dt \geq 0$$

for all admissible u , and a standard argument using Lusin's theorem yields (3.5). In the same way, (3.6) is obtained. This completes the proof of Theorem 3.3. \square

5. The semilinear wave equation. The theorems of § 3 can be applied to control problems for a hyperbolic system of order n if the system can be reduced to a first-order system with regular characteristic matrix and if the boundary conditions are such that admissible controls produce a state which is L_∞ up to derivatives of order $n-1$. We illustrate this with the following control problem for the semilinear wave equation in $\Omega = (0, T) \times (0, 1)$:

Minimize

$$(5.1) \quad J(w) = \int_0^1 \tilde{L}^T(x, w(T, x), w_t(T, x)) dx$$

subject to

$$(5.2) \quad \begin{aligned} w_{tt} - w_{xx} &= \tilde{f}(t, x, u, w, w_t, w_x) \quad \text{in } \Omega, \\ w(0, x) &= w_0(x), \quad w_t(0, x) = w_1(x), \end{aligned}$$

$$(5.3) \quad \begin{aligned} \frac{\partial w}{\partial n}(t, X) &= \tilde{g}_X(t, u_X^B, w(t, X)), \quad X = 0, 1, \\ u(t, x) &\in U, \quad u_0^B(t) \in U_0^B, \quad u_1^B(t) \in U_1^B \quad \text{a.e.} \end{aligned}$$

where n is the exterior normal. We write (5.2) as a first-order system

$$z_t + Az_x = f(t, x, u, z)$$

with the four variables

$$z_1 \equiv w_t - w_x, \quad z_2 \equiv w_t + w_x, \quad z_3 \equiv w, \quad z_4 \equiv w$$

and $A = \text{diag}(1, -1, 1, -1)$,

$$f = \begin{pmatrix} \tilde{f}(t, x, u, z_3, \frac{1}{2}(z_2 + z_1), \frac{1}{2}(z_2 - z_1)) \\ \tilde{f}(t, s, u, z_4, \frac{1}{2}(z_2 + z_1), \frac{1}{2}(z_2 - z_1)) \\ z_2 \\ z_1 \end{pmatrix}.$$

The boundary conditions (5.3) become

$$(5.4) \quad \begin{aligned} &\left. \begin{aligned} z_1 &= z_2 + 2\tilde{g}_0(t, u_0^B, z_4) \\ z_3 &= z_4 \end{aligned} \right\} \text{at } X = 0, \\ &\left. \begin{aligned} z_2 &= z_1 + 2\tilde{g}_1(t, u_1^B, z_3) \\ z_4 &= z_3 \end{aligned} \right\} \text{at } X = 1. \end{aligned}$$

The cost functional can be written as

$$(5.5) \quad J(z) = \int_0^1 \tilde{L}^T\left(x, \frac{z_3 + z_4}{2}(T, x), \frac{z_1 + z_4}{2}(T, x)\right) dx.$$

From Theorem 3.4 we obtain the adjoint system

$$(5.6) \quad \begin{aligned} p_{1t} + p_{1x} &= -\frac{1}{2}(C_1 - C_2)(p_1 + p_2) - p_4, & p_{2t} - p_{2x} &= -\frac{1}{2}(C_1 + C_2)(p_1 + p_2) - p_3, \\ p_{3t} + p_{3x} &= -Bp_1, & p_{4t} - p_{4x} &= -Bp_2 \end{aligned}$$

where we abbreviate $B = D_w \tilde{f}$, $C_1 = D_{wt} \tilde{f}$, $C_2 = D_{wx} \tilde{f}$. One verifies directly that $\tilde{p} = p_1 + p_2$ is a weak solution of

$$(5.7) \quad \tilde{p}_{tt} - \tilde{p}_{xx} = B\tilde{p} - (C_1 \tilde{p})_t - (C_2 \tilde{p})_x,$$

which is the adjoint equation one obtains formally from the linearization of (5.2). The end conditions are

$$p_1(T, x) = p_2(T, x) = \frac{1}{2} D_w \tilde{L}, \quad p_3(T, x) = p_4(T, x) = \frac{1}{2} D_w \tilde{L},$$

which can be written in terms of \tilde{p} as

$$(5.8) \quad \tilde{p}(T, x) = D_w \tilde{L}, \quad \tilde{p}_t(T, x) = -C_1 \tilde{p}(T, x) - D_w \tilde{L}.$$

The adjoint boundary conditions are (note $|a_i| = 1$, $b'_j = 0$)

$$\left. \begin{aligned} p_2 &= p_1 \\ p_4 &= p_3 + 2D_w \tilde{g}_0 \cdot p_1 \end{aligned} \right\} \text{at } X = 0,$$

$$\left. \begin{aligned} p_1 &= p_2 \\ p_3 &= p_4 + 2D_w \tilde{g}_1 \cdot p_2 \end{aligned} \right\} \text{at } X = 1,$$

which are expressed by \tilde{p} as

$$(5.9) \quad \frac{\partial \tilde{p}}{\partial n}(t, X) = D_w \tilde{g}_X \cdot \tilde{p}(t, X) \pm C_2 \cdot \tilde{p}(t, X)$$

with the plus sign at $X = 1$ and the minus sign at $X = 0$. The maximum conditions (3.10), (3.11) become

$$(5.10) \quad \begin{aligned} \langle \tilde{p}(t, x) D_u \tilde{f}(t, x), u - u_*(t, x) \rangle &\geq 0 && \text{a.e. in } \Omega, \\ \langle \tilde{p}(t, 0) D_u \tilde{g}_0(t), u - u_0^B(t) \rangle &\geq 0 && \text{a.e. in } (0, T), \\ \langle \tilde{p}(t, 1) D_u \tilde{g}_1(t), u - u_1^B(t) \rangle &\geq 0 && \text{a.e. in } (0, T) \end{aligned}$$

for all u in U , U_0^B , U_1^B , respectively. The system (5.7)–(5.10) constitutes the differential version of the maximum principle.

We might also consider boundary conditions of the form

$$\tilde{g}(t, w) = 0.$$

These are treated by taking the time derivative and writing down conditions analogous to (5.4). However, boundary conditions of form

$$w = \tilde{g}(t, u^B)$$

fall outside the scope of Theorem 3.4. Here, for nondifferentiable boundary controls the first-order system does not have L_∞ solutions.

REFERENCES

- [1] L. CESARI, *Optimization with partial differential equations in Dieudonné-Rashevsky form and conjugate problems*, Arch. Rational Mech. Anal., 33 (1969), pp. 339–357.

- [2] A. G. BUTKOVSKY, *Distributed Control Systems*, Elsevier, Amsterdam, New York, 1969.
- [3] A. G. BUTKOVSKY, A. I. EGOROV AND K. A. LURIE, *Optimal control of distributed systems (A survey of Soviet publications)*, SIAM J. Control, 6 (1968), pp. 437-476.
- [4] J. L. LIONS, *Optimal Control of Systems Governed by Partial Differential Equations*, Springer-Verlag, Berlin, New York, 1971.
- [5] K. G. CHOO, K. L. TEO AND Z. S. WU, *On an optimal control problem involving first order hyperbolic systems with boundary controls*, Numer. Funct. Anal. Optim., 4 (1981-1982), pp. 171-190.
- [6] ———, *On an optimal control problem involving second order hyperbolic systems with boundary controls*, Bull. Austral. Math. Soc., 27 (1983), pp. 139-148.
- [7] Z. S. WU AND K. L. TEO, *A convex optimal control problem involving a class of linear hyperbolic systems*, J. Optim. Theory Appl., 39 (1983), pp. 541-560.
- [8] A. I. EGOROV, *Necessary optimality conditions for distributed parameter systems*, SIAM J. Control, 5 (1967), pp. 352-408.
- [9] M. B. SURYANARAYANA, *Necessary conditions for optimization problems with hyperbolic partial differential equations*, SIAM J. Control, 11 (1973), pp. 130-147.
- [10] V. I. PLOTNIKOV AND V. I. SUMIN, *The optimization of objects with distributed parameters described by Goursat-Darboux systems*, U.S.S.R. Comput. Math. and Math. Phys., 12 (1972), pp. 73-92.
- [11] L. BITTNER, *On optimal control of processes governed by abstract functional, integral and hyperbolic differential equations*, Math. Operationsforsch. Statist., 6 (1975), pp. 107-134.
- [12] L. V. WOLFERSDORF, *Zum Maximum-Prinzip von Pontrjagin für eine Klasse partieller Differentialgleichungssysteme 1, Ordnung*, Z. Angew. Math. Mech., 58 (1978), pp. 261-269.
- [13] N. U. AHMED AND K. L. TEO, *Optimal Control of Distributed Parameter Systems*, Elsevier, Amsterdam, New York, 1981.
- [14] L. T. ASHCHEPKOV AND O. V. VASILEV, *On the optimality of singular controls in Goursat-Darboux systems*, U.S.S.R. Comput. Math. and Math. Phys., 15 (1975), pp. 63-73.
- [15] S. V. ZUBAREV, *On an analogue of Kelly's condition for systems of hyperbolic equations*, Ukrainian Math. J., 33 (1981), pp. 1-5.
- [16] L. T. ASHCHEPKOV, O. V. VASILEV AND I. L. KOVALENOK, *Strengthened condition for the optimization of singular controls in a Goursat-Darboux system*, Differential Equations, 16 (1980), pp. 665-668.
- [17] N. U. AHMED, *Necessary conditions of optimality for a class of second-order hyperbolic systems with spatially dependent controls in the coefficients*, J. Optim. Theory Anal., 38 (1982), pp. 423-446.
- [18] L. V. PETUKHOV AND V. A. TROITSKII, *Variational optimization problems for equations of hyperbolic type*, J. Appl. Math. Mech., 36 (1972), pp. 545-555.
- [19] ———, *Some optimal problems of the theory of longitudinal vibrations of rods*, J. Appl. Math. Mech., 36 (1972), pp. 842-851.
- [20] ———, *Variational problems of optimization for equations of the hyperbolic type in the presence of boundary controls*, J. Appl. Math. Mech., 39 (1975), pp. 244-253.
- [21] L. V. PETUKHOV, *Optimal control of processes described by equations of hyperbolic type*, J. Appl. Math. Mech., 41 (1977), pp. 385-396.
- [22] S. V. ZUBAREV, *Necessary conditions for optimality for certain systems with distributed parameters*, Ukrainian Math. J., 31 (1979), pp. 344-346.
- [23] A. YA. DUBOVITSKII AND A. A. MILYUTIN, *Extremum problems in the presence of restrictions*, U.S.S.R. Comput. Math. and Math. Phys., 5 (1965), pp. 1-80.
- [24] I. V. GIRSANOV, *Lectures on Mathematical Theory of Extremum Problems*, Springer-Verlag, Berlin, New York, 1972.
- [25] A. D. IOFFE AND V. M. TICHOMIROV, *Theory of extremal problems*, North-Holland, Amsterdam, New York, 1979.
- [26] F. COLONIUS AND D. HINRICHSSEN, *Optimal control of functional differential systems*, this Journal, 16 (1978), pp. 861-879.
- [27] A. KOWALEWSKI AND W. KOTARSKI, *Application of Milutin-Dubovicki's method to solving an optimal control problem for hyperbolic systems*, Probl. Control Inform. Theory, 9 (1980), pp. 183-193.
- [28] L. V. WOLFERSDORF, *A counter example to the maximum principle of Pontryagin for a class of distributed parameter systems*, Z. Angew. Math. Mech., 60 (1980), p. 204.
- [29] F. JOHN, *Partial Differential Equations*, 3rd edition, Springer, New York, 1978.
- [30] J. ZOWE AND S. KURCYSZ, *Regularity and stability for the mathematical programming problem in Banach spaces*, Appl. Math. Optim., 5 (1979), pp. 49-62.
- [31] S. M. ROBINSON, *Regularity and stability for convex multifunctions*, Math. Oper. Res., 1 (1976), pp. 130-143.
- [32] ———, *Stability theory for systems of inequalities, Part II: differentiable nonlinear systems*, SIAM J. Numer. Anal., 13 (1976), pp. 497-513.

- [33] W. ALT, *Stabilität mengenwertiger Abbildungen mit Anwendungen auf nichtlineare Optimierungsprobleme*, Dissertation, Bayreuther Mathematische Schriften, 3, Bayreuth, West Germany, 1979.
- [34] K. O. FRIEDRICHS, *Nonlinear hyperbolic differential equations for functions of two independent variables*, Amer. J. Math., 70 (1948), pp. 555–589.
- [35] R. COURANT AND D. HILBERT, *Methods of Mathematical Physics II*, Interscience, New York, 1962.
- [36] K. YOSIDA AND E. HEWITT, *Finitely additive measures*, Trans. Amer. Math. Soc., 52 (1952), pp. 46–66.
- [37] M. B. SURYANARAYANA, *Existence theorems for optimization problems concerning linear, hyperbolic partial differential equations without convexity conditions*, J. Optim. Theory Appl., 19 (1976), pp. 47–61.
- [38] N. U. AHMED, *Properties of relaxed trajectories for a class of nonlinear evolution equations on a Banach space*, this Journal, 21 (1983), pp. 953–967.

SUFFICIENT CONDITIONS FOR LOCAL CONTROLLABILITY WITH UNBOUNDED CONTROLS*

J. BASTO GONÇALVES†

Abstract. Local controllability of analytic affine control systems Σ with an arbitrary number of controls is studied, without any restriction on the dimension of the Lie algebra T' generated by the input vector fields.

A sufficient condition for local controllability at a point is obtained, first for scalar input systems and then generalized in two different ways to the multi-input case. In all versions, this involves the computation of Lie brackets at the point under consideration and the verification of a criterion of the type considered by Sussmann (this Journal, 16 (1978), pp. 790–802).

As an application, controllability of systems defined on \mathbb{R}^n and with constant input vector fields is studied.

Key words. nonlinear analytic systems, affine systems, local controllability, Lie brackets

AMS(MOS) subject classifications. 93B05, 93C10, 93C15, 49E15, 34H05

Introduction. This article studies the local controllability of affine control systems Σ with dynamics $\dot{x} = f(x, u) = X(x) + \sum_{i=1}^s u_i X^i(x)$ on the n -dimensional analytic manifold M . The codimension of Σ is the codimension of the Lie algebra T' generated by the input vector fields at every point in M .

Local controllability of nonlinear systems is one of the basic problems of the geometric theory of control systems, but besides the importance it has on its own, in [3], [8], [13]–[15] global controllability has been shown to depend on the existence of points where the system is locally controllable in some sense,

There are not many easy methods of verifying local controllability, besides the linear test, and all sources with the exception of [19] assume Σ to be a scalar input or a codimension one system; the conditions presented in this paper attempt to provide criteria as simple to use as that test and valid for any codimension.

It would be nicer to be able to use bounded controls, but the lack of criteria for local controllability and specially for global controllability in that situation fully justifies the approach followed here.

The methods now presented are a development of the ideas introduced by Hermes [9]–[12], Sussmann [19], [20] and Hunt [13]–[15] and already studied by Bacciotti and Stefani [3] and the author [5], [6]; in particular, the sufficient conditions for local controllability proved here are natural generalizations to arbitrary codimension of the computational sufficient condition presented in [5]. For recent and related work see [17] and [21].

Basically the method works like this: around p we can choose coordinates on a neighbourhood U so that U looks like a product of open sets $S_1 \times S_2$ in $\mathbb{R}^{n-k} \times \mathbb{R}^k$ with S_1 the integral submanifold of T' through p .

As controls appear linearly and are unbounded, the trajectories of the input vector fields X^i will be assumed to be admissible trajectories of Σ [13]–[15]. We can move in S_1 as we please, and therefore if local controllability at p along directions tangent to S_2 is proved, we have local controllability of Σ at p .

* Received by the editors November 25, 1985; accepted for publication September 1, 1986. This work was supported by the Instituto Nacional de Investigação Científica and the Calouste Gulbenkian Foundation.

† Grupo de Matemática Aplicada, Faculdade de Ciências da Universidade do Porto, 4000, Porto, Portugal.

In the above coordinates and if $X(x_1, x_2) = (X_1(x_1, x_2), X_2(x_1, x_2))$, we can consider in S_2 a family of vector fields with parameter x_1 given by $X_2(x_1, x_2)$; considering the movement in S_2 , we have to prove that family to be locally controllable (to be rigorous the correct family is not that one, but it takes the same values at p , and the criterion we are going to use involves just that.)

It is well known that given a family of vector fields a sufficient condition for local controllability at a point p is that 0 belongs to the interior of the convex hull of the values of the vector fields at that point.

If $k = 1$ this last condition applied to the family above is fulfilled if arbitrarily near p we have X pointing to opposite sides of S_1 ; this condition already appears in one form or another in [3], [5], [6], [10].

To obtain a sufficient condition when $k > 1$ we consider the trajectory of X through p and project it in S_1 parallel to S_2 ; Let X' be a vector field in M belonging to T' such that its trajectory through p is that projection. X' is not uniquely defined, but that does not alter the situation.

Let us consider the vectors in $T_p M$ defined by $v_i = (\text{ad}^i X', X)(p)$ and let W_i be the projection onto $T_p S_2$ parallel to $T_p S_1$ of the subspace generated by v_j with $j = 1, \dots, i$.

Assume that $X(p) = X'(p) \neq 0$ and (i) $W = \bigcup W_i = T_p S_2$; (ii) if i is even $W_i = W_{i-1}$; then the system Σ is locally controllable at p .

In fact, if β is the trajectory of X' through p , the components of $X(\beta(t))$ along S_2 can be written in a suitable basis as $a(t) = a_1(t)e_{n-k+1} + \dots + a_n(t)e_n$, where the least order term in t of each $a_i(t)$ is odd, using the Campbell-Baker-Hausdorff formula [9]; it is quite easy to verify that 0 belongs to the interior of the convex hull of the set $\{a(t), |t| \text{ small and nonzero}\}$.

For scalar input systems ($k = n - 1$) we obtain a better condition, with the input vector field substituted for X' , generalizing a condition already obtained for systems defined on a plane ($n = 2$) in [16].

For general systems we also obtain a sufficient condition involving only the input vector fields, an extension to $n > 2$ of the previous one.

Applying those sufficient conditions to systems defined on \mathbb{R}^n with constant input vector fields, we obtain sufficient conditions for their local controllability at a point; if the codimension of the systems is one, we have then a sufficient condition for global controllability which reduces to Theorem 2 in [2] for systems in the plane.

1. Basic results and definitions. We shall study the local controllability of affine control systems Σ with dynamics $\dot{x} = f(x, u) = X(x) + \sum_{i=1}^s u_i X^i(x)$, defined on an n -dimensional connected analytic manifold M . The system Σ is said to have codimension k if the Lie algebra T' generated by the input vector fields X^1, \dots, X^s has codimension k at every point in M .

Denote by T the Lie algebra generated by X, X^1, \dots, X^s and let T_0 be the Lie algebra containing $(\text{ad}^k X, X^i)$ with $k \geq 0, i = 1, \dots, s$ and $(\text{ad}^j X, Y)$ defined by Y if $j = 0$ and inductively by $(\text{ad}^j X, Y) = [X, (\text{ad}^{j-1} X, Y)]$ if $j > 0$.

The definition of local controllability adopted here will be the following: Σ is said to be locally controllable at $p \in M$ if for every neighbourhood U of p the set $A(p, U)$ of the points attainable from p in positive time without leaving U contains p in its interior.

$A(p, U)$ is defined as the set of points p such that there exists a piecewise C^1 continuous map $c: [0, T] \rightarrow U$ verifying

- (i) $c(0) = p, c(T) = p'$ with $T \in \mathbb{R}^+$,

(ii) There exists t_0, t_1, \dots, t_m such that $0 = t_0 < t_1 < \dots < t_m = T$ and c in $]t_{i-1}, t_i[$ is an integral curve γ_i of $f(\cdot, u_i)$ for some $u_i \in \mathbb{R}^s$.

If the map $(\tau_1, \dots, \tau_m) \rightarrow x = \gamma_m(\tau_m) \circ \dots \circ \gamma_1(\tau_1)p$ has rank n at the point (t_1, \dots, t_m) , the corresponding point p' is said to be normally reachable from p [18].

The set of attainability from p , denoted by $A(p)$, is defined by $A(p) = A(p, M)$, and the set $A(p, T, U)$ of attainability from p in positive time T without leaving U is defined as $A(p, U)$, but with fixed T .

Σ is locally accessible at p if $A(p, U)$ has a nonempty interior for any neighbourhood U of p ; it is obviously a necessary condition for local controllability at p . A necessary and sufficient condition for local accessibility in the analytic case is that the Lie algebra \mathbf{T} has dimension n at p [22].

Σ is strongly accessible at p if $A(p, T) = A(p, T, M)$ with positive T has a nonempty interior; in the analytic case a necessary and sufficient condition is that the Lie algebra \mathbf{T}_0 has dimension n at p [22].

Denote by L_p the leaf of \mathbf{T}' containing p , and by $X_t(p)$ the trajectory of X through p at $t = 0$; strong accessibility at p is a necessary condition for local controllability of Σ at p , as well as $X(p) \in \mathbf{T}'(p)$ [5].

The constant codimension assumption is important in the last statement, as the following example shows.

Example (Bacciotti [4]). $M = \mathbb{R}^2$, $s = 1$, $p = 0$; $X = e_1$, $X^1 = x_2 e_1 + x_1 e_2$, $X^2 = x_1 e_1 + x_2 e_2$. It is easy to see that \mathbf{T}' is two-dimensional everywhere except at p , where the dimension is zero, and that we have local controllability at all points of M including $p = 0$.

On the other hand, $X(0) \neq 0$ and therefore the conclusions of Theorem 1 and Proposition 2 of [5] are not verified.

In the proof of Theorem 1 in [5] and when $X(p)$ is not tangent to L_p we need to construct an $(n-1)$ -dimensional submanifold N , containing $L_p \cap V$ and such that $X(p) \notin T_p N$. The important point missed there is that N has to be foliated by the leaves of \mathbf{T}' ; we describe briefly the alterations needed to correct the situation.

If \mathbf{T}' has constant dimension, we can use the Frobenius theorem to construct N and then $\dim \mathbf{T}' = n$ on U ; otherwise we can still use that theorem around points of maximal local dimension.

THEOREM 1.1. *If Σ is locally controllable at every point of an open set $U \subset M$, then $\dim \mathbf{T}' = n$ on an open dense set U' in U .*

Proof. In view of the above remarks it is enough to show that the set U' of the points in U where \mathbf{T}' has maximal dimension is open and dense in U . That U' is open is obvious.

Assume $U - U'$ has a nonempty interior A and choose a point p' in it where \mathbf{T}' has maximal dimension when restricted to A . We can now apply the Frobenius theorem in a neighbourhood of p' inside A to construct the submanifold N already mentioned; therefore Σ cannot be locally controllable at p' . \square

Now we have, as in [5], the following proposition.

PROPOSITION 1.2. *If Σ is locally controllable at p and \mathbf{T}' has constant dimension or has maximal local dimension at p , then $X(p)$ belongs to $\mathbf{T}'(p)$.*

The other results in [5] are not affected by these changes; note that if the system is not analytic we have to consider the smallest integrable distribution containing the associated vector fields, instead of \mathbf{T}' .

The definitions above can also be made for families D of vector fields, instead of control systems; we just require that c is an integral curve of some vector field belonging to D in (ii) when defining the set $A(p, U)$.

Of particular interest are the families $D_{\Sigma} = \{X, X^1, \dots, X^s, -X^1, \dots, -X^s\}$ and $D_{\Sigma^-} = \{-X, X^1, \dots, X^s, -X^1, \dots, -X^s\}$ as the next lemma shows.

LEMMA 1.3. *If D_{Σ} and D_{Σ^-} are locally controllable, so is Σ .*

Proof. Local controllability of Σ follows from the local controllability of the family of vector fields $D' = \{X, X + uX^1, \dots, X + uX^s, X - uX^1, \dots, X - uX^s, u \in \mathbb{R}^+\}$ or equivalently of the family $D'' = \{X, \varepsilon X + X^1, \dots, \varepsilon X + X^s, \varepsilon X - X^1, \dots, \varepsilon X - X^s, \varepsilon \in \mathbb{R}^+\}$.

From the local controllability of D_{Σ^-} it follows that there exists a neighbourhood V of p such that $V \subset A(p, U; D_{\Sigma^-})$, therefore $p \in A(p', U; D_{\Sigma})$ for every $p' \in V$.

As D_{Σ} has to be locally accessible, its restriction to V has the normal reachability property from p [18], i.e., some y in $A(p, V; D_{\Sigma})$ is normally reachable from p . Since p is attainable from y , it follows that p (and every point in $A(p, U; D_{\Sigma})$) is normally reachable from p for the family D_{Σ} .

From the stability of normal reachability for small perturbations [7, Prop. 4.2] we can say that there exists a neighbourhood W of p attainable from p without leaving U using vector fields in D'' corresponding to small values of ε ; this means the use of large values for the control, but as there are no a priori bounds, we can say that Σ is locally controllable. \square

The criteria of local controllability presented in this paper are independent of the sign of X ; if they are verified for a system Σ with dynamics $\dot{x} = f(x, u) = X(x) + \sum_{i=1}^s u_i X^i(x)$ they are also verified for the system Σ^- with dynamics $\dot{x} = -f(x, u) = -X(x) - \sum_{i=1}^s u_i X^i(x)$. Thus if they are proved for the D_{Σ} they are automatically true for Σ , in view of the previous lemma; this justifies that the trajectories of the input vector fields X^i will be assumed to be admissible trajectories of Σ .

2. Scalar input systems. Let Σ be a scalar input analytic affine system, of the form $\dot{x} = X(x) + uX^1(x)$; we are going to study the local controllability of Σ around a point p , where the following conditions are verified: Σ is strongly accessible at p , $X(p) \in T'(p)$ and moreover $X^1(p) \neq 0$. With these assumptions S_1 is a one-dimensional submanifold of M .

We denote by $W_i(x)$ the subspace of $T_x M$ generated by the vectors $(\text{ad}^j X^1, X(x))$ with $j = 1, \dots, i$, and by W_{∞} the union of the W_i for all i .

THEOREM 2.1. *If $W_i(p) = W_{i-1}(p)$ for i even and $T_p M = W_{\infty}(p) + T_p S_1$ then the system Σ is locally controllable at p .*

Proof. First note that we can assume $X(p) \neq 0$ without any loss of generality, since otherwise we can substitute any other associated vector field for X without changing the hypothesis; also, in convenient local coordinates we can have $p = 0$ and $X^1 = e_1$.

The system being analytic and strongly accessible, X and X^1 are linearly independent at every point $x(t) = te_1$ with $0 < |t| < \varepsilon$ if $\varepsilon \in \mathbb{R}^+$ is sufficiently small, and by continuity we have $T_{x(t)} M = W_{\infty}(x(t)) + T_{x(t)} S_1$ for small $|t|$.

We can then apply Proposition 4.2 in [9] to obtain the following lemma.

LEMMA 2.2. *For each $|t| \neq 0$ sufficiently small, there exist $n - 2$ analytic vector fields $Y^i (i = 1, 2, \dots, n - 2)$, commuting between themselves and with X , and such that $X(x(t)), X^1(x(t)), Y^1(x(t)), \dots, Y^{n-2}(x(t))$ are linearly independent.*

Therefore there exists an integral submanifold N_t of the distribution generated by the vector fields Y^i and X containing $x(t)$; moreover $X^1(x(t))$ is not tangent to it.

Around $x(t)$ the projection $\pi: (x_1, x_2, \dots, x_n) \rightarrow (0, x_2, \dots, x_n)$ is a diffeomorphism when restricted to N_t ; that projection takes the restriction of X to N_t into a vector field Z' defined on a neighbourhood of the origin in $\{x_1 = 0\}$, and $Z'(0) = \underline{X}(x(t))$, where \underline{X} is obtained from X by substituting 0 for the first component.

We have thus defined on a neighbourhood of the origin in $\{x_1 = 0\}$ a family Ψ of vector fields Z^t , with $|t| \neq 0$ small. We claim that if that family is locally controllable at the origin, so is Σ .

As previously noted, we can assume the trajectories of X^1 and also of $-X^1$ to be admissible trajectories of Σ , therefore if (x_1, x_2, \dots, x_n) is reachable from the origin so is the set $(]x_1 - \delta, x_1 + \delta[, x_2, \dots, x_n)$ for $|\delta|$ small.

Thus, proving that if a point in $\{x_1 = 0\}$ is attainable from the origin for the family Ψ , it is also attainable for Σ , is equivalent to proving the above claim.

Suppose $y = (0, x_2, \dots, x_n)$ is attainable for Ψ using only one vector field Z^t ; to reach that point following trajectories of Σ we just have to go from the origin to $x(t)$ along X^1 , follow X from there until the point in N_t which projects on y , and then follow $-X^1$ until y .

It is easy to see that the situation is not altered if there are several vector fields in Ψ involved, and we can be careful and not leave any fixed neighbourhood of the origin in M , therefore our claim is proved.

It remains to be shown that Ψ is locally controllable; using the criterion in [19] concerning the values of the vector fields at the point under consideration, it is enough to show that 0 belongs to the interior in $\{x_1 = 0\}$ of the convex hull of the values of $Z_t(0) = \underline{X}(x(t))$ for $|t| \neq 0$ and small.

The assumption $W_i(p) = W_{i-1}(p)$ for i even means that $v_i = (\text{ad}^i X^1, X)(p)$ can be written as a linear combination of v_j with j odd and smaller than i ; this is also true for the corresponding projections \underline{v}_i and \underline{v}_j .

Using the Campbell-Baker-Hausdorff formula [9] we can write $X(x(t)) = X(0) - tv_1 + t^2/2!v_2 - \dots$, for small $|t|$; projecting on $\{0\} \times \mathbb{R}^{n-1}$ and bearing in mind the remark above, we obtain $Z^t = \underline{X}(x(t)) = a_1(t)\underline{v}_1 + \dots + a_j(t)\underline{v}_j + \dots + a_m(t)\underline{v}_m$, where j is odd, m is the lowest j for which $W_j(p) = W_\infty(p)$.

In the power series $a_j(t)$ the lowest degree in t is of course odd, and we have either $a_j(t) = -t^j/j! + o(t^j)$ or $a_j(t) = o(t^j)$; define $b_j(t)$ as $-t^j/j!$ or 0 so that $a_j(t) = b_j(t) + o(t^j)$, and let $z(t) = b_1(t)\underline{v}_1 + \dots + b_j(t)\underline{v}_j + \dots + b_m(t)\underline{v}_m$.

It is now trivial to verify that if we take enough values of t in the interval $]0, \varepsilon[$ with ε positive and as small as we want, we obtain $n-1$ linearly independent vectors z_1, \dots, z_{n-1} among the corresponding $z(t)$, and we can get their symmetric vectors by taking the symmetric times; it follows that 0 belongs to the interior in $\{0\} \times \mathbb{R}^{n-1}$ of the convex hull of

$$\{z_1, \dots, z_{n-1}, z_n = -z_1, \dots, z_{2n-2} = -z_{n-1}\}.$$

There exists then a positive δ such that if $|y_i - z_i| < \delta$ the origin belongs to the interior in $\{0\} \times \mathbb{R}^{n-1}$ of the convex hull of $\{y_1, \dots, y_{2n-2}\}$; if we take ε small enough, we have $|Z^t - z(t)| < \delta$ if $|t| < \varepsilon$, and thus the interior of the convex hull of $\{Z^t, |t| \neq 0 \text{ and small}\}$ contains the origin.

Therefore the family Ψ is locally controllable at the origin and so is Σ . \square

The following example is due to Jakubczyk and was presented by Sussmann in [20].

Example. Let $M = \mathbb{R}^3$, and $X(x, y, z) = (0, x, x^3 + y^2)$, $X^1(x, y, z) = e_1$; the local controllability of this system, proved in [20], cannot be established with the linear test nor using the sufficient condition presented there; the condition of Theorem 2.1 is very easy to apply and gives a positive answer:

$$v_1 = (\text{ad}^1 X^1, X)(x) = (0, 1, 3x^2),$$

$$v_2 = (\text{ad}^2 X^1, X)(x) = (0, 0, 6x),$$

$$v_3 = (\text{ad}^3 X^1, X)(x) = (0, 0, 6).$$

At $(x, y, z) = 0, v_1 = e_2, v_2 = 0, v_3 = 6e_3$, and therefore $W_2 = W_1 = \{0\} \times \mathbb{R} \times \{0\}$, $W_\infty = W_3 = \{0\} \times \mathbb{R} \times \mathbb{R}$, and $T_0 S_1 = \mathbb{R} \times \{0\} \times \{0\}$; we have $\mathbb{R}^3 = T_0 S_1 + W_\infty$, and the system is locally controllable at the origin.

3. Multi-input systems. The system Σ will now be supposed to have codimension k and $s > 1$, and we assume the necessary conditions $\dim T_0(p) = n$ and $X(p) \in T'(p)$ to be verified.

For multi-input systems the first version of the sufficient condition involves a vector field Y that will play the role X^1 has played for scalar input systems. The construction of Y has already been explained in [5], and goes as follows: locally, around p , M looks like $\mathbb{R}^{n-k} \times \mathbb{R}^k$, where we can take p as the origin and $X(0) = e_1$, with the leaves of T' defined by $(x_{n-k+1}, \dots, x_n) = a \in \mathbb{R}^k$; let $\alpha(t)$ be the trajectory of X through p , and $\beta(t)$ its projection on the leaf S_1 of T' containing the origin, $\beta(t) = (\alpha_1(t), \dots, \alpha_{n-k}(t), 0, \dots, 0)$.

Now we can define a vector field Y belonging to T' such that its trajectory through $p = 0$ is β : $Y(x_1, \dots, x_n) = \underline{X}(\alpha(t))$, where \underline{X} is the projection of X on the first $n-k$ coordinates, and $x_1 = \alpha_1(t)$; note that t can be defined from $x_1 = \alpha_1(t)$ near the origin because $X(0) = e_1$.

We denote by $W'_i(x)$ the subspace of $T_x M$ generated by the vectors $(\text{ad}^j Y, X)(x)$ with $j = 1, \dots, i$, and by W'_∞ the union of the W'_i for all i .

We can now state a sufficient condition for local controllability at p , similar to Theorem 2.1.

THEOREM 3.1. *If $W'_i(p) = W'_{i-1}(p)$ for i even and $T_p M = W'_\infty(p) + T_p S_1$ then the system Σ is locally controllable at p .*

Proof. The proof is the repetition, with obvious adaptations, of the proof of Theorem 2.1; we just have to substitute Y for X^1 , except that the rank condition that allows the use of Proposition 4.2 in [9] to obtain Lemma 2.2 is not valid in this case.

We have to construct N_t , a k -dimensional submanifold transverse to $\mathbb{R}^{n-k} \times \{0\}$ and passing through $\beta(t)$, such that X is tangent to it at every point. If the last component of $X(\beta(t))$ is nonzero, for instance, let H be the plane defined by the $k-1$ vectors $e_{n-k+1}, \dots, e_{n-1}$ containing $\beta(t)$; near this point, let us say on a neighbourhood H' of $\beta(t)$ in H , X is transverse to H , and therefore $N_t = \{X_t(H'), |t| \text{ small}\}$ is the required submanifold. \square

For $k = 1$ this theorem reduces to the computational sufficient condition proved in [5] for codimension one control systems.

The construction of the vector field Y can be quite difficult in particular examples, and it would be much nicer to have a condition just in terms of the associated vector fields of the system Σ , as in Theorem 2.1.

We denote by ${}_r W_i(x)$ the subspace of $T_x M$ generated by the vectors $(\text{ad}^j X^r, X)(x)$ with $j = 1, \dots, i$, and $r = 1, \dots, s$, by ${}_r W_\infty$ the union of the ${}_r W_i$ for all i and by W_∞ the subspace spanned by the union of the ${}_r W_\infty$ for all r .

We can state a new sufficient condition for local controllability at p .

THEOREM 3.2. *If ${}_r W_i(p) = {}_r W_{i-1}(p)$ for i even and any r , and $T_p M = W_\infty(p) + T_p S_1$, then the system Σ is locally controllable at p .*

Proof. Reasoning according to the proof of Theorem 2.1, we can construct families Ψ_r , symmetric at p and spanning ${}_r W_\infty(p)$; if we define Ψ as the family of all vector fields belonging to some Ψ_r , we see that 0 belongs to the interior (in the complement of S_1) of the convex hull of the values of Ψ at p ; as in the proof of Theorem 2.1, the local controllability of Ψ implies that of Σ , and we can then conclude that Σ is locally controllable at p . \square

Remarks. (i) It is clear from the proof that we do not need ${}_rW_i(p) = {}_rW_{i-1}(p)$ for i even to be true for every r , if the subspace W spanned by the union of the ${}_rW_\infty$ for those r for which that property is verified is such that $T_pM = W + T_pS_1$.

(ii) It is not necessary to have ${}_rW_i(p) = {}_rW_{i-1}(p)$ if this is true for their projections on a complement of T_pS_1 .

As an example of the usefulness of this condition, we are going to consider systems with $M = \mathbb{R}^n$, $X^i = e_i$ ($i = 1, 2, \dots, s = n - k$); their local and global controllability has already been studied in [2].

A very easy computation shows that

$$(\text{ad}^j X^r, X)(x) = (\partial^j / \partial x_r^j \varphi_1, \dots, \partial^j / \partial x_r^j \varphi_n)$$

where φ_i is the i th component of X .

Therefore Σ verifies ${}_rW_i(p) = {}_rW_{i-1}(p)$ for i even and any r , if all functions φ_j have an odd (or infinite) order zero with respect to x_r at p (i.e. $\varphi_j(p) = 0$ and the first nonzero derivative at p with respect to x_r is odd) for every $r = 1, \dots, s = n - k$. In fact we have only to consider $j \geq n - k + 1$ as noted in Remark (ii).

Let $d(i, r)$ be the order of the zero p of φ_i with respect to x_r , the smallest $j \geq 0$ for which $\partial^j / \partial x_r^j \varphi_i$ is nonzero at p , with $r = 1, \dots, s = n - k$ and $i = n - k + 1, \dots, n$; the dimension of the projection of ${}_rW_\infty$ is at least the number of functions φ_i , $i \geq n - k + 1$, having different order at p with respect to x_r , the number of different $d(i, r)$.

Denote by ${}^{\#}d$ the maximum number of different $d(i, r)$ corresponding to different values of i (i.e. ${}^{\#}d$ is the maximum of the cardinals of all possible sets A such that the elements of A are numbers $d(i, r)$, all different and corresponding to the different i). It is clear that if ${}^{\#}d \geq k$ then $T_pM = W_\infty(p) + T_pS_1$ is verified.

This completes the proof of Proposition 3.3 below.

PROPOSITION 3.3. Σ is locally controllable at p if every $d(i, r)$ is odd and ${}^{\#}d \geq k$.

A slightly more general statement can be obtained if in the definition of ${}^{\#}d$ we only consider those r for which $d(i, r)$ is odd irrespective of i , taking in account the remark following Theorem 3.2; ${}^{\#}d \geq k$ is then a sufficient condition for local controllability.

Theorem 2 in [2] is a consequence of this proposition for $k = 1$ and $n = 2$, noting that if the hypothesis above is verified at one point (at least) of every leaf of T' , then global controllability follows from Theorem 4.2 in [3].

With $k = 1$ and for any dimension n , if $d(n, r)$ is odd for some r on at least a point in every leaf of T' (in this case $\{x_n = \text{const.}\}$), one has global controllability, noting that in this case ${}^{\#}d \geq 1$.

PROPOSITION 3.4. Let Σ be a codimension one system; if for any x_n the function φ_n has an odd order zero with respect to some x_i ($i = 1, \dots, s = n - 1$) then Σ is globally controllable.

Proof. It follows from Theorem 3.2 that Σ is locally controllable at the zero where the hypothesis is verified; since to each x_n corresponds a leaf of T' , to prove global controllability is enough to invoke Theorem 4.2 of [3]. \square

For a more general treatment of local and global controllability of codimension one systems, see [6].

Acknowledgments. Andrea Bacciotti pointed out the mistake in [5] and gave me the example used in this paper; he looked through the correction as well, and I am very thankful for his interest and cooperation.

Thanks are also due to an anonymous referee for many suggestions to improve the rigour of the presentation, especially the treatment of the trajectories of the input vector fields as admissible trajectories.

REFERENCES

- [1] D. AYELS, *Global controllability for smooth nonlinear systems; a geometric approach*, this Journal, 23 (1985), pp. 452-465.
- [2] ———, *Local and global controllability for nonlinear systems*, Systems Control Lett., 5 (1984), pp. 19-26.
- [3] A. BACCIOTTI AND G. STEFANI, *On the relationship between global and local controllability*, Math. Systems Theory, 16 (1983), pp. 79-91.
- [4] A. BACCIOTTI, Private communication, 1986.
- [5] J. BASTO GONÇALVES, *Local controllability of nonlinear systems*, Systems & Control Lett., 6 (1985), pp. 213-217.
- [6] ———, *Controllability in codimension one*, J. Differential Equations, to appear.
- [7] K. GRASSE, *On accessibility and normal accessibility: the openness of controllability in the fine C^0 controllability*, J. Differential Equations, 53 (1984), pp. 387-414.
- [8] ———, *A condition equivalent to global controllability in systems of vector fields*, J. Differential Equations, 56 (1985), pp. 263-269.
- [9] H. HERMES, *Local controllability and sufficient conditions in singular problems*, J. Differential Equations, 20 (1976), pp. 213-232.
- [10] ———, *Controlled stability*, Ann. Mat. Pura Appl. (4), 114 (1977), pp. 103-119.
- [11] ———, *On local controllability*, this Journal, 20 (1982), pp. 211-220.
- [12] ———, *Control systems which generate decomposable Lie algebras*, J. Differential Equations, 44 (1982), pp. 166-187.
- [13] R. HUNT, *Controllability of general nonlinear systems*, Math. Systems Theory, 12 (1979), pp. 361-370.
- [14] ———, *Global controllability of nonlinear systems in two dimensions*, Math. Systems Theory, 13 (1980), pp. 361-376.
- [15] ———, *n -Dimensional controllability with $n-1$ controls*, IEEE Trans. Automat. Control, AC-27 (1982), pp. 113-117.
- [16] R. STANGELAND, *Some aspects of systems and control*, M.Sc. thesis, Texas Technological University, Lubbock, TX.
- [17] G. STEFANI, *Polynomial approximations to control systems and local controllability*, Proc. 24th IEEE Conference on Decision and Control, 1985, pp. 33-38.
- [18] H. SUSSMANN, *Some properties of vector field systems that are not altered by small perturbations*, J. Differential Equations, 20 (1976), pp. 292-315.
- [19] ———, *A sufficient condition for local controllability*, this Journal, 16 (1978), pp. 790-802.
- [20] ———, *Lie brackets and local controllability: a sufficient condition for scalar input systems*, this Journal, 21 (1983), pp. 686-713.
- [21] ———, *A general theorem on symmetries and local controllability*, Proc. 24th IEEE Conference on Decision and Control, 1985, pp. 27-32.
- [22] H. SUSSMANN AND V. JURDJEVIC, *Controllability of nonlinear systems*, J. Differential Equations, 12 (1972), pp. 95-116.

LEGENDRE-TAU APPROXIMATION FOR FUNCTIONAL DIFFERENTIAL EQUATIONS PART II: THE LINEAR QUADRATIC OPTIMAL CONTROL PROBLEM*

KAZUFUMI ITO[†] AND RUSSELL TEGLAS[‡]

Abstract. A numerical scheme based on the Legendre-tau approximation is proposed to approximate the feedback solution to the linear quadratic optimal control problem for hereditary differential systems. The convergence property is established using the Trotter-Kato theorem. The method yields very good approximations at low orders and provides an approximation technique for computing closed-loop eigenvalues of the feedback system. A comparison with existing methods (based on "averaging" and "spline" approximations) is made.

Key words. Legendre-tau approximations, hereditary differential equations, linear quadratic regulator problem

AMS(MOS) subject classifications. 34K35, 65N35, 93C20

1. Introduction. This paper is the continuation of the study [9] on the use of Legendre-tau approximation for functional differential equations (FDE) and concerns the problem of constructing feedback solutions to linear quadratic regulator problems for hereditary systems. This problem has received a rather extensive study and we refer to [15], [2] and [4] for the summary of the earlier contributions. Our approach is based upon the pioneering work of Banks and Burns [2] who clarified the idea of approximating FDE by systems of finite-dimensional ordinary differential equations and applied it to optimal control problems, i.e., the convergence of a particular numerical scheme (so called "averaging" approximation) is established, using the Trotter-Kato Theorem of linear semigroups. Recently, Gibson [8] has developed the approximation theory for the Riccati equations associated with a hereditary system and applied it to the averaging approximation scheme.

The purposes of this paper are (i) to apply the basic idea developed in [9] to the linear quadratic regulator problem, (ii) to prove convergence of numerical approximations of the feedback control laws, and (iii) to demonstrate the feasibility of our numerical schemes.

For the multiple point delay case, the solution to the algebraic Riccati equation (ARE) has jump discontinuities as shown in [8]. With this consideration, an extended version of the scheme described in [9] is developed for such a case in § 3.

It has been shown in [8] that if a sequence $S^N(t)$ of approximating semigroups converges to the solution semigroup $S(t)$, the solution Π^N to the algebraic Riccati equation corresponding to the N th approximation is uniformly bounded, and the solution Π to ARE is unique, then Π^N converges weakly to Π . This implies the strong

* Received by the editors May 20, 1985; accepted for publication (in revised form) September 28, 1986. This research was supported by the National Aeronautics and Space Administration under contracts NAS1-17070 and NAS1-17130 while the first author was in residence at The Institute for Computer Applications in Science and Engineering, NASA Langley Research Center, Hampton, Virginia 23665.

[†] Institute for Computer Applications in Science and Engineering, NASA Langley Research Center, Hampton, Virginia 23665. Present address, Division of Applied Mathematics, Brown University, Providence, Rhode Island 02192.

[‡] University of Vermont, Burlington, Vermont 05405.

convergence of approximating optimal feedback gains. In § 3, we show the strong convergence of $S^N(t)$ to $S(t)$ using Trotter–Kato theorem. In Lemma 5.2 we show the uniform boundedness of $\|\Pi^N\|$ for certain special cases.

Moreover, Gibson has shown that if (i) the exponential stability is preserved under approximation, and (ii) the adjoint of approximate semigroups converges strongly to the adjoint semigroup, then Π^N converges strongly to Π . For the single point delay case, (i) and (ii) hold for the Legendre-tau approximation (see our paper [10, Part III]), and the numerical computations for several examples indicate the strong convergence of Π^N . Moreover, we show a rather interesting result in Theorem 5.1. It says that if a sequence of approximate solutions to ARE converges weakly to the solution to ARE, then the closed loop system which results from the approximate feedback control law is exponentially stable for sufficiently large orders of approximation.

As will be discussed in § 6, the tau method may offer considerable improvements over other methods (e.g. those discussed in [4], [8]) and it gives a good approximation to the closed loop eigenvalue.

The following is a brief summary of the contents of this paper. In § 2 we review the equivalence results between FDE and abstract Cauchy problems on the product space $\mathbb{R}^n \times L_2$ and results on the regulator problem for hereditary differential systems. In § 3 we introduce the numerical scheme based on the Legendre-tau approximation for the multiple point delay case and discuss the basic convergence of approximate semigroups using the Trotter–Kato theorem. In § 4 we show how one can use the numerical scheme described in § 3 to obtain the feedback solutions. In § 5 we state the basic convergence property of approximate solutions to ARE. Finally, in § 6 we present numerical results and compare these results with those obtained by other methods [4], [8].

Throughout this paper the following notation will be used. $r > 0$ stands for the largest delay time appearing in the FDE. The Hilbert space of \mathbb{R}^n -valued square integrable functions on the interval $[a, b]$ is denoted by $L_2([a, b]; \mathbb{R}^n)$. When the underlying space and interval can be understood from the context, we will abbreviate the notation and simply write L_2 , $L_2^{\text{loc}}([0, \infty); \mathbb{R}^n)$, or L_2^{loc} , is the space of \mathbb{R}^n -valued locally square integrable functions on the semi-infinite interval $[0, \infty)$. H^k is the Sobolev space of \mathbb{R}^n -valued functions f on a compact interval with $f^{(k-1)}$ absolutely continuous and $f^{(k)} \in L_2$. We denote by Z the product space $\mathbb{R}^n \times L_2([-r, 0]; \mathbb{R}^n)$. Given an element $z \in Z$, $\eta \in \mathbb{R}^n$ and $\phi \in L_2$ denote the two coordinates of z : $z = (\eta, \phi)$. The bracket $\langle \cdot, \cdot \rangle_H$ stands for the inner product in the Hilbert space H and the subscript for the underlying Hilbert space will be omitted when understood from the context. $\|\cdot\|$ denotes the norm for elements of a Banach space and for operators between Banach spaces, while $|\cdot|$ denotes the Euclidean norm in \mathbb{R}^n .

If X and Y are Banach spaces, then the space of bounded operators from X to Y is denoted by $\mathcal{L}(X, Y)$. $\mathcal{D}(\mathcal{A})$ denotes the domain of a linear operator \mathcal{A} . χ_I denotes the characteristic function of the interval I . Finally, for any function ϕ of independent variable θ , we shall use ϕ or $(\partial/\partial\theta)\phi$ to denote the derivative of ϕ with respect to θ .

2. Riccati equations. In this section, we state the type of problems to be considered and recall some results on the linear quadratic regulator problem for hereditary differential systems.

Given $(\eta, \phi) \in Z$ and $u \in L_2^{\text{loc}}([0, \infty), \mathbb{R}^n)$, we consider the initial value problem

$$(2.1) \quad \frac{d}{dt}x(t) = \int_{-r}^0 d\mu(\theta)x(t+\theta) + Bu(t), \quad x(0) = \eta, \quad x(\theta) = \phi(\theta), \quad \theta \in [-r, 0)$$

where μ is a matrix-valued function of bounded variation on $[-r, 0]$ with the form

$$(2.2) \quad \mu(\theta) = \sum_{i=0}^l A_i \chi_{(-\theta_i, 0]}(\theta) + \int_{-r}^{\theta} A(s) ds$$

with $0 = \theta_0 < \theta_1 < \dots < \theta_l = r$. A_i and $A(\cdot)$ are $n \times n$ matrices, the elements of the latter being square integrable on $[-r, 0]$. Alternatively, for $t \geq 0$

$$\int_{-r}^0 d\mu(\theta)x(t+\theta) = \sum_{i=0}^l A_i x(t-\theta_i) + \int_{-r}^0 A(\theta)x(t+\theta) d\theta.$$

It is well known [2], [5], [6] that for $(\eta, \phi) \in Z$ and $u \in L_2^{\text{loc}}$, (2.1) admits a unique solution $x \in L_2([-r, T]; \mathbb{R}^n) \cap H^1([0, T]; \mathbb{R}^n)$ for any $T \geq 0$, and that (2.1) can be formulated as an evolution equation on Z

$$(2.3) \quad \frac{d}{dt}z(t) = \mathcal{A}z(t) + \mathcal{B}u(t), \quad t \geq 0$$

where $z(t) = (x(t), x(t+\cdot)) \in Z$, $t \geq 0$ and $\mathcal{B}u = (Bu, 0) \in Z$ for $u \in \mathbb{R}^m$. The infinitesimal generator \mathcal{A} is defined by

$$(2.4) \quad \mathcal{D}(\mathcal{A}) = \{(\eta, \phi) \in Z \mid \eta = \phi(0) \text{ and } \dot{\phi} \in L_2\}$$

and for $(\phi(0), \phi) \in \mathcal{D}(\mathcal{A})$

$$(2.5) \quad \mathcal{A}(\phi(0), \phi) = \left(\int_{-r}^0 d\mu(\theta)\phi(\theta), \dot{\phi} \right).$$

The C_0 -semigroup generated by \mathcal{A} on Z will be denoted by $\{S(t) \mid t \geq 0\}$.

Consider the optimal control problem on a finite interval $[0, T]$: for given initial $(\eta, \phi) \in Z$,

$$(2.6) \quad \text{minimize } J(u; [0, T]) = \int_0^T (|Cx(t)|^2 + |u(t)|^2) dt + |Rx(T)|^2,$$

over $u \in L_2([0, T]; \mathbb{R}^m)$ subject to (2.1). Here C and R are $p \times n$ matrices. Within the framework of (2.3), (2.6) can be written as

$$\mathcal{J}(u) = \int_0^T (|\mathcal{C}z(t)|^2 + |u(t)|^2) dt + |\mathcal{R}z(T)|^2$$

where $\mathcal{C}(\eta, \phi) = C\eta$ and $\mathcal{R}(\eta, \phi) = R\eta$ for $(\eta, \phi) \in Z$. It then follows from [1], [7] that the optimal solution u^0 to (2.6) is given by

$$(2.7) \quad u^0(t) = -\mathcal{B}^*\Pi(t)z^0(t), \quad t \geq 0$$

where $\Pi(\cdot)$ is the unique solution, within a class of nonnegative (definite) selfadjoint operators for which $\langle \Pi(t)z, z \rangle$ is absolutely continuous on $[0, T]$, of the Riccati equation

$$(2.8) \quad \frac{d}{dt}\langle \Pi(t)z, z \rangle = -2\langle \mathcal{A}z, \Pi(t)z \rangle + \langle \mathcal{B}^*\Pi(t)z, \mathcal{B}^*\Pi(t)z \rangle - \langle \mathcal{C}z, \mathcal{C}z \rangle$$

for all $z \in \mathcal{D}(\mathcal{A})$,

$$\Pi(T) = \mathcal{R}^*\mathcal{R},$$

and $z^0(t)$ satisfies the evolution equation

$$(2.9) \quad \frac{d}{dt}z^0(t) = (\mathcal{A} - \mathcal{B}\mathcal{B}^*\Pi(t))z^0(t), \quad t \geq 0,$$

$$z^0(0) = (\eta, \phi).$$

Now we consider the optimal control problem on the infinite interval. For given initial data $z = (\eta, \phi)$, minimize the cost functional

$$(2.10) \quad J(u, z) = \int_0^\infty (|\mathcal{C}z(t)|^2 + |u(t)|^2) dt,$$

subject to (2.3).

DEFINITION 2.1. (i) $(\mathcal{A}, \mathcal{B})$ is stabilizable if there exists a bounded operator \mathcal{K} such that $\mathcal{A} - \mathcal{B}\mathcal{K}$ generates a uniformly exponentially stable semigroup.

(ii) $(\mathcal{C}, \mathcal{A})$ is detectable if $(\mathcal{A}^*, \mathcal{C}^*)$ is stabilizable.

Remark 2.2 (e.g. [16]). For hereditary differential systems, condition (ii) is equivalent to

$$z \in \mathcal{D}(\mathcal{A}), \quad \mathcal{A}z = \lambda z, \quad \mathcal{C}z = 0$$

for $\lambda \in \mathbb{C}^+$ imply that $z \equiv 0$. Moreover, (ii) holds if and only if

$$\text{rank} [\Delta(\lambda)^T, C^T] = n \quad \text{for all } \lambda \in \mathbb{C}^+$$

where $\Delta(\lambda) = \lambda I - \int_{-r}^0 d\mu(\theta) e^{\lambda\theta}$.

An operator $\Pi \in \mathcal{L}(Z)$ is a solution of the algebraic Riccati equation (ARE) if

$$(ARE) \quad 2\langle \mathcal{A}z, \Pi z \rangle - \langle \mathcal{B}^* \Pi z, \mathcal{B}^* \Pi z \rangle + \langle \mathcal{C}z, \mathcal{C}z \rangle = 0 \quad \text{for all } z \in \mathcal{D}(\mathcal{A}).$$

The next theorem follows from [7], [19].

THEOREM 2.3. (i) If $(\mathcal{A}, \mathcal{B})$ is stabilizable, then (ARE) has a self-adjoint, nonnegative solution.

(ii) If $(\mathcal{C}, \mathcal{A})$ is detectable, then (ARE) has at most, one selfadjoint, nonnegative solution. Moreover, if Π denotes the said solution, then $\mathcal{A} - \mathcal{B}\mathcal{B}^*\Pi$ generates a uniformly exponentially stable semigroup.

(iii) If $(\mathcal{A}, \mathcal{B})$ is stabilizable and $(\mathcal{C}, \mathcal{A})$ is detectable, then (ARE) has a unique selfadjoint, nonnegative solution and the optimal control to (2.10) is given

$$(2.11) \quad u^0(t) = -\mathcal{B}^* \Pi z^0(t),$$

where $z^0(t)$ is the mild solution to

$$\frac{d}{dt} z^0(t) = (\mathcal{A} - \mathcal{B}\mathcal{B}^*\Pi) z^0(t), \quad z^0(0) = z.$$

In what follows, we assume that condition (iii) in Theorem 2.3 holds and recall some of the important results due to Gibson [8].

THEOREM 2.4. If Π is the selfadjoint, nonnegative solution to (ARE), then

$$\Pi Z \subset \mathcal{D}(\mathcal{A}^*).$$

Note that $\mathcal{D}(\mathcal{A}^*)$ consists of elements $(y, \psi) \in Z$ for which $z(\theta) = \psi(\theta) - \sum_{i=1}^l A_i^T \chi_{(-\theta_i, 0]} y$ is absolutely continuous on $[-r, 0]$ with $z(-r) = 0$, [18]. If we write Π as a matrix of operators on $Z = \mathbb{R}^n \times L_2$,

$$(2.12) \quad \Pi = \begin{bmatrix} \Pi^{00} & \Pi^{01} \\ \Pi^{10} & \Pi^{11} \end{bmatrix}$$

where Π^{00} is a nonnegative, symmetric $n \times n$ matrix, $\Pi^{10}(\cdot)$ is a square integrable matrix function on $[-r, 0]$, $\Pi^{01} = \Pi^{10*}$ and

$$\Pi^{01} \phi = \int_{-r}^0 \Pi^{10}(\theta)^T \phi(\theta) d\theta, \quad \phi \in L_2,$$

and Π^{11} is a nonnegative, selfadjoint operator on L_2 , then from (2.11) the optimal control u^0 may be written as follows:

$$u^0(t) = -B^T \left(\Pi^{00}x(t) + \int_{-r}^0 \Pi^{10}(\theta)^T x(t+\theta) d\theta \right).$$

From Theorem 2.4 we have the following.

THEOREM 2.5. $\Pi^{10}(\cdot)$ is piecewise absolutely continuous on $[-r, 0]$ with the jump conditions at $-\theta_i$, $1 \leq i \leq l-1$

$$(2.13) \quad \Pi^{10}((-\theta_i)^+) - \Pi^{10}((-\theta_i)^-) = A_i^T \Pi^{00}.$$

Also,

$$(2.14) \quad \Pi^{10}(-r) = A_l^T \Pi^{00}.$$

Let us define an operator on $Z \times Z$

$$\mathcal{H} = \begin{bmatrix} \mathcal{A} & -\mathcal{B}\mathcal{B}^* \\ -\mathcal{C}^*\mathcal{C} & -\mathcal{A}^* \end{bmatrix}$$

with $\mathcal{D}(\mathcal{H}) = \mathcal{D}(\mathcal{A}) \times \mathcal{D}(\mathcal{A}^*)$. Then we have the following.

THEOREM 2.6. \mathcal{H} is closed and densely defined and has compact resolvent. For a complex number λ with $\text{Re } \lambda < 0$,

$$\lambda \in \sigma(\mathcal{A} - \mathcal{B}\mathcal{B}^*\Pi) \quad \text{if } \lambda \in \sigma(\mathcal{H}).$$

The algebraic and geometric multiplicities of λ as an eigenvalue of $\mathcal{A} - \mathcal{B}\mathcal{B}^*\Pi$ are finite and are identical to the respective multiplicities of λ as an eigenvalue of \mathcal{H} . Moreover, λ is an eigenvalue of \mathcal{H} if $\det \hat{\Delta}(\lambda) = 0$ where

$$(2.15) \quad \hat{\Delta}(\lambda) = \lambda I - \begin{bmatrix} \int_{-r}^0 d\mu(\theta) e^{\lambda\theta} & -BB^T \\ -C^TC & -\int_{-r}^0 d\mu(\theta)^T e^{-\lambda\theta} \end{bmatrix}.$$

3. Legendre-tau approximations. As pointed out in § 2, for the multiple point delay case, $\Pi^{10}(\cdot)$ has jump discontinuities. If we were to try to approximate the solution to $\Pi^{10}(\cdot)$ using a series of polynomials on $[-r, 0]$, we would observe the so-called Gibbs phenomenon. To avoid this difficulty, we proceed as follows. For simplicity of exposition we deal with the system of the form

$$(3.1) \quad \begin{aligned} \frac{d}{dt}x(t) &= A_0x(t) + A_1x(t + (-\theta_1)^+) + A_2x(t - \theta_2) \\ &\quad + \int_{-r}^0 A(\theta)x(t+\theta) d\theta + Bu(t), \end{aligned}$$

with $-r = -\theta_2 < -\theta_1 < 0$.

Alternatively, if $z(t, \theta) = x(t + \theta)$, then

$$(3.1a) \quad \frac{\partial}{\partial t}z(t, \theta) = \frac{\partial}{\partial \theta}z(t, \theta), \quad -r \leq \theta \leq 0,$$

$$(3.1b) \quad \frac{d}{dt}z(t, 0) = A_0z(t, 0) + A_1z(t, -\theta_1) + A_2z(t, -r) + \int_{-r}^0 A(\theta)z(t, \theta) d\theta + Bu(t).$$

The approximate solution $z^N(t, \theta)$ is assumed to be represented as

$$(3.2) \quad z^N(t, \theta) = \sum_{k=0}^N a_k^N(t) p_k^{(1)}(\theta) \chi_{(-\theta_1, 0]}(\theta) + \sum_{k=0}^N b_k^N(t) p_k^{(2)}(\theta) \chi_{[-r, -\theta_1]}(\theta)$$

where

$$p_k^{(1)}(\theta) = P_k((2\theta + \theta_1)/\theta_1), \quad p_k^{(2)}(\theta) = P_k((2(\theta + \theta_1) + \theta_2 - \theta_1)/(\theta_2 - \theta_1)),$$

for $0 \leq k \leq N$ and $\{P_k\}_{k \geq 0}$ are the Legendre polynomials on $[-1, 1]$. Note that $(\partial/\partial\theta)z^N$ is given by the following as an element in H^{-1} .

$$\begin{aligned} \frac{\partial}{\partial\theta} z^N(t, \theta) &= \sum_{k=0}^N (2/\Delta_1) a_k^N(t) \dot{p}_k^{(1)}(\theta) \chi_{(-\theta_1, 0]}(\theta) \\ &\quad + \sum_{k=0}^N (2/\Delta_2) b_k^N(t) \dot{p}_k^{(2)}(\theta) \chi_{[-r, -\theta_1]}(\theta) \\ &\quad + \left(\sum_{k=0}^N a_k^N(t) p_k^{(1)}(\theta) - \sum_{k=0}^N b_k^N(t) p_k^{(2)}(\theta) \right) \delta(\theta - \theta_1) \end{aligned}$$

where $\Delta_1 = \theta_1$, $\Delta_2 = \theta_2 - \theta_1$ and $\delta(\cdot)$ is the delta function. The underlying ideas of the tau method for approximating (3.1) are: (i) equating (3.1a) in the sense that

$$(3.3) \quad \left\langle \frac{\partial}{\partial t} z^N(t, \theta) - \frac{\partial}{\partial\theta} z^N(t, \theta), f \right\rangle_{L_2} = 0$$

for all

$$f \in \left\{ f \in L_2 \left| f = \sum_{k=0}^{N-1} \alpha_k p_k^{(1)} \chi_{(-\theta_1, 0]} + \sum_{k=0}^N \beta_k p_k^{(2)} \chi_{[-r, -\theta_1]}, \alpha_k, \beta_k \in \mathbb{R}^n \right. \right\},$$

and (ii) imposing (3.1b) on the approximate solution $z^N(t, \theta)$. From (i) we obtain $(2N+1)$ equations:

$$(3.4) \quad \begin{aligned} \frac{d}{dt} a_k^N(t) &= (2/\Delta_1) (S^N a^N)_k, \quad 0 \leq k \leq N-1, \\ \frac{d}{dt} b_k^N(t) &= (2/\Delta_2) (S^N b^N)_k + \sum_{i=0}^N ((-1)^i a_i^N - b_i^N) (2k+1)/\Delta_2, \quad 0 \leq k \leq N \end{aligned}$$

where S^N is the matrix representation of the derivative $\partial/\partial\theta$ (i.e. if the vector a is associated with a series of Legendre polynomials whose coefficients are the components of a , then the components of $S^N a$ give the Legendre coefficients of the derived series), and is given by

$$(3.5) \quad S^N = \begin{bmatrix} 0 & 1 & 0 & 1 & 0 & \cdots & 1 & 0 \\ 0 & 0 & 3 & 0 & 3 & \cdots & 0 & 3 \\ 0 & 0 & 0 & 5 & 0 & \cdots & 5 & 0 \\ \vdots & & & & & & \vdots & \vdots \\ 0 & 0 & 0 & 0 & 0 & \cdots & 2N-3 & 0 \\ 0 & 0 & 0 & 0 & 0 & \cdots & 0 & 2N-1 \end{bmatrix} \otimes I$$

for N even. Here I is the $n \times n$ identity matrix and \otimes denotes Kronecker product. From (ii) we obtain an equation for a_N^N , i.e.,

$$\frac{d}{dt} \left(\sum_{k=0}^N a_k^N(t) \right) = \int_{-r}^0 d\mu(\theta) z^N(t, \theta) + Bu(t)$$

or

$$(3.6) \quad \frac{d}{dt} a_N^N(t) = - \sum_{k=0}^{N-1} \frac{d}{dt} a_k^N(t) + \int_{-r}^0 d\mu(\theta) z^N(t, \theta) + Bu(t).$$

Here one needs the following modification in order to ensure the numerical stability of our scheme (see the proof of Lemma 3.3 for detailed discussions). Equation (3.6) is replaced by

$$(3.6)' \quad \frac{d}{dt} a_N^N(t) = - \sum_{k=0}^{N-1} \frac{d}{dt} a_k^N + \int_{-r}^0 d\mu^N(\theta) z^N(t, \theta) + Bu(t)$$

where

$$\int_{-r}^0 d\mu^N(\theta) \phi(\theta) = A_0 \phi(0) + A_1 \phi((- \theta_1)^+) + A_2 \phi(-r) + \int_{-r}^0 A^N(\theta) \phi(\theta) d\theta$$

with

$$A^N(\theta) = \sum_{k=0}^{N-1} \left[\frac{2k+1}{\theta_1} \int_{-\theta_1}^0 A(\xi) p_k^{(1)}(\xi) d\xi \right] p_k^{(1)}(\theta) \quad \text{on } [-\theta_1, 0]$$

and

$$A^N(\theta) = A(\theta) \quad \text{on } [-r, -\theta_1].$$

Hence, from (3.4) and (3.6)' we obtain a system of ordinary differential equations for $\text{col}(b_0^N, \dots, b_N^N, a_0^N, \dots, a_N^N)$:

$$(3.7) \quad \begin{aligned} \frac{d}{dt} \begin{bmatrix} \beta^N \\ \alpha^N \end{bmatrix} &= A^N \begin{bmatrix} \beta^N \\ \alpha^N \end{bmatrix} + B^N u(t), \\ B^N &= (e_{2N+2} \otimes I) B \end{aligned}$$

where $\alpha^N = \text{col}(a_0^N, a_1^N, \dots, a_N^N)$, $\beta^N = \text{col}(b_0^N, b_1^N, \dots, b_N^N)$ and $e_{2N+2} = \text{col}(0, 0, \dots, 1) \in \mathbb{R}^{2N+2}$. If we define the matrices $J^{(1)}$, $J^{(2)}$ and \tilde{S} by

$$\begin{aligned} J^{(1)} &= \overbrace{[u, u, \dots, u]}^{N+1} \otimes I, \\ J^{(2)} &= \overbrace{[u, -u, \dots, u, (-1)^N u]}^{N+1} \otimes I, \end{aligned}$$

with $u = \text{col}(1, 3, \dots, 2N+1)$, and

$$\tilde{S}^N = \left[\begin{array}{c} S^N \\ \left(0, -1, -3, \dots, -\frac{N(N+1)}{2} \right) \otimes I \end{array} \right],$$

then A^N is given by $A^N = A_0^N + A_\mu^N$, where

$$A_0^N = \begin{bmatrix} \frac{1}{\Delta_2} (2S^N - J^{(1)}) & \frac{1}{\Delta_2} J^{(2)} \\ 0 & \frac{2}{\Delta_1} \tilde{S}^N \end{bmatrix}$$

and

$$A_{\mu}^N = \begin{bmatrix} 0 & 0 \\ 0 & 0 \\ F_0 \cdots F_N & D_0^N \cdots D_N^N \end{bmatrix}$$

with

$$D_k^N = (A_0 + (-1)^k A_1) + (1 - \delta_{k,N}) \int_{-\theta_1}^0 A(\theta) p_k^{(1)}(\theta) d\theta$$

and

$$F_k = (-1)^k A_2 + \int_{-r}^{-\theta_1} A(\theta) p_k^{(2)}(\theta) d\theta \quad \text{for } 0 \leq k \leq N$$

where $\delta_{k,N}$ is the Kronecker delta.

Note that in the case when $\theta_1 = r$, the approximation scheme described above for (3.1) is exactly the same as that given in [9]. Let us introduce the orthogonal projection Q^N on Z . For any $z = (\eta, \phi)$, Q^N is defined by

$$Q^N z = \left(\eta, \sum_{k=0}^{N-1} \alpha_k p_k^{(1)}(\theta) \chi_{(-\theta_1, 0]} + \sum_{k=0}^N \beta_k p_k^{(2)}(\theta) \chi_{[-r, -\theta_1]} \right),$$

where

$$(3.8) \quad \alpha_k = \frac{2k+1}{\Delta_2} \int_{-r}^{-\theta_1} \phi(\theta) p_k^{(2)}(\theta) d\theta, \quad 0 \leq k \leq N-1,$$

$$(3.9) \quad \beta_k = \frac{2k+1}{\Delta_1} \int_{-\theta_1}^0 \phi(\theta) p_k^{(1)}(\theta) d\theta, \quad 0 \leq k \leq N,$$

and we define the projection operator L^N on Z by

$$(3.10) \quad \begin{aligned} L^N z &= Q^N z + a_N(0, p_N^{(1)} \chi_{[-\theta_1, 0]}), \\ a_N &= \eta - \sum_{k=0}^{N-1} \alpha_k p_k^{(1)}(0) = \eta - \sum_{k=0}^{N-1} \alpha_k. \end{aligned}$$

Immediately, we can obtain the following lemma.

LEMMA 3.1. *If $(\eta^N, \phi^N) = L^N(\eta, \phi)$, then $\phi^N(0) = \eta^N = \eta$. For $N \geq 1$,*

$$(3.11) \quad L^N Q^N = L^N \quad \text{and} \quad Q^N L^N = Q^N.$$

Moreover, if

$$L^N z = \sum_{k=0}^N a_k(p_k^{(1)}(0), p_k^{(1)} \chi_{(-\theta_1, 0]}) + \sum_{k=0}^N b_k(0, p_k^{(2)} \chi_{[-r, -\theta_1]})$$

and

$$Q^N z = \sum_{k=0}^{N-1} \alpha_k(0, p_k^{(1)} \chi_{(-\theta_1, 0]}) + \eta(1, 0) + \sum_{k=0}^N \beta_k(0, p_k^{(2)} \chi_{[-r, -\theta_1]}),$$

then we have

$$a_k = \alpha_k, \quad 0 \leq k \leq N-1, \quad b_k = \beta_k, \quad 0 \leq k \leq N,$$

and $a_N = \eta - \sum_{k=0}^{N-1} \alpha_k$; i.e.,

$$\Omega^N(b_0, \dots, b_N, a_0, \dots, a_{N-1}, a_N)^T = (\beta_0, \dots, \beta_N, \alpha_0, \dots, \alpha_{N-1}, \eta)^T$$

where

$$\Omega^N = \left[\begin{array}{c|cccc} I & & & & 0 \\ \hline & & & & 0 \\ & & I & & \vdots \\ 0 & & & & 0 \\ \hline & 1 & 1 & \cdots & 1 \end{array} \right].$$

As shown in [9], the tau method without the modification of (3.6) can be interpreted as follows. Let

$$(3.12) \quad z^N(t) = (z^N(t, 0), z^N(t, \cdot)) \in Z$$

where $z^N(t, \cdot)$ is given by (3.2). Then $z^N(t)$, $t \geq 0$ satisfies

$$(3.13) \quad \frac{d}{dt} z^N(t) = L^N \mathcal{A} z^N - L^N \mathcal{B} u(t), \quad z^N(0) = L^N z.$$

From (3.11), the function

$$(3.14) \quad \tilde{z}^N(t) = Q^N z^N(t), \quad z \geq 0$$

satisfies

$$\frac{d}{dt} \tilde{z}^N(t) = Q^N \mathcal{A} L^N \tilde{z}^N(t) + \mathcal{B} u(t), \quad \tilde{z}^N(0) = Q^N z.$$

Similarly, for the modified scheme with (3.6)' the function defined by (3.14) satisfies

$$\frac{d}{dt} \tilde{z}^N(t) = \mathcal{A}^N \tilde{z}^N(t) + \mathcal{B} u(t)$$

where

$$(3.15) \quad \mathcal{A}^N(\eta, \phi) = Q^N \left(\int_{-r}^0 d\mu^N(\theta) \phi^N(\theta), \dot{\phi}^N \right)$$

with

$$(\eta, \phi^N) = L^N(\eta, \phi) \quad \text{for } (\eta, \phi) \in Z.$$

To establish convergence for the tau approximation, we will use the Trotter-Kato theorem (see [13, Chap. III, Thm. 4.6]).

THEOREM 3.2. *Let $S(t)$ and $S^N(t)$, $N \geq 1$ be C_0 -semigroups acting on a Banach space X with infinitesimal generators \mathcal{A} and \mathcal{A}^N respectively. Assume that the following conditions are satisfied:*

(i) (stability). *There exists a constant ω such that*

$$\|S(t)\|_X \leq e^{\omega t} \quad \text{and} \quad \|S^N(t)\|_X \leq e^{\omega t}, \quad t \geq 0.$$

(ii) (consistency). *There exists a subset \mathcal{D} contained in $\mathcal{D}(\mathcal{A}) \cap \bigcap_{N=1}^{\infty} \mathcal{D}(\mathcal{A}^N)$ which together with $\text{Range}(\lambda I - \mathcal{A})$ for some $\lambda > 0$ is dense in X and such that $\mathcal{A}^N \phi \rightarrow \mathcal{A} \phi$ for all $\phi \in \mathcal{D}$ as $N \rightarrow \infty$. Then for all $\phi \in X$,*

$$\|S^N(t)\phi - S(t)\phi\| \rightarrow 0$$

uniformly on bounded t -intervals.

The following lemma concerns the question of stability of the tau approximation for (3.1). Following an idea in [2], we define the norm $\|\cdot\|_g$ on Z by

$$\|z\|_g = |\eta|^2 + \int_{-r}^0 |\phi(\theta)|^2 g(\theta) d\theta \quad \text{for } z = (\eta, \phi) \in Z,$$

where g is the piecewise constant function on $[-r, 0]$ defined by

$$(3.16) \quad g(\theta) = \begin{cases} 1, & \theta \in [-r, -\theta_1], \\ 2, & \theta \in (-\theta_1, 0]. \end{cases}$$

LEMMA 3.3. *Let $\{S^N(t), t \geq 0\}$ be the semigroup on Z generated by \mathcal{A}^N defined by (3.15). Then there exists a positive constant ω such that*

$$\|S^N(t)\|_g \leq e^{\omega t}, \quad t \geq 0,$$

i.e., for all $z \in Z$, $\langle A^N z, z \rangle_g \leq \omega \|Q^N z\|_g^2 \leq \omega \|z\|_g^2$, where $\langle \cdot, \cdot \rangle_g$ denotes the inner product on Z :

$$(3.17) \quad \langle (\eta^1, \phi^1), (\eta^2, \phi^2) \rangle_g = \langle \eta^1, \eta^2 \rangle_{\mathbb{R}^n} + \int_{-r}^0 \langle \phi^1, \phi^2 \rangle_{\mathbb{R}^n} g(\theta) d\theta.$$

Proof. Let

$$(\eta, \phi) = L^N z = \left(\eta, \sum_{k=0}^N a_k p_k^{(1)} \chi_{(-\theta_1, 0]} + \sum_{k=0}^N b_k p_k^{(2)} \chi_{[-r, -\theta_1]} \right)$$

where a_k, b_k are given in Lemma 3.1 and let $(\eta, \tilde{\phi}) = Q^N z$. Since p_N is orthogonal to all polynomials of degree at most $N-1$, it follows from (3.4), (3.6)', and (3.7) that

$$(3.18) \quad \begin{aligned} \langle \mathcal{A}^N z, z \rangle_g &= \left\langle \int_{-r}^0 d\mu^N(\theta) \phi(\theta), \eta \right\rangle + \int_{-r}^{-\theta_1} \langle \dot{\phi}(\theta), \phi(\theta) \rangle d\theta \\ &\quad + 2 \int_{-\theta_1}^0 \langle \dot{\phi}(\theta), \phi(\theta) \rangle d\theta + \left\langle \sum_{i=0}^N ((-1)^i a_i - b_i), \sum_{k=0}^N b_k \right\rangle. \end{aligned}$$

Note that

$$\phi(0) = \eta, \quad \phi((-\theta_1)^-) = \sum_{k=0}^N b_k, \quad \phi((-\theta_1)^+) = \sum_{k=0}^N (-1)^k a_k.$$

Then the right-hand side of (3.18) becomes

$$\begin{aligned} &= \left\langle \int_{-r}^0 d\mu^N(\theta) \phi(\theta), \phi(0) \right\rangle + \frac{1}{2} |\phi((-\theta_1)^-)|^2 - |\phi(-r)|^2 + |\phi(0)|^2 - |\phi((-\theta_1)^+)|^2 \\ &\quad + \langle \phi((-\theta_1)^+), \phi((-\theta_1)^-) \rangle - |\phi((-\theta_1)^-)|^2 \\ &= \left\langle A_0 \phi(0) + A_1 \phi((-\theta_1)^+) + A_2 \phi(-r) + \int_{-r}^0 A(\theta) \tilde{\phi}(\theta) d\theta, \phi(0) \right\rangle \\ &\quad - \frac{1}{2} |\phi(-r)|^2 - \frac{1}{2} |\phi((-\theta_1)^+) - \phi((-\theta_1)^-)|^2 - \frac{1}{2} |\phi((-\theta_1)^+)|^2 + |\phi(0)|^2, \end{aligned}$$

where the term $\int_{-r}^0 A(\theta) \tilde{\phi}(\theta) d\theta$ is a result of the modification (3.6)'. Without the modification, instead, we have $\int_{-r}^0 A(\theta) \phi(\theta) d\theta$. But in general, the absolute value of

this cannot be bounded by $(\text{constant}) \times \|Q^N z\|$ since $\|Q^N z\|^2 = |\eta|^2 + \|\tilde{\phi}\|_{L_2}^2$. Now, the right-hand side of (3.18) is bounded by

$$\begin{aligned} & \leq \left(1 + |A_0| + \frac{1}{2}|A_1^T A_1| + \frac{1}{2}|A_2^T A_2|\right) |\phi(0)|^2 \\ & \quad + \left(\int_{-r}^0 |A(\theta)|^2 d\theta\right)^{1/2} \left(\int_{-r}^0 |\tilde{\phi}(\theta)|^2 d\theta\right)^{1/2} |\phi(0)| \\ & \leq \omega \|Q^N z\|_g^2 \leq \omega \|z\|_g^2, \end{aligned}$$

where

$$\omega = 1 + |A_0| + \frac{1}{2}|A_1^T A_1| + \frac{1}{2}|A_2^T A_2| + \frac{1}{2}\|A(\cdot)\|_{L_2},$$

and we used the relation $2\langle x, y \rangle \leq |x|^2 + |y|^2$ for $x, y \in \mathbb{R}^n$, and the fact that Q^N is symmetric w.r.t. $\langle \cdot, \cdot \rangle_g$ —inner product. Q.E.D.

Next we will prove the consistency of the tau approximation. Let us denote by \mathcal{D}^k , $k \geq 1$, the domain of the k th power of \mathcal{A} . Then \mathcal{D}^k is dense in Z . Let us introduce the graph norm on \mathcal{D}^k :

$$\|z\|_{\mathcal{D}^k} = \left(\sum_{i=0}^k \|\mathcal{A}^i z\|_Z^2 \right)^{1/2} \quad \text{for } z \in \mathcal{D}^k.$$

Note that $\|\phi\|_{H^k} \leq \|z\|_{\mathcal{D}^k}$ for all $z = (\phi(0), \phi) \in \mathcal{D}^k$.

LEMMA 3.4. $\|(\mathcal{A}^N - \mathcal{A})z\| \rightarrow 0$ as $N \rightarrow \infty$ for all $z \in \mathcal{D}^k$, $k \geq 2$.

To prove this lemma, we need the following technical lemma.

LEMMA 3.5. Let us define the projection operator P^N of $L_2[-1, 1]$ by

$$P^N f = \sum_{k=0}^N f_k P_k, \quad f_k = \frac{2k+1}{2} \int_{-1}^1 f(x) P_k(x) dx.$$

Then for any positive integer m , there exists a constant K such that

$$|P^N f(\pm 1) - f(\pm 1)| \leq KN^{-m+1/2} \|f\|_{H^m}$$

and

$$\left| \frac{d}{dx} (P^N f)(\pm 1) - \frac{d}{dx} f(\pm 1) \right| \leq KN^{-m+5/2} \|f\|_{H^m}.$$

Proof. Note that for $k \geq 1$, P_k satisfies

$$\mathcal{L}P_k + k(k+1)P_k = 0$$

where \mathcal{L} is the differential operator:

$$(\mathcal{L}f)(x) = \frac{d}{dx} \left((1-x^2) \frac{d}{dx} f \right).$$

Thus for $k \geq 1$ and $f \in H^1$,

$$f_k = -\frac{2k+1}{2k(k+1)} \int_{-1}^1 f(x) \mathcal{L}P_k dx = \frac{2k+1}{2k(k+1)} \int_{-1}^1 (1-x^2) \frac{d}{dx} P_k \frac{d}{dx} f dx.$$

Using the relation

$$(3.19) \quad (1-x^2) \frac{d}{dx} P_k = \frac{k(k+1)}{2k+1} (P_{k+1} - P_{k-1}),$$

we obtain

$$f_k = \frac{1}{2} \int_{-1}^1 (P_{k+1} - P_{k-1}) \frac{d}{dx} f dx.$$

It then follows that

$$(P^N f)(\pm 1) = \sum_{k=0}^N (\pm 1)^k a^k = a_0 + \left(\frac{1}{2} \sum_{k:\text{even}} \pm \frac{1}{2} \sum_{k:\text{odd}} \right) \int_{-1}^1 (P_{k+1} - P_{k-1}) \dot{f} dx.$$

If N is even, then

$$\begin{aligned} (P^N f)(\pm 1) &= a_0 - \frac{1}{2} \int_{-1}^1 (P_1 \pm P_0) \dot{f} dx + \frac{1}{2} \int_{-1}^1 (P_{N+1} \pm P_N) \dot{f} dx \\ (3.20) \quad &= \frac{1}{2} \int_{-1}^1 f dx - \frac{1}{2} \int_{-1}^1 (x \pm 1) \dot{f} dx + \frac{1}{2} \int_{-1}^1 (P_{N+1} \pm P_N) \dot{f} dx \\ &= f(\pm 1) + \frac{1}{2} \int_{-1}^1 (P_{N+1} \pm P_N) \dot{f} dx. \end{aligned}$$

Similarly, for N odd,

$$(3.21) \quad (P^N f)(\pm 1) = f(\pm 1) + \frac{1}{2} \int_{-1}^1 (P_N \pm P_{N+1}) \dot{f} dx.$$

If $m = 2k + 1$, $k \geq 0$, then

$$\int_{-1}^1 P_N \dot{f} dx = \left(-\frac{1}{N(N+1)} \right)^k \int_{-1}^1 (\mathcal{L}^k P_N) \dot{f} dx = \left(-\frac{1}{N(N+1)} \right)^k \int_{-1}^1 P_N (\mathcal{L}^k \dot{f}) dx.$$

And, if $m = 2k + 2$, $k \geq 0$, then

$$\int_{-1}^1 P_N \dot{f} dx = \left(-\frac{1}{N(N+1)} \right)^{k+1} \int_{-1}^1 (1-x^2) \frac{d}{dx} P_N \frac{d}{dx} (\mathcal{L}^k \dot{f}) dx$$

and from (3.19)

$$= \left(-\frac{1}{N(N+1)} \right)^k \frac{1}{2N+1} \int_{-1}^1 (P_{N+1} - P_{N-1}) \frac{d}{dx} (\mathcal{L}^k \dot{f}) dx.$$

Since \mathcal{L}^k is a differentiable operator of order $2k$ with polynomial coefficients on $[-1, 1]$, there exists a constant c_k for $k \geq 0$ such that

$$\|\mathcal{L}^k \dot{f}\| \leq c_k \|f\|_{H^{2k+1}} \quad \text{and} \quad \left\| \frac{d}{dx} (\mathcal{L}^k \dot{f}) \right\| \leq c_k \|f\|_{H^{2k+2}}.$$

Now, the first inequality of the lemma follows from (3.20) and (3.21).

To prove the second inequality, we note that

$$\frac{d}{dx} (P^N f)(\pm 1) = \sum_{k=0}^N (\mp 1)^k \frac{k(k+1)}{2} f_k = \frac{1}{2} \sum_{k=0}^N (\mp 1)^k \left(\frac{2k+1}{2} \right) \int_{-1}^1 (\mathcal{L} f) P_k dx.$$

Then the same arguments as above enable us to obtain the second inequality. Q.E.D.

Proof of Lemma 3.4. From the definition (3.10) of L^N

$$z^N = L^N z = (\phi^N(0), \phi^N), \quad \phi^N = \phi^{(1)} \chi_{(-\theta_1, 0]} + \phi^{(2)} \chi_{[-r, -\theta_1]}$$

where

$$\phi^{(1)} = \sum_{k=0}^N a_k p_k^{(1)} \quad \text{on } (-\theta_1, 0], \quad \phi^{(2)} = \sum_{k=0}^N b_k p_k^{(2)} \quad \text{on } [-r, -\theta_1]$$

and $\{a_k\}$ and $\{b_k\}$ are as given in Lemma 3.1. It then follows from (3.4) and (3.6)' that

$$\mathcal{A}^N z = (\eta^N, \psi^N)$$

with

$$\begin{aligned} \eta^N &= \int_{-r}^0 d\mu^N(\theta) \phi^N, \\ \psi^N &= \dot{\phi}^{(1)} \chi_{(-\theta_1, 0]} + \dot{\phi}^{(2)} \chi_{[-r, -\theta_1]} \\ &\quad + (\phi^{(1)}(-\theta_1) - \phi^{(2)}(-\theta_1)) \sum_{k=0}^N \frac{2k+1}{\Delta_2} p_k^{(2)} \chi_{[-r, -\theta_1]}. \end{aligned}$$

Thus, for $z \in \mathcal{D}(\mathcal{A})$,

$$\begin{aligned} \delta &= \|(\mathcal{A}^N - \mathcal{A})z\| \leq \|\dot{\phi}^{(1)} - \dot{\phi}\|_{L_2[-\theta_1, 0]} + \|\dot{\phi}^{(2)} - \dot{\phi}\|_{L_2[-r, -\theta_1]} \\ &\quad + |\phi^{(1)}(-\theta_1) - \phi^{(2)}(-\theta_1)| \left(\sum_{k=0}^N \frac{2k+1}{\Delta_2} \right)^{1/2} \\ &\quad + \left| \int_{-r}^0 d\mu^N(\theta) \phi^N(\theta) - \int_{-r}^0 d\mu(\theta) \phi(\theta) \right| \\ &= \delta_1 + \delta_2 + \delta_3 + \delta_4. \end{aligned} \tag{3.22}$$

Here, we note that

$$\phi^{(1)} = \tilde{\phi}^{(1)} + (\phi(0) - \tilde{\phi}^{(1)}(0)) p_N^{(1)} \quad \text{on } [-\theta_1, 0], \tag{3.23}$$

where

$$\tilde{\phi}^{(1)} = \sum_{k=0}^{N-1} a_k p_k^{(1)} \quad \text{on } [-\theta_1, 0].$$

From (3.6)', for $z = (\phi(0), \phi) \in \mathcal{D}(\mathcal{A})$,

$$\begin{aligned} &\int_{-r}^0 d\mu^N(\theta) \phi^N(\theta) - \int_{-r}^0 d\mu(\theta) \phi(\theta) \\ &= A_1(\phi^{(1)}(-\theta_1) - \phi(-\theta_1)) + A_2(\phi^{(2)}(-r) - \phi(-r)) \\ &\quad + \int_{-\theta_1}^0 A(\theta)(\tilde{\phi}^{(1)} - \phi) d\theta + \int_{-r}^{-\theta_1} A(\theta)(\phi^{(2)} - \phi) d\theta \end{aligned}$$

where from (3.23)

$$\phi^{(1)}(-\theta_1) = \tilde{\phi}^{(1)}(-\theta_1) + (-1)^N (\phi(0) - \tilde{\phi}^{(1)}(0)).$$

It then follows that

$$\begin{aligned} \delta_4 &\leq |A_1| (|\tilde{\phi}^{(1)}(-\theta_1) - \phi(-\theta_1)| + |\tilde{\phi}^{(1)}(0) - \phi(0)|) + |A_2| |\phi^{(2)}(-r) - \phi(-r)| \\ &\quad + \|A(\cdot)\|_{L_2} (\|\tilde{\phi}^{(1)} - \phi\|_{L_2[-\theta_1, 0]} + \|\phi^{(2)} - \phi\|_{L_2[-r, -\theta_1]}). \end{aligned}$$

It now follows from Lemma 3.5 and Lemmas 3.1 and 3.2 in [9] that

$$\begin{aligned} |\delta_4| &\leq K((2|A_1| + |A_2|) N^{-k+1/2} + 2\|A(\cdot)\|_{L_2} N^{-k}) \|z\|_{\mathcal{D}^k} \\ &\leq K_4 N^{-k+1/2} \|z\|_{\mathcal{D}^k}. \end{aligned}$$

From Lemma 3.2 in [9]

$$\delta_2 \leq K_2 N^{-k+3/2} \|z\|_{\mathcal{D}^k}.$$

From (3.23)

$$\delta_1 \leq \|\dot{\tilde{\phi}}^{(1)} - \dot{\phi}\|_{L_2[-\theta_1, 0]} + \sqrt{N(N+1)} |\tilde{\phi}^{(1)}(0) - \phi(0)|$$

where we used the fact that

$$\int_{-1}^1 |\dot{P}_N(\theta)|^2 d\theta = N(N+1).$$

It then follows from Lemma 3.5 and Lemma 3.2 in [9] that

$$\delta_1 \leq K_1 N^{-k+3/2} \|z\|_{\mathcal{D}^k}.$$

Since

$$\begin{aligned} |\phi^{(1)}(-\theta_1) - \phi^{(2)}(-\theta_1)| &\leq |\phi^{(1)}(-\theta_1) - \phi(-\theta_1)| + |\phi^{(2)}(-\theta_1) - \phi(-\theta_1)| \\ &\leq |\tilde{\phi}^{(1)}(-\theta_1) - \phi(-\theta_1)| + |\tilde{\phi}^{(1)}(0) - \phi(0)| + |\phi^{(2)}(-\theta_1) - \phi(-\theta_1)|, \end{aligned}$$

it follows from Lemma 3.5 that

$$\delta_3 \leq K_3 N^{-k+3/2} \|z\|_{\mathcal{D}^k}.$$

Hence from (3.22)

$$\|(\mathcal{A}^N - \mathcal{A})z\|_Z \leq \tilde{K} N^{-k+3/2} \|z\|_{\mathcal{D}^k}, \quad k \geq 2$$

where \tilde{K} is independent of N . Q.E.D.

Now we state the convergence result for the tau approximation.

THEOREM 3.6. *Let $S^N(t)$, $t \geq 0$ be the semigroup on Z generated by \mathcal{A}^N defined by (3.15). Then for all $z \in Z$*

$$\|S^N(t)z - S(t)z\| \rightarrow 0$$

uniformly on bounded t -intervals.

Proof. The theorem follows from Theorem 3.2. X is the Hilbert space Z equipped with the inner product (3.17). The stability (i) follows from Lemma 3.3. Note that the weighted norm $\|\cdot\|_g$ and $\|\cdot\|_Z$ norm are equivalent, i.e.,

$$|z|_Z^2 \leq \|z\|_g^2 \leq 2\|z\|_Z^2 \quad \text{for } z = (\eta, \phi) \in Z.$$

Thus, from Lemma 3.4

$$\|\mathcal{A}^N z - \mathcal{A}z\|_g \rightarrow 0 \quad \text{for } z \in \mathcal{D}^2.$$

Since \mathcal{D}^2 and $(\lambda I - \mathcal{A})\mathcal{D}^2$ for λ sufficiently large are dense in Z , the statement (ii) in Theorem 3.2 holds if we choose $\mathcal{D} = \mathcal{D}^2$. Q.E.D.

Remark. Although we will not pursue the details here, one can prove that the adjoint semigroups $S^N(t)^*$ also converge strongly to $S^*(t)$ uniformly on bounded t -intervals.

4. An approximation scheme for the Riccati equation. In this section, we discuss an approximation scheme for the regulator problem (2.10) based upon the Legendre-tau approximation.

Let us consider the N th approximate problem to (2.10)

$$(4.1) \quad \text{Minimize } J^N(u, \tilde{z}) = \int_0^\infty (|\tilde{z}^N(t)|^2 + |u(t)|^2) dt,$$

subject to (3.15):

$$\frac{d}{dt}\tilde{z}^N(t) = \mathcal{A}^N\tilde{z}^N(t) + \mathcal{B}u(t), \quad z^N(0) = \tilde{z} = Q^N z.$$

It follows from Theorem 2.3 that if $(\mathcal{A}^N, \mathcal{B})$ is stabilizable and $(\mathcal{C}, \mathcal{A}^N)$ is detectable, then there exists a unique solution Π^N to $(\text{ARE})^N$:

$$(\text{ARE})^N \quad (\mathcal{A}^N)^* \Pi^N + \Pi^N \mathcal{A}^N - \Pi^N \mathcal{B} \mathcal{B}^* \Pi^N + \mathcal{C}^* \mathcal{C} = 0,$$

and the optimal solution to (4.1) is given by

$$(4.2) \quad u^N(t) = -\mathcal{B}^* \Pi^N \tilde{z}^N(t)$$

where $\tilde{z}^N(t)$, $t \geq 0$ satisfies

$$\frac{d}{dt}\tilde{z}^N(t) = (\mathcal{A}^N - \mathcal{B} \mathcal{B}^* \Pi^N) \tilde{z}^N(t), \quad z^N(0) = \tilde{z}.$$

Remark. For the single point delay case, we are able to prove that if $(\mathcal{A}, \mathcal{B})$ is stabilizable (respectively, $(\mathcal{C}, \mathcal{A})$ is detectable), then for N sufficiently large $(\mathcal{A}^N, \mathcal{B})$ is stabilizable (respectively, $(\mathcal{C}, \mathcal{A}^N)$ is detectable) [10]. The proof is based upon the characterization of detectability in Remark 2.2.

In terms of the Legendre coordinate system,

$$\begin{aligned} \tilde{z}^N(t) = & \sum_{k=0}^{N-1} a_k^N(t)(0, p_k^{(1)} \chi_{(-\theta_1, 0]}) \\ & + \eta^N(t)(1, 0) + \sum_{k=0}^N b_k^N(t)(0, p_k^{(2)} \chi_{[-r, -\theta_1]}). \end{aligned}$$

It then follows from Lemma 3.1 and (3.7) that

$$\tilde{\xi}^N(t) = (b_0^N, \dots, b_N^N, a_0^N, \dots, a_{N-1}^N, \eta^N)^T, \quad t \geq 0$$

satisfies

$$(4.3) \quad \frac{d}{dt}\tilde{\xi}^N(t) = \tilde{A}^N \tilde{\xi}^N(t) + B^N u(t), \quad \tilde{\xi}^N(0) = \tilde{\xi}$$

where $\tilde{A}^N = \Omega^N A^N (\Omega^N)^{-1}$ and $\tilde{\xi}$ is the vector representation of $Q^N z$ in terms of Legendre coordinates. A^N , B^N and Ω^N are given in § 3. Thus we can write (4.1) as

$$(4.4) \quad \text{Minimize } J^N(u, \tilde{\xi}) = \int_0^\infty (|\tilde{C}^N \tilde{\xi}^N(t)|^2 + |u(t)|^2) dt$$

subject to (4.3), where $\tilde{C}^N = C^N (\Omega^N)^{-1}$ with

$$C^N = \begin{pmatrix} 0 & |C|C| \cdots |C \end{pmatrix} \in \mathbb{R}^{p \times 2n(N+1)}$$

Hence the optimal solution u^N to (4.1) can be also given by

$$u^N(t) = -(B^N)^T \Sigma^N \tilde{\xi}^N(t),$$

where Σ^N satisfies the matrix Riccati equation

$$(4.5) \quad (\tilde{A}^N)^T \Sigma^N + \Sigma^N \tilde{A}^N - \Sigma^N B^N (B^N)^T \Sigma^N + (\tilde{C}^N)^T \tilde{C}^N = 0.$$

If $(\mathcal{A}^N, \mathcal{B})$ is stabilizable and $(\mathcal{C}, \mathcal{A}^N)$ is detectable, then (4.5) has a unique, symmetric, nonnegative definite solution. Σ^N can be computed effectively by Potter's method (e.g. [14], [11]), which involves the eigenvalue-eigenvector decomposition of the matrix

$$(4.6) \quad H^N = \begin{bmatrix} \tilde{A}^N & -B^N(B^N)^T \\ -(\tilde{C}^N)^T \tilde{C}^N & -(\tilde{A}^N)^T \end{bmatrix}.$$

Let us define the matrix $\hat{\Sigma}^N$ by

$$\hat{\Sigma}^N = \Lambda^N \Sigma^N \Lambda^N$$

where Λ^N is a diagonal matrix:

$$\Lambda^N = \text{diag} \left(\frac{\Delta_2}{1}, \dots, \frac{\Delta_2}{2k+1}, \dots, \frac{\Delta_2}{2N+1}, \frac{\Delta_1}{1}, \dots, \frac{\Delta_1}{2k+1}, \dots, \frac{\Delta_1}{2N-1}, 1 \right) \otimes I,$$

and define the $n \times n$ matrices σ_{ij} , $0 \leq i, j \leq 2N+1$ by

$$(4.7) \quad \sigma_{ij} = (e_{i+1} \otimes I)^T \hat{\Sigma}^N (e_{j+1} \otimes I)$$

where e_i is the i th unit vector in \mathbb{R}^{2N+2} , i.e.,

$$e_i = (\underbrace{0, \dots, 0}_{i-1}, 1, 0 \dots 0)^T.$$

LEMMA 4.1. Suppose $(\mathcal{A}^N, \mathcal{B})$ is stabilizable and $(\mathcal{C}, \mathcal{A}^N)$ is detectable. Then for $z = (\eta, \phi) \in Z$, $\Pi^N z = (y, \psi)$ with

$$\begin{aligned} y &= \sigma \eta + \int_{-\theta_1}^0 \sum_{k=0}^{N-1} (\sigma_{\cdot, k+N+1}) p_k^{(1)}(\theta) \phi(\theta) d\theta \\ &\quad + \int_{-r}^{-\theta_1} \sum_{k=0}^M (\sigma_{\cdot, k}) p_k^{(2)}(\theta) \phi(\theta) d\theta, \end{aligned}$$

and

$$\begin{aligned} \psi(\theta) &= \sum_{i=0}^{N-1} \left[(\sigma_{i+N+1, \cdot}) \eta + \int_{-\theta_1}^0 \sum_{k=0}^{N-1} (\sigma_{i+N+1, k+N+1}) p_k^{(1)}(\theta) \phi(\theta) d\theta \right. \\ &\quad \left. + \int_{-r}^{-\theta_1} \sum_{k=0}^N (\sigma_{i+N+1, k}) p_k^{(2)}(\theta) \phi(\theta) d\theta \right] p_i^{(1)}(\theta) \chi_{(-\theta_1, 0]} \\ &\quad + \sum_{i=0}^N \left[(\sigma_{i, \cdot}) \eta + \int_{-\theta_1}^0 \sum_{k=0}^{N-1} (\sigma_{i, k+N+1}) p_k^{(1)}(\theta) \phi(\theta) d\theta \right. \\ &\quad \left. + \int_{-r}^{-\theta_1} \sum_{k=0}^N (\sigma_{i, k}) p_k^{(2)}(\theta) \phi(\theta) d\theta \right] p_i^{(2)}(\theta) \chi_{[-r, -\theta_1]} \end{aligned}$$

where the symbol (\cdot) stands for $2N+1$, $\sigma = \sigma_{2N+1, 2N+1}$.

Proof. It is known [19] that

$$\langle \Pi^N z, z \rangle_Z = \langle \Sigma^N \tilde{\xi}, \tilde{\xi} \rangle_{\mathbb{R}^{2n(N+1)}} = \min J^N(u)$$

for all $z \in Z$, where $\tilde{\xi}$ is the vector representation of $Q^N z$. Since Σ^N and Π^N are symmetric,

$$(4.8) \quad \langle \Pi^N z^1, z^2 \rangle_Z = \langle \Sigma^N \tilde{\xi}^1, \tilde{\xi}^2 \rangle_{\mathbb{R}^{2n(N+1)}}$$

for all $z^i = (\eta^i, \phi^i) \in Z$, $i = 1, 2$, where $\tilde{\xi}^i$ is the vector representation of $Q^N z^i$ for $i = 1, 2$. Note that

$$\Lambda^N \tilde{\xi}^i = (\underline{\beta}^i, \underline{\alpha}^i, \gamma^i)^T, \quad i = 1, 2$$

where for $i = 1, 2$

$$(4.9) \quad \beta_k^i = \int_{-r}^{-\theta_1} \phi^i(\theta) p_k^{(2)}(\theta) d\theta, \quad \alpha_k^i = \int_{-\theta_1}^0 \phi^i(\theta) p_k^{(1)}(\theta) d\theta, \quad \gamma^i = \eta^i.$$

Then

$$\langle \Sigma^N \tilde{\xi}^1, \tilde{\xi}^2 \rangle = (\underline{\beta}^1, \underline{\alpha}^1, \gamma^1)^T \hat{\Sigma}^N (\underline{\beta}^2, \underline{\alpha}^2, \gamma^2).$$

Now, if $\Pi^N(\eta^1, \phi^1) = (y, \psi)$, then

$$\langle \Pi^N(\eta^1, \phi^1), (\eta^2, \phi^2) \rangle_2 = \langle \eta^2, y \rangle + \int_{-\theta_1}^0 \langle \phi^2(\theta), \psi(\theta) \rangle d\theta + \int_{-r}^{-\theta_1} \langle \phi^2(\theta), \psi(\theta) \rangle d\theta.$$

Equating (4.8), we obtain

$$y = \sigma \eta^1 + \sum_{k=0}^{N-1} (\sigma_{\cdot, k+N+1}) \alpha_k^1 + \sum_{k=0}^N (\sigma_{\cdot, k}) \beta_k^1$$

and

$$\begin{aligned} \psi(\theta) = & \sum_{i=0}^{N-1} \left[(\sigma_{i+N+1, \cdot}) \eta^1 + \sum_{i=0}^{N-1} (\sigma_{i+N+1, k+N+1}) \alpha_k^1 + \sum_{k=0}^N (\sigma_{i+N+1, k}) \beta_k^1 \right] p_i^{(1)} \chi_{(-\theta_1, 0]} \\ & + \sum_{i=0}^N \left[(\sigma_{i, \cdot}) \eta^1 + \sum_{k=0}^{N-1} (\sigma_{i, k+N+1}) \alpha_k^1 + \sum_{k=0}^N (\sigma_{i, k}) \beta_k^1 \right] p_i^{(2)} \chi_{[-r, -\theta_1]}, \end{aligned}$$

which completes the proof along with (4.9). Q.E.D.

COROLLARY 4.2. *The optimal solution u^N to (4.1) can be written in the operator form:*

$$u^N(t) = -\mathcal{K}^N z^N(t).$$

$\mathcal{K}^N \in \mathcal{L}(Z, \mathbb{R}^m)$ is given by

$$\mathcal{K}^N z = B^T \left(\Pi_N^{00} \eta + \int_{-r}^0 \Pi_N^{10}(\theta)^T \phi(\theta) d\theta \right) \quad \text{for } z = (\eta, \phi) \in Z$$

where

$$\Pi_N^{00} = \sigma,$$

and

$$\begin{aligned} \Pi_N^{10}(\theta) = & \sum_{i=0}^{N-1} (\sigma_{i+N+1, \cdot}) p_i^{(1)}(\theta) \chi_{(-\theta_1, 0]} \\ & + \sum_{i=0}^N (\sigma_{i, \cdot}) p_i^{(2)}(\theta) \chi_{[-r, -\theta_1]}, \quad -r \leq \theta \leq 0. \end{aligned}$$

Proof. Since $\mathcal{B}^*(\eta, \phi) = B^T \eta$, the corollary follows from Lemma 4.1 and (4.2). Q.E.D.

5. Convergence proof. In this section we discuss the convergence property of Π^N .

THEOREM 5.1. Suppose Π^N satisfies the N th algebraic Riccati equation $(\text{ARE})^N$. If $\{\Pi^N\}$ is uniformly bounded on Z and $(\mathcal{C}, \mathcal{A})$ is detectable, then

- (i) Π^N converges weakly to Π which is the unique solution to ARE;
- (ii) there exists an integer N_0 such that if $N \geq N_0$, then

$$(5.1) \quad \|S^N(t)\| \leq M e^{-\omega t}$$

for some positive constants M and ω , where $\{S^N(t), t \geq 0\}$ is the semigroup on Z generated by $\mathcal{A} - \mathcal{B}\mathcal{B}^*\Pi^N$.

Proof of (i). Although (i) follows from Theorems 3.6 and 6.7 in [8], we will give an alternative proof here. Since $\{\Pi^N\}$ is uniformly bounded on Z , by Theorem 6.5 in [8], there exists a subsequence $\{\Pi^{N_j}\}$ which converges weakly to some nonnegative, selfadjoint operator Π . If $(\mathcal{C}, \mathcal{A})$ is detectable, then from Theorem 2.3, ARE has at most one nonnegative, selfadjoint solution. Hence, we only need to show Π satisfies ARE. Without loss of generality we can assume that Π^N converges weakly to Π . Note that for $N \geq 1$, Π^N satisfies $(\text{ARE})^N$. Since $\dim(R^m) < \infty$, $\mathcal{B}^*\Pi^N$ converges strongly to $\mathcal{B}^*\Pi$. It now follows from Lemma 3.4

$$(5.2) \quad 2\langle \mathcal{A}z, \Pi z \rangle - \langle \mathcal{B}^*\Pi z, \mathcal{B}^*\Pi z \rangle + \langle \mathcal{C}z, \mathcal{C}z \rangle = 0 \quad \text{for all } z \in \mathcal{D}^2.$$

Since $\mathcal{D}^2 = \mathcal{D}(\mathcal{A}^2)$ is dense in $\mathcal{D}(\mathcal{A})$, a simple limit argument shows that (5.2) holds for all $z \in \mathcal{D}(\mathcal{A})$, i.e., Π is a solution to ARE.

Proof of (ii). First of all, we note that $\tilde{\mathcal{A}} = \mathcal{A} - \mathcal{B}\mathcal{B}^*\Pi$ generates a uniformly exponentially stable semigroup $\{\tilde{S}(t), t \geq 0\}$ on Z , i.e., there exist positive constants \tilde{M} and $\tilde{\omega}$ such that

$$(5.3) \quad \|\tilde{S}(t)\| \leq \tilde{M} e^{-\tilde{\omega}t}, \quad t \geq 0.$$

For $z \in \mathcal{D}(\mathcal{A})$

$$(\mathcal{A} - \mathcal{B}\mathcal{B}^*\Pi^N)z = \tilde{\mathcal{A}}z - \mathcal{B}(\mathcal{B}^*\Pi - \mathcal{B}^*\Pi^N)z.$$

Thus,

$$(5.4) \quad S^N(t)z = \tilde{S}(t)z + \int_0^t \tilde{S}(t-s)\mathcal{B}(\mathcal{B}^*\Pi - \mathcal{B}^*\Pi^N)S^N(s)z \, ds \quad \text{for all } z \in Z.$$

For $t \geq r$, we may write (5.4) as

$$(5.5) \quad S^N(t)z = \tilde{S}(t-r)\tilde{z} + \int_r^t \tilde{S}(t-s)\mathcal{B}(\mathcal{B}^*\Pi - \mathcal{B}^*\Pi^N)S^N(s)z \, ds,$$

with

$$(5.6) \quad \tilde{z} = \tilde{S}(r)z + \int_0^r \tilde{S}(r-s)\mathcal{B}(\mathcal{B}^*\Pi - \mathcal{B}^*\Pi^N)S^N(s)z \, ds.$$

From [5] we have that $\tilde{S}(r)z \in \mathcal{D}(\tilde{\mathcal{A}})$ for all z and $\|\tilde{S}(r)z\|_{\mathcal{D}(\tilde{\mathcal{A}})} \leq \gamma_1 \|z\|_Z$ for some positive constant γ_1 . If

$$z(t) = (x(t), x(t+\cdot)) = \int_0^t \tilde{S}(t-s)\mathcal{B}u(s) \, ds, \quad t \geq 0,$$

then $x(t) \in H^1([-r, T]; \mathbb{R}^n)$ for any $T \geq 0$ and satisfies

$$\begin{aligned} \frac{d}{dt}x(t) &= \int_{-r}^0 d\mu(\theta)x(t+\theta) - BB^T \Pi^{00}x(t) - BB^T \int_{-r}^0 \Pi^{10}(\theta)^T x(t+\theta) d\theta + Bu(t) \\ &\equiv \int_{-r}^0 d\tilde{\mu}(\theta)x(t+\theta) + Bu(t). \end{aligned}$$

Hence for $u \in L_2^{\text{loc}}([0, \infty); \mathbb{R}^n)$, $z(t) \in \mathcal{D}(\tilde{\mathcal{A}})$ and

$$(5.7) \quad \tilde{\mathcal{A}} \int_0^t \tilde{S}(t-s) \mathcal{B}u(s) ds = \left(\int_{-r}^0 d\tilde{\mu}(\theta)x(t+\theta), \psi(t+\cdot) \right)$$

where

$$\psi(t) = \int_{-r}^0 d\tilde{\mu}(\theta)x(t+\theta) + Bu(t), \quad t \geq 0,$$

and

$$\psi(t) = 0 \quad \text{for } t < 0.$$

Here we note that

$$(5.8) \quad \left\| \int_{-r}^0 d\tilde{\mu}(\theta)x(\cdot + \theta) \right\|_{L_2([a, b], \mathbb{R}^n)} \leq \gamma_2 \|x\|_{L_2([a-r, b], \mathbb{R}^n)}$$

for $b \geq a \geq 0$, where $\gamma_2 = \int_{-r}^0 |d\tilde{\mu}|$. Since $\{\Pi^N\}$ is uniformly bounded, it now follows from (5.6) and (5.7) that $\tilde{z} \in \mathcal{D}(\tilde{\mathcal{A}})$ for all $z \in Z$ and

$$(5.9) \quad \|\tilde{z}\|_{\mathcal{D}(\tilde{\mathcal{A}})} \leq \gamma_3 \|z\|_Z,$$

for some positive constant γ_3 . From (5.5) and (5.7), $S^N(t)z \in \mathcal{D}(\tilde{\mathcal{A}})$, $t \geq r$ for $z \in Z$ and

$$(5.10) \quad \tilde{\mathcal{A}}S^N(t)z = \tilde{S}(t-r)\tilde{\mathcal{A}}\tilde{z} + \tilde{\mathcal{A}} \int_r^t \tilde{S}(t-s) \mathcal{B}F^N(\tilde{\mathcal{A}}S^N(s)z) ds$$

where $F^N : Z \rightarrow \mathbb{R}^m$ is given by

$$F^N = (\mathcal{B}^*\Pi - \mathcal{B}^*\Pi^N)\tilde{\mathcal{A}}^{-1}.$$

Since $0 \notin P\sigma(\tilde{\mathcal{A}})$, $(\tilde{\mathcal{A}})^{-1}$ exists and moreover, it is compact [18]. Note that

$$(F^N)^* = (\tilde{\mathcal{A}}^*)^{-1}(\Pi\mathcal{B} - \Pi^N\mathcal{B}) \in \mathcal{L}(\mathbb{R}^m, Z).$$

Since $\Pi^N\mathcal{B}$ converges weakly to $\Pi(\mathcal{B})$ as $N \rightarrow \infty$ and $(\tilde{\mathcal{A}})^{-1}$ is compact, $(F^N)^*$ converges strongly to zero. Hence, the finite dimensionality of \mathbb{R}^m implies

$$\|(F^N)^*\| = \|F^N\| \rightarrow 0 \quad \text{as } N \rightarrow \infty,$$

i.e., for any $\varepsilon > 0$ there exists an integer $N_0(\varepsilon)$ such that $\|F^N\| \leq \varepsilon$ for $N \geq N_0$.

For $z \in Z$, let us define the Z -valued function $\beta^N(t)$, $t \geq r$ by

$$\beta^N(t) = \tilde{\mathcal{A}}S^N(t)z.$$

Then from (5.7) and (5.10)

$$\beta^N(t) = \tilde{S}(t-r)\tilde{\mathcal{A}}\tilde{z} + \left(\int_{-r}^0 d\tilde{\mu}(\theta)x(t+\theta), \psi(t+\cdot) \right)$$

where for $t \geq 0$

$$\psi(t) = \int_{-r}^0 d\tilde{\mu}(\theta)x(t+\theta) + BF^N\beta^N(t)$$

and

$$(5.11) \quad (x(t), x(t+\cdot)) = \int_r^t \tilde{S}(t-s)\mathcal{B}F^N\beta^N(s) ds, \quad t \geq r$$

with $x(t) = 0$, $t \leq r$. Now from (5.8), for $T > r$

$$\begin{aligned} \left(\int_r^T \|\beta^N(t)\|^2 dt \right)^{1/2} &\leq \|\tilde{\mathcal{A}}\tilde{z}\| \left(\int_r^T \|\tilde{S}(t-r)\|^2 dt \right)^{1/2} + \gamma_2 \left(\int_0^T |x(t)|^2 dt \right)^{1/2} \\ &\quad + \gamma_2 \left(\int_r^T \int_{t-2r}^t |x(s)|^2 ds dt \right)^{1/2} \\ &\quad + |B| \|F^N\| \left(\int_r^T \int_{t-r}^t \|\beta^N(s)\|^2 ds dt \right)^{1/2}, \end{aligned}$$

and from (5.3) and (5.11),

$$\begin{aligned} &\leq (\tilde{M}/(2\tilde{\omega})^{1/2}) \|\tilde{\mathcal{A}}\tilde{z}\| \\ &\quad + |B| \|F^N\| \left(\frac{\tilde{M}}{\tilde{\omega}} \gamma_2(1+(2r)^{1/2}) + (r)^{1/2} \right) \left(\int_r^T \|\beta^N(s)\|^2 ds \right)^{1/2} \end{aligned}$$

where we used Fubini's theorem and Young's inequality. Thus, from (5.9)

$$\int_r^T \|\beta^N(t)\|_N^2 dt \leq \left(\frac{\tilde{M}^2}{\tilde{\omega}} \right) \gamma_3^2 \|z\|_Z^2 + 2 \|F^N\|^2 \gamma^2 \int_r^T \|\beta^N(s)\|^2 ds,$$

where

$$\gamma = |B| \left(\frac{\tilde{M}}{\tilde{\omega}} \gamma_2(1+(2r)^{1/2}) + (r)^{1/2} \right).$$

If we choose ε such that $2\varepsilon^2\gamma^2 \leq \frac{1}{2}$, then it follows that for $T \geq r$

$$\int_r^T \|\beta^N(t)\|_N^2 dt \leq 2 \left(\frac{\tilde{M}^2}{\tilde{\omega}} \right) \gamma_3^2 \|z\|_Z^2.$$

Note that $S^N(t)z = \tilde{\mathcal{A}}^{-1}\beta^N(t)$, $t \geq r$ and $\tilde{\mathcal{A}}^{-1} \in \mathcal{L}(Z)$. Hence, for $T \geq r$

$$\int_r^T \|S^N(t)z\|^2 dt \leq 2 \left(\frac{\tilde{M}^2}{\tilde{\omega}} \right) \gamma_3^2 \|\tilde{\mathcal{A}}^{-1}\|^2 \|z\|_Z^2.$$

It now follows from Lemma 7.4 in [8] that there exist positive constants M and ω such that

$$\|S^N(t)\| \leq M e^{-\omega t}, \quad t \geq 0 \quad \text{for } N \geq N_0(\varepsilon). \quad \text{Q.E.D.}$$

The next lemma concerns the uniform boundedness of $\{\Pi^N\}$ in Theorem 5.1(i).

LEMMA 5.2. *Consider the system with the form*

$$(5.12) \quad \frac{d}{dt}x(t) = \sum_{i=0}^l A_i x(t-\theta_i) + Bu(t).$$

If the pair (A_0, B) is controllable and the range of B contains the range of A_i , $1 \leq i \leq l$, then $\{\Pi^N\}$ is a uniformly bounded sequence on Z .

Proof. For simplicity of exposition we consider the case, $l = 2$. The approximate solution $z^N(t) = (z^N(t, 0), z^N(t, \cdot)) \in Z$ of initial value problem (5.17) satisfies

$$(5.13) \quad \begin{aligned} \frac{d}{dt} z^N(t, 0) &= A_0 z^N(t, 0) + A_1 z^N(t, (-\theta_1)^+) + A_2 z^N(t, -r) + Bu(t), \\ \frac{\partial}{\partial t} z^N(t, \theta) &= \frac{\partial}{\partial \theta} z^N(t, \theta), \quad -r \leq \theta \leq 0 \end{aligned}$$

where the second equation holds in the sense of (3.3). Since (A_0, B) is controllable, then there exists an $m \times n$ matrix K such that the matrix $(A_0 - BK)$ has distinct negative real eigenvalues λ_i , $1 \leq i \leq n$ with $\max_{1 \leq i \leq n} \lambda_i \leq -3/2$. Since the eigenvalues of $(A_0 - BK)$ are distinct, there exists a nonsingular matrix P such that

$$P^{-1}(A_0 - BK)P = \Lambda = \text{diag}(\lambda_1, \dots, \lambda_n).$$

Let us consider the feedback control law to (4.1):

$$(5.14) \quad \tilde{u}(t) = -Kz^N(t, 0) - (B^T B)^{-1} B^T (A_1 z^N(t, (-\theta_1)^+) + A_2 z^N(t, -r)).$$

Then (5.13) has the closed loop equation

$$(5.15) \quad \begin{aligned} \frac{d}{dt} z^N(t, 0) &= (A_0 - BK)z^N(t, 0), \quad \frac{\partial}{\partial t} z^N(t, \theta) = \frac{\partial}{\partial \theta} z^N(t, \theta). \end{aligned}$$

If $\hat{z}^N(t) = (P^{-1}z^N(t, 0), P^{-1}z^N(t, \cdot))$, $t \geq 0$, then $\hat{z}^N(\cdot)$ satisfies

$$\frac{d}{dt} \hat{z}^N(t, 0) = \Lambda \hat{z}^N(t, 0), \quad \frac{\partial}{\partial t} \hat{z}^N = \frac{\partial}{\partial \theta} \hat{z}^N.$$

By using the same arguments given in the proof of Lemma 3.3, we obtain

$$\begin{aligned} \frac{1}{2} \frac{d}{dt} \|Q^N \hat{z}^N(t)\|_g^2 &\leq \langle \Lambda \hat{z}^N(t, 0), \hat{z}^N(t, 0) \rangle + |\hat{z}^N(t, 0)|^2 \\ &\quad - \frac{1}{2} |\hat{z}^N(t, (-\theta_1)^+)|^2 - \frac{1}{2} |\hat{z}^N(t, -r)|^2 \\ &\leq -\frac{1}{2} (|\hat{z}^N(t, 0)|^2 + |\hat{z}^N(t, (-\theta_1)^+)|^2 + |\hat{z}^N(t, -r)|^2) \end{aligned}$$

where we used the fact that $\Lambda \leq -3I/2$. Integration of this with respect to t yields

$$\|Q^N \hat{z}^N(t)\|_g^2 - \|Q^N \hat{z}^N(0)\|_g^2 \leq - \int_0^t (|\hat{z}^N(s, 0)|^2 + |\hat{z}^N(s, (-\theta_1)^+)|^2 + |\hat{z}^N(s, -r)|^2) ds$$

for all $t \geq 0$. Thus, for all $t \geq 0$,

$$\int_0^t |\hat{z}^N(s, 0)|^2 ds, \quad \int_0^t |\hat{z}^N(s, (-\theta_1)^+)|^2 ds,$$

and

$$\int_0^t |\hat{z}^N(s, -r)|^2 ds \leq \|Q^N \hat{z}^N(0)\|_g^2 \leq \|\hat{z}\|_g^2 \leq \sigma_{\max}(P^{-T}P^{-1}) \|z\|_g^2.$$

Since $\|z\|_g^2 \leq 2\|z\|_2^2$, it now follows from (5.14) and (5.15) that

$$\langle \Pi^N z, z \rangle \leq J^N(\tilde{u}, z) = \int_0^\infty (|Cz^N(t, 0)|^2 + |\tilde{u}(t)|^2) dt \leq \beta \|z\|_2^2$$

for some positive constant β . Since Π^N is nonnegative and selfadjoint, for $N \geq 1$, $\Pi^N \leq \beta I$. Q.E.D.

6. Numerical examples and conclusions. In this section, we discuss some numerical examples which demonstrates the feasibility of the Legendre-tau method for approximating the optimal feedback solution. We only consider examples of optimal control on the infinite interval. We solved the Riccati equation (4.5) for the matrix Σ^N using Potter's method. All computations were performed using MATLAB developed by Cleve Moler [12] which provides easy access to matrix software developed by LINPACK and EISPACK projects.

The N th feedback control is given by

$$(6.1) \quad u^N(t) = -B^T \left(\Pi_N^{00} x(t) + \int_{-r}^0 \Pi_N^{10}(\theta)^T x(t+\theta) d\theta \right)$$

where Π_N^{00} and Π_N^{10} are given in terms of the coefficients of Σ^N in Corollary 4.2. The strong convergence of Π^N to Π implies $\Pi_N^{00} \rightarrow \Pi^{00}$ and $\Pi_N^{10} \rightarrow \Pi^{10}$ in $L_2([-r, 0]; \mathbb{R}^{n \times n})$. We also discuss below how closely Π_N^{00} and Π_N^{10} approximate the conditions described in Theorem 2.5 and how closely the eigenvalues of the N th Hamiltonian matrix H^N in (4.6) approximate the closed-loop eigenvalues of $\mathcal{A} - \mathcal{B}\mathcal{B}^*\Pi$.

Example 6.1 (Gibson [8, Example 8.1]). Consider the scalar differential equation

$$(6.2) \quad \frac{d}{dt} x(t) = x(t) + x(t-1) + u(t);$$

the performance index of (2.6) is

$$(6.3) \quad J(u, (\eta, \phi)) = \int_0^\infty (x^2(t) + u^2(t)) dt.$$

For each N , Π_N^{00} is a scalar and $\Pi_N^{10}(\cdot)^T \in L_2([-1, 0]; \mathbb{R})$ and $B \equiv 1$ in (6.1). Table 1 shows the numerical results for Π_N^{00} and the expansion coefficients of Π_N^{10} , i.e.,

$$\Pi_N^{10}(\theta) = \sum_{k=0}^{N-1} a_k^N P_k(2\theta+1), \quad -1 \leq \theta \leq 0,$$

and how closely we have approximated the boundary condition (2.14).

TABLE 1

N	2	4	6	8
Π_N^{00}	2.8139	2.8094	2.8094	2.8094
$k=0$	1.44222	1.4267	1.4267	1.4267
1	-1.0844	-1.0438	-1.0438	-1.0438
2		0.2919	0.2919	0.2919
3		-0.0424	-0.0420	-0.0420
$\{a_k^N\}$ 4			0.0046	0.0046
5			-0.0004	-0.0004
6				2.3×10^{-5}
7				1.2×10^{-6}
$ \Pi_N^{00} - \Pi_N^{10}(-1) $	0.3074	0.0046	2.3×10^{-5}	5.3×10^{-8}

For comparison, the following are obtained using the average (AVE) scheme [8] and the linear spline (SPL) scheme [4].

$$\Pi_{74}^{00}(\text{AVE}) = 2.8130, \quad \Pi_{32}^{00}(\text{SPL}) = 2.8091.$$

Note that both schemes have not fully converged yet. However, for the Legendre-tau method, the result for $N = 4$ appears to give a fairly good approximation of the optimal feedback, e.g.,

$$|\Pi_4^{00} - \Pi_8^{00}| = 4.4 \times 10^{-7}, \quad \|\Pi_4^{10} - \Pi_8^{10}\|_{L_2[-1,0]} = 1.5 \times 10^{-3}.$$

Table 2 compares $\Pi_{74}^{10}(\theta)(\text{AVE})$ and $\Pi_4^{10}(\theta)(\text{L-T})$, where L-T denotes the Legendre-tau approximation.

TABLE 2

$\ \Pi_{74}^{10}(\theta)(\text{AVE}) - \Pi_4^{10}(\theta)(\text{L-T})\ _{L_2} = 1.9 \times 10^{-2}$		
θ	$\Pi_{74}^{10}(\theta)(\text{AVE})$	$\Pi_4^{10}(\theta)(\text{L-T})$
0.0	0.6435	0.6323
-0.1	0.7273	0.7225
-0.2	0.8258	0.8273
-0.3	0.9607	0.9519
-0.4	1.1023	1.1013
-0.5	1.2694	1.2807
-0.6	1.4965	1.4951
-0.7	1.7315	1.7497
-0.8	2.0480	2.0494
-0.9	2.3748	2.3994
-1.0	2.7541	2.8048

The oscillatory behavior exhibited by the spline approximation to Π^{10} [4] has not been observed for the Legendre-tau approximation.

Table 3 shows the eigenvalues λ_i^N of H^N which give the relatively small equation error $|\det \hat{\Delta}(\lambda_i^N)|$ where $\hat{\Delta}(\lambda)$ is given by (2.15), i.e., in this example

$$\hat{\Delta}(\lambda) = (\lambda - 1 - e^{-\lambda})(\lambda + 1 + e^{\lambda}) - 1.$$

In Table 3, the numbers inside () stand for the corresponding equation errors $|\det \hat{\Delta}(\lambda)|$ to the eigenvalues λ_i^N .

Example 6.2 (Gibson [8, Example 8.3]). We consider the problem of minimizing

$$(6.4) \quad J(u) = \int_0^\infty (y^2(t) + \dot{y}^2(t) + u^2(t)) dt,$$

subject to the harmonic oscillator with delayed restoring force and delayed damping given by

$$(6.5) \quad \frac{d^2}{dt^2} y(t) + \frac{d}{dt} y(t-1) + y(t-1) = u(t).$$

If we define $x(t) \in \mathbb{R}^2$ by

$$x(t) = \left(y(t), \frac{d}{dt} y(t) \right)^T,$$

TABLE 3

<i>N</i>		2	4
{λ _{<i>i</i>} ^{<i>N</i>} }	<i>i</i> = 1	−1.4032 (.019)	−1.4011 (1.9 × 10 ^{−6})
	2		−1.6351 ± 4.1627 <i>i</i> (.38)
<i>N</i>		6	8
{λ _{<i>i</i>} ^{<i>N</i>} }	<i>i</i> = 1	−1.4011 (3.2 × 10 ^{−11})	−1.4011 (2.4 × 10 ^{−15})
	<i>i</i> = 2	−1.6343 ± 4.1827 <i>i</i> (8.2 × 10 ^{−4})	−1.6343 ± 4.1827 <i>i</i> (4.5 × 10 ^{−7})
	<i>i</i> = 3		−2.4284 ± 10.6698 <i>i</i> (2.3)
<i>N</i>		16	
	<i>i</i> = 1	−1.4011 (1.2 × 10 ^{−14})	
	<i>i</i> = 2	−1.6344 ± 4.1827 <i>i</i> (6.7 × 10 ^{−14})	
	<i>i</i> = 3	−2.4256 ± 10.6890 <i>i</i> (2.1 × 10 ^{−10})	
	<i>i</i> = 4	−3.1695 ± 23.3811 <i>i</i> (4.5 × 10 ^{−4})	

then (6.4) and (6.5) are equivalent to

$$J(u; (\eta, \phi)) = \int_0^\infty (|x(t)|^2 + u^2(t)) \, dt$$

and

$$\frac{d}{dt}x(t) = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix} x(t) + \begin{bmatrix} 0 & 0 \\ -1 & -1 \end{bmatrix} x(t-1) + \begin{bmatrix} 0 \\ 1 \end{bmatrix} u(t),$$

respectively.

The optimal control in feedback form is

(6.6)

$$u(t) = -\Pi_{,21}^{00}x_1(t) - \Pi_{,22}^{00}x_2(t) - \int_{-1}^0 (\Pi_{,12}^{10}(\theta)x_1(t+\theta) + \Pi_{,22}^{10}(\theta)x_2(t+\theta)) \, d\theta$$

where $\Pi_{,i,j}^{00}$ and $\Pi_{,i,j}^{10}(\theta)$ are the (i, j) -elements of the matrix Π^{00} and $\Pi^{10}(\theta)$, respectively. The N th feedback control law is

(6.7)

$$u^N(t) = -\Pi_{N,21}^{00}x_1(t) - \Pi_{N,22}^{00}x_2(t) - \int_{-1}^0 (\Pi_{N,12}^{10}(\theta)x_1(t+\theta) + \Pi_{N,22}^{10}(\theta)x_2(t+\theta)) \, d\theta.$$

Note that if we define $\xi(t) \in \mathbb{R}^2$ by

$$\xi(t) = \left(y(t), \frac{d}{dt}y(t) + y(t) \right)^T,$$

then (6.4) and (6.5) are equivalent to

$$(6.8) \quad J(u; (\tilde{\eta}, \tilde{\phi})) = \int_0^\infty (\xi(t)^T Q \xi(t) + u^2(t)) dt,$$

with

$$Q = \begin{bmatrix} 2 & -1 \\ -1 & 1 \end{bmatrix},$$

and

$$(6.9) \quad \frac{d}{dt}\xi(t) = \begin{bmatrix} -1 & 1 \\ -1 & 1 \end{bmatrix} \xi(t) + \begin{bmatrix} 0 & 0 \\ 0 & -1 \end{bmatrix} \xi(t-1) + \begin{bmatrix} 0 \\ 1 \end{bmatrix} u(t),$$

respectively. Here, the initial conditions

$$(6.10) \quad \tilde{\eta} = (0, 0)^T \quad \text{and} \quad \xi_2(\theta) = 0, \quad -1 \leq \theta \leq 0$$

yield $\xi(t) \equiv 0$, $t \geq 0$, regardless of the initial functions $\xi_1(\theta)$, $-1 \leq \theta \leq 0$. Hence, for the initial conditions in (6.10) and any initial history $\xi_1(\cdot)$, the optimal control is $u(t) = 0$, $t \geq 0$. Therefore, the optimal control $u(t)$ must have the form

$$(6.11) \quad u(t) = -\tilde{\Pi}_{,21}^{00}\xi_1(t) - \tilde{\Pi}_{,22}^{00}\xi_2(t) - \int_{-1}^0 \tilde{\Pi}_{,22}^{10}(\theta)\xi_1(t+\theta) d\theta$$

where $\tilde{\Pi}$ corresponds to the minimization problem to (6.8) and (6.9).

Note that $\xi_2(t) = x_1(t) + x_2(t)$, $t \geq 1$. Hence, it follows from (6.6) and (6.11) that

$$\Pi_{,12}^{10} = \Pi_{,22}^{10}.$$

Similarly,

$$\Pi_{N,12}^{10} = \Pi_{N,22}^{10}, \quad N \geq 1.$$

Numerically, we have the results in Table 4.

TABLE 4

	$\Pi_2^{00} = \begin{bmatrix} 2.1407 & 1.2988 \\ 1.2988 & 1.8611 \end{bmatrix}$	$\Pi_4^{00} = \begin{bmatrix} 2.1387 & 1.2963 \\ 1.2963 & 1.8579 \end{bmatrix}$		
	$\Pi_6^{00} = \begin{bmatrix} 2.1387 & 1.2963 \\ 1.2963 & 1.8579 \end{bmatrix}$	$\Pi_8^{00} = \begin{bmatrix} 2.1387 & 1.2963 \\ 1.2963 & 1.8579 \end{bmatrix}$		
N	2	4	6	8
$k=0$	-0.8846	-0.8821	-0.8821	-0.8821
1	0.8971	0.8969	0.8969	0.8969
2	-0.0835	-0.0835	-0.0835	-0.0835
3	-0.0031	-0.0030	-0.0030	-0.0030
$\{a_k^N\}$	4	0.0014	0.0014	0.0014
5		-0.0001	-0.0001	-0.0001
6				2.4×10^{-6}
7				2.4×10^{-7}
$ \Pi_N^{10}(-1) - A_1^T \Pi_N^{00} $	0.2182	0.0024	1.3×10^{-5}	3.5×10^{-8}

In Table 4, $\{a_k^N\}_{k=0}^{N-1}$ are the expansion coefficients of $\Pi_{N,12}^{10}(\theta)$, i.e.,

$$\Pi_{N,12}^{10}(\theta) = \sum_{k=0}^{N-1} a_k^N P_k(2\theta + 1), \quad -1 \leq \theta \leq 0.$$

Moreover, we have

$$|\Pi_4^{00} - \Pi_8^{00}| = 6.3 \times 10^{-7}, \quad \|\Pi_{4,12}^{10} - \Pi_{8,12}^{10}\|_{L_2} = 4.8 \times 10^{-4}.$$

Again, one can see that the result for $N = 4$ gives a fairly good approximation. For comparison, the following are obtained by AVE and SPLINE schemes:

(AVE)
$$\Pi_{22}^{00}(\text{AVE}) = \begin{bmatrix} 2.1034 & 1.2574 \\ 1.2574 & 1.8123 \end{bmatrix},$$

(SPL)
$$\Pi_{16}^{00}(\text{SPL}) = \begin{bmatrix} 2.1389 & 1.2963 \\ 1.2963 & 1.8576 \end{bmatrix}.$$

Table 5 compares $\Pi_{22,1,2}^{10}(\theta)(\text{AVE})$ [8, p. 137] and $\Pi_{4,1,2}^{10}(\theta)(\text{L-T})$.

In this example, the closed-loop eigenvalues of $\mathcal{A} - \mathcal{B}\mathcal{B}^*\Pi$ are roots of the characteristic equation $\det \hat{\Delta}(\lambda) = 0$, where

$$\begin{aligned} \hat{\Delta}(\lambda) &= \begin{bmatrix} \lambda I - A_0 - e^{-\lambda} A_1 & -B B^T \\ -I & \lambda I + A_0 + e^{\lambda} A_1 \end{bmatrix}, \\ A_0 &= \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}, \quad A_1 = \begin{bmatrix} 0 & 0 \\ -1 & -1 \end{bmatrix}, \quad B = \begin{bmatrix} 0 \\ 1 \end{bmatrix}. \end{aligned}$$

Table 6 lists the eigenvalues λ_i^N of H^N which lies in the left half plane of C and give the relatively small equation error $|\det \hat{\Delta}(\lambda_i^N)|$.

Example 6.3. Here we deal with the equation which has multiple point delays

(6.12)
$$\frac{d}{dt} x(t) = x(t) + 2x(t-1) + x(t-2) + u(t),$$

with the cost functional

$$J(u, (\eta, \phi)) = \int_0^\infty (x^2(t) + u^2(t)) \, dt.$$

TABLE 5

$\ \Pi_{22,1,2}^{10}(\theta)(\text{AVE}) - \Pi_{4,1,2}^{10}(\theta)(\text{L-T})\ _{L_2} = 7.0 \times 10^{-2}$		
θ	$\Pi_{22,1,2}^{10}(\theta)(\text{AVE})$	$\Pi_{4,1,2}^{10}(\theta)(\text{L-T})$
0.0	-0.1152	-0.0719
-0.1	-0.2247	-0.2033
-0.2	-0.3449	-0.3462
-0.3	-0.4750	-0.5003
-0.4	-0.6147	-0.6652
-0.5	-0.7631	-0.8404
-0.6	-1.0013	-1.0257
-0.7	-1.1698	-1.2206
-0.8	-1.3455	-1.4247
-0.9	-1.5278	-1.6378
-1.0	-1.7160	-1.8593

TABLE 6

<i>N</i>		2	4
{λ _{<i>i</i>} ^{<i>N</i>} }	<i>i</i> = 1	1.3983 (0.0342)	−1.3893 (3.3 × 10 ^{−6})
	<i>i</i> = 2	−0.7358 ± 1.2207 <i>i</i> (.0277)	−0.7339 ± 1.2235 <i>i</i> (3.4 × 10 ^{−6})
<i>N</i>		6	8
{λ _{<i>i</i>} ^{<i>N</i>} }	<i>i</i> = 1	−1.3893 (5.4 × 10 ^{−11})	−1.3893 (3.9 × 10 ^{−14})
	<i>i</i> = 2	−0.7339 ± 1.2235 <i>i</i> (6.2 × 10 ^{−11})	−0.7339 ± 1.2235 <i>i</i> (2.1 × 10 ^{−14})
	<i>i</i> = 3	−2.0927 ± 7.4395 <i>i</i> (78.1)	−2.0890 ± 7.4619 <i>i</i> (.474)

For each *N*, the *N*th feedback control law is

$$u^N(t) = -\Pi_N^{00}x(t) - \int_{-2}^0 \Pi_N^{10}(\theta)x(t + \theta) \, d\theta$$

where Π_N^{00} is a scalar and $\Pi_N^{10}(\cdot) \in L_2([-2, 0]; R)$ is given by

$$\Pi_N^{10}(\theta) = \sum_{k=0}^N b_k^N P_k(2\theta + 3)\chi_{[-2, -1]}(\theta) + \sum_{k=0}^{N-1} a_k^N P_k(2\theta + 1)\chi_{(-1, 0]}(\theta), \quad -2 \leq \theta \leq 0.$$

TABLE 7

<i>N</i>	2	4	6	8
Π _{<i>N</i>} ⁰⁰	3.2159	3.2074	3.2073	3.2074
<i>b</i> ₀	1.5306	1.5246	1.5244	1.5243
<i>b</i> ₁	−1.220	−1.2205	−1.2214	−1.2216
<i>b</i> ₂	0.3295	0.3990	0.3972	0.3969
<i>b</i> ₃		−0.0583	−0.0590	−0.0595
<i>b</i> ₄		−0.0002	−0.0050	−0.0049
<i>b</i> ₅			−0.0008	−0.0001
<i>b</i> ₆			−0.0011	−0.0001
<i>b</i> ₇				−0.0005
<i>b</i> ₈				−0.0002
<i>a</i> ₀	3.3767	3.3911	3.3914	3.3914
<i>a</i> ₁	−2.8081	−2.6999	−2.7004	−2.7006
<i>a</i> ₂		0.8479	0.8477	0.8478
<i>a</i> ₃		−0.1119	−0.1092	−0.1094
<i>a</i> ₄			0.0083	0.0080
<i>a</i> ₅			−0.0018	−0.0009
<i>a</i> ₆				−0.0002
<i>a</i> ₇				0.0004
Π _{<i>N</i>} ¹⁰ (−2) − Π _{<i>N</i>} ⁰⁰	0.1352	0.0047	6.0 × 10 ^{−4}	6.2 × 10 ^{−5}
Π _{<i>N</i>} ¹⁰ ((−1) ⁺) − Π _{<i>N</i>} ¹⁰ ((−1) [−]) − 2Π ₀₀ ^{<i>N</i>}	0.8865	0.0090	2.0 × 10 ^{−4}	8.7 × 10 ^{−5}

TABLE 8

θ	$\Pi_2^{10}(\theta)$	$\Pi_4^{10}(\theta)$	$\Pi_6^{10}(\theta)$	$\Pi_8^{10}(\theta)$
-2.0	3.0807	3.2026	3.2067	3.2074
-1.9	2.6586	2.6892	2.6879	2.6877
-1.8	2.2761	2.2518	2.2492	2.2498
-1.7	1.9331	1.8834	1.8830	1.8831
-1.6	1.6297	1.5769	1.5769	1.5784
-1.5	1.3658	1.3252	1.3280	1.3277
-1.4	1.1415	1.1213	1.1229	1.1233
-1.3	0.9567	0.9583	0.9574	0.9581
-1.2	0.8114	0.8292	0.8266	0.8265
-1.0	0.6396	0.6451	0.6443	0.6440
-1.0	6.1848	7.0508	7.0587	7.0587
-0.9	5.6232	5.9500	5.9477	5.9482
-0.8	5.0615	5.0047	5.0025	5.0031
-0.7	4.4999	4.2014	4.2026	4.2025
-0.6	3.9383	3.5267	3.5303	3.5298
-0.5	3.3767	2.9671	2.9706	2.9704
-0.4	2.8150	2.5094	2.5103	2.5106
-0.3	2.2534	2.1399	2.1375	2.1378
-0.2	1.6918	1.8454	1.8413	1.8411
-0.1	1.1301	1.6123	1.6111	1.6107
0.0	0.5685	1.4272	1.4360	1.4362

TABLE 9

N	2	4
λ_1^N	-1.5217 (.2172)	-1.5174 (3.0×10^{-5})
λ_2^N	$0.9524 \pm 2.4826i$ (2.090)	$-0.9028 \pm 2.5445i$ (.0031)
λ_3^N		$-0.6103 \pm 5.0272i$ (.8349)
N	6	8
λ_1^N	-1.5174 (6.9×10^{-10})	-1.5174 (6.0×10^{-14})
λ_2^N	$-0.9029 \pm 2.5445i$ (7.0×10^{-7})	$-0.9029 \pm 2.5445i$ (4.6×10^{-11})
λ_3^N	$-0.5890 \pm 5.0114i$ (.0030)	$-0.5889 \pm 5.0114i$ (2.7×10^{-6})
λ_4^N	$-1.3588 \pm 8.7500i$ (10.18)	$-1.3159 \pm 8.7703i$ (0.1018)
λ_5^N		$-1.0595 \pm 11.4781i$ (4.108)

Table 7 shows the numerical results for Π_N^{00} and $\Pi_N^{10}(\theta)$ and how closely we have approximated the jump condition (2.13) and the boundary condition (2.14).

We have the function values of $\Pi_N^{10}(\theta)$ in Table 8. In this example, the closed loop characteristic equation is given by

$$\hat{\Delta}(\lambda) = (\lambda - 1 - 2e^{-\lambda} - e^{-2\lambda})(\lambda + 1 + 2e^{\lambda} + e^{2\lambda}) - 1 = 0.$$

Table 9 shows the eigenvalues of H^N in the same manner as before.

The numerical results presented here reveal that numerically one has strong convergence of Π^N for the Legendre-tau approximation. We have not proved the strong convergence of Π^N in the general case (except for single point delay case [10]). It requires a careful study of the asymptotic behavior of the spectra of \mathcal{A}^N . However, the efficiency of the numerical schemes is most important from the practical point of view. We observe, from the numerical results of this section, that the Legendre-tau method provides faster convergence and better approximation at low orders (i.e. small N) than the AVE and SPLINE schemes. In the above examples, the results corresponding to $N = 4$ give a fairly good approximation of the optimal feedback gain.

As further evidence of the usefulness of the Legendre-tau approximation, one can use it as an approximation technique for computing closed-loop eigenvalues of the feedback system. Note that eigenvalues close to the origin are approximated quite well at low orders on the above examples.

From these observations, we believe the Legendre-tau approximation scheme offers one of the favorable methods for construction of feedback gains. In future investigations, our efforts for constructing feedback gains for delay systems will be combined with the approach to finite-order compensator design for distributed parameter systems [17], developed by J. M. Schumacher to develop a design procedure for the construction of compensators for delay systems.

REFERENCES

- [1] A. V. BALAKRISHNAN, *Applied Functional Analysis*, 2nd ed., Springer-Verlag, Berlin, 1981.
- [2] H. T. BANKS AND J. A. BURNS, *Hereditary control problems: Numerical methods on averaging approximations*, this Journal, 16 (1978), pp. 169-208.
- [3] H. T. BANKS, J. A. BURNS AND E. M. CLIFF, *Parameter estimation and identification for systems with delays*, this Journal, 19 (1981), pp. 791-828.
- [4] H. T. BANKS, I. G. ROSEN AND K. ITO, *A spline-based technique for computing Riccati operators and feedback controls in regulator problems and feedback equations*, Institute for Computer Applications in Science and Engineering Report 82-31, NASA Langley Research Center, Hampton, VA, September 1982; SIAM J. Sci. Statist. Comput., 5 (1984), pp. 830-855.
- [5] C. BERNIER AND A. MANITIUS, *On semigroups in $\mathbb{R}^n \times L^p$ corresponding to differential equations with delays*, Canad. J. Math., 30 (1980), pp. 969-978.
- [6] J. G. BORISOVIC AND A. S. TURBABIN, *On the Cauchy problem for linear nonhomogeneous differential equations with retarded argument*, Soviet Math. Dokl., 10 (1969), pp. 401-405.
- [7] R. F. CURTAIN AND A. J. PRITCHARD, *Infinite dimensional linear system theory*, Lecture Notes in Control and Information Sci., 8, Springer-Verlag, Berlin, 1978.
- [8] J. S. GIBSON, *Linear-quadratic optimal control of hereditary differential systems: Infinite dimensional Riccati equations and numerical approximations*, this Journal, 21 (1983), pp. 95-139.
- [9] K. ITO AND R. TEGLAS, *Legendre-tau approximation for functional differential equations*, this Journal, 24 (1986), pp. 737-759.
- [10] K. ITO, *Legendre-tau approximation for functional differential equations, Part III: Eigenvalue approximations and uniform stability*, Lecture Notes in Control and Information Sci., 75; Distributed Parameter Systems Proc. 2nd International Conference, Vorau, Austria, 1984, pp. 191-212.
- [11] H. KWAKERNAK AND R. SIVAN, *Linear Optimal Control Systems*, Wiley-Interscience, New York, 1972.

- [12] C. MOLER, *Matlab Users' Guide*, Technical Report CS-81-1, Univ. of New Mexico, Dept. of Computer Science, August 1982.
- [13] A. PAZY, *Semigroups of Linear Operators and Applications to Partial Differential Equations*, Springer-Verlag, New York, 1983.
- [14] J. E. POTTER, *Matrix quadratic solutions*, SIAM J. Appl. Math., 14 (1966), pp. 496–501.
- [15] D. W. ROSS, *Controller design for time lag systems via a quadratic criterion*, IEEE Trans. Automat. Control, AC-16 (1971), pp. 644–672.
- [16] D. SALAMON, *Control and Observation of Neutral Systems*, Res. Notes in Math., 91, Pitman, London, 1984.
- [17] J. M. SCHUMACHER, *A direct approach to compensator design for distributed parameter systems*, this Journal, 21 (1983), pp. 823–836.
- [18] R. B. VINTER, *On the evolution of the state of linear differential delay equations in M^2 : Properties of the generator*, J. Inst. Math. Appl., 21 (1978), pp. 13–23.
- [19] ———, *Filter stability of stochastic evolution equations*, this Journal, 15 (1977), pp. 465–485.

STABILITY IN TWO-STAGE STOCHASTIC PROGRAMMING*

STEPHEN M. ROBINSON† AND ROGER J.-B. WETS‡

Abstract. We analyze the effect of changes in problem functions and/or distributions in certain two-stage stochastic programming problems with recourse. Under reasonable assumptions the locally optimal value of the perturbed problem will be continuous and the corresponding set of local optimizers will be upper semicontinuous with respect to the parameters (including the probability distribution in the second stage).

Key words. stochastic programming, recourse, stability, sensitivity analysis, weak convergence

AMS(MOS) subject classifications. 90C42, 90C31

1. Introduction. This paper analyzes the effect of perturbations in problem data on the optimal value and optimal policy set of certain two-stage stochastic programming problems. We show that under reasonable assumptions, such as are likely to be satisfied in practice, the optimal value will be continuous in the perturbations and the optimal policy set will be an upper semicontinuous multifunction. The problem data being perturbed may include the probability distribution of the right-hand side in the second-stage problem, and therefore a particular contribution of this paper is the identification of appropriate classes of probability distributions for which such stability results can be proved, as well as a suitable topology for those distributions.

Several previous works, including [3]–[5] and [12], have dealt with various aspects of stability in stochastic optimization, often by applying general theorems about stability in nonlinear optimization involving, e.g., differentiability conditions. We do not require any differentiability for the results of this paper, but rather employ the continuity results of [10]. Those results are also the basis for the recent work of Kall [8] of which the authors became aware after having completed the research for this paper. Although the approach of [8] differs somewhat in emphasis from that adopted here, the results for two-stage stochastic programming are closely related to ours.

In the remainder of this section we state precisely the problem that we shall consider, and the ways in which it can be perturbed. Then, in § 2, we identify an appropriate combination of topologies for a space of functions and a space of probability measures, so that the expectation operator will be a continuous function on the product of the two spaces. We then show how to apply this knowledge, together with results from [10], to analyze the stability of the stochastic programming problem under consideration. In § 3 we show how to apply these results to the commonly-used model of two-stage stochastic linear programming with complete fixed recourse, and we present two examples to show how the continuity results we obtain can yield information important in practical applications.

We shall be concerned here with minimizing (in x) the function defined by

$$(1.1) \quad c(p, x) + \int Q(p, x, \xi) P(d\xi),$$

* Received by the editors June 2, 1986; accepted for publication October 28, 1986. This work was sponsored by the National Science Foundation under grants DCR-8502202 and ECS-8542328, and the U.S. Army under contract DAAG29-80-C-0041. An earlier version of this work was presented by invitation at the IIASA Workshop on Numerical Methods for Stochastic Optimization, December 1983.

† Department of Industrial Engineering, University of Wisconsin-Madison, Madison, Wisconsin 53706.

‡ International Institute for Applied Systems Analysis (IIASA), and University of California, Davis, California 95616.

where p is a perturbation parameter, c is an extended real valued function, Q is a function of p, x and a random variable ξ , and P is a probability measure. Detailed technical assumptions about these objects will be stated in § 2.

Problem (1.1) is called a *two-stage* problem because the function Q is generally not explicitly prescribed, but instead is defined by means of an “inner” optimization problem. Often this is a linear programming problem of the form

$$(1.2) \quad Q(p, x, \xi) := \inf \{ \langle h^*, y \rangle \mid Wy = t(p, x, \xi), y \geq 0 \}.$$

This class of problems has a long history; see e.g. [2], [7], [13] and further references therein.

It is clear from the form of (1.1) that in order to deal with our problem we have to gain information about how an expectation behaves when both the function involved, and the probability measure with respect to which the expectation is taken, can vary. We then have to apply this information to analyze continuity of the optimal value and optimal solutions of (1.1) as functions of the pair (p, P) . This analysis is done in the next section.

2. Continuity analysis. Let \mathcal{P} be a family of probability measures on $(\mathbb{R}^m, \mathcal{B})$, where \mathcal{B} is the σ -algebra of Borel sets, and let \mathcal{F} be a family of continuous real-valued functions on \mathbb{R}^m . In this section we first exhibit appropriate conditions on, and topologies for, \mathcal{F} and \mathcal{P} so that the quantity

$$\langle f, P \rangle := \int f(\xi) P(d\xi)$$

will be a well-defined, continuous function on $\mathcal{F} \times \mathcal{P}$. We then use these conditions to analyze the stability of the optimization problem (1.1).

For our purposes convenient topologies will be those of uniform convergence on compact sets for \mathcal{F} and weak convergence for \mathcal{P} . The latter is defined as follows: if $Z(\mathbb{R}^m)$ represents the class of all probability measures on $(\mathbb{R}^m, \mathcal{B})$, if $\{P_n\} \subset Z(\mathbb{R}^m)$ and $P \in Z(\mathbb{R}^m)$, then we say P_n *converges weakly to* P (written $P_n \Rightarrow P$) if $\langle f, P_n \rangle$ converges to $\langle f, P \rangle$ for every f belonging to the class $C(\mathbb{R}^m)$ of bounded, continuous real-valued functions on \mathbb{R}^m . This notion of weak convergence amounts to convergence in the topological space $(Z(\mathbb{R}^m), \tau)$, where the topology τ has for its base at an element $P \in Z(\mathbb{R}^m)$ the sets of the form

$$\{s \in Z(\mathbb{R}^m) \mid |\langle f_i, s \rangle - \langle f_i, P \rangle| < \varepsilon, i = 1, \dots, k\},$$

where $\varepsilon > 0$, k is a positive integer, and the f_i all belong to $C(\mathbb{R}^m)$. For more information about weak convergence, see [1].

To see why we need some restrictions on \mathcal{F} and \mathcal{P} to make $\langle \cdot, \cdot \rangle$ continuous, consider the example in which P_0 is the probability measure on \mathbb{R} having mass 1 at $x = 0$, and P is that having mass $6/(\pi k)^2$ at $x = k$ for $k = 1, 2, \dots$, (P is a probability measure because $\sum_{k=1}^{\infty} k^{-2} = \pi^2/6$). For $n = 1, 2, \dots$, let $P_n := (1 - n^{-1})P_0 + n^{-1}P$. Now for any bounded function f on \mathbb{R} , one has

$$|\langle f, P_n \rangle - \langle f, P_0 \rangle| \leq 2\beta n^{-1}$$

where β is the bound on f . Therefore $P_n \Rightarrow P_0$. However, with $g(x) := x$ we have $\langle g, P_n \rangle = +\infty$ for each n , but $\langle g, P_0 \rangle = 0$. Thus even $\langle g, \cdot \rangle$ for a fixed g is not continuous.

The difficulty in this example arose from the fact that, intuitively speaking, P_n had too much mass in regions where g was large. To avoid this problem we shall use a slight extension of the idea of uniform integrability [1]. If \mathcal{P} is a family of probability measures on $(\mathbb{R}^m, \mathcal{B})$ we shall say that the family \mathcal{F} is *uniformly integrable* with respect

to the family \mathcal{P} if for each $\varepsilon > 0$ there is a compact set K in \mathbb{R}^m such that for each $f \in \mathcal{F}$ and each $P \in \mathcal{P}$,

$$\int_{cK} |f(\xi)| P(d\xi) < \varepsilon$$

where cK denotes the complement of K . With this definition we can then prove the following theorem.

THEOREM 2.1. *Let \mathcal{F} and \mathcal{P} be endowed with the topologies of uniform convergence on compact sets and weak convergence, respectively. If \mathcal{F} is uniformly integrable with respect to \mathcal{P} , then $\langle \cdot, \cdot \rangle$ is continuous on $\mathcal{F} \times \mathcal{P}$.*

Proof. Let $f_0 \in \mathcal{F}$ and $P_0 \in \mathcal{P}$, and choose $\varepsilon > 0$. We shall exhibit neighborhoods U of f_0 in \mathcal{F} and V of P_0 in \mathcal{P} , such that if $(f, P) \in U \times V$ then $|\langle f, P \rangle - \langle f_0, P_0 \rangle| < 4\varepsilon$.

First, use the hypothesis to find a compact set K such that for each $f \in \mathcal{F}$ and each $P \in \mathcal{P}$, $\int_{cK} |f(\xi)| P(d\xi) < \varepsilon$. Let $d(\xi, K) := \inf \{\|\xi - k\| \mid k \in K\}$, the point-to-set distance function, and let $K_1 := \{\xi \in \mathbb{R}^m \mid d(\xi, K) \leq 1\}$. Define $\theta(\xi) := \max \{0, 1 - d(\xi, K)\}$; note that θ is a continuous function taking the values 1 on K and 0 on cK_1 . Now define

$$U := \{f \in \mathcal{F} \mid \sup \{|\langle f(\xi) - f_0(\xi) | \xi \in K_1 \rangle| \} < \varepsilon\};$$

this is a neighborhood of f_0 in the topology of \mathcal{F} . Let

$$V := \{P \in \mathcal{Z}(\mathbb{R}^m) \mid |\langle f_0 \theta, P \rangle - \langle f_0 \theta, P_0 \rangle| < \varepsilon\};$$

note that this is a neighborhood of P_0 since $(f_0 \theta)(\xi) := f_0(\xi) \theta(\xi)$ is a continuous bounded function. Choose $(f, P) \in U \times V$ and write

$$(2.1) \quad \begin{aligned} |\langle f, P \rangle - \langle f_0, P_0 \rangle| &\leq |\langle f \theta - f_0 \theta, P \rangle| + |\langle f_0 \theta, P \rangle - \langle f_0 \theta, P_0 \rangle| \\ &\quad + |\langle f(1 - \theta), P \rangle| + |\langle f_0(1 - \theta), P_0 \rangle|. \end{aligned}$$

The first term on the right in (2.1) is less than ε since $|f(\xi) \theta(\xi) - f_0(\xi) \theta(\xi)| \leq |f(\xi) - f_0(\xi)|$ and since $f \in U$; recall that θ is zero off K_1 , and P is a probability measure. The second term is less than ε by definition of V , and the last two terms are each less than ε because $(1 - \theta)$ is zero on K so that those terms represent integrals over cK . This completes the proof.

With the continuity result of Theorem 2.1, we are able to analyze the stability properties of the optimization problem (1.1). For a family \mathcal{P} of probability measures on $(\mathbb{R}^m, \mathcal{B})$ and a topological space Π , let Q be a real-valued function on $\Pi \times \mathbb{R}^n \times \mathbb{R}^m$. For $p \in \Pi$, $x \in \mathbb{R}^n$ and $P \in \mathcal{P}$, let

$$(2.2) \quad I(p, P, x) := \int Q(p, x, \xi) P(d\xi),$$

provided that the integral exists and is not $-\infty$. Then the optimization problem described in § 1 can be expressed as

$$(2.3) \quad \inf_x h(p, P, x)$$

where

$$(2.4) \quad h(p, P, x) := c(p, x) + I(p, P, x)$$

and where c is a given extended-real-valued function from $\Pi \times \mathbb{R}^n$ to $(-\infty, +\infty]$.

We shall establish continuity properties of (2.3) by using results from [10]. Since in general the function $h(p, P, \cdot)$ may not be convex, we deal with local minimizers; in the convex case there will of course also be global minimizers. To specify precisely the objects we shall study, we recall from [10] the concept of *complete local minimizing set* (CLM set): for given values $p_0 \in \Pi$ and $P_0 \in \mathcal{P}$, a set $M \subset \mathbb{R}^n$ is called a CLM set for $h(p_0, P_0, \cdot)$ with respect to an open set $G \subset \mathbb{R}^n$, if (i) $M \subset G$, and (ii) the set of

minimizers of $h(p, P_0, \cdot)$ on $\text{cl } G$ is M . Roughly speaking, to say M is a CLM set is to say that it contains all of the nearby minimizers of the function being studied.

We shall suppose that an open bounded set G has been found such that if for $(p, P) \in \Pi \times \mathcal{P}$ we define

$$(2.5) \quad \theta(p, P) := \inf \{h(p, P, x) \mid x \in \text{cl } G\}$$

and

$$(2.6) \quad \Theta(p, P) := \{x \in \text{cl } G \mid h(p, P, x) = \theta(p, P)\},$$

then $\Theta(p_0, P_0)$ is a CLM set for $h(p_0, P_0, \cdot)$ with respect to G . We shall then study the behavior of θ and Θ as the pair (p, P) varies near (p_0, P_0) .

To use the results of [10] we need to introduce the additional idea of epi-upper semicontinuity (epi-usc) at a point. We say that the function c appearing in (2.4) is epi-usc at $x_0 \in \mathbb{R}^n$ (as $p \rightarrow p_0$) if

$$c(p_0, x_0) \geq \sup_{V \in \mathcal{N}(x_0)} \limsup_{p \rightarrow p_0} \inf_{x \in V} c(p, x),$$

where $\mathcal{N}(x_0)$ denotes the neighborhood system of x_0 . The property of epi-usc at x_0 holds in particular if $c(\cdot, x_0)$ is usc at p_0 . However, it can hold in many other situations too, including some in which $c(p, x_0) = +\infty$ for $p \neq p_0$ but $c(p_0, x_0)$ is finite, so that $c(\cdot, x_0)$ is obviously not usc at p_0 . For more details see [10].

The next theorem employs some terms that we now define. An extended-real-valued function f is *proper* if it never takes $-\infty$ and it is not identically $+\infty$. The *effective domain* of f is the set of arguments for which f is not $+\infty$. Finally, the multivalued function Θ is *Berge-usc* at (p_0, P_0) if for each open set W in \mathbb{R}^n with $W \supset \Theta(p_0, P_0)$, there is a neighborhood U of (p_0, P_0) in $\Pi \times \mathcal{P}$ such that for each $(p, P) \in U$, $\Theta(p, P) \subset W$.

THEOREM 2.2. *Let Q and P be as previously defined, let I be defined by (2.2) and h by (2.4), and let G be an open bounded set in \mathbb{R}^n . Assume the following hypotheses:*

- (i) Q is continuous on $\Pi \times (\text{cl } G) \times \mathbb{R}^m$.
- (ii) The collection $\mathcal{F} := \{Q(p, x, \cdot) \mid p \in \Pi, x \in \text{cl } G\}$ is uniformly integrable with respect to the family \mathcal{P} .
- (iii) The function c is lsc on $\Pi \times \text{cl } G$, and is epi-usc at some point $x_0 \in \Theta(p_0, P_0)$, with $c(p_0, x_0)$ finite.
- (iv) $\Theta(p_0, P_0)$ is a CLM set for $h(p_0, P_0, \cdot)$ with respect to G .

Then we have

- (a) $\theta(p_0, P_0)$ is finite and θ is continuous at (p_0, P_0) .
- (b) Θ is Berge-usc at (p_0, P_0) .
- (c) There is a neighborhood U of (p_0, P_0) such that for each $(p, P) \in U$, the function $h(p, P, \cdot)$ restricted to $\text{cl } G$ is proper and $\Theta(p, P)$ is a nonempty, compact CLM set for $h(p, P, \cdot)$ with respect to G .

Proof. We apply Theorem 4.3 of [10] to problem (2.3) to obtain conclusions (a)–(c). Thus the only proof required is verification that the hypotheses needed in [10] hold for (2.3). We list these hypotheses below, indicating for each why it holds here.

- (1) $G \cap \text{dom } h(p_0, P_0, \cdot) \neq \emptyset$: finiteness of $c(p_0, x_0)$ (Hypothesis (iii)) and the fact that I is continuous on $\mathcal{F} \times \mathcal{P}$ (Theorem 2.1 with Hypotheses (i) and (ii)).
- (2) $\Theta(p_0, P_0)$ is a CLM set with respect to G : Hypothesis (iv).
- (3) h is epi-usc at $x_0 \in \Theta(p_0, P_0)$ as $(p, P) \rightarrow (p_0, P_0)$: Proposition 2.8 of [10] together with epi-usc of c at x_0 (Hypothesis (iii)), continuity of I on $\mathcal{F} \times \mathcal{P}$ (see (1) above), and continuity of Q (Hypothesis (i)).

(4) h is lsc on $\Pi \times \mathcal{P} \times \text{cl } G$: lsc of c on $\Pi \times \text{cl } G$ (Hypothesis (iii)) with continuity of I on $\mathcal{F} \times \mathcal{P}$ and continuity of Q .

This completes the proof of Theorem 2.2.

Theorem 2.2 gives a general continuity result for (1.1) under Hypotheses (i)–(iv). In the next section we specialize this result to two-stage stochastic linear programming with complete fixed recourse, showing how to verify the hypotheses in that case.

3. Application to stochastic linear programming. We are concerned in this section with the particular case of (1.1) for which

$$(3.1) \quad Q(p, x, \xi) = \inf_y \{ \langle h^*, y \rangle \mid Wy = t(p, x, \xi), y \geq 0 \},$$

where W is a linear transformation from \mathbb{R}^k to \mathbb{R}^l , $h^* \in \mathbb{R}^k$, and t is a function from $\Pi \times \text{cl } G \times \mathbb{R}^m$ to \mathbb{R}^l , with G an open set in \mathbb{R}^n (the same G as in Theorem 2.2). The special form of Q in (3.1) reflects a situation in which realizations of the random variable ξ lead to adjustment processes whose costs can be modeled by linear programming. The sum of the expected adjustment (or recourse) cost and the cost $c(p, x)$ of adopting the policy x then comprises the total cost $h(p, P, x)$ to be minimized. We shall first show how to apply Theorem 2.2 to (3.1), and then discuss two practical examples in which the conclusions of that theorem yield important information.

Recall that in order to apply Theorem 2.2 to (3.1) we need to know that Q is continuous on $\Pi \times (\text{cl } G) \times \mathbb{R}^m$, that the collection

$$\mathcal{F} := \{ Q(p, x, \cdot) \mid p \in \Pi, x \in \text{cl } G \}$$

is uniformly integrable with respect to a given family \mathcal{P} of probability measures, and that for some fixed elements $p_0 \in \Pi$ and $P_0 \in \mathcal{P}$, the set $\Theta(p_0, P_0)$ defined by (2.5) is a CLM set for $h(p_0, P_0, \cdot)$ with respect to G . There is also a continuity condition on c , detailed in Hypothesis (iii) of Theorem 2.2. Here we concentrate on the first two requirements with the form of Q given by (3.1).

In all of our analysis of (3.1) we shall assume the *complete recourse condition*

$$(3.2) \quad \{ Wy \mid y \geq 0 \} = \mathbb{R}^l.$$

This condition says that no matter what value t assumes in (3.1), the programming problem has a feasible point. We shall also assume that the dual of that problem is feasible:

$$(3.3) \quad \{ u^* \mid W^* u^* \leq h^* \} \neq \emptyset.$$

If we define a function v on \mathbb{R}^l by

$$(3.4) \quad v(z) := \inf_y \{ \langle h^*, y \rangle \mid Wy = z, y \geq 0 \},$$

then v has the following (well-known) properties.

LEMMA 3.1. *Under Hypotheses (3.2) and (3.3), the function v defined by (3.4) is a finite, Lipschitzian, positively homogeneous convex function on \mathbb{R}^l , and for each z there is a $y(z)$ attaining the infimum in (3.4).*

Proof. The existence of $y(z)$ follows from the elementary duality theory of linear programming. This shows that v is finite everywhere. It is clearly convex and positively homogeneous, so we need only prove Lipschitz continuity. Since the cone on the left side of (3.2) is assumed to be \mathbb{R}^l , its polar $\{ u^* \mid W^* u^* \leq 0 \}$ must be the origin. But this polar is the recession cone of the dual feasible set on the left side of (3.3), so that set (for any fixed h^*) is bounded [11, Thm. 8.4]. However, the conjugate function v^* is just the indicator of that set: thus v^* has a bounded effective domain, so v is Lipschitzian [11, Cor. 13.3.3]. This proves Lemma 3.1.

We can now state conditions for the properties of Q in which we are interested.

THEOREM 3.2. Assume that (3.2) and (3.3) hold, and that in addition

(a) t is continuous on $\Pi \times (\text{cl } G) \times \mathbb{R}^m$, and

(b) The collection $\{\|t(p, x, \cdot)\| \mid p \in \Pi, x \in \text{cl } G\}$ is uniformly integrable with respect to \mathcal{P} .

Then the function Q given by (3.1) satisfies Hypotheses (i) and (ii) of Theorem 2.2.

Proof. Continuity of Q is obvious from assumption (a) and Lemma 3.1. For uniform integrability we note that since $v(0) = 0$, we have for any z

$$(3.5) \quad |v(z)| = |v(z) - v(0)| \leq \lambda \|z\|$$

where λ is the Lipschitz constant given by Lemma 3.1. As Q is the composition of v with t , its uniform integrability follows from assumption (b) and (3.5). This completes the proof.

In practice one often finds t given by an expression of the form

$$(3.6) \quad t(p, x, \xi) = a(p, x) + S(p)\xi + T(p, \xi)x$$

where $a: \Pi \times \text{cl } G \rightarrow \mathbb{R}^l$, $S: \Pi \rightarrow \mathcal{L}(m, l)$ (the space of linear transformations from \mathbb{R}^m to \mathbb{R}^l) and $T: \Pi \times \mathbb{R}^m \rightarrow \mathcal{L}(n, l)$. If all of these functions are continuous then assumption (a) of Theorem 3.2 is satisfied. Further, if Π and $\text{cl } G$ are compact, if T satisfies a growth condition of the form

$$\|T(p, \xi)\| \leq \alpha \|\xi\|^\gamma$$

where $\gamma \geq 1$, and if the members of \mathcal{P} happens to come from a family of density functions dominated for $\|\xi\| \geq \delta$ by some function of the form $\beta \|\xi\|^{-\eta}$, where β and δ are any positive numbers and $\eta > \gamma + m$, then assumption (b) will also hold. Many common probability density functions fall into this category.

We conclude by considering two examples. Although these are mostly for illustrative purposes, the first of them is of great importance in the everyday use of mathematical programming techniques to analyze decision making in an uncertain environment.

Example 3.3. Scenario analysis. When mathematical programming models are used as a tool to help in decision making at the policy level (see for example [6], [9]) the “practitioner” approach to handling uncertainty is through reliance on a technique known as *scenario analysis*. Typically, the situation is as follows: a policy must be laid out for the next T years. It corresponds to a sequence of decisions x^1, \dots, x^T , with x^t the decision in period t . In the PILOT model [9], these decisions are related to energy planning and involve such questions as the number and the type of generating plants to support or to start in order to supply the energy needed for industrial and individual consumption. The model is usually of the type

$$(3.7) \quad \begin{aligned} & \text{minimize} && \sum_{t=1}^T f_t(x^t) \\ & \text{subject to} && \sum_{j=1}^l A_{tj} x^j = b_t, \quad t = 1, \dots, T, \\ & && x^t \in C_t \subset \mathbb{R}^n, \quad t = 1, \dots, T, \end{aligned}$$

assuming that the dynamics are linear (the A_{tj} are m by n matrices, $b_t \in \mathbb{R}^m$). The constraints $x^t \in C_t$ represent the limitations imposed on the possible alternatives available in period t . The vectors b_t include the projections for supply and demand in period t . The associated costs (not necessarily purely monetary) are modeled by introducing the objective function

$$f(x) := \sum_{t=1}^T f_t(x^t)$$

that may, as is the case here, or may not be time-separable. Of course, the only decision that is crucial is x^1 . The subsequent decisions x^2, \dots, x^T are only included in the model to measure the implications that a poor or good choice of x^1 may have on the remainder of the planning period. In some sense it would be more appropriate to formulate the problem as

$$(3.8) \quad \begin{aligned} & \underset{x^1 \in C_1}{\text{minimize}} && f_1(x^1) + R(x^1) \\ & \text{subject to} && A_{11}x^1 = b_1 \end{aligned}$$

where

$$(3.9) \quad R(x^1) = \inf_{\substack{x^2, \dots, x^T \\ x^t \in C_t}} \left\{ \sum_{t=2}^T f_t(x^t) \mid \sum_{j=2}^t A_{tj}x^j = b_t - A_{t1}x^1, t=2, \dots, T \right\}.$$

If demands and supplies may be sufficiently well known for period (or stage) 1, so that we may consider b^1 to be given, there is usually much uncertainty as to the values to assign to b_2, \dots, b_T ; obviously we are painfully hampered by lack of data to make reliable predictions for the faraway future. To overcome this obstacle, one technique is to rely on scenario analysis. Instead of solving the optimization problem (3.8) for one choice of

$$\xi := (b_2, \dots, b_T),$$

the problem is solved for a number of possible scenarios

$$\xi^l := (b_2^l, \dots, b_T^l) \quad \text{for } l = 1, \dots, L,$$

generating a collection of solutions $\{x^{1l}, l = 1, \dots, L\}$. These solutions are then analyzed: how does solution x^{1l} “hold up” if instead of scenario ξ^l , the actual outcome turns out to be ξ^k , with $k \neq l$; what is the “best”/“worst” solution; how should one combine the “optimal” solutions to reach a desirable mix; and so on. This involves solving problem (3.8) a very large number of times and nonetheless being left with no more than an educated guess about a potentially good choice of x^1 .

A second approach is to rely on the model of stochastic programming with recourse to analyze such a situation. As before, we accept the future as describable by a range of scenarios, that we again denote by ξ^1, \dots, ξ^L . But this time we attach probabilities p_1, \dots, p_L to these scenarios and instead of solving (3.8) for each possible scenario ξ^1, \dots, ξ^L , we solve the following stochastic optimization problem:

$$(3.10) \quad \begin{aligned} & \underset{x^1 \in C_1}{\text{minimize}} && f_1(x^1) + Q(x^1) \\ & \text{subject to} && A_{11}x^1 = b_1, \end{aligned}$$

with

$$(3.11) \quad \begin{aligned} Q(x_1) &= E \left\{ \inf_{\substack{x^2, \dots, x^T \\ x^t \in C_t}} \sum_{t=2}^T f_t(x^t) \mid \sum_{j=2}^t A_{tj}x^j = \mathbf{b}_t - A_{t1}x^1, t=2, \dots, T \right\} \\ &= \inf_{x^1 \in C_1} \left\{ \sum_{l=1}^L p_l \sum_{t=2}^T f_t(x^{1l}) \mid \sum_{j=2}^t A_{tj}x^{1l} = b_t^l - A_{t1}x^1, t=2, \dots, T; l=1, \dots, L \right\}; \end{aligned}$$

we write \mathbf{b}_t in boldface to suggest that we now view it as a random vector whose possible realizations are b_t^1, \dots, b_t^L . The solution generated by (3.10) takes into account not just one scenario, but *all* of them and if one can argue that the (somewhat arbitrarily) assigned probabilities p_1, \dots, p_L do correspond to the best available information, then the optimal solution x^1 would provide a reliable basis for arriving at a decision in period 1.

The advantage of relying on a stochastic programming model, instead of on scenario analysis, does not stop at the observation that a solution of (3.10) will take into account all eventualities, whereas each individual solution of (3.8) only takes into account one scenario. In fact Theorem 2.2 provides the convincing argument. It informs us that, under perturbations of the probability measure that we have assigned somewhat arbitrarily to the scenarios, the set of solutions to (3.10) as it depends on these "probabilities" is "continuous" (in the sense of upper semicontinuity). Hence a parametric analysis of the behavior of x^1 , with respect to p_1, \dots, p_L , will reveal a "continuous" behavior that allows us to zero in on the stable characteristics of the optimal solutions. Scenario analysis only considers as possible values for the vector (p_1, \dots, p_L) the extreme points of the simplex of probabilities, and thus it is very difficult, if not impossible, to recognize the built-in stabilities.

Example 3.4. Approximation schemes. To solve stochastic programming problems one must usually rely on a discretization of the probability measure, in order to be able to calculate the integral that appears in the objective function (for a general discussion cf. [14]). The discretization is refined, in the appropriate fashion, to guarantee the convergence of the solutions of the approximating problems to the solution of the given problem. Theorem 2.2 allows us to work out the conditions that would guarantee local convergence. Earlier results [13, Thm. 3.9] required global convergence and were, for all practical purposes, only useful in the convex setting. Similarly, Theorem 2.2 allows us to deal with a situation when we have only partial information about the probability measure associated with the random elements of the problem. We can think of this last case as a refined version of "scenario analysis."

REFERENCES

- [1] P. BILLINGSLEY, *Convergence of Probability Measures*, John Wiley, New York, 1968.
- [2] M. DEMPSTER, *Introduction to stochastic programming*, in Stochastic Programming, M. Dempster, ed., Academic Press, New York, 1980, pp. 3-59.
- [3] J. DUPAČOVÁ, *Stability in stochastic programming with recourse*, Acta Univ. Carolin.—Math. Phys., 24 (1983), pp. 23-34.
- [4] ———, *Stability in stochastic programming with recourse—estimated parameters*, Math. Programming, 28 (1984), pp. 72-83.
- [5] ———, *Stability in stochastic programming with recourse. Contaminated distributions*, preprint, Charles Univ., Prague, Czechoslovakia, December 1983.
- [6] T. HIGGINS AND H. JENKINS-SMITH, *Analysis of the economic effect of the Alaskan oil export ban*, Oper. Res., 33 (1985), pp. 1173-1202.
- [7] P. KALL, *Stochastic Linear Programming*, Springer-Verlag, Berlin, 1976.
- [8] ———, *On approximations and stability in stochastic programming*, Report, Institut für Operations Research der Universität Zürich, Zürich, Switzerland, February 1986; in Parametric Optimization and Related Topics, J. Guddat, ed., Akademie-Verlag, Berlin, 1987, to appear.
- [9] P. MCALLISTER, J. STONE, G. DANTZIG AND B. AVI-ITZHAK, *The PILOT 1983 model*, Technical Report SOL 85-12, Systems Optimization Laboratory, Dept. of Operations Research, Stanford Univ., Stanford, CA, 1985.
- [10] S. M. ROBINSON, *Local epi-continuity and local optimization*, Math. Programming, 37 (1987), pp. 208-222.
- [11] R. T. ROCKAFELLAR, *Convex Analysis*, Princeton Univ. Press, Princeton, NJ, 1970.
- [12] J. WANG, *Distribution sensitivity analysis for stochastic programs with complete recourse*, Math. Programming, 31 (1985), pp. 286-297.
- [13] R. J.-B. WETS, *Stochastic programming: solution techniques and approximation schemes*, in Mathematical Programming, Bonn 1982: The State of the Art, A. Bachem, M. Grötschel and B. Korte, eds., Springer-Verlag, Berlin, 1983, pp. 566-603.
- [14] ———, *Algorithmic procedures for stochastic optimization*, in Computational Mathematical Programming, K. Schittkowski, ed., Springer-Verlag (NATO ASI Series), Berlin, 1985, pp. 309-322.

BOUNDARY CONTROL OF THE TIMOSHENKO BEAM*

JONG UHN KIM† AND YURIKO RENARDY‡

Abstract. It is shown that the Timoshenko beam can be uniformly stabilized by means of a boundary control. A numerical study on the spectrum is also presented.

Key words. Timoshenko beam, uniform stabilization, exponential decay, boundary control, energy method, C_0 -semigroup, linear stability, eigenvalues, spectral method

AMS(MOS) subject classifications. 35B37, 35L15, 73K05, 93C20, 93D15, 65F15, 65N25

0. Introduction. The purpose of this paper is to investigate uniform stabilization of the Timoshenko beam with boundary control. The motion of a beam can be described by the Euler beam equation when the cross-sectional dimensions are small in comparison with the length of the beam. If the cross-sectional dimensions are not negligible, the effect of the rotatory inertia should be considered and the motion is better described by the Rayleigh beam equation. If the deflection due to shear is also taken into account in addition to the rotatory inertia, we arrive at a still more accurate model, which is called the Timoshenko beam. Its motion is described by the following system of equations:

$$(0.1) \quad \rho \frac{\partial^2 w}{\partial t^2} - K \frac{\partial^2 w}{\partial x^2} + K \frac{\partial \phi}{\partial x} = 0,$$

$$(0.2) \quad I_p \frac{\partial^2 \phi}{\partial t^2} - EI \frac{\partial^2 \phi}{\partial x^2} + K \left(\phi - \frac{\partial w}{\partial x} \right) = 0.$$

Here, t is the time variable and x is the space coordinate along the beam in its equilibrium position. We denote by $w(x, t)$ the deflection of the beam from the equilibrium line, which is described by $w = 0$, and by $\phi(x, t)$ the slope of the deflection curve when the shearing force is neglected; for the precise meaning of ϕ , see Timoshenko [11] or Traill-Nash and Collar [12]. We assume that the motion occurs in the w - x -plane and that $0 \leq x \leq L$. The coefficients ρ , I_p , E and I are the mass per unit length, the mass moment of inertia of the cross section, Young's modulus and the moment of inertia of the cross section, respectively. The coefficient K is equal to kGA , where G is the modulus of elasticity in shear, A is the cross sectional area and k is a numerical factor depending on the shape of the cross section. The boundary condition we employ at $x = 0$ is

$$(0.3) \quad w(0, t) = 0, \quad \phi(0, t) = 0,$$

which is for the clamped end at $x = 0$, and the boundary control at $x = L$ is of the form

$$(0.4) \quad K\phi(L, t) - K \frac{\partial w}{\partial x}(L, t) = \alpha \frac{\partial w}{\partial t}(L, t),$$

$$(0.5) \quad EI \frac{\partial \phi}{\partial x}(L, t) = -\beta \frac{\partial \phi}{\partial t}(L, t)$$

* Received by the editors January 27, 1986; accepted for publication November 9, 1986.

† Department of Mathematics, Virginia Polytechnic Institute and State University, Blacksburg, Virginia 24061. The work of this author was supported by the Air Force Office of Scientific Research under grant AFOSR-86-0085 and by the National Science Foundation grant DMS-8521848.

‡ Department of Mathematics, Virginia Polytechnic Institute and State University, Blacksburg, Virginia 24061. The work of this author was supported by National Science Foundation grant DMS-8615203 under the National Science Foundation Research Opportunities for Women Program.

where α and β are positive constants depending on the control device. This boundary control corresponds to a control mechanism which monitors $\partial w/\partial t$ and $\partial \phi/\partial t$ at $x = L$ and transforms them into the lateral force and moment applied at $x = L$, respectively. Russell [10] and Washizu [13] derived (0.1) and (0.2) through the energy principle by using the natural energy of the beam given by

$$(0.6) \quad \varepsilon(t) = \frac{1}{2} \int_0^L \left\{ \rho \left[\frac{\partial w}{\partial t} \right]^2 + I_\rho \left[\frac{\partial \phi}{\partial t} \right]^2 + K \left(\phi - \frac{\partial w}{\partial x} \right)^2 + EI \left[\frac{\partial \phi}{\partial x} \right]^2 \right\} dx.$$

One can also derive an equivalent fourth-order equation in terms of w ; see Timoshenko [11] and Traill-Nash and Collar [12]. In particular, [12] discusses various boundary conditions.

This paper consists of two main parts. In the first part, we show that the energy $\varepsilon(t)$ decays exponentially fast under (0.3), (0.4) and (0.5). For the one-dimensional wave equation with boundary control, Quinn and Russell [9] established the exponential decay of solutions. Later, Chen [1], [2] obtained the same result for a wave equation in any space dimension under some geometrical conditions on the domain. A very restrictive part of these conditions was eliminated by Lagnese [5] with the aid of a new energy estimate. Lagnese [6] also extended the result to linear elastodynamic systems. In contrast to the above works, Lasiecka and Triggiani [7] employed boundary feedback acting in the Dirichlet boundary condition to achieve exponential decay of solutions to the wave equation. More recently, Chen et al. [3] discussed the case of a chain of Euler beams and obtained a similar result. Our result for the Timoshenko beam is most closely related to [3]. We use the energy method combined with C_0 -semigroup theory as in [1]–[3], [5] and [6]. The essence of the method is to construct a suitable energy functional associated with $\varepsilon(t)$. Details are given in § 2.

The second part of this paper is concerned with a numerical study. Since the nature of the spectrum is an important question in the investigation of the stability of a linear system, we carried out numerical experiments on the spectrum of (0.1) and (0.2) under (0.3)–(0.5). We express the temporal variation of the eigenfunction in normal modes of the form $e^{\lambda t}$, transforming (0.1)–(0.5) to ordinary differential equations with boundary conditions. The Chebyshev-tau method [4], [8] is used to discretize the spatial variation of the eigenfunctions, thus yielding a matrix eigenvalue problem with discrete complex-valued eigenvalues λ . These are computed using a NAG routine in quadruple precision on a VAX 11/785. Results of numerical experiments are presented in § 3.

1. Notation and preliminaries. We shall use the notation

$$f_t = \partial_t f = \frac{\partial f}{\partial t}, \quad f_x = \partial_x f = \frac{\partial f}{\partial x}, \quad f_{xx} = \partial_{xx} f = \frac{\partial^2 f}{\partial x^2}, \quad \text{etc.}$$

L^2 always denotes $L^2(0, 1)$ and we write

$$H^m = \left\{ f: f, \left[\frac{d}{dx} \right]^k f \in L^2, k = 1, \dots, m \right\}.$$

Our basic function space \mathcal{G} is the set of all quadruplets

$$z = \begin{bmatrix} w_1 \\ w_2 \\ \phi_1 \\ \phi_2 \end{bmatrix}$$

satisfying

$$\begin{aligned} w_1 &\in H^1, \quad w_1(0) = 0, \quad w_2 \in L^2, \\ \phi_1 &\in H^1, \quad \phi_1(0) = 0, \quad \phi_2 \in L^2, \end{aligned}$$

equipped with the inner product

$$(1.1) \quad \langle z, \tilde{z} \rangle_{\mathcal{G}} = \int_0^1 \left\{ \frac{K}{\rho} (\partial_x w_1)(\partial_x \tilde{w}_1) + w_2 \tilde{w}_2 + \frac{EI}{I_\rho} (\partial_x \phi_1)(\partial_x \tilde{\phi}_1) + \phi_2 \tilde{\phi}_2 \right\} dx.$$

We shall also use the function space \mathcal{S} which is the set of all quadruplets

$$z = \begin{bmatrix} w_1 \\ w_2 \\ \phi_1 \\ \phi_2 \end{bmatrix}$$

satisfying

$$\begin{aligned} w_1 &\in H^2, \quad w_1(0) = 0, \quad w_2 \in H^1, \quad w_2(0) = 0, \\ \phi_1 &\in H^2, \quad \phi_1(0) = 0, \quad \phi_2 \in H^1, \quad \phi_2(0) = 0, \\ K\phi_1(1) - K\partial_x w_1(1) &= \alpha w_2(1), \quad EI\partial_x \phi_1(1) = -\beta \phi_2(1), \end{aligned}$$

equipped with the inner product induced by $H^2 \times H^1 \times H^2 \times H^1$. Here, K , α , E , I and β are the same as in the previous section. It is easy to show that \mathcal{S} is dense in \mathcal{G} .

We define the operator Λ in \mathcal{G} :

$$(1.2) \quad \Lambda = \begin{bmatrix} 0 & \text{id} & 0 & 0 \\ (K/\rho)\partial_{xx} & 0 & -(K/\rho)\partial_x & 0 \\ 0 & 0 & 0 & \text{id} \\ (K/I_\rho)\partial_x & 0 & (EI/I_\rho)\partial_{xx} - (K/I_\rho)\text{id} & 0 \end{bmatrix}$$

where id is the identity mapping and the domain of Λ is taken to be \mathcal{S} . It is then easy to see that (0.1), (0.2) with (0.3), (0.4) and (0.5) can be put in the abstract form:

$$(1.3) \quad \frac{dz}{dt} = \Lambda z \quad \text{where } z = \begin{bmatrix} w \\ w_t \\ \phi \\ \phi_t \end{bmatrix} \text{ and } L \text{ is taken to be } 1.$$

We also observe the following.

LEMMA 1.1. Λ is an infinitesimal generator of a C_0 -semigroup in \mathcal{G} .

Proof. Let us write $\Lambda = \Lambda_0 + \Lambda_1$, where

$$(1.4) \quad \Lambda_0 = \begin{bmatrix} 0 & \text{id} & 0 & 0 \\ (K/\rho)\partial_{xx} & 0 & 0 & 0 \\ 0 & 0 & 0 & \text{id} \\ 0 & 0 & (EI/I_\rho)\partial_{xx} & 0 \end{bmatrix}$$

with $\mathcal{D}(\Lambda_0) = \mathcal{S}$ and

$$(1.5) \quad \Lambda_1 = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & -(K/\rho)\partial_x & 0 \\ 0 & 0 & 0 & 0 \\ (K/I_\rho)\partial_x & 0 & -(K/I_\rho)\text{id} & 0 \end{bmatrix}.$$

It is apparent that Λ_1 is a bounded linear operator on \mathcal{G} . Thus, it is enough to show that Λ_0 is an infinitesimal generator of a C_0 -semigroup. By virtue of the Lumer-Phillips theorem, it is enough to show that

$$(1.6) \quad \langle \Lambda_0 z, z \rangle_{\mathcal{G}} \leq c \|z\|_{\mathcal{G}}^2 \quad \text{for all } z \in \mathcal{S},$$

where c is a positive constant, and that

$$(1.7) \quad \text{Range of } (\lambda \text{ id} - \Lambda_0) = \mathcal{G} \quad \text{for some } \lambda > c.$$

By integration by parts using the boundary condition of $z \in \mathcal{S}$, we find that

$$\begin{aligned} \langle \Lambda_0 z, z \rangle_{\mathcal{G}} &= \frac{K}{\rho} w_2(1) \partial_x w_1(1) + \frac{EI}{I_\rho} \phi_2(1) \partial_x \phi_1(1) \\ &= \frac{1}{\rho} (K \phi_1(1) - \alpha w_2(1)) w_2(1) - \frac{\beta}{I_\rho} \phi_2(1)^2 \\ (1.8) \quad &\leq \frac{1}{4\rho\alpha} K^2 \phi_1(1)^2 \\ &\leq \frac{1}{4\rho\alpha} K^2 \int_0^1 (\partial_x \phi_1)^2 dx \\ &\leq \frac{K^2}{4\rho\alpha EI} I_\rho \|z\|_{\mathcal{G}}^2, \end{aligned}$$

from which (1.6) follows. We next prove (1.7) for any $\lambda > 0$. Let $\lambda > 0$ and

$$\begin{bmatrix} f_1 \\ f_2 \\ g_1 \\ g_2 \end{bmatrix} \in \mathcal{G}.$$

Then, we have to find

$$\begin{bmatrix} w_1 \\ w_2 \\ \phi_1 \\ \phi_2 \end{bmatrix} \in \mathcal{S}$$

such that

$$(1.9) \quad \lambda w_1 - w_2 = f_1,$$

$$(1.10) \quad \lambda w_2 - \frac{K}{\rho} \partial_{xx} w_1 = f_2,$$

$$(1.11) \quad \lambda \phi_1 - \phi_2 = g_1,$$

$$(1.12) \quad \lambda \phi_2 - \frac{EI}{I_\rho} \partial_{xx} \phi_1 = g_2.$$

We can find $\phi_1 \in H^2$ such that

$$(1.13) \quad \left(\lambda^2 - \frac{EI}{I_\rho} \partial_{xx} \right) \phi_1 = g_2 + \lambda g_1,$$

$$(1.14) \quad \phi_1(0) = 0, \quad \beta \lambda \phi_1(1) + EI \partial_x \phi_1(1) = \beta g_1(1).$$

In fact, ϕ_1 is given by

$$(1.15) \quad \phi_1(x) = c \sinh \mu x - \frac{I_p}{\mu EI} \int_0^x (g_2(s) + \lambda g_1(s)) \sinh \mu(x-s) ds$$

where $\mu = \lambda(I_p/EI)^{1/2}$ and c is uniquely determined from $\beta\lambda\phi_1(1) + EI \partial_x \phi_1(1) = \beta g_1(1)$.

Then, ϕ_2 is determined by (1.13). It is obvious that $\phi_2 \in H^1$, $\phi_2(0) = 0$ and $EI \partial_x \phi_1(1) = -\beta\phi_2(1)$. Similarly, we can find $w_1 \in H^2$ such that

$$(1.16) \quad \left(\lambda^2 - \frac{K}{\rho} \partial_{xx} \right) w_1 = f_2 + \lambda f_1,$$

$$(1.17) \quad w_1(0) = 0, \quad \alpha\lambda w_1(1) + K \partial_x w_1(1) = K\phi_1(1) + \alpha f_1(1).$$

Then, w_2 is determined from (1.19). It is easy to see that

$$\begin{bmatrix} w_1 \\ w_2 \\ \phi_1 \\ \phi_2 \end{bmatrix} \in \mathcal{S}$$

and (1.9)–(1.12) hold.

We shall use the following elementary inequality later on:

$$(1.18) \quad \int_0^1 w_x^2 dx \leq 2 \int_0^1 (\phi - w_x)^2 dx + 2 \int_0^1 \phi_x^2 dx$$

for all $\phi \in H^1$, $w \in H^1$ satisfying $\phi(0) = 0$.

2. Statement and proof of the main result. In this section, we take $L = 1$ without loss of generality. Let $S(t)$ be the C_0 -semigroup in \mathcal{G} generated by Λ in the previous section. We assert the following.

THEOREM 2.1. *The operator norm of $S(t)$ satisfies*

$$(2.1) \quad \|S(t)\| \leq M e^{-rt} \quad \text{for all } t \geq 0$$

where M and r are positive constants.

Before giving details of the proof, we shall outline our arguments. Let us fix any $z_0 \in \mathcal{S}$. Using $\varepsilon(t)$ associated with $S(t)z_0$, we define

$$(2.2) \quad F(t) = \mu t \varepsilon(t) + G(S(t)z_0)$$

where μ is a positive constant depending only on the coefficients of (0.1), (0.2), and $G(\cdot)$ is a suitable functional on \mathcal{G} such that

$$(2.3) \quad |G(z)| \leq C \|z\|_{\mathcal{G}}^2 \quad \text{for all } z \in \mathcal{G}.$$

With the aid of (1.18), we can derive that

$$(2.4) \quad d_1 \|S(t)z_0\|_{\mathcal{G}}^2 \leq \varepsilon(t) \leq d_2 \|S(t)z_0\|_{\mathcal{G}}^2$$

holds for all $t \geq 0$, where d_1 and d_2 are positive constants depending only on the coefficients of (0.1) and (0.2). We then show that

$$(2.5) \quad F(t) \leq M_1 \|z_0\|_{\mathcal{G}}^2$$

holds for all $t \geq 0$, where M_1 is a positive constant depending only on α , β and the coefficients of (0.1) and (0.2). Formulae (2.3) and (2.4) imply

$$(2.6) \quad \|S(t)z_0\|_{\mathcal{G}}^2 \leq \frac{1}{t} M_2 \|z_0\|_{\mathcal{G}}^2 \quad \text{for all } t > 0$$

where M_2 is a positive constant depending only on α , β and the coefficients of (0.1) and (0.2). Since \mathcal{S} is dense in \mathcal{G} , (2.6) implies

$$(2.7) \quad \|S(t)z\|_{\mathcal{G}} \leq \left[\frac{M_2}{t} \right]^{1/2} \|z_0\|_{\mathcal{G}} \quad \text{for all } z \in \mathcal{G}.$$

Finally, we use the semigroup property of $S(t)$ to arrive at (2.1).

Proof of Theorem 2.1. Fix any $z_0 \in \mathcal{S}$. Then,

$$(2.8) \quad S(t)z_0 \in C([0, \infty); \mathcal{S}) \cap C^1([0, \infty); \mathcal{G})$$

and

$$(2.9) \quad \frac{d}{dt} S(t)z_0 = \Lambda S(t)z_0 \quad \text{for every } t \geq 0.$$

Hence, we can write

$$(2.10) \quad S(t)z_0 = \begin{bmatrix} w(x, t) \\ \partial_t w(x, t) \\ \phi(x, t) \\ \partial_t \phi(x, t) \end{bmatrix}$$

where $w(x, t)$ and $\phi(x, t)$ satisfy (0.1)–(0.5).

We now construct $F(t)$:

$$(2.11) \quad \begin{aligned} F(t) = & \frac{\mu t}{2} \int_0^1 \{ \rho w_t^2 + I_\rho \phi_t^2 + K(\phi - w_x)^2 + EI\phi_x^2 \} dx \\ & + \rho \int_0^1 x w_t w_x dx + I_\rho \int_0^1 x \phi_t \phi_x dx + \frac{1}{2+\eta} I_\rho \int_0^1 \phi \phi_t dx \\ & - \frac{1}{2+\eta} \rho \int_0^1 w w_t dx \end{aligned}$$

where μ and η are positive constants which will be determined later on. By virtue of (2.8) and (2.10), we can differentiate (2.11) to obtain

$$(2.12) \quad \frac{dF}{dt} = \mu J_1 + \mu t J_2 + \sum_{n=3}^9 J_n$$

where

$$\begin{aligned} J_1 = \varepsilon(t) &= \frac{1}{2} \int_0^1 \{ \rho w_t^2 + I_\rho \phi_t^2 + K(\phi - w_x)^2 + EI\phi_x^2 \} dx, \\ J_2 &= \int_0^1 \{ \rho w_t w_{tt} + I_\rho \phi_t \phi_{tt} + K(\phi - w_x)(\phi_t - w_{xt}) + EI\phi_x \phi_{xt} \} dx, \end{aligned}$$

$$\begin{aligned}
J_3 &= \rho \int_0^1 x w_t w_{xt} dx, & J_6 &= I_\rho \int_0^1 x \phi_{tt} \phi_x dx, \\
J_4 &= \rho \int_0^1 x w_{tt} w_x dx, & J_7 &= \frac{1}{2+\eta} I_\rho \int_0^1 \phi \phi_{tt} dx, \\
J_5 &= I_\rho \int_0^1 x \phi_t \phi_{xt} dx, & J_8 &= \frac{1}{2+\eta} I_\rho \int_0^1 \phi_t^2 dx, \\
J_9 &= -\frac{\rho}{2+\eta} \int_0^1 w_t^2 dx - \frac{\rho}{2+\eta} \int_0^1 w w_{tt} dx.
\end{aligned}$$

Using (0.1)–(0.5), we can integrate by parts to arrive at

$$(2.13) \quad J_2 = -\alpha w_t(1, t)^2 - \beta \phi_t(1, t)^2,$$

$$(2.14) \quad J_3 = \frac{1}{2} \rho w_t(1, t)^2 - \frac{1}{2} \rho \int_0^1 w_t^2 dx,$$

$$\begin{aligned}
(2.15) \quad J_4 &= \int_0^1 x w_x (K w_{xx} - K \phi_x) dx \\
&= \frac{1}{2} K w_x(1, t)^2 - \frac{1}{2} K \int_0^1 w_x^2 dx - K \int_0^1 x w_x \phi_x dx,
\end{aligned}$$

$$(2.16) \quad J_5 = \frac{1}{2} I_\rho \phi_t(1, t)^2 - \frac{1}{2} I_\rho \int_0^1 \phi_t^2 dx,$$

$$\begin{aligned}
(2.17) \quad J_6 &= \int_0^1 x \phi_x (EI \phi_{xx} - K \phi + K w_x) dx \\
&= \frac{1}{2} EI \phi_x(1, t)^2 - \frac{1}{2} EI \int_0^1 \phi_x^2 dx - \frac{1}{2} K \phi(1, t)^2 \\
&\quad + \frac{1}{2} K \int_0^1 \phi^2 dx + K \int_0^1 x \phi_x w_x dx,
\end{aligned}$$

$$\begin{aligned}
(2.18) \quad J_7 &= \frac{1}{2+\eta} \int_0^1 \phi (EI \phi_{xx} - K \phi + K w_x) dx \\
&= \frac{1}{2+\eta} EI \phi(1, t) \phi_x(1, t) - \frac{1}{2+\eta} EI \int_0^1 \phi_x^2 dx \\
&\quad - \frac{1}{2+\eta} K \int_0^1 \phi^2 dx + \frac{1}{2+\eta} K \phi(1, t) w(1, t) \\
&\quad + \frac{1}{2+\eta} \rho \int_0^1 w_{tt} w dx - \frac{1}{2+\eta} K w_x(1, t) w(1, t) \\
&\quad + \frac{1}{2+\eta} K \int_0^1 w_x^2 dx.
\end{aligned}$$

Let us choose $\eta > 0$ such that

$$(2.19) \quad K \left(\frac{1}{2} - \frac{1}{2+\eta} \right) \leq \frac{1}{4} EI,$$

and then, choose $\mu > 0$ such that

$$(2.20) \quad \mu(2K + EI) \leq \left(\frac{1}{4} + \frac{1}{2+\eta}\right) EI,$$

$$(2.21) \quad 2\mu \leq \frac{1}{2} - \frac{1}{2+\eta}.$$

Then, we find that

$$(2.22) \quad \begin{aligned} \frac{dF}{dt} \leq & -\frac{1}{2} \left(\frac{1}{4} + \frac{1}{2+\eta}\right) EI \int_0^1 \phi_x^2 dx - \frac{K}{2} \left(\frac{1}{2} - \frac{1}{2+\eta}\right) \int_0^1 w_x^2 dx \\ & - \alpha \mu t w_t(1, t)^2 - \beta \mu t \phi_t(1, t)^2 + \frac{1}{2} \rho w_t(1, t)^2 \\ & + \frac{1}{2} K w_x(1, t)^2 + \frac{1}{2} I_\rho \phi_t(1, t)^2 + \frac{1}{2} EI \phi_x(1, t)^2 \\ & - \frac{1}{2} K \phi(1, t)^2 + \frac{1}{2+\eta} EI \phi(1, t) \phi_x(1, t) \\ & + \frac{1}{2+\eta} K \phi(1, t) w(1, t) - \frac{1}{2+\eta} K w_x(1, t) w(1, t). \end{aligned}$$

By means of (0.4) and (0.5) and the inequalities

$$(2.23) \quad \phi(1, t)^2 \leq \int_0^1 \phi_x(x, t)^2 dx,$$

$$(2.24) \quad w(1, t)^2 \leq \int_0^1 w_x(x, t)^2 dx$$

we deduce from (2.22) that for all $t \geq T$,

$$(2.25) \quad \frac{dF}{dt} \leq 0$$

where T is a positive constant depending only on α, β and the coefficients of (0.1) and (0.2). Consequently, we arrive at (2.5). The argument following (2.5) completes the proof.

Remark 2.2. Finally, we remark that our arguments with the same energy functional also yield exponential stabilization for a hinged boundary condition at $x = 0$:

$$(2.26) \quad w(0, t) = 0, \quad \frac{\partial \phi}{\partial x}(0, t) = 0.$$

However, in this case it seems necessary to impose the zero mean condition $\int_0^1 \phi dx = 0$ in order to avoid some technical difficulties (see [1]).

3. Numerical study of the spectrum. We present numerical results on the linear stability of our system. We use normal mode analysis and set

$$(3.1) \quad w(x, t) = e^{\lambda t} P(x),$$

$$(3.2) \quad \phi(x, t) = e^{\lambda t} Q(x).$$

Thus, (0.1)–(0.5) become the following system of ordinary differential equations with boundary conditions:

$$(3.3) \quad -KP_{xx} + KQ_x + \lambda^2 \rho P = 0,$$

$$(3.4) \quad -EIQ_{xx} + K(Q - P_x) + \lambda^2 I_\rho Q = 0,$$

$$(3.5) \quad P = Q = 0 \quad \text{at } x = 0,$$

$$(3.6) \quad EIQ_x + \lambda \beta Q = 0 \quad \text{at } x = L,$$

$$(3.7) \quad K(Q - P_x) - \lambda \alpha P = 0 \quad \text{at } x = L.$$

This system is of the form $A_0 + A_1 \lambda + A_2 \lambda^2 = 0$, where

$$(3.8) \quad A_0 = \begin{bmatrix} -KP_{xx} + KQ_x \\ -EIQ_{xx} + K(Q - P_x) \\ P(0) \\ Q(0) \\ EIQ_x(L) \\ KQ(L) - KP_x(L) \end{bmatrix},$$

$$(3.9) \quad A_1 = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ \beta Q(L) \\ -\alpha P(L) \end{bmatrix},$$

$$(3.10) \quad A_2 = \begin{bmatrix} \rho P \\ I_\rho Q \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}.$$

We rewrite this in the form

$$(3.11) \quad \det \begin{bmatrix} \lambda & -A_0 \\ 1 & A_1 + \lambda A_2 \end{bmatrix} = 0$$

so that our system takes on the customary form $A = \lambda B$, where

$$(3.12) \quad A = \begin{bmatrix} 0 & -A_0 \\ 1 & A_1 \end{bmatrix},$$

$$(3.13) \quad B = \begin{bmatrix} -1 & 0 \\ 0 & -A_2 \end{bmatrix}.$$

We discretize $P(x)$ and $Q(x)$ by the Chebyshev-tau method [4], [8]. This is a spectral method where the expansion functions are the Chebyshev polynomials $T_n(z)$ defined by $T_n(\cos \theta) = \cos n\theta$ when $z = \cos \theta$. This method approximates discrete eigenvalues belonging to C^∞ eigenfunctions with infinite-order accuracy.

We rescale the spatial variable to $z = (2x)/L - 1$, so that $-1 \leq z \leq 1$. We set

$$(3.14) \quad P(z) = \sum_{n=0}^N p_n T_n(z),$$

$$(3.15) \quad Q(z) = \sum_{n=0}^N q_n T_n(z)$$

and substitute into (3.3)–(3.7). Thus, there are $2N+2$ unknowns. In the differential equations, we equate coefficients of like powers of the Chebyshev polynomials. Since the first equation (3.3) contains P_{xx} , it yields equations for the coefficients up to degree $N-2$ in the polynomials, thus giving $N-1$ equations. Similarly, (3.4) yields $N-1$ equations. In the tau approximation, the expansion functions $T_n(z)$ are not required to satisfy the boundary conditions individually. The four boundary conditions are imposed as part of the conditions determining the coefficients p_n and q_n . The total number of equations is $2N+2$. The size of the final matrix equation $A = \lambda B$ is $4N+4$ square. Our computer program uses the NAG routine F02GJF to compute the eigenvalues in complex quadruple precision on a VAX 11/785.

The accuracy of our numerical results was established in the following way. The eigenvalues must satisfy the characteristic equation:

$$(3.16) \quad \det \begin{bmatrix} m_{11}(\lambda) & m_{12}(\lambda) & m_{13}(\lambda) & m_{14}(\lambda) \\ m_{21}(\lambda) & m_{22}(\lambda) & m_{23}(\lambda) & m_{24}(\lambda) \\ m_{31}(\lambda) & m_{32}(\lambda) & m_{33}(\lambda) & m_{34}(\lambda) \\ m_{41}(\lambda) & m_{42}(\lambda) & m_{43}(\lambda) & m_{44}(\lambda) \end{bmatrix} = 0$$

where, for $j = 1, 2, 3, 4$

$$(3.17) \quad \begin{aligned} m_{1j} &= 1, \quad m_{2j} = I_p \lambda^2 \eta_j - EI \eta_j^3, & m_{3j} &= (EI \eta_j + \beta \lambda) e^{\eta_j L}, \\ m_{4j} &= \left(EI \lambda \eta_j^2 - I_p \lambda^3 - \frac{\alpha}{\rho} I_p \lambda^2 \eta_j + \frac{\alpha}{\rho} EI \eta_j^3 \right) e^{\eta_j L}, \end{aligned}$$

and η_j are the roots of

$$(3.18) \quad \eta^2 = \frac{1}{2EI} \left[\lambda^2 \left(I_p + \frac{\rho EI}{K} \right) \pm \left(\lambda^4 \left(I_p - \frac{\rho EI}{K} \right)^2 - 4\rho EI \lambda^2 \right)^{1/2} \right].$$

In order to check that our computed eigenvalues satisfy the determinant equation (3.16), we have chosen moderate-sized parameters so that the evaluation of the determinant avoids cancellation between large numbers. We choose $\rho = 1$, $K = 1.5$, $I_p = 2$, $E = 2.5$, $I = 3$, $L = 0.1$, $\alpha = 3.5$ and $\beta = 4.1$. Computations at $N = 15, 20, 25$ and 30 showed that a few eigenvalues are already converged to about 15 digits at $N = 15$. About 12 eigenvalues at $N = 15$ are converged to at least 5 digits, and satisfy (3.16) to that accuracy. All converged eigenvalues have negative real parts and are either real or complex conjugates. The number of digits to which each pair is a complex conjugate is an indication of the amount of roundoff error present. The eigenvalues consist of two groups. One group is lined up approximately along the line -4.47 and the imaginary parts are almost multiples, starting with $\lambda = -4.4709$ and then $\lambda = -4.4769 \pm 38.475i$, $-4.4770 \pm 76.951i$ and so on. The other group is lined up approximately along -34.45 , and the imaginary parts are almost multiples, starting with $\lambda = -34.505$ and then $\lambda = -34.446 \pm 60.829i$, $-34.452 \pm 121.68i$ and so on.

Computations were done at the following set of parameters to model a solid aluminum bar: $\rho = 400 \text{ g/cm}$, $K = 2.8 \times 10^{13} \text{ g} \cdot \text{cm/sec}^2$, $I_p = 3,332 \text{ g} \cdot \text{cm}$, $E = 7.6 \times 10^{11} \text{ g/cm/sec}^2$, $I = 833 \text{ cm}^4$ and $L = 200 \text{ cm}$. We allow α and β to be

- (i) $\alpha = 10^3 \text{ g/sec}$, $\beta = 10^7 \text{ g} \cdot \text{cm}^2/\text{sec}$;
- (ii) $\alpha = 2 \times 10^3 \text{ g/sec}$, $\beta = 2 \times 10^7 \text{ g} \cdot \text{cm}^2/\text{sec}$;
- (iii) $\alpha = 5 \times 10^3 \text{ g/sec}$, $\beta = 5 \times 10^7 \text{ g} \cdot \text{cm}^2/\text{sec}$.

By comparing the results of $N = 40$ and $N = 45$, we conclude that about 20 complex conjugate pairs have converged to 5 digits at $N = 40$, and there are no real-valued eigenvalues. The results for case (i) are plotted in Figs. 1 and 2. Figure 2 is a magnification of Fig. 1 close to the origin. All eigenvalues have negative real parts. Figure 1 does not indicate that the ratio $\text{Im}(\lambda)/\text{Re}(\lambda)$ approaches a constant for large $|\lambda|$. Results of cases (i)–(iii) are displayed in Table 1. Essentially, the real parts of λ are approximately proportional to α or β and the imaginary parts of the three cases

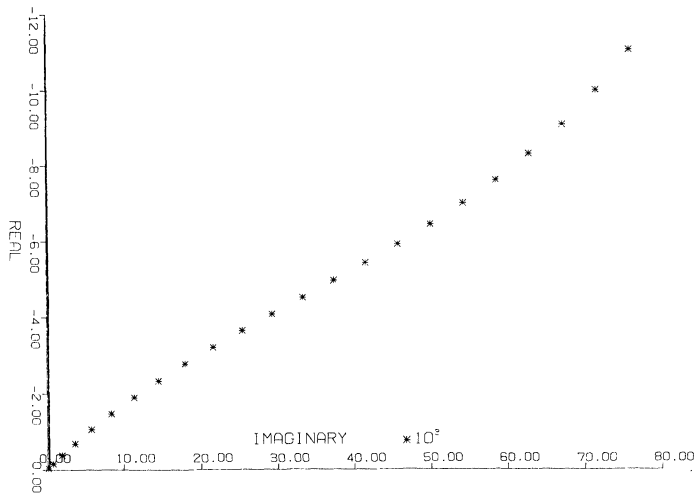


FIG. 1. The graph displays the upper half quadrant of the first 23 complex conjugate pairs of case (i).

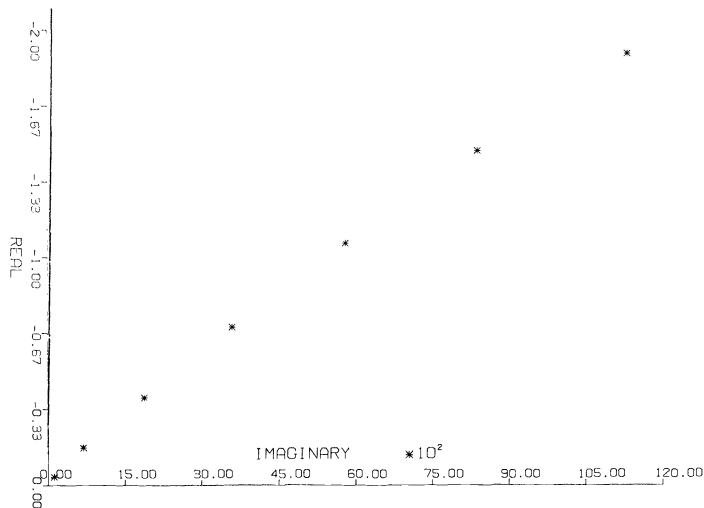


FIG. 2. The graph is a magnification of Fig. 1 close to the origin to clarify the location of the eigenvalues there.

TABLE 1

This table displays to 5 digits the first 22 to 23 complex conjugate pairs for cases (i)–(iii) described in § 3.

Case (i)		Case (ii)		Case (iii)	
−0.36717E−01	±0.11039E+03	−0.73434E−01	±0.11039E+03	−0.18359E+00	±0.11039E+03
−0.16364E+00	±0.68450E+03	−0.32729E+00	±0.68450E+03	−0.81822E+00	±0.68450E+03
−0.38376E+00	±0.18853E+04	−0.76753E+00	±0.18853E+04	−0.19188E+01	±0.18853E+04
−0.69283E+00	±0.36108E+04	−0.13857E+01	±0.36108E+04	−0.34641E+01	±0.36108E+04
−0.10607E+01	±0.58051E+04	−0.21214E+01	±0.58051E+04	−0.53033E+01	±0.58051E+04
−0.14663E+01	±0.84010E+04	−0.29326E+01	±0.84010E+04	−0.73313E+01	±0.84011E+04
−0.18925E+01	±0.11335E+05	−0.37851E+01	±0.11335E+05	−0.94624E+01	±0.11335E+05
−0.23278E+01	±0.14548E+05	−0.46555E+01	±0.14548E+05	−0.11638E+02	±0.14548E+05
−0.27652E+01	±0.17990E+05	−0.55303E+01	±0.17990E+05	−0.13825E+02	±0.17990E+05
−0.32017E+01	±0.21620E+05	−0.64033E+01	±0.21620E+05	−0.16007E+02	±0.21620E+05
−0.36372E+01	±0.25401E+05	−0.72743E+01	±0.25401E+05	−0.18184E+02	±0.25402E+05
−0.40737E+01	±0.29306E+05	−0.81472E+01	±0.29306E+05	−0.20366E+02	±0.29306E+05
−0.45146E+01	±0.33310E+05	−0.90290E+01	±0.33311E+05	−0.22570E+02	±0.33311E+05
−0.49647E+01	±0.37395E+05	−0.99293E+01	±0.37395E+05	−0.24820E+02	±0.37396E+05
−0.54301E+01	±0.41544E+05	−0.10860E+02	±0.41544E+05	−0.27146E+02	±0.41545E+05
−0.59180E+01	±0.45745E+05	−0.11836E+02	±0.45745E+05	−0.29584E+02	±0.45745E+05
−0.64370E+01	±0.49986E+05	−0.12874E+02	±0.49986E+05	−0.32178E+02	±0.49987E+05
−0.69979E+01	±0.54258E+05	−0.13995E+02	±0.54258E+05	−0.34982E+02	±0.54259E+05
−0.76138E+01	±0.58553E+05	−0.15227E+02	±0.58553E+05	−0.38061E+02	±0.58554E+05
−0.83016E+01	±0.62865E+05	−0.16603E+02	±0.62865E+05	−0.41499E+02	±0.62866E+05
−0.90827E+01	±0.67186E+05	−0.18165E+02	±0.67186E+05	−0.45405E+02	±0.67187E+05
−0.99853E+01	±0.71509E+05	−0.19970E+02	±0.71509E+05	−0.49919E+02	±0.71510E+05
−0.11047E+02	±0.75828E+05	−0.22093E+02	±0.75828E+05		

are approximately equal. This indicates that when α and β are zero, the real parts are zero and the eigenvalues are purely imaginary. The relative importance of the damping terms in our computations is seen from the dimensionless equations. There are 6 dimensionless parameters: $C_1 = L\bar{Q}/\bar{P}$, where \bar{Q} is the scale of Q and \bar{P} is the scale of P , $C_2 = \lambda^2 \rho L^2 / K$, $C_3 = KL^2 / EI$, $C_4 = \lambda^2 I_p L^2 / EI$, $C_5 = \lambda \beta L / EI$ and $C_6 = \lambda \alpha L / K$. Our values for case (i) are: $C_2 = \lambda^2 \times 6 \times 10^{-7}$, $C_3 = 1.8 \times 10^3$, $C_4 = \lambda^2 \times 2 \times 10^{-7}$, $C_5 = \lambda \times 3 \times 10^{-6}$ and $C_6 = \lambda \times 7 \times 10^{-9}$. The dimensionless equations are

$$(3.19) \quad -P_{xx} + C_1 Q_x + C_2 P = 0,$$

$$(3.20) \quad -Q_{xx} + C_3 Q - \frac{C_3}{C_1} P + C_4 Q = 0,$$

$$(3.21) \quad P = Q = 0 \quad \text{at } x = 0,$$

$$(3.22) \quad Q_x + C_5 Q = 0 \quad \text{at } x = 1,$$

$$(3.23) \quad C_1 Q - P_x - C_6 P = 0 \quad \text{at } x = 1$$

where P , Q and x have been made dimensionless. The largest eigenvalues in Fig. 1 are $O(10^5)$ so that C_5 is $O(10^{-1})$ and C_6 is $O(10^{-3})$, indicating that the damping terms are not large. For the smallest eigenvalues, the damping terms are small so that the property of proportionality of $\text{Re}(\lambda)$ to α or β may be an asymptotic behavior for small damping.

We note that the qualitative features obtained for the computation with moderate data are very different from those of the model of an aluminium bar. This is reminiscent of the qualitative differences in Figs. 5 and 7 of Chen et al. [3].

Acknowledgments. We are very grateful to Professors K. Hannsgen and R. Wheeler for suggesting this problem and for useful discussions. We are also much indebted to Professor J. Burns whose help was indispensable in revising the previous manuscript.

REFERENCES

- [1] G. CHEN, *Energy decay estimates and exact boundary value controllability for the wave equation in a bounded domain*, J. Math. Pures Appl., 58 (1979), pp. 249–273.
- [2] ———, *A note on the boundary stabilization of the wave equation*, this Journal, 19 (1981) pp. 106–113.
- [3] G. CHEN, M. C. DELFOUR, A. M. KRALL AND G. PAYRE, *Modeling, stabilization and control of serially connected beams*, this Journal, 25 (1987), pp. 527–546.
- [4] D. GOTTLIEB AND S. A. ORSZAG, *Numerical Analysis of Spectral Methods: Theory and Applications*, CBMS-NSF Regional Conference Series in Applied Mathematics, 26, Society for Industrial and Applied Mathematics, Philadelphia, PA, 1983.
- [5] J. LAGNESE, *Decay of solutions of wave equations in a bounded region with boundary dissipation*, J. Differential Equations, 50 (1983), pp. 163–182.
- [6] ———, *Boundary stabilization of linear elastodynamic systems*, this Journal, 21 (1983), pp. 968–984.
- [7] I. LASIECKA AND R. TRIGGIANI, *Exponential uniform stabilization of the wave equation with $L_2(0, \infty; L_2(\Gamma))$ boundary feedback acting in the Dirichlet boundary conditions*, Proc. of the 24th IEEE Conference on Decision and Control, Fort Lauderdale, FL, 1985.
- [8] S. A. ORSZAG, *Accurate solution of the Orr–Sommerfeld stability equation*, J. Fluid Mech., 50, (1971), pp. 689–703.
- [9] J. P. QUINN AND D. L. RUSSELL, *Asymptotic stability and energy decay rates for solutions of hyperbolic equations with boundary damping*, Proc. Roy. Soc. Edinburgh Sect. A, 77, 1977/78, pp. 97–127.
- [10] D. L. RUSSELL, *Mathematical models for the elastic beams and their control-theoretic implications*, preprint.
- [11] S. TIMOSHENKO, *Vibration Problems in Engineering*, Van Nostrand, New York, 1955.
- [12] R. W. TRAILL-NASH AND A. R. COLLAR, *The effects of shear flexibility and rotatory inertia on the bending vibrations of beams*, Quart. J. Mech. Appl. Math., 6 (1953), pp. 186–222.
- [13] K. WASHIZU, *Variational Methods in Elasticity and Plasticity*, Pergamon, London–New York, 1968.

FINITE DIMENSIONAL FILTERS FOR A CLASS OF NONLINEAR SYSTEMS AND IMMERSION IN A LINEAR SYSTEM*

J. LEVINE†

Abstract. We study the finite dimensionality of the filter associated with the following system:

$$(\Sigma) \begin{cases} dx_t = f(x_t) dt, & x_0: \text{random with probability measure } \mu_0, \\ dy_t = h(x_t) dt + J dv_t, & y_0 = 0. \end{cases}$$

Σ is distinguished by the absence of noise in the dynamics.

A necessary and sufficient condition for the existence of a finite dimensional filter is obtained, providing an explicit formula for the minimal filter. Furthermore, this condition is proved to be equivalent to each of the following properties:

- The finite dimensionality of the estimation algebra (the Lie algebra associated to the D-M-Z equation);
- The immersibility of Σ in a linear system (input-output equivalence with a linear system of larger dimension).

Key words. finite dimensional filtering, realization of nonlinear systems, immersion into linear systems, estimation algebra

AMS(MOS) subject classification. 93E11

1. Introduction. After R. Brockett [5], V. Benes [3], Mitter [19], Ocone [20], and others (see also the beautiful survey paper by Marcus [18]), most of the works on finite dimensional filters (FDF) make explicit use of the finite dimensionality of the estimation algebra (the formal Lie algebra generated by the drift and diffusion operators of the Duncan-Mortensen-Zakai (D-M-Z) equation, considered as an infinite dimensional bilinear system).

A general method, outlined by Brockett [5], consists of finding sufficient conditions for finite dimensionality of the estimation algebra and of obtaining a filter by the Wei-Norman integration technique. Nevertheless, its complexity is such that finding new classes of systems admitting FDF constitutes a challenge to virtuosity (see for example [22]).

General necessity results can also be found in [7], and seem to be as difficult to exploit as Brockett's method.

Another stream [2], [6], [18], [19], consists in classifying the transformations that leave the estimation algebra unchanged. Diffeomorphic changes of coordinates in the state-space and gauge transformations are among those invariant transformations.

A natural question then arises as to whether one can characterize the existence of FDF directly from the system's input-output mapping rather than from the estimation algebra. Note that for discrete-time systems the input-output approach is useful (see [15]), whereas Lie algebraic ideas actually do not have a discrete-time counterpart.

The purpose of this paper is precisely to give a complete answer to this question for the particular class of systems without dynamical noise.

* Received by the editors September 11, 1985; accepted for publication (in revised form) November 18, 1986. This work has been partly supported by the French Army.

† Centre d'Automatique et Informatique, Ecole Nationale Supérieure des Mines de Paris, 35, rue St. Honoré, 77305 Fontainebleau, France.

More precisely, we consider a system of the following form:

$$(\Sigma) \begin{cases} dx_t = f(x_t) dt, & x_0: \text{random with probability measure } \mu_0, \\ dy_t = h(x_t) dt + J dv_t, & y_0 = 0. \end{cases}$$

In applications, it may be interesting to modelize systems without dynamical noises in at least two situations:

—Processes with short life duration, or short observation histories.

—Processes with fast-slow time scales, without noise in the fast components and small noises in the slow ones.

By FDF, we mean a *universal finite dimensional filter* (see [7], [18]), or, roughly speaking, a finite dimensional diffusion ξ_t driven by the observation process of Σ such that the unnormalized conditional measure of x_t knowing the observations history up to time t is obtained as a function of ξ_t only.

Our main result is that Σ admits FDF if and only if there exists an immersion from Σ into a linear system.

This result turns out to be quite negative since, on the one hand, the systems that are immersible into one that is linear are rare and on the other hand since, at least for our restricted class of systems without dynamical noise, finite dimensionality of the filter reduces in some sense to linearity.

The same result was obtained by Hijab [12] for a slightly different problem: Hijab considered systems Σ with scalar observations, and addressed the existence problem of finite dimensional realizations of the scalar input-output map $Y_t \rightarrow E(h(x_t)|Y_t)$, where Y_t denotes any continuous path of observations between 0 and t . Using formal arguments, he showed that the immersion property of Σ into a linear system is necessary and sufficient for the existence of a finite dimensional realization. We use here similar input-output realization ideas.

Coming back to the FDF problem, our results generalize those of Roth and Loparo [21] who proved that a sufficient condition for Σ to admit FDF is that the estimation algebra is nilpotent.

The paper is organized as follows. In § 2, we give an explicit formula for the unnormalized conditional measure. The necessary and sufficient condition for the existence of FDF is proved in § 3 and the minimal filter is explicitly obtained. Finally, § 4 is devoted to the study of the equivalence between existence of FDF and immersibility into a linear system.

2. The unnormalized conditional measure. Let us assume that $x_t \in \mathbf{R}^n$, $y_t \in \mathbf{R}^p$, x_0 : random with probability measure μ_0 on \mathbf{R}^n , and v_t is a standard Brownian motion of \mathbf{R}^p , dv_t being its Stratonovitch differential. J is an invertible $p \times p$ matrix.

We denote x' for x transpose, $Y_t = \{y_s | 0 \leq s \leq t\}$, and we also assume that

$$(2.1) \quad \begin{aligned} f &\in C^\infty(\mathbf{R}^n; \mathbf{R}^n), & h &\in C^\infty(\mathbf{R}^n; \mathbf{R}^p), \\ \exists c_1 > 0 \text{ s.t. } |x'f(x)| &\leq c_1(1 + \|x\|^2) & \forall x \in \mathbf{R}^n, \\ \exists c_2 > 0 \text{ s.t. } |\operatorname{div} f(x)| &\leq c_2 & \forall x \in \mathbf{R}^n, \end{aligned}$$

with $\operatorname{div} f(x) = \sum_{i=1}^n (\partial f_i / \partial x_i)(x)$.

Formulae (2.1) ensure that the flow $(x, t) \rightarrow X_t(x)$ of the differential equation $\dot{x} = f(x)$, with $X_0(x) = x$, is such that X_t is a C^∞ diffeomorphism of \mathbf{R}^n , for all $t \geq 0$. We also denote $(X_t)_* \mu_0$ the image of μ_0 by X_t .

THEOREM 1. Suppose that (2.1) holds. Then the unnormalized conditional measure of x_t knowing Y_t , noted π_t^Y , exists and is given by:

(2.2)

$$\pi_t^Y(x) = \exp \left[\sum_{i,j=1}^p \int_0^t h_i(X_{s-t}(x))(JJ')_{i,j}^{-1} dy_s^j - \frac{1}{2} \int_0^t \|J^{-1}h(X_{s-t}(x))\|^2 ds \right] \cdot (X_t)_* \mu_0(x) \\ \forall t \geq 0 \quad \text{a.e. } x \in \mathbf{R}^n, \quad Y \text{ a.s.}$$

Furthermore, π_t^Y solves the D-M-Z equation (in Stratonovitch form and in the sense of distributions):

$$(2.3) \quad d\pi_t^Y = \left[-\operatorname{div}(f\pi_t^Y) - \frac{1}{2} \|J^{-1}h\|^2 \pi_t^Y \right] dt + \sum_{i,j=1}^p h_i(JJ')_{i,j}^{-1} \pi_t^Y dy_t^j, \\ \pi_0^Y = \mu_0.$$

Proof. Let Q_t be the law of (x_t, Y_t) , and let us define Q_0^t by:

$$\frac{dQ_t}{dQ_0^t} = Z_t$$

with

$$Z_t = \exp \left[\int_0^t h'(X_{s-t}(x))(JJ')^{-1} dy_s - \frac{1}{2} \int_0^t \|J^{-1}h(X_{s-t}(x))\|^2 ds \right].$$

By Girsanov's Theorem, y_t is a Wiener process independent of x_t for Q_0^t .

On the other hand, the probability measure of x_t is the image by the flow X_t of μ_0 , denoted by:

$$\mu_t(x) = (X_t)_* \mu_0(x) \quad \text{a.e. } x \in \mathbf{R}^n.$$

Thus $Q_0^t = \mu_t \otimes P_{F_t}$, with P the Wiener measure on $(C^0(\mathbf{R}_+; \mathbf{R}^p), F_t)$. Hence, $Q_t = Z_t \cdot \mu_t \otimes P_{F_t}$ and $\pi_t^Y = Z_t \cdot \mu_t$, which proves the result.

Equations (2.3) are trivially obtained by the Itô–Stratonovitch formula for $\langle \phi, \pi_t^Y \rangle$ for every $\phi \in C^\infty$ with compact support in \mathbf{R}^n .

3. Existence of finite dimensional filters and their minimal realization.

DEFINITION 1. An FDF for Σ is a stochastic system on \mathbf{R}^r :

$$(3.1) \quad d\xi_t = a(\xi_t) dt + \sum_{i=1}^p b_i(\xi_t) dy_t^i, \\ \pi_t^Y = \sigma(\xi_t, \cdot) \cdot \nu_t \quad \text{in the sense of measures on } \mathbf{R}^n, \quad Y \text{ a.s.,}$$

with $a, b_1, \dots, b_p \in C^\infty(\mathbf{R}^r; \mathbf{R}^r)$, a having linear growth, b_i bounded for all i , ν_t bounded positive measure on \mathbf{R}^n , for all $t \geq 0$, $\sigma \in C^\infty(\mathbf{R}^r \times \mathbf{R}^n; \mathbf{R}_+)$, and ξ_0 s.t. $\mu_0 = \sigma(\xi_0, \cdot) \cdot \nu_0$.

Our FDF's definition is, apart from integrability details, equivalent to the definition of universal filters of [7] (see also [18]), but appears more naturally with (2.2).

The FDF problem can be seen as a special case of realization theory for nonlinear systems in the following sense: one can see the mapping $dY \rightarrow \pi^Y$ as an input-output mapping where the Stratonovitch differentials of the observations play the role of the inputs and where the unnormalized conditional measure plays the role of the output. A filter is thus a realization of this input-output mapping, and it is finite dimensional if the dimension of its state-space is finite.

Of course, by construction, this mapping is causal. Nevertheless, the input set which is considered in realization theory, namely the set of piecewise constant controls,

has P -measure 0. Furthermore, the output is infinite dimensional (for each t and Y , π_t^Y is a bounded positive measure on \mathbf{R}^n). Consequently, the standard theory does not apply.

Let us introduce the following notation. If h_i denotes the i th component of the function h in Σ , let us denote $L_f h_i(x)$ the Lie derivative of h_i along f :

$$(3.2) \quad L_f h_i(x) = \sum_{j=1}^n f_j(x) \frac{\partial h_i}{\partial x_j}(x), \quad L_f^k h_i = L_f(L_f^{k-1} h_i) \quad \forall k \geq 1, \quad L_f^0 h_i = h_i.$$

We also need the following space:

$$(3.3) \quad \mathbf{H} = \mathbf{R}\text{-Span} \{L_f^k h_i | 1 \leq i \leq p, \forall k \geq 0\}.$$

In terms of system theory, \mathbf{H} is the observations space of Σ [10].

THEOREM 2. *The three following assertions are equivalent:*

- (i) Σ has an FDF.
- (ii) $\dim \mathbf{H} < \infty$.
- (iii) $\exists r \in \mathbf{N}, \exists \theta_1, \dots, \theta_r \in \mathbf{H}, \exists A \in \mathbf{R}^{r \times r}, \exists B_1, \dots, B_p \in \mathbf{R}^{1 \times r}$ such that

$$(3.4) \quad \begin{aligned} L_f \theta &= A\theta, \\ h_i &= B_i \theta \quad \forall i = 1, \dots, p, \end{aligned}$$

with $\theta(x) = (\theta_1(x), \dots, \theta_r(x))'$.

Proof. Assertion (i) implies (ii): Suppose that an r -dimensional filter exists, given by (3.1). Comparing (2.2) and (3.1), we obtain:

$$\left(\exp \int_0^t h'(X_{s-t}(\cdot))(JJ')^{-1} dy_s \right) \eta_t = \sigma(\xi_t, \cdot) \cdot \nu_t$$

with

$$(3.5) \quad \eta_t = \left(\exp -\frac{1}{2} \int_0^t \|J^{-1} h(X_{s-t}(\cdot))\|^2 ds \right) (X_t)_* \mu_0.$$

Thus, η_t is absolutely continuous with respect to ν_t , and if we denote $d\eta_t/d\nu_t$ its Radon-Nikodým derivative (which is necessarily a nonnegative C^∞ function), we have:

$$(3.6) \quad \sum_{i,j=1}^p \int_0^t h_i(X_{s-t}(x))(JJ')_{ij}^{-1} dy_s^j = \log \sigma(\xi_t, x) - \log \frac{d\eta_t}{d\nu_t}(x).$$

On the other hand, since $Y \rightarrow \pi_t^Y$ and $Y \rightarrow \sigma(\xi_t)$ are two descriptions of the same input-output map, they must coincide for P -almost every input in $C^0(\mathbf{R}_+; \mathbf{R}^p)$.

But, in virtue of the following integration by parts formula:

$$\int_0^t h'(X_s(x))(JJ')^{-1} dy_s = h'(X_t(x))(JJ')^{-1} y_t - \int_0^t L_f h'(X_s(x))(JJ')^{-1} y_s ds$$

it is immediately seen from (2.2) that π_t^Y is continuous from $C^0(\mathbf{R}_+; \mathbf{R}^p)$ (endowed with the uniform topology) to $M_+(\mathbf{R}^n)$ (endowed with the strong topology). Thus, the equality of the two aforementioned input-output maps, and therefore (3.6), must hold for every observation path in $C^0(\mathbf{R}_+; \mathbf{R}^p)$. In particular, (3.6) holds for every observation path of the form: $y_t + \varepsilon \int_0^t u_s ds$, with $\varepsilon \geq 0$, $u \in L^2(\mathbf{R}_+; \mathbf{R}^p)$, and $Y \in C^0(\mathbf{R}_+; \mathbf{R}^p)$.

Now, with $u \in L^2(\mathbf{R}_+; \mathbf{R}^p)$, let us compute the directional derivative of both sides of (3.6) with respect to y in the direction $\int_0^t u_s ds$, in the quadratic mean sense. Precisely,

changing y_t into $y_t + \varepsilon \int_0^t u_s ds$ in (3.6), we get:

$$(3.7) \quad \int_0^t h'(X_{s-t}(x))(JJ')^{-1}(dy_s + \varepsilon u_s ds) = \log \sigma(\xi_t^\varepsilon, x) - \log \frac{d\eta_t}{d\nu_t}(x)$$

where ξ_t^ε is the solution of

$$(3.8) \quad d\xi_t^\varepsilon = a(\xi_t^\varepsilon) dt + \sum_{i=1}^p b_i(\xi_t^\varepsilon)(dy_t^i + \varepsilon u_t^i dt)$$

with $\xi_0^\varepsilon = \xi_0$.

Denoting "l.i.m." the limit in the mean-square sense, we obviously have:

$$(3.9) \quad \begin{aligned} \text{l.i.m.}_{\varepsilon \rightarrow 0} \frac{1}{\varepsilon} & \left(\int_0^t h'(X_{s-t}(x))(JJ')^{-1}(dy_s + \varepsilon u_s ds) - \int_0^t h'(X_{s-t}(x))(JJ')^{-1} dy_s \right) \\ &= \int_0^t h'(X_{s-t}(x))(JJ')^{-1} u_s ds. \end{aligned}$$

On the other hand, it is well known (see for example [1]) that the mapping $\varepsilon \rightarrow \xi_t^\varepsilon$ is mean-square differentiable around $\varepsilon = 0$ and, denoting $\zeta_t = \text{l.i.m.}_{\varepsilon \rightarrow 0} (1/\varepsilon)(\xi_t^\varepsilon - \xi_t)$, it is straightforward to check that ζ satisfies:

$$d\zeta_t = \frac{\partial a}{\partial \xi}(\xi_t) \zeta_t dt + \sum_{i=1}^p \frac{\partial b_i}{\partial \xi}(\xi_t) \zeta_t dy_t^i + \sum_{i=1}^p b_i(\xi_t) u_t^i dt$$

with $\zeta_0 = 0$. Thus, introducing $T(t, s)$ the fundamental matrix solution of

$$dT(t, s) = \frac{\partial a}{\partial \xi}(\xi_t) T(t, s) dt + \sum_{i=1}^p \frac{\partial b_i}{\partial \xi}(\xi_t) T(t, s) dy_t^i,$$

$$T(s, s) = I,$$

we easily get

$$\zeta_t = \sum_{i=1}^p \int_0^t T(t, s) b_i(\xi_s) u_s^i ds.$$

Hence

$$\text{l.i.m.}_{\varepsilon \rightarrow 0} \frac{1}{\varepsilon} (\log \sigma(\xi_t^\varepsilon, x) - \log \sigma(\xi_t, x)) = \left\langle \frac{\partial \log \sigma}{\partial \xi}(\xi_t, x), \sum_{i=1}^p \int_0^t T(t, s) b_i(\xi_s) u_s^i ds \right\rangle$$

for every t, x , and every u in $L^2(\mathbf{R}_+; \mathbf{R}^p)$, where $\langle \cdot, \cdot \rangle$ denotes the scalar product of \mathbf{R}^r . Thus, grouping (3.6), (3.7) and (3.9), we have proved that

$$(3.10) \quad \int_0^t h'(X_s(x))(JJ')^{-1} u_s ds = \left\langle \frac{\partial \log \sigma}{\partial \xi}(\xi_t, X_t(x)), \sum_{i=1}^p \int_0^t T(t, s) b_i(\xi_s) u_s^i ds \right\rangle$$

$$\forall u \in L^2(\mathbf{R}_+; \mathbf{R}^p), \quad \forall t \geq 0, \quad \forall x \in \mathbf{R}^n, \quad \text{and with probability 1.}$$

Now, i being fixed, let us denote 1_i the p -dimensional vector whose i th component is 1, all the others being 0.

Let us choose $\tau \in [0, t]$, and

$$u_s = 1_{[0, \tau]}(s)(JJ')1_i \quad \forall s \in \mathbf{R}_+.$$

Equation (3.10) becomes

$$\int_0^\tau h_i(X_s(x)) ds = \left\langle \rho(\xi_t, X_t(x)), \sum_{j=1}^p \int_0^\tau T(t, s) b_j(\xi_s) (JJ')_{j,i} ds \right\rangle$$

$$\forall \tau \in [0, t], \quad \forall i = 1, \dots, p,$$

with

$$\rho(\xi_t, X_t(x)) = \frac{\partial \log \sigma}{\partial \xi}(\xi_t, X_t(x)).$$

Differentiating with respect to τ , we obtain:

$$(3.11) \quad h_i(X_\tau(x)) = \left\langle \rho(\xi_t, X_t(x)), \sum_{j=1}^p T(t, \tau) b_j(\xi_\tau) (JJ')_{j,i} \right\rangle$$

$$\forall \tau \in [0, t] \text{ and } \forall i = 1, \dots, p.$$

Let us denote

$$[a, b] = \frac{\partial b}{\partial \xi} a - \frac{\partial a}{\partial \xi} b$$

the Lie bracket of the vector fields a and b , and

$$ad_a^k b = [a, ad_a^{k-1} b] \quad \forall k \geq 1 \quad \text{with } ad_a^0 b = b.$$

Applying the Itô-Stratonovitch formula to (3.11) (see [4]), we obtain:

$$L_f h_i(X_\tau(x)) d\tau = \sum_{j=1}^p \left\langle \rho(\xi_t, X_t(x)), T(t, \tau) \left([a, b_j](\xi_\tau) d\tau + \sum_{k=1}^p [b_k, b_j](\xi_\tau) dy_\tau^k \right) (JJ')_{j,i} \right\rangle.$$

Thus

$$(3.12) \quad L_f h_i(X_\tau(x)) = \sum_{j=1}^p \langle \rho(\xi_t, X_t(x)), T(t, \tau) ad_a b_j(\xi_\tau) (JJ')_{j,i} \rangle,$$

$$\sum_{j=1}^p \langle \rho(\xi_t, X_t(x)), T(t, \tau) [b_k, b_j](\xi_\tau) (JJ')_{j,i} \rangle = 0 \quad \forall k = 1, \dots, p \quad \forall \tau.$$

From (3.12) it is not difficult to prove by induction that for every $\alpha = 0, 1, \dots$

$$(3.13) \quad L_f^\alpha h_i(X_\tau(x)) = \sum_{j=1}^p \langle \rho(\xi_t, X_t(x)), T(t, \tau) ad_a^\alpha b_j(\xi_\tau) (JJ')_{j,i} \rangle,$$

$$\sum_{j=1}^p \langle \rho(\xi_t, X_t(x)), T(t, \tau) [b_k, ad_a^{\alpha-1} b_j](\xi_\tau) (JJ')_{j,i} \rangle = 0 \quad \forall k = 1, \dots, p.$$

Therefore, at $\tau = t$, and using the fact that X_t is a diffeomorphism

$$(3.14) \quad L_f^k h_i(x) = \sum_{j=1}^p \langle \rho(\xi_t, x), ad_a^k b_j(\xi_t) (JJ')_{j,i} \rangle \quad \forall k \geq 0 \quad \forall i = 1, \dots, p \quad \forall x$$

and, since $ad_a^k b_j(\xi_t)$ for every j, k , is independent of x ,

$$\mathbf{H} = \mathbf{R}\text{-Span} \{L_f^k h_i | 1 \leq i \leq p, k \geq 0\} \subset \mathbf{R}\text{-Span} \{\rho_1(\xi_t, \cdot), \dots, \rho_r(\xi_t, \cdot)\}$$

which proves that $\dim \mathbf{H} \leq r$.

Assertion (ii) implies (iii): Let $\theta_1, \dots, \theta_r$ be a basis of \mathbf{H} . Since $h_i \in \mathbf{H}$ and $L_f \mathbf{H} \subset \mathbf{H}$, the assertion is proved.

Assertion (iii) implies (i): Let us consider the r -dimensional diffusion (3.1) with

$$a(\xi) = -A'\xi, \quad b_i(\xi) = (B'(JJ')^{-1})_i \quad \forall i = 1, \dots, p$$

where $(B'(JJ')^{-1})_i$ denotes the i th column of $B'(JJ')^{-1}$, and the output functions

$$\nu_t = \eta_t \quad (\text{given by (3.5)}), \quad \sigma(\xi, x) = \exp \sum_{i=1}^r \xi^i \theta_i(x).$$

We have the following:

$$\begin{aligned} \xi'_t \theta(X_t(x)) &= \int_0^t \theta(X_s(x)) d\xi'_s + \int_0^t \xi'_s L_f \theta(X_s(x)) ds \\ &= \sum_{i,j=1}^p \int_0^t B_i \theta(X_s(x)) (JJ')_{i,j}^{-1} dy_s^j \\ &= \int_0^t h'(X_s(x)) (JJ')^{-1} dy_s. \end{aligned}$$

Thus

$$\sigma(\xi_t, x) \nu_t = \left(\exp \int_0^t h'(X_{s-t}(x)) (JJ')^{-1} dy_s \right) \nu_t = \pi_t^Y(x)$$

which proves that (a, b, ν, σ) is an r -dimensional filter for Σ . \square

In fact, we have almost proved the following minimal realization result.

THEOREM 3. *If $\dim \mathbf{H} = r < \infty$, the minimal realization of Σ 's filter is linear of dimension r , and explicitly given by:*

$$(3.15) \quad d\xi_t = -A'\xi_t dt + \sum_{i=1}^p (B'(JJ')^{-1})_i dy_t^i, \quad \xi_0 = 0,$$

$$\pi_t^Y = \exp \left(\sum_{i=1}^r \xi_t^i \theta_i \right) \cdot \nu_t,$$

with $A, B_1, \dots, B_p, \theta_1, \dots, \theta_r$ given in (iii) of Theorem 2, $\nu = \eta$ (given in (3.5)), and $(B'(JJ')^{-1})_i$ being the i th column of $B'(JJ')^{-1}$.

Proof. In (i) implies (ii), we have proved that $\dim \mathbf{H}$ is less than or equal to the dimension of any realization. On the other hand, in (iii) implies (i), the system (3.15) is proved to be a realization; and since it has the same dimension as \mathbf{H} , the result is proved. \square

Remark 1. If the unnormalized conditional measure has a density still denoted π_t^Y , the relation (3.10) can be interpreted as the gradient in the sense of Malliavin (see for example [14]) of the logarithm of π_t^Y and, by (3.13), the differentiability of this gradient with respect to τ can be seen as a robustness result in the spirit of [8].

Remark 2. In the finite dimensional deterministic nonlinear realization theory, the dimension of the minimal realization is determined by the rank of Jakubczyk (see [13]) of the input-output map. This result still holds in our infinite dimensional framework since it can be checked that this rank is precisely equal to $\dim \mathbf{H}$ and, by Theorem 3, to the dimension of the minimal filter.

Remark 3. In [21], Roth and Loparo propose to exploit the fact that the D-M-Z equation (2.3) is an ordinary first order partial differential equation after transformation into robust form, and to compute a filter via the method of characteristics. Our method has the advantages of being simpler and of providing explicitly the minimal filter.

Notice also that this filter is obtained without using Wei-Norman equations. The reader can check with the results of the next section that this method applies and produces a nonminimal and not completely explicit filter.

On the other hand, the filter (3.15) is linear and this fact is made clear in the next section, where we identify the class of systems for which (ii) of Theorem 2 holds as the class of systems that are immersible into a linear one.

4. FDF, immersion in a linear system and estimation algebra. Let us first recall what is meant by immersion in a linear system [9], [12], and by estimation algebra [5]:

DEFINITION 2. Σ is said to be immersible in the linear system Λ :

$$(\Lambda) \begin{cases} dz_t = Az_t dt, & z_0 \in \mathbf{R}^m \\ dy_t = Bz_t dt + J dv_t, & y_0 = 0 \end{cases}$$

if there exists a C^∞ mapping θ from \mathbf{R}^n to \mathbf{R}^m (m arbitrary) satisfying:

$$(4.1) \quad h(X_t(x)) = B e^{At} \theta(x) \quad \forall x \in \mathbf{R}^n \quad \forall t \geq 0.$$

The mapping θ is called the immersion from Σ to Λ .

Thus, the concept of immersion simply means that a nonlinear system has the same input-output behavior as a linear system, which has possibly a larger dimension.

DEFINITION 3. Let us denote by L_0, \dots, L_p the following operators:

$$(4.2) \quad L_0 \phi = L_f \phi - \frac{1}{2} \|J^{-1} h\|^2 \phi, \quad L_i \phi = \left(\sum_{j=1}^p (JJ')^{-1}_{i,j} h_j \right) \phi$$

$$\forall i = 1, \dots, p, \quad \forall \phi \in C^\infty(\mathbf{R}^n; \mathbf{R}).$$

The estimation algebra of Σ , denoted $E(\Sigma)$ is the Lie algebra generated by L_0, \dots, L_p .

THEOREM 4. *The following conditions are equivalent:*

- (i) Σ has an FDF.
- (ii) Σ is immersible in a linear system.
- (iii) $E(\Sigma)$ is finite dimensional.

Furthermore, if there is an immersion from Σ to the linear system Λ , then one can find an isomorphism of Lie algebras from $E(\Sigma)$ to $E(\Lambda)$.

Proof. Condition (i) equivalent to (ii): Assume first that $\dim \mathbf{H} = r < \infty$. By (iii) of Theorem 2, there exist r functions $\theta_1, \dots, \theta_r$, satisfying:

$$L_f \theta = A \theta, \quad h_i = B_i \theta, \quad i = 1, \dots, p$$

and $\mathbf{H} = \mathbf{R}\text{-Span} \{ \theta_1, \dots, \theta_r \}$.

Thus, denoting $z_t = \theta(x_t)$, we have:

$$\begin{aligned} dz_t &= L_f \theta(x_t) dt = Az_t dt, \\ dy_t &= h(x_t) dt + J dv_t = Bz_t dt + J dv_t, \end{aligned}$$

with B the matrix whose lines are B_1, \dots, B_p , and this proves that θ is the desired immersion. Conversely, if θ is an immersion from Σ to Λ , using (4.1), we have for every t, x and $i = 1, \dots, p$:

$$h_i(X_t(x)) = B_i e^{At} \theta(x)$$

and, differentiating with respect to t , it is easy to obtain for every t, x, i and every $k \geq 0$:

$$L_f^k h_i(X_t(x)) = B_i A^k e^{At} \theta(x).$$

Thus, at $t = 0$, we have:

$$\mathbf{H} = \mathbf{R}\text{-Span} \{L_f^k h_i | i = 1, \dots, p, k \geq 0\} \subset \mathbf{R}\text{-Span} \{\theta_1, \dots, \theta_r\}$$

and thus $\dim \mathbf{H} \leq r$. The conclusion follows from Theorem 2.

Condition (i) equivalent to (iii): It is not difficult to check that for every $k \geq 0$, $i, j = 1, \dots, p$, and every $\phi \in C^\infty(\mathbf{R}^n; \mathbf{R})$:

$$ad_{L_0}^k L_i(\phi) = \left(\sum_{j=1}^p (JJ')_{ij}^{-1} L_f^k h_j \right) \phi, \quad [L_j, ad_{L_0}^k L_i](\phi) = 0$$

which proves that the estimation algebra of Σ is:

$$E(\Sigma) = \mathbf{R} \cdot L_0 \oplus \mathbf{R}\text{-Span} \{(L_f^k h_i) \cdot | k \geq 0, i = 1, \dots, p\}$$

where \cdot denotes the multiplication operator. Thus, clearly, we have $\dim \mathbf{H} \leq \dim E(\Sigma) \leq \dim \mathbf{H} + 1$, which proves the equivalence.

Isomorphism of estimation algebras. Let $E(\Sigma)$ be defined as in the preceding paragraph. By the same arguments, we can easily check that the estimation algebra $E(\Lambda)$ of the corresponding linear system Λ in which Σ is immersed is:

$$E(\Lambda) = \mathbf{R} \cdot M_0 \oplus \mathbf{R}\text{-Span} \{(B_i A^k z) \cdot | k \geq 0, i = 1, \dots, p\}$$

with M_0 defined by:

$$M_0(\psi) = L_{Az} \psi - \frac{1}{2} \|J^{-1} Bz\|^2 \psi \quad \forall \psi \in C^\infty(\mathbf{R}^m; \mathbf{R}).$$

Let Θ be the linear mapping from $E(\Sigma)$ to $E(\Lambda)$ defined as follows:

$$\Theta(L_0) = M_0, \quad \Theta((L_f^k h_i) \cdot) = (B_i A^k z) \cdot, \quad \forall i = 1, \dots, p, \quad \forall k \geq 0.$$

It is a consequence of the immersion that $E(\Sigma)$ and $E(\Lambda)$ have the same dimension and thus Θ is one to one. Also it is straightforward to check that

$$\Theta([L, M]) = [\Theta(L), \Theta(M)] \quad \forall L, M \in E(\Sigma)$$

which achieves the proof. \square

Remark 4. In [12], the equivalence between (ii) and (iii) is proved for one-dimensional outputs ($p = 1$), and in [21] a special case of Theorem 4 is obtained: If $E(\Sigma)$ is nilpotent, then Σ has an FDF.

Remark 5. As a consequence of Theorems 3 and 4, it appears that there is a duality, in the classical meaning of the theory of linear systems, between the minimal filter (3.15) and the linear system Λ in which Σ is immersed.

Remark 6. One might try to take profit of the linearity of Λ by remarking that a filter can be obtained via Makowski's method [17] for linear systems with non-Gaussian initial measure. Unfortunately, this filter has twice the minimal dimension and is not completely explicit since it involves an inverse Fourier transform.

Remark 7. In the language of statistics, the state ξ of the minimal filter is a finite dimensional vector of sufficient statistics, and it must be noted that, similarly to the static parameter estimation problem, the conditional measure belongs to an exponential family.

Remark 8. The results of this section prove that the existence of FDF is basically a property of linear systems when expressed in suitable (nonminimal) coordinates, and thus is quite restrictive. Real applications exist however (see [16] where the problem is solved in discrete time).

Acknowledgments. The author is indebted to Professors E. Pardoux, M. Fliess and R. Marino for very helpful suggestions.

REFERENCES

- [1] L. ARNOLD, *Stochastic Differential Equations: Theory and Applications*, John Wiley, New York, 1973.
- [2] J. S. BARAS, *Group invariance methods in nonlinear filtering of diffusion processes*, Proc. 19th CDC Conference, Albuquerque, NM, 1980.
- [3] V. E. BENES, *Exact finite dimensional filters for certain diffusions with nonlinear drifts*, Stochastics, 5 (1981), pp. 65-92.
- [4] J. M. BISMUT, *Mécanique aléatoire*, Lecture Notes in Mathematics 866, Springer, Berlin-New York, 1981.
- [5] R. W. BROCKETT, *Remarks on finite dimensional nonlinear estimation*, in *Analyse des systèmes*, C. Lobry, ed., Astérisque 75, 76 (SMF), pp. 47-56.
- [6] ———, *Classification and equivalence in estimation theory*, Proc. 18th CDC Conference, Ft. Lauderdale, FL, 1979.
- [7] M. CHALEYAT-MAUREL AND D. MICHEL, *Un théorème de non existence de filtre de dimension finie*, C. R. Acad. Sci., Paris, 296 (1983), pp. 933-936.
- [8] ———, *Une propriété de robustesse en filtrage non linéaire*, to appear.
- [9] D. CLAUDE, M. FLIESS AND A. ISIDORI, *Immersion, directe et par bouclage, d'un système non linéaire dans un linéaire*, C.R. Acad. Sci., Paris, Sér. I 296 (1983), pp. 237-240.
- [10] M. FLIESS AND I. KUPKA, *A finiteness criterion for nonlinear input-output differential systems*, this Journal, 21 (1983), pp. 721-728.
- [11] M. HAZEWINKEL, S. I. MARCUS AND H. J. SUSSMANN, *Nonexistence of exact finite dimensional filters for conditional statistics of the cubic sensor problem*, Systems Control Lett., 3 (1983), pp. 331-340.
- [12] O. HIJAB, *A class of infinite dimensional filters*, Proc. 19th CDC Conference, Albuquerque, NM, 1980.
- [13] B. JAKUBCZYK, *Existence and uniqueness of realizations of nonlinear systems*, this Journal, 18 (1980), pp. 455-471.
- [14] S. KUSUOKA AND D. W. STROOCK, *Applications of the Malliavin calculus, Part I*, to appear.
- [15] J. LEVINE AND G. PIGNIE, *Exact finite dimensional filters via systems realization for a class of discrete-time nonlinear systems*, Systems Control Lett., 5 (1985), pp. 403-412.
- [16] ———, *The finite dimensional filtering problem for a class of nonlinear discrete-time systems*, Proc. 9th IFAC World Congress, Budapest, Hungary, 1984.
- [17] A. M. MAKOWSKI, *Results on the filtering problem for linear systems with non-Gaussian initial conditions*, Proc. 21st CDC, Orlando, FL, December, 1982.
- [18] S. I. MARCUS, *Algebraic and geometric methods in nonlinear filtering*, this Journal, 22 (1984), pp. 817-844.
- [19] S. K. MITTER, *Geometric theory of nonlinear filtering*, in *Outils et modèles mathématiques en automatique, analyse des systèmes et traitement du signal*, Part 3, I. Landau, ed., Editions du CNRS, Paris, 1982, pp. 37-60.
- [20] D. L. OCONE, *Topics in nonlinear filtering theory*, Ph.D. thesis, Massachusetts Institute of Technology, Cambridge, MA, 1980.
- [21] Z. S. ROTH AND K. A. LOPARO, *Optimal filter realization for a class of nonlinear systems with finite dimensional estimation algebra*, Systems Control Lett., 4 (1984), pp. 23-26.
- [22] W. WONG, *New classes of finite dimensional filters*, Systems Control Lett., 3 (1983), pp. 155-164.

STATE CONSTRAINTS IN OPTIMAL CONTROL: A CASE STUDY IN PROXIMAL NORMAL ANALYSIS*

FRANK H. CLARKE[†] AND PHILIP D. LOEWEN[‡]

Abstract. We consider an optimal control problem on a given interval $[0, T]$ whose trajectories must satisfy the state constraint $g(t, x(t)) \leq 0$ a.e. Infinite-dimensional perturbations of this constraint give rise to a value function V , whose epigraph is a closed set containing sensitivity information, controllability and penalization results and even necessary conditions for optimality. Studying $\text{epi } V$ when the domain of V is $L^2[0, T]$ and $AC^2[0, T]$ allows the derivation of a variety of necessary conditions, each with its own merits, and provides a concrete illustration of the intricacies of infinite-dimensional proximal normal analysis.

Key words. optimal control, state constraints, necessary conditions, value function, proximal analysis, proximal normal analysis

AMS(MOS) subject classification. 49B10

Foreword. Value functions have a long and distinguished history in the study of dynamic optimization. For instance, they provide the cornerstone of the Hamilton–Jacobi sufficiency theory. Until recently, however, the role of such functions in the study of necessary conditions has been limited to their appearance in rather informal attempts to motivate and interpret multiplier rules like the maximum principle. The passage from informal arguments to the rigorous and systematic use of value functions in the study of necessary conditions is not an easy one, the main obstacle being lack of smoothness. Value functions are defined by minimization, an operation which may produce corners, discontinuities, and even infinite values. Functions with such bad behaviour, although beyond the capacity of classical analysis, can be successfully studied using generalized gradients. The modern theory of nonsmooth analysis is now capable of giving value functions a rigorous treatment which does justice to their intuitive significance.

The present article is designed to substantiate the assertions above in the case of a well-known problem: the derivation of necessary conditions for optimal control problems with state constraints. We consider a differential inclusion formulation:

$$P(0) \quad \min \{f(x(0), x(T)) : x'(t) \in F(t, x(t)) \text{ a.e., } x(0) \in C, g(t, x(t)) \leq 0\}.$$

Additive perturbations of the state constraint $g(t, x) \leq 0$ lead to a value function V which dominates the discussion. Besides its obvious sensitivity implications, a first-order analysis of the behaviour of V near 0 permits the simultaneous derivation and interpretation of necessary conditions and controllability results. This completely new approach leads to necessary conditions comparable to those in the literature [11], [14], [16]–[18], with the considerable advantage of an interpretation which is both intuitive and rigorous.

1. Introduction. Let H be a Hilbert space on which a lower semicontinuous function $V: H \rightarrow \mathbf{R} \cup \{+\infty\}$ is given; assume $V(0) < +\infty$. The *generalized gradient* of V at 0, denoted $\partial V(0)$, is a closed convex set of points in $H^* = H$ which describes the local behaviour of V near 0. Although this set can be defined in terms of limits of

* Received by the editors June 16, 1986; accepted for publication (in revised form) December 3, 1986. This work was supported by the Natural Sciences and Engineering Research Council of Canada.

[†] Centre de Recherches Mathématiques, Université de Montréal, Montréal, Québec, Canada, H3C 3J7.

[‡] Electrical Engineering Department, Imperial College, London SW7 2BT, England.

difference quotients, a geometrical definition based on the closed set $\text{epi } V = \{(\alpha, r): r \geq V(\alpha)\}$ has recently proven extremely useful. We define

$$(1.1) \quad \begin{aligned} \partial V(0) &= \{\zeta: (\zeta, -1) \in N_{\text{epi } V}(0, V(0))\}, \\ \partial^\infty V(0) &= \{\zeta: (\zeta, 0) \in N_{\text{epi } V}(0, V(0))\}, \end{aligned}$$

where $N_{\text{epi } V}(0, V(0))$ denotes the normal cone to the set $\text{epi } V$ at $(0, V(0))$, a set for which the next paragraph provides a formula.

Let C be a closed subset of H containing a point c . The set of *proximal normals* to C at c , denoted $PN_C(c)$, consists of all $v \in H$ with the following property:

$$(1.2) \quad \exists M > 0 \text{ s.t. } \langle v, c' - c \rangle \leq M \|c' - c\|^2 \quad \forall c' \in C.$$

The set $PN_C(c)$ is a convex cone containing 0. The normal cone to C at c can be recovered from the proximal normal cones nearby:

$$(1.3) \quad N_C(c) = \text{cl co} \{w - \lim_{i \rightarrow \infty} v_i: v_i \in PN_C(c_i), c_i \rightarrow c \text{ in } C\}.$$

This is the "proximal normal formula" due to Clarke [6], whose validity in infinite-dimensional Banach spaces has recently been established by Borwein and Strojwas [3] (see also [13]). Its relevance to us is that it essentially reduces the study of N_C to the study of vectors v obeying (1.2). And (1.2) can be viewed as an optimization problem in its own right, since it says that c minimizes the functional $\langle -v, c' \rangle + M \|c' - c\|^2$ over C . When C is the epigraph of a suitably chosen value function V , we find that $\partial V(0)$ can thus be studied in terms of a sequence of auxiliary minimization problems of a particularly simple form. This approach has been used in a broad range of finite-dimensional applications (see [6], [9], [5], [8], [15] for example); and infinite-dimensional applications are multiplying quickly ([7], [12]).

In our application, the proximal normal inequality (1.2) leads to an auxiliary optimization problem of the same form as $P(0)$, but without state constraints:

$$(Q) \quad \min \{f(x(0), x(T)): x'(t) \in F(t, x(t)) \text{ a.e., } x(0) \in C\}.$$

Necessary conditions for optimality in (Q) are well known, and involve the *distance function* $d_C(x) = \inf \{ \|x - c\|: c \in C \}$ and the *Hamiltonian* $H(t, x, p) = \sup \{ p \cdot v: v \in F(t, x) \}$.

PROPOSITION 1.1 [5]. *Let the data in problem (Q) obey (H1)–(H4) listed in §2. If x solves (Q), and $\rho > (2K_f + 2)(1 + K \ln K)$, then there exists an arc $p \in AC([0, T]; \mathbf{R}^n)$ such that*

- (a) $(-p'(t), x'(t)) \in \partial H(t, x(t), p(t))$ a.e.;
- (b) $(p(0), -p(T)) \in (\delta_0, \delta_T) + \rho |(1, E)| \partial d_C(x(0)) \times \{0\}$ for some $(\delta_0, \delta_T) \in \partial f(x(0), x(T))$.

Here ∂H signifies the generalized gradient of H in the (x, p) variables only, and $E = \delta_0 - p(0)$.

We conclude this introduction by noting that when the domain H of a lower semicontinuous (l.s.c.) function $V: H \rightarrow \mathbf{R} \cup \{+\infty\}$ is infinite dimensional, complications unknown when $H = \mathbf{R}^n$ may arise. For example, we may have $\partial V(0) = \emptyset$ and $\partial^\infty V(0) = \{0\}$ even for a lower semicontinuous convex function V . J. M. Borwein gives an example of such a situation in [2], where he also identifies a class of comparatively well-behaved functions. He calls a lower semicontinuous map $V: H \rightarrow \mathbf{R} \cup \{+\infty\}$ *Lipschitz-like at 0* if $V(0) < +\infty$ and if there exist an open set Ω containing $(0, V(0))$, a closed convex set C with locally weakly compact polar, and a constant $\varepsilon > 0$ such that

$$(1.4) \quad \Omega \cap \text{epi } V + \lambda C \subseteq \text{epi } V \quad \forall \lambda \in (0, \varepsilon).$$

This condition is easy to verify whenever V is directionally Lipschitzian at 0. Moreover, it holds for any lower semicontinuous function V when $H = \mathbf{R}^n$, since one may then choose $\Omega = \mathbf{R}^n \times \mathbf{R}$ and $C = \{(0, 0)\}$. (In particular, many Lipschitz-like functions on \mathbf{R}^n fail to be directionally Lipschitzian.)

PROPOSITION 1.2 [2]. *If a lower semicontinuous map $V: H \rightarrow \mathbf{R} \cup \{+\infty\}$ is Lipschitz-like at 0, then the following are equivalent:*

- (a) V is Lipschitz near 0;
- (b) $\partial^\infty V(0) = \{0\}$;
- (c) $\partial V(0)$ is bounded and nonempty.

Consequently, any such V obeys either $\partial V(0) \neq \emptyset$ or $\partial^\infty V(0) \neq \{0\}$.

PROPOSITION 1.3. *Let $P \subseteq H$ be a closed convex cone with vertex at 0 and nonempty interior. If an l.s.c. map $V: H \rightarrow \mathbf{R} \cup \{+\infty\}$ is finite at 0 and obeys*

$$(1.5) \quad \beta - \alpha \in P \Rightarrow V(\alpha) \leq V(\beta),$$

then V is Lipschitz-like at 0.

Proof. To verify (1.4), take $\varepsilon = 1$, $\Omega = H \times \mathbf{R}$, and $C = -P \times \{0\}$. Any point in the left side of (1.4) has the form $(\alpha, v) + (-\beta, 0)$ for $v \geq V(\alpha)$ and $\beta \in P$. Then (1.5) implies $V(\alpha - \beta) \leq V(\alpha) \leq v$, so $(\alpha - \beta, v) \in \text{epi } V$ as required. The local weak compactness of C^0 follows because $\text{int } P \neq \emptyset$. \square

2. The value function. Let B denote the closed unit ball in \mathbf{R}^n . The following hypotheses concerning the data of problem $P(0)$ are assumed throughout this paper:

- (H1) The multifunction $F: [0, T] \times \mathbf{R}^n \rightarrow \mathbf{R}^n$ has nonempty compact convex values. For each $x \in \mathbf{R}^n$, $F(\cdot, x)$ is measurable.
- (H2) There are nonnegative functions $\varphi(t), k(t) \in L^2[0, T]$ such that
 - (a) $F(t, x) \subseteq \varphi(t)B \quad \forall t \in [0, T], \quad \forall x \in \mathbf{R}^n$;
 - (b) $F(t, y) \subseteq F(t, x) + k(t)|y - x|B \quad \forall t \in [0, T], \quad \forall x, y \in \mathbf{R}^n$.
 We define $K_F = \exp(\int_0^T k(t) dt)$.

(H3) The set C is compact.

(H4) The function $f: \mathbf{R}^n \times \mathbf{R}^n \rightarrow \mathbf{R}$ is globally Lipschitz of rank K_F .

These are standard assumptions for problem $P(0)$, identical to those in [5] except for the assumption $k \in L^2$ in (H2) instead of $k \in L^1$ in [5]. This condition assures that any F -trajectory (i.e. any absolutely continuous function x obeying $x'(t) \in F(t, x(t))$ a.e.) lies in the Hilbert space

$$(2.1) \quad AC^2([0, T], \mathbf{R}^n) = \{x = AC([0, T]; \mathbf{R}^n): x' \in L^2([0, T]; \mathbf{R}^n)\}.$$

This space, also known as $H^1(0, T)$ or $W^{1,2}(0, T)$, comes with inner product

$$(2.2) \quad \langle x, y \rangle = x(0) \cdot y(0) + \int_0^T x'(t) \cdot y'(t) dt;$$

evidently AC^2 is isometric to $\mathbf{R}^n \times L^2$, with $x(t)$ and $(x_0, v(t))$ being identified iff $x(t) = x_0 + \int_0^t v(s) ds$. In particular, $x_i(t)$ converges weakly in AC^2 to $x(t)$ iff $x_i(0) \rightarrow x(0)$ and $x'_i \rightarrow x'$ weakly in L^2 . When this is the case, one has $\sup_i \|x'_i\|_2 < +\infty$, and the Arzela-Ascoli Theorem implies that $x_i \rightarrow x$ uniformly on $[0, T]$.

The conditions governing the state constraint function $g: [0, T] \times \mathbf{R}^n \rightarrow \mathbf{R}$ depend on our choice of the perturbation scheme and the interpretation of the state constraint. In general terms, let H be a Hilbert space of functions $\alpha: [0, T] \rightarrow \mathbf{R}$, and let $P \subseteq H$ be a closed convex cone with vertex at 0 defining the "positive" elements of H : write

$\alpha \leq_P 0$ iff $-\alpha \in P$. Then additive perturbations of the state constraint by elements $\alpha \in H$ give rise to a value function $V: H \rightarrow \mathbf{R} \cup \{+\infty\}$ as follows:

$$P(\alpha) \quad V(\alpha) := \min \{f(x(0), x(T)): x'(t) \in F(t, x(t)) \text{ a.e., } x(0) \in C, \\ g(t, x(t)) + \alpha(t) \leq_P 0\}.$$

One possible choice of H is $L^2[0, T]$, with P consisting of all functions which are nonnegative almost everywhere; another possibility is $H = AC^2[0, T]$ with P consisting of functions which are nonnegative everywhere. In any case, $\alpha(\cdot)$ and $g(\cdot, x(\cdot))$ must be comparable, so we cannot hope to get far without the following assumption.

- (h5) For all F -trajectories x , the map $t \rightarrow g(t, x(t))$ lies in H ; also, if a sequence of F -trajectories $\{x_i\}$ converges to x weakly in AC^2 , then $g(t, x_i(t)) \rightarrow g(t, x(t))$ weakly in H .

The hypothesis will be strengthened considerably when we come to consider specific choices for H and P . For our present purposes, however, it is enough to establish a satisfactory existence property and prove the lower semicontinuity of V .

PROPOSITION 2.1. (a) *Let $\{x_i\}$ be a sequence of F -trajectories with $x_i(0) \in C$. Then $\{x_i\}$ has a subsequence converging weakly in AC^2 to an F -trajectory x such that $x(0) \in C$.*

(b) *Suppose that, along with the sequence $\{x_i\}$ in part (a), there is a sequence $\{\alpha_i\} \subseteq H$ such that $g(t, x_i(t)) + \alpha_i(t) \in -P$ for all i , and such that $\alpha_i \rightarrow \alpha$ weakly in H . Then the subsequence in (a) can be chosen so that one obtains the limiting relationship $g(t, x(t)) + \alpha(t) \in -P$.*

(c) *If $V(\alpha) < +\infty$ for some α , then $P(\alpha)$ has a solution.*

(d) *The function V is weakly sequentially lower semicontinuous. (In particular, V is norm lower semicontinuous.)*

Proof. Proposition 2.1(a) is a minor revision of [5, Thm. 3.1.7, p. 118]; in view of (h5), (b) follows immediately from (a).

To prove (c), let $\{x_i\}$ be a minimizing sequence for $P(\alpha)$. Taking $\alpha_i = \alpha$ in (b), we find that there is a subsequence along which x_i tends weakly in AC^2 (hence uniformly) to an F -trajectory x obeying both $x(0) \in C$ and $g(t, x(t)) + \alpha(t) \in -P$. By (H4), $f(x(0), x(T)) = \lim f(x_i(0), x_i(T)) = V(\alpha)$.

As for (d), let $\alpha_i \rightarrow \alpha$ weakly in H and $V(\alpha_i) \rightarrow v$. We must show $v \geq V(\alpha)$. If $v = +\infty$ this is trivial, so assume $v < +\infty$. Then (c) gives a solution x_i for each $P(\alpha_i)$: in particular,

$$V(\alpha_i) = f(x_i(0), x_i(T)) \quad \text{and} \quad g(t, x_i(t)) + \alpha_i(t) \in -P.$$

Mimicking the arguments of (c) leads to the conclusion that $v = f(x(0), x(T))$ for some limiting F -trajectory x admissible for $P(\alpha)$. Hence $V(\alpha) \leq v$ as required. \square

Without specifying the space H further, we can state the basic features of a proximal normal to the closed set $\text{epi } V$ in $H \times \mathbf{R}$.

THEOREM 2.2. *Let $(\zeta, -\varepsilon) \in PN_{\text{epi } V}(\alpha^*, v^*)$. Then there is a constant $M > 0$ and a solution x^* for $P(\alpha^*)$ such that*

- (a) $\varepsilon \geq 0$ and $\varepsilon[v^* - f(x^*(0), x^*(T))] = 0$;
 (b) $-\zeta \in N_P(r^*)$, where $r^*(t) = -g(t, x^*(t)) - \alpha^*(t)$;
 (c) among all F -trajectories x starting in C , x^* minimizes

$$\varepsilon f(x(0), x(T)) + \langle \zeta(t), g(t, x(t)) \rangle + M \|g(t, x(t)) - g(t, x^*(t))\|^2 \\ + M |f(x(0), x(T)) - f(x^*(0), x^*(T))|^2.$$

Proof. For any F -trajectory x starting in C , we have

$$V(-g(t, x(t))) \leq f(x(0), x(T)).$$

Moreover, this inequality is preserved if a scalar $s \geq 0$ is added to the right side and a function $r \in P$ is subtracted from the argument of V . Thus

$$(-g(t, x(t)) - r(t), f(x(0), x(T)) + s) \in \text{epi } V.$$

From the definition of a proximal normal (1.2), there exists $M > 0$ such that (suppressing the t -dependence of x for simplicity)

$$(2.3) \quad \begin{aligned} &\langle (\zeta, -\varepsilon), (-g(t, x) - r, f(x) + s) - (\alpha^*, v^*) \rangle \\ &\leq M \|(-g(t, x) - r, f(x) + s) - (\alpha^*, v^*)\|^2. \end{aligned}$$

Now since $V(\alpha^*) \leq v^* < +\infty$, Proposition 2.1(c) implies that $P(\alpha^*)$ has a solution x^* . Taking $x = x^*$ and $r = r^* = -\alpha^* - g(t, x^*)$ in (2.3) gives

$$0 \leq \varepsilon(s - v^* + f(x^*(0), x^*(T))) + M|s - v^* + f(x^*(0), x^*(T))|^2 \quad \forall s \geq 0.$$

This implies that $s^* = v^* - f(x^*(0), x^*(T))$ gives a minimum to the right-hand side over the set $[0, +\infty)$. Hence the right derivative of this expression must vanish if $s^* > 0$, or at least be nonnegative if $s^* = 0$. This is conclusion (a).

If we now take $x = x^*$ and $s = s^*$ in (2.3), we get

$$\langle -\zeta, r - r^* \rangle \leq M \|r - r^*\|^2 \quad \forall r \in P.$$

This says $-\zeta \in PN_P(r^*)$, which implies $-\zeta \in N_P(r^*)$ by the proximal normal formula. This is conclusion (b).

Finally, taking $r = r^*$ and $s = s^*$ in (2.3) gives conclusion (c). \square

The conclusions of Theorem 2.2 illustrate the general claims made in § 1: the proximal normals which will ultimately characterize $\partial V(0)$ can be understood in terms of the differential inclusion problem (c) *without state constraints*. Conclusion (a) simply captures the geometrically obvious fact that a proximal normal to an epigraph must point downward. Since P is a closed convex cone, conclusion (b) may be decomposed as follows:

$$(2.4) \quad -\zeta \in N_P(r^*) \Leftrightarrow \langle \zeta, r^* \rangle = 0 \quad \text{and} \quad -\zeta \in N_P(0).$$

Note that none of our calculations so far have precluded the study of equality state constraints of the form $g(t, x(t)) = 0$. The choice $P = \{0\}$ allows Proposition 2.1 and Theorem 2.2 to be used in this case also.

3. The choice $H = L^2[0, T]$. A Hilbert space of functions very common in analysis is $L^2[0, T]$, with the inner product

$$\langle \alpha, \beta \rangle = \int_0^T \alpha(t) \beta(t) dt.$$

Given a closed subset J of $[0, T]$, a natural choice for the positive cone is

$$P = \{\alpha \in L^2[0, T]: \alpha(t) \geq 0 \text{ a.e. } t \in J\}.$$

In this case the state constraint $g(t, x(t)) \leq_P 0$ must be interpreted as

$$g(t, x(t)) \leq 0 \quad \text{a.e. } t \in J.$$

With these specifications, the value function $V: L^2[0, T] \rightarrow \mathbf{R} \cup \{+\infty\}$ of § 2 is well defined and permits a straightforward application of proximal normal analysis. Loewen has given a detailed account of this procedure in [12]. We sketch his main results in this section, paying particular attention to the subtleties of the infinite-dimensional context.

We proceed under the following hypothesis.

- (H5) The state constraint function $g: [0, T] \times \mathbf{R}^n \rightarrow \mathbf{R}$ is Lebesgue measurable in t and Lipschitz in x , with $|g(t, y) - g(t, x)| \leq K_g |y - x|$ for all $t \in [0, T]$, $x, y \in \mathbf{R}^n$. Also $\int_0^T |g(t, 0)|^2 < +\infty$.

(Note that (H5) \Rightarrow (h5), and that any function g defined only on $J \times \mathbf{R}^n$ but satisfying (H5) on J is easily extended to $[0, T] \times \mathbf{R}^n$ in a way which preserves (H5).)

LEMMA 3.1. Let $r^* \in P$ and $\zeta \in L^2[0, T]$ be given. The following are equivalent:

- (a) $-\zeta \in N_P(r^*)$;
- (b) $\zeta(t) \geq 0$ a.e., $\zeta(t)r^*(t) = 0$ a.e., $\zeta(t) = 0$ a.e. $t \notin J$.

Proof. Clearly (b) \Rightarrow (a). Conversely, $-\zeta \in N_P(r^*)$ implies both $\langle \zeta, r^* \rangle = 0$ and $\langle \zeta, r \rangle \geq 0$ for all $r \in P$ by (2.4). The second of these statements implies $\zeta(t) \geq 0$ a.e., with $\zeta(t) = 0$ a.e. $t \notin J$. The first then forces $\zeta(t)r^*(t) = 0$ a.e. \square

PROPOSITION 3.2. Let $(\zeta, -\varepsilon) \in PN_{\text{epi } V}(\alpha^*, v^*)$. Then $P(\alpha^*)$ has a solution x^* to which there corresponds an arc $p \in AC^2([0, T]; \mathbf{R}^n)$ and a measurable selection $\gamma(t) \in \partial_x g(t, x^*(t))$ a.e. such that

- (a) $\varepsilon \geq 0$, $-\zeta \in N_P(-g(t, x^*(t)) - \alpha^*(t))$;
- (b) $(-p'(t) + \zeta(t)\gamma(t), x^{*'}(t)) \in \partial H(t, x^*(t), p(t))$ a.e.;
- (c) $(p(0), -p(T)) \in \varepsilon \partial f(x^*(0), x^*(T)) + N_C(x^*(0)) \times \{0\}$.

Proof. Conclusion (a) follows from Theorem 2.2(a). Items (b) and (c) arise when Proposition 1.1 is applied to the minimization problem described in Theorem 2.2(c). Details are given in [12]. \square

PROPOSITION 3.3. Assume $V(0) < +\infty$. Suppose $(\zeta, -\varepsilon) = w\text{-}\lim (\zeta_i, -\varepsilon_i)$ for a weakly convergent sequence $(\zeta_i, -\varepsilon_i)$ of vectors proximal normal to $\text{epi } V$ at base points $(\alpha_i, v_i) \rightarrow (0, V(0))$. Then $P(0)$ has a solution x to which there corresponds an arc p and a measurable selection $\gamma(t) \in \partial_x g(t, x(t))$ a.e. on $\{t \in [0, T]: \zeta(t) \neq 0\}$ such that

- (a) $\varepsilon \geq 0$, $-\zeta \in N_P(-g(t, x(t)))$;
- (b) $(-p'(t) + \zeta(t)\gamma(t), x'(t)) \in \partial H(t, x(t), p(t))$ a.e.;
- (c) $(p(0), -p(T)) \in \varepsilon \partial f(x(0), x(T)) + N_C(x(0)) \times \{0\}$.

Proof. Proposition 3.2 asserts that to each α_i there corresponds a solution x_i to $P(\alpha_i)$ together with corresponding quantities p_i and γ_i such that conclusions (a)–(c) hold for each i . We are now concerned with their validity in the limit.

The arguments of Proposition 2.1 show that $\{x_i\}$ has a subsequence converging uniformly to a limiting arc x which actually solves $P(0)$. Hence conclusions (c) _{i} show that $\{p_i(T)\}$ is a bounded sequence, while $\{p'_i\}$ is a sequence of functions bounded in $L^2([0, T]; \mathbf{R}^n)$ by (b) _{i} . So along a further subsequence, p_i converges uniformly to some arc p , and $p'_i \rightarrow p'$ weakly in L^2 . Taking limits in (c) _{i} with the aid of [5, Prop. 2.1.5(b)] and [5, Prop. 2.4.2] yields conclusion (c). (To be completely precise, we recover (c) from the slightly stronger versions of (c) _{i} in which $N_C(x_i(0))$ is replaced by $R\partial d_C(x_i(0))$ for a large constant R independent of i , as in Proposition 1.1(b).)

As for (a), it is clear that the weak limit of points $\varepsilon_i \geq 0$ and $\zeta_i \in P$ obeys $\varepsilon \geq 0$ and $\zeta \in P$. Also, the uniform convergence of x_i to x implies that $g(t, x_i(t)) \rightarrow g(t, x(t))$ strongly in $L^2[0, T]$. Consequently the relationship $\zeta_i(t)[g(t, x_i(t)) + \alpha_i(t)] = 0$ a.e. $t \in J$ implies $\zeta(t)g(t, x(t)) = 0$ a.e. $t \in J$. Finally, $\zeta(t) = 0$ a.e. $t \notin J$ is obvious, so we have $-\zeta \in N_P(-g(t, x(t)))$ as (a) asserts.

Finally, it remains to verify (b). By a modification of [17, Lemma 4.5], we are assured that the bounded sequence $\gamma_i(t)\zeta_i(t)$ in $L^2([0, T]; \mathbf{R}^n)$ has a subsequence converging weakly to $\gamma(t)\zeta(t)$, where ζ is the given weak limit of $\{\zeta_i\}$ and where $\gamma(t) \in \partial_x g(t, x(t))$ a.e. on $\{t: \zeta(t) \neq 0\}$. Therefore the left-hand sides $(-p'_i + \gamma_i\zeta_i, x'_i)$ converge weakly to $(-p' + \gamma\zeta, x')$. At the same time, the Hamiltonian arguments

(x_i, p_i) converge uniformly to (x, p) . By [5, Prop. 3.1.7], the limiting relationship (b) follows. \square

Proposition 3.3 enables us to give a succinct characterization of $N_{\text{epi } V}(0, V(0))$. Let us write Σ for the set of solutions to $P(0)$; for every $x \in \Sigma$ and $\varepsilon > 0$, we write $M^\varepsilon(x)$ for the collection of all vectors ζ satisfying Proposition 3.3(a)–(c). We then have

$$(3.1) \quad N_{\text{epi } V}(0, V(0)) = \text{cl co } [N \cup N^\infty],$$

where

$$(3.2) \quad \begin{aligned} N &= \{\lambda(\zeta, -1) : \lambda \geq 0, \zeta \in M^1(\Sigma) \cap \partial V(0)\}, \\ N^\infty &= \{(\zeta, 0) : \zeta \in M^0(\Sigma) \cap \partial^\infty V(0)\}. \end{aligned}$$

Lines (3.1) and (3.2) carry information about the sensitivity of problem $P(\alpha)$, and about its relation to multiplier rules. For example, if $N_{\text{epi } V}(0, V(0)) \neq \{0\}$ then (3.1) implies that $N \cup N^\infty$ contains a nonzero point. Definitions (3.2) then imply that either $M^1(\Sigma)$ or $M^0(\Sigma) \setminus \{0\}$ is nonempty. This fact, which we may write as

$$(3.3) \quad M^1(\Sigma) \cup [M^0(\Sigma) \setminus \{0\}] \neq \emptyset,$$

affirms that some solution x to $P(0)$ has either a nontrivial multiplier with $\varepsilon = 0$ or else a multiplier with $\varepsilon = 1$.

We emphasize that the multiplier rule (3.3) is valid only under the assumption that $N_{\text{epi } V}(0, V(0)) \neq \{0\}$. If the domain of V were finite-dimensional, this assumption would hold automatically. But with $H = L^2[0, T]$ the situation is considerably more delicate, and it is possible for every weakly convergent sequence appearing in the proximal normal formula (1.3) to converge to zero. J. M. Borwein [2] presents a comparatively simple instance of $P(\alpha)$, meeting all our hypotheses, for which this pathology occurs. In the remainder of this report we will discuss three ways to avoid this problem. The first is to introduce a certain constraint qualification (“calmness”) to guarantee the nontriviality of $N_{\text{epi } V}(0, V(0))$. Alternatively, one can use different limiting techniques along a sequence of proximal normals to derive a nontrivial multiplier rule without regard for monitoring the sensitivity of $P(\alpha)$. Finally, one can start over at the beginning with a different choice of H , in the hopes that a “smaller” infinite-dimensional space might give better results. Each of these alternatives has a lesson to teach: the dense validity of the calmness condition shows that the analysis above is not trivial; taking weak limits of measures leads to the most widely applicable necessary conditions given in this paper; and the analysis when $H = AC^2[0, T]$ shows how strongly the outcome of the proximal normal approach depends upon the choice of V ’s domain, while giving sensitivity information unknown until now.

Calmness. Problem $P(\alpha)$ is *calm* at α_0 if $V(\alpha_0) < +\infty$ and

$$\liminf_{\alpha \rightarrow \alpha_0} \frac{V(\alpha) - V(\alpha_0)}{\|\alpha - \alpha_0\|} > -\infty.$$

Since $V : L^2[0, T] \rightarrow \mathbf{R} \cup \{+\infty\}$ is lower semicontinuous, [1, Thm. VI.65, p. 281] implies that the points at which $P(\alpha)$ is calm are dense in $\text{Dom } V = \{\alpha \in L^2[0, T] : V(\alpha) < +\infty\}$.

Calmness is a well-respected constraint qualification in other settings—the calculus of variations and mathematical programming, for example [5]. Here, too, it guarantees the nontriviality and normality of the necessary conditions proposed in (3.3).

THEOREM 3.4. *Suppose $P(\alpha)$ is calm at 0. Then for any $x \in \Sigma$ there is a vector $\zeta \in L^2[0, T]$, an arc p , and a measurable selection $\gamma(t) \in \partial_x g(t, x(t))$ a.e. on $\{t: \zeta(t) \neq 0\}$ such that*

$$(a) \quad \zeta(t) \geq 0 \text{ a.e., } \zeta(t)g(t, x(t)) = 0 \text{ a.e., } \zeta(t) = 0 \text{ a.e. } t \notin J;$$

$$(b) \quad (-p'(t) + \zeta(t)\gamma(t), x'(t)) \in \partial H(t, x(t), p(t)) \text{ a.e.};$$

$$(c) \quad (p(0), -p(T)) \in \partial f(x(0), x(T)) + N_C(x(0)) \times \{0\}.$$

If x is the only element of Σ , then ζ may be chosen to lie in $\partial V(0)$; in any case, we have

$$(3.4) \quad \partial V(0) = \text{cl co } [M^1(\Sigma) \cap \partial V(0) + M^0(\Sigma) \cap \partial^\infty V(0)].$$

Proof. Formula (3.4) is valid even without the assumption of calmness, being a direct consequence of (3.1)–(3.2) ([13, Prop. 4.2]). But when $P(\alpha)$ is calm at 0, [12, Prop. III.4.2] implies $\partial V(0) \neq \emptyset$, so (3.4) has nontrivial content. Indeed, one of its consequences is that $M^1(\Sigma) \cap \partial V(0) \neq \emptyset$. Consequently there exists $x \in \Sigma$ and $\zeta \in \partial V(0)$ such that (a)–(c) hold. If Σ is a singleton then the proof is complete. If Σ contains several points, then we may fix our attention on any $x^* \in \Sigma$ and consider the modified problem $P^*(\alpha)$, which is identical to $P(\alpha)$ in every way except that the cost functional $f(x(0), x(T))$ is replaced by $f(x(0), x(T)) + (\int_0^T |x(t) - x^*(t)|^2 dt)^2$. Now $\Sigma^* = \{x^*\}$ and

$$V^*(\alpha) \geq V(\alpha) \quad \forall \alpha, \quad V^*(0) = V(0).$$

Consequently $P^*(\alpha)$ is calm at 0, and the conclusions derived above allow one to choose $\zeta^* \in \partial V^*(0)$ such that (a*)–(c*) hold. These conclusions reduce immediately to the desired results (a)–(c). \square

In order to compare the conclusions of Theorem 3.4 with the known necessary conditions of Vinter and Pappas [17], let us define an arc

$$q(t) = p(t) - \int_0^t \gamma(s)\zeta(s) ds.$$

Then conditions (a)–(c) become

$$(a) \quad \zeta(t) \geq 0 \text{ a.e., } \zeta(t)g(t, x(t)) = 0 \text{ a.e., } \zeta(t) = 0 \text{ a.e. } t \notin J;$$

$$(b) \quad (-q'(t), x'(t)) \in \partial H\left(t, x(t), q(t) + \int_0^t \gamma(s)\zeta(s) ds\right);$$

$$(c) \quad (q(0), -q(T) - \int_0^T \gamma(s)\zeta(s) ds) \in \partial f(x(0), x(T)) + N_C(x(0)) \times \{0\}.$$

To recover the measure conditions of [17], it suffices to introduce the nonnegative measure μ via $d\mu = \zeta(t) dt$. Conclusion (a) then states that μ is supported on $J \cap \text{cl } \{t: g(t, x(t)) = 0\}$, while conclusions (b) and (c) assume the forms given in [17]. Thus Theorem 3.4 is very close in spirit to the necessary conditions proposed in the literature, with one very significant advantage: it shows that the measure μ may be taken to be absolutely continuous with respect to Lebesgue measure, with a square-integrable density ζ . This advantage is due both to the method of proof via proximal normals, and to the assumption of calmness—a condition known to hold arbitrarily near to any problem of interest.

Remark. The proximal normal approach used here adapts very easily to the treatment of *equality state constraints*, which have the form $g(t, x(t)) = 0$ a.e. Indeed, it suffices to take the positive cone $P = \{0\}$ to find that Theorem 3.4 remains valid in this case, except that conclusion (a), which states that $-\zeta \in N_P(-g(t, x(t)))$, becomes trivial. In the remainder of this paper, however, we make much more use of the detailed properties of N_P . To go beyond Theorem 3.4 in the case of equality state constraints would therefore require separate study.

Weak* convergence of measures. The scope of proximal normal analysis goes considerably beyond the simple application of (1.3). For example, the theory guarantees that for any closed subset C of a Hilbert space, the boundary of C contains a dense subset of points c where $PN_C(c) \neq \{0\}$. So even in the case where $N_C(c) = \{0\}$, there is no shortage of nontrivial proximal normal cones based near c . In terms of our problem, we can certainly find a weakly convergent sequence of unit vectors $(\zeta_i, -\varepsilon_i) \in PN_{\text{epi } v}(\alpha_i, v_i)$ at points $(\alpha_i, v_i) \rightarrow (0, V(0))$. Despite the possibility that every such sequence might tend weakly to zero in $L^2 \times \mathbf{R}$, nontrivial results are still available if we take limits in a different topology. The cost of doing so is the loss of sensitivity information about $P(\alpha)$; the benefit is a proof of known necessary conditions which is closely related to the proximal approach, but which requires no constraint qualification.

Given a weakly convergent sequence of unit vectors $(\zeta_i, -\varepsilon_i)$ as described above, Proposition 3.2 asserts that there is a sequence x_i of solutions to $P(\alpha_i)$ such that certain conclusions hold for some p_i and γ_i . Let us define constants m_i , arcs q_i , and measures μ_i via

$$(3.5) \quad m_i = \varepsilon_i + \|\zeta_i\|_1,$$

$$(3.6) \quad q_i(t) = \left(p_i(t) - \int_0^t \gamma_i(s) \zeta_i(s) ds \right) / m_i,$$

$$(3.7) \quad d\mu_i = (\zeta_i(t)/m_i) dt.$$

We then have

$$-m_i q_i'(t) = -p_i'(t) + \gamma_i(t) \zeta_i(t) \quad \text{a.e.},$$

and the conclusions of Proposition 3.2 imply

$$(3.8) \quad \varepsilon_i \geq 0, \quad \zeta_i(t) \geq 0 \quad \text{a.e.}, \quad \zeta_i(t) = 0 \quad \text{a.e. } t \notin J,$$

$$(3.9) \quad (-q_i'(t), x_i'(t)) \in \partial H\left(t, x_i(t), q_i(t) + \int_0^t \gamma_i(s) d\mu_i(s)\right),$$

$$(3.10) \quad \left(q_i(0), -q_i(T) - \int_0^T \gamma_i(s) d\mu_i(s) \right) \in \left(\frac{\varepsilon_i}{m_i} \right) \partial f(x_i(0), x_i(T)) + N_C(x_i(0)) \times \{0\}.$$

Definitions (3.5) and (3.7), together with condition (3.8), imply that each μ_i is a nonnegative measure supported on $J \subseteq [0, T]$ with $\mu_i(J) \leq 1$. Hence some subsequence of $\{\mu_i\}$ converges weakly* to a nonnegative measure μ supported on J with

$$\mu(J) = \lim_{i \rightarrow \infty} \mu_i(J).$$

Defining $\lambda_i = \varepsilon_i/m_i \in [0, 1]$ implies that (perhaps along a further subsequence) $\lambda_i \rightarrow \lambda$ for some $\lambda \in [0, 1]$, and that $\lambda + \mu(J) = 1$. Now exactly the same limiting argument used in [5, p. 129] establishes the following result, in which we use the notation

$$\overline{N_C}(c) = \{\lim n_i : n_i \in N_C(c_i), c_i \rightarrow c \text{ in } C\},$$

$$\overline{\partial_x g}(t, x) = \{\lim z_i : z_i \in \partial_x g(t_i, x_i), (t_i, x_i) \rightarrow (t, x)\}.$$

PROPOSITION 3.5. *Assume (H1)–(H5). If $V(0) < +\infty$ then $P(0)$ has a solution x to which there correspond an arc q , a measurable function γ , a constant λ and a nonnegative measure μ such that*

(a) $\lambda \in [0, 1]$, μ is supported on J , and $\lambda + \mu(J) = 1$;

(b) $\gamma(t) \in \overline{\partial_x g}(t, x(t)) \mu$ -a.e.;

- (c) $(-q'(t), x'(t)) \in \partial H(t, x(t), q(t) + \int_0^t \gamma(s) d\mu(s));$
 (d) $(q(0), -q(T) - \int_0^T \gamma(s) d\mu(s)) \in \lambda \partial f(x(0), x(T)) + \overline{N_C}(x(0)) \times \{0\}.$

The only substantial difference between Proposition 3.5 and [17, Thm. 3.1] concerns the support of the measure μ . In [17], Vinter and Pappas assume that g is upper semicontinuous in t and deduce that μ is supported on the set of "active" constraint times

$$J \cap \{t: g(t, x(t)) = 0\}.$$

Taking $J = [0, T]$ for simplicity, we now derive their result from Proposition 3.5.

THEOREM 3.6 [17]. *In addition to (H1)–(H5), assume that $J = [0, T]$ and that $g(\cdot, x)$ is upper semicontinuous (u.s.c.) for each fixed x . Then for every solution x to $P(0)$, there correspond an arc q , a measurable function γ , a constant λ , and a nonnegative measure μ such that*

- (a) $\lambda \in [0, 1]$, μ is supported on $\{t: g(t, x(t)) = 0\}$, and $\lambda + \mu[0, T] = 1;$
 (b) $\gamma(t) \in \overline{\partial_x g}(t, x(t))$ μ -a.e.;
 (c) $(-q'(t), x'(t)) \in \partial H(t, x(t), q(t) + \int_0^t \gamma(s) d\mu(s));$
 (d) $(q(0), -q(T) - \int_0^T \gamma(s) d\mu(s)) \in \lambda \partial f(x(0), x(T)) + \overline{N_C}(x(0)) \times \{0\}.$

Proof. Let us assume for the moment that x is the unique solution to $P(0)$. This temporary hypothesis will be dispensed with later. For any positive integer i , consider the problem $P_i(0)$ which is identical to $P(0)$ in every way except that the closed set J is replaced by

$$J_i = \{t: g(t, x(t)) \geq -1/i\}.$$

(J_i is closed because g is u.s.c. in t .) Now x remains the unique local solution for $P_i(0)$, since any arc y which is admissible for $P_i(0)$ and obeys $\|y - x\|_\infty \leq (iK_g)^{-1}$ has

$$g(t, y(t)) \leq 0 \quad \text{for } t \in J_i,$$

$$g(t, y(t)) \leq g(t, x(t)) + K_g \|y - x\|_\infty < (-1/i) + i^{-1} = 0 \quad \text{for } t \notin J_i.$$

Thus y is admissible for $P(0)$ and hence performs strictly worse than x . We may apply Proposition 3.5 to $P_i(0)$ to find q_i , γ_i , λ_i , and μ_i , such that

- (a)_i $\lambda_i \in [0, 1]$, μ_i is supported on J_i , and $\lambda_i + \mu_i[0, T] = 1;$
 (b)_i $\gamma_i(t) \in \overline{\partial_x g}(t, x(t))$ μ_i -a.e.;

(c)_i $(-q'_i(t), x'(t)) \in \partial H\left(t, x(t), q_i(t) + \int_0^t \gamma_i(s) d\mu_i(s)\right);$

(d)_i $\left(q_i(0), -q_i(T) - \int_0^T \gamma_i(s) d\mu_i(s)\right) \in \lambda_i \partial f(x(0), x(T)) + \overline{N_C}(x(0)) \times \{0\}.$

Once again, there is a subsequence along which $\mu_i \rightarrow \mu$ weak*, $\lambda_i \rightarrow \lambda$, $q_i \rightarrow q$ uniformly, $\gamma_i d\mu_i \rightarrow \gamma d\mu$ weak* for some $\gamma(t) \in \overline{\partial_x g}(t, x(t))$ μ -a.e., and so on. Taking limits in (b)_i–(d)_i yields (b)–(d) of the theorem. As for conclusion (a), the support of μ_i is the closed set J_i . Since $J_1 \supseteq J_2 \supseteq \dots$, it follows that the support of the limiting measure μ lies in

$$\bigcap_{i=1}^{\infty} J_i = \{t: g(t, x(t)) = 0\},$$

as required.

Now if x is one of several solutions to $P(0)$, then it is easy to define a modified problem $P^*(0)$ for which x is the unique solution and the conclusions derived above imply that x obeys (a)–(d). The idea here is to modify the objective function $f(y(0), y(T))$ by adding a term of the form $\int_0^T |y(t) - x(t)|^2 dt$. Details are given in the proof of [5, Thm. 3.5.2]. \square

4. The choice $H = AC^2[0, T]$. In the special case when the state constraint function $g(t, x)$ is smooth, the mapping $t \rightarrow g(t, x(t))$ actually lies in $AC^2[0, T]$ for any F -trajectory x . The space AC^2 , defined in (2.1), (2.2), has certain advantages over L^2 as the domain of V . The most striking of these is that the natural positive cone

$$P = \{\alpha \in AC^2[0, T]: \alpha(t) \geq 0 \forall t \in [0, T]\}$$

actually has interior. Hence Propositions 1.2 and 1.3 imply that either $\partial V(0) \neq \emptyset$ or $\partial^\infty V(0) \neq \{0\}$, and consequently that weak limits of proximal normals will yield both nontrivial necessary conditions and meaningful sensitivity results without any constraint qualifications at all. These facts make it worthwhile to pursue the analysis when $H = AC^2[0, T]$ even though this space is more difficult to work with than $L^2[0, T]$ and despite the extra regularity assumptions which must be imposed on F and g to complete the analysis.

Throughout this section, (H1)–(H4) of § 2 are augmented by

(H5) The function $g: \mathbf{R}^n \rightarrow \mathbf{R}$ is autonomous and of class C^2 on \mathbf{R}^n .

(H6) The nonnegative function φ appearing in (H2)(a) is actually constant.

The smoothness of g required by (H5) will appear as a necessary evil in the proof of Proposition 4.4. Explicit time dependence of g can be allowed, provided that $\partial g / \partial t$ is smooth in x , but we have chosen to simplify the presentation by taking $\partial g / \partial t \equiv 0$.

Let us now interpret the conclusions of Theorem 2.2 in the space $H = AC^2[0, T]$. The first step is to consider the normal cone $N_P(r^*)$.

PROPOSITION 4.1. *Let $r^* \in P$ and $\zeta \in AC^2[0, T]$ be given. The following are equivalent:*

(a) $-\zeta \in N_P(r^*)$;

(b) *there is a finite nonnegative measure μ with support in $Z[r^*] = \{t \in [0, T]: r^*(t) = 0\}$ such that*

$$\zeta(t) = \mu[0, T] + t\mu(0, T) - \int_0^t \mu(0, s) ds.$$

Proof. The arguments of Hestenes [10, p. 50] show that $-\zeta \in N_P(0)$ if and only if $\zeta'(t)$ is a nonincreasing function which obeys

$$(4.1) \quad \zeta(0) \geq \zeta'(0+), \quad \zeta'(T-) \geq 0.$$

Recalling (2.4), one may verify that a function of this sort obeys $\langle \zeta, r^* \rangle = 0$ iff ζ is constant on all relatively open subintervals of $\Omega[r^*] = \{t \in [0, T]: r^*(t) > 0\}$ and equality holds in (4.1) iff $r^*(0) > 0$, respectively, $r^*(T) > 0$. Certainly the ζ defined in (b) meets these criteria; conversely, any ζ in $-N_P(r^*)$ gives rise to a measure as in (b) when we set

$$\mu\{0\} = \zeta(0) - \zeta'(0+), \quad \mu\{T\} = \zeta'(T-), \quad \mu(s, t) = \zeta'(s) - \zeta'(t). \quad \square$$

The conclusions of Proposition 4.1 support the following remarkable fact.

PROPOSITION 4.2. *Let $\{\zeta_i\}$ be a sequence in $-N_P(0)$ converging weakly to ζ . Then $\{\zeta_i\}$ actually converges strongly to ζ ; in fact, $\{\zeta'_i\}$ converges to ζ' in every $L^p[0, T]$ for $p \in [1, \infty)$.*

Proof. Since $\{\zeta_i\}$ is weakly convergent, it is norm bounded. In particular, there exists $M > 0$ such that $\zeta_i(0) \in [0, M]$ for all i . From (4.1) it follows that $\zeta'_i(t) \in [0, M]$ for all i a.e. Moreover, each ζ'_i is nonincreasing (see the proof of Proposition 4.1). Thus Helly's selection theorem implies that any subsequence of ζ'_i has a further subsequence converging pointwise to a nonincreasing limit function, which must be

ζ' . This shows that $\zeta'_i \rightarrow \zeta'$ a.e., so the bounded convergence theorem implies that $\zeta'_i \rightarrow \zeta'$ in any $L^p[0, T]$, $p \in [1, \infty)$. In particular $\zeta'_i \rightarrow \zeta'$ in L^2 and it follows that $\zeta_i \rightarrow \zeta$ strongly in $AC^2[0, T]$. \square

COROLLARY 4.3. *Let $\{r_i\} \subseteq P$ be a sequence converging weakly to r , and let $\{\zeta_i\}$ be a sequence converging weakly to ζ while obeying $-\zeta_i \in N_P(r_i)$ for all i . Then one has $-\zeta \in N_P(r)$.*

Proof. We have $\zeta_i \in -N_P(r_i) \subseteq -N_P(0)$ for all i by (2.4). Since $-N_P(0)$ is convex and strongly closed, it is also weakly closed and $\zeta \in -N_P(0)$. Also by (2.4), we have $\langle \zeta_i, r_i \rangle = 0$ for all i . Now $r_i \rightarrow r$ weakly by assumption and $\zeta_i \rightarrow \zeta$ strongly by Proposition 4.2, so we obtain $\langle \zeta, r \rangle = 0$ in the limit. From (2.4) it now follows that $-\zeta \in N_P(r)$. \square

Now we turn to the remaining conclusions of Theorem 2.2.

PROPOSITION 4.4. *Let $(\zeta, -\varepsilon) \in PN_{\text{epi } v}(\alpha^*, v^*)$. Then $P(\alpha^*)$ has a solution x^* to which there corresponds an arc $p \in AC^2([0, T]; \mathbf{R}^n)$ such that*

- (a) $\varepsilon \geq 0$ and $-\zeta \in N_P(-g(x^*(t)) - \alpha^*(t))$;
- (b) $(-p'(t) + \zeta'(t)g_{xx}(x^*(t))x^{*'}(t), x^{*'}(t)) \in \partial H(t, x^*(t), p(t) - \zeta'(t)g_x(x^*(t)))$ a.e.;
- (c) $(p(0) - \zeta(0)g_x(x^*(0)), -p(T)) \in \varepsilon \partial f(x^*(0), x^*(T)) + N_C(x^*(0)) \times \{0\}$.

Proof. Conclusion (a) follows from Theorem 2.2(a, b); we will deduce conclusions (b) and (c) from Theorem 2.2(c). We know that among all F -trajectories x starting in C , x^* gives the lowest value to the functional

$$\begin{aligned} \varepsilon f(x(0), x(T)) + \zeta(0)g(x(0)) + \int_0^T \zeta'(t)g(x(t))' dt \\ + M|g(x(0)) - g(x^*(0))|^2 + M|f(x(0), x(T)) - f(x^*(0), x^*(T))|^2 \\ + M \int_0^T |g(x(t))' - g(x^*(t))'|^2 dt. \end{aligned}$$

Now a careful review of Clarke's proof of [5, Thm. 3.2.6] (the result we have quoted as Proposition 1.1) shows that the necessary conditions for this minimization problem are the same as those when $M = 0$. The intuitive reason for this is that the conclusions of Proposition 1.1 are first-order conditions, and the terms multiplied by M here are all of second order. The technical justification of this fact relies on (H6).

To write the first-order conditions when $M = 0$ we introduce the auxiliary state $y \in AC[0, T]$, governed by

$$y'(t) = \zeta'(t)g_x(x(t))x'(t), \quad y(0) = 0.$$

Then the arc (x^*, y^*) , where $y^*(t) = \int_0^t \zeta'(s)g_x(x^*(s))x^{*'}(s) ds$, satisfies the conclusions of Proposition 1.1 for the problem of minimizing

$$\varepsilon f(x(0), x(T)) + \zeta(0)g(x(0)) + y(T)$$

over all trajectories $(x(t), y(t))$ for the multifunction

$$G(t, x, y) = \{(v, \zeta'(t)g_x(x)v) : v \in F(t, x)\}.$$

The Hamiltonian for this problem, evaluated at (x, y, p, q) , equals

$$\sup \{pv + q\zeta'(t)g_x(x)v : v \in F(t, x)\} = H(t, x, p + q\zeta'(t)g_x(x)).$$

We deduce the existence of arcs $p(t)$ and $q(t) \equiv -1$ such that

$$\begin{aligned} (-p'(t), x^{*'}(t)) \in \{(d - \zeta'(t)g_{xx}(x^*(t))e, e) : (d, e) \in \partial H(t, x^*(t), p(t) - \zeta'(t)g_x(x^*(t)))\}, \\ (p(0), -p(T)) \in \varepsilon \partial f(x^*(0), x^*(T)) + (\zeta(0)g_x(x^*(0)), 0) + N_C(x^*(0)) \times \{0\}. \end{aligned}$$

These reduce easily to the desired conclusions (b) and (c). \square

Remark 4.5. The crux of the preceding proof is our ability to take $M = 0$ when writing down necessary conditions for the auxiliary problem solved by x^* . This is the only place in the development where (H6) is required. For smooth relaxed problems where $F(t, x) = \psi(t, x, U)$, even (H6) can be removed. Indeed, let $\psi: [0, T] \times \mathbf{R}^n \times \mathbf{R}^n \rightarrow \mathbf{R}^n$ be measurable in t , differentiable in x , with $\psi_x(t, \cdot, \cdot)$ jointly continuous, and suppose U is compact. We also require $F(t, x) = \psi(t, x, U)$ to obey (H1), (H2). In this situation the analysis of the auxiliary problem solved by x^* involves the additional arc y given by

$$\begin{aligned} y'(t) &= \zeta'(t)g_x(x(t))x'(t) + M|g_x(x(t))x'(t) - g_x(x^*(t))x^{*'}(t)|^2, \\ y(0) &= 0, \end{aligned}$$

and the arc (x^*, y^*) really minimizes

$$\begin{aligned} \varepsilon f(x(0), x(T)) + \zeta(0)g(x(0)) + y(T) + M|g(x(0)) - g(x^*(0))|^2 \\ + M|f(x(0), x(T)) - f(x^*(0), x^*(T))|^2 \end{aligned}$$

over all trajectories $(x(t), y(t))$ for the multifunction

$$G(t, x, y) = \{(v, \zeta'(t) \cdot g_x(x)v + M|g_x(x)v - g_x(x^*(t))x^{*'}(t)|^2): v \in \psi(t, x, U)\}.$$

This multifunction is not convex, but it does satisfy all other hypotheses of Proposition 1.1. And since we are dealing with a free-endpoint problem, it follows [5, p. 117] that (x^*, y^*) continues to solve the problem above when $\text{co } G$ is used instead of G . The latter problem satisfies the hypotheses of Proposition 1.1, in whose conclusions the Hamiltonian is now effectively

$$\mathcal{H}_M(t, x, p) = \sup \{(p - \zeta'(t)g_x(x)) \cdot v - M|g_x(x)v - g_x(x^*(t))x^{*'}(t)|^2: v \in \psi(t, x, U)\}.$$

Under the above conditions on ψ , $\mathcal{H}_M(t, \cdot, \cdot)$ can be viewed as the maximum over the fixed compact set U of a family of functions whose (x, p) -gradients are continuous in x, p , and u . Consequently this function is regular [5, Thm. 2.8.2]. Since $\mathcal{H}_M(t, x, p) \leq \mathcal{H}_0(t, x, p)$ with equality when $x = x^*(t)$, it follows that

$$\partial \mathcal{H}_M(t, x^*(t), p) \subseteq \partial \mathcal{H}_0(t, x^*(t), p) \quad \forall p \in \mathbf{R}^n.$$

Thus the Hamiltonian inclusion for the auxiliary problem, when $M > 0$, implies the Hamiltonian inclusion we would obtain for the problem when $M = 0$, and the proof of Proposition 4.4 proceeds as above. Clearly, any hypotheses ensuring that $\mathcal{H}_M(t, \cdot, \cdot)$ is regular may be used in place of $F(t, x) = \psi(t, x, U)$ in this argument. \square

To simplify our study of the limiting stability of Proposition 4.4(a)–(c), we introduce some notation. The set $\Sigma(\alpha)$ will consist of all solutions $x(\cdot)$ of problem $P(\alpha)$: by Proposition 2.1(c), $V(\alpha) < +\infty \Rightarrow \Sigma(\alpha) \neq \emptyset$. If $x \in \Sigma(\alpha)$ and $\varepsilon \geq 0$, the *index ε multiplier set for x* , denoted $M^\varepsilon(x)$, consists of all $\zeta \in AC^2[0, T]$ for which there is an arc $p(\cdot)$ such that

- (a) $\varepsilon \geq 0$ and $-\zeta \in N_P(-g(x(\cdot)) - \alpha(\cdot))$;
- (b) $(-p'(t) + \zeta'(t)g_{xx}(x(t))x'(t), x'(t)) \in \partial H(t, x(t), p(t) - \zeta'(t)g_x(x(t)))$ a.e.;
- (c) $(p(0) - \zeta(0)g_x(x(0)), -p(T)) \in \varepsilon \partial f(x(0), x(T)) + \overline{N_C}(x(0)) \times \{0\}$.

(Recall the definition of $\overline{N_C}$ given just before Proposition 3.5.)

The multiplier sets have a useful scaling property:

$$(4.2) \quad \zeta \in M^\varepsilon(x) \Leftrightarrow \forall \lambda > 0, \quad \lambda \zeta \in M^{\lambda \varepsilon}(x).$$

In case $\Sigma(\alpha)$ contains several points, we define $M^\varepsilon(\Sigma(\alpha))$ to be the union of $M^\varepsilon(x)$ over x in $\Sigma(\alpha)$. This definition allows Proposition 4.4 to be restated as follows:

$$(4.3) \quad (\zeta, -\varepsilon) \in PN_{\text{epi } V}(\alpha^*, v^*) \Rightarrow \zeta \in M^\varepsilon(\Sigma(\alpha^*)).$$

To emphasize the relationship between the multiplier sets defined here and those current in the literature ([17], [5]), we introduce the arc q and the measure μ via

$$(4.4) \quad q(t) = p(t) - \zeta(0)g_x(x(0)) - \int_0^t \zeta'(s)g_{xx}(x(s))x'(s) ds,$$

$$(4.5) \quad \zeta(t) = \mu[0, T] + t\mu(0, T) - \int_0^t \mu(0, s) ds.$$

(Compare Proposition 4.1.)

We may then reduce $p(t) - \zeta'(t)g_x(x(t))$ to

$$\begin{aligned} q(t) + \mu[0, T]g_x(x(0)) + \int_0^t \mu[s, T]g_x(x(s))' ds - \mu[t, T]g_x(x(t)) \\ = q(t) + \int_0^t g_x(x(s)) d\mu(s). \end{aligned}$$

It follows that $M^\varepsilon(x)$ can also be written as the set of all ζ of the form (4.5) arising from a nonnegative measure μ on $[0, T]$ and an arc q such that

- (a') $\varepsilon \geq 0$ and $\text{supp } (\mu) \subseteq \{t \in [0, T]: g(x(t)) + \alpha(t) = 0\}$;
- (b') $(-q'(t), x'(t)) \in \partial H(t, x(t), q(t) + \int_0^t g_x(x(s)) d\mu(s))$;
- (c') $(q(0), -q(T) - \int_0^T g_x(x(s)) d\mu(s)) \in \varepsilon \partial f(x(0), x(T)) + \overline{N_C}(x(0)) \times \{0\}$.

In view of (4.3), the following result makes a statement about weak limits of proximal normals.

PROPOSITION 4.6. *Let $\{\alpha_i\} \subseteq AC^2[0, T]$ be a sequence converging strongly to α , along which $V(\alpha_i) \rightarrow V(\alpha)$. If a sequence $\{(\zeta_i, -\varepsilon_i)\} \subseteq AC^2[0, T] \times \mathbf{R}$ converges weakly to $(\zeta, -\varepsilon)$ and obeys $\zeta_i \in M^{\varepsilon_i}(\Sigma(\alpha_i))$ for all i , then we have the limiting relationship $\zeta \in M^\varepsilon(\Sigma(\alpha))$. Moreover, $\zeta_i \rightarrow \zeta$ in norm.*

Proof. Since $\zeta_i \in M^{\varepsilon_i}(\Sigma(\alpha_i))$ for all i , it follows that ζ_i is a weakly convergent sequence of elements in $-N_P(0)$. Proposition 4.2 implies that $\zeta_i \rightarrow \zeta$ strongly in $AC^2[0, T]$.

Implicit in the hypotheses of this result is the existence of a solution x_i to each $P(\alpha_i)$, with a corresponding arc p_i , for which the conditions (a)–(c) defining $\zeta_i \in M^{\varepsilon_i}(x_i)$ hold. Now the growth and Lipschitz conditions (H2) on the multifunction F imply similar restrictions on ∂H (see [5, Prop. 3.2.4, p. 121]), so some subsequence of (x_i, p_i) converges weakly in AC^2 to a limiting arc (x, p) . This follows from Proposition 2.1(a); the proof of Proposition 2.1 also shows that one has $g(x(t)) + \alpha(t) \in -P$ and $f(x(0), x(T)) \leq \limsup f(x_i(0), x_i(T)) \leq V(\alpha)$. Hence $x \in \Sigma(\alpha)$.

Now since $\zeta_i'(t)g_{xx}(x_i(t))$ is bounded and converges pointwise to $\zeta'(t)g_{xx}(x(t))$ by Proposition 4.2 and (H5), we infer that

$$(-p_i'(t) + \zeta_i'(t)g_{xx}(x_i(t))x_i'(t), x_i'(t)) \rightarrow (-p'(t) + \zeta'(t)g_{xx}(x(t))x'(t), x'(t))$$

weakly in $L^2([0, T], \mathbf{R}^n \times \mathbf{R}^n)$. Also

$$(x_i(t), p_i(t) - \zeta_i'(t)g_x(x_i(t))) \rightarrow (x(t), p(t) - \zeta'(t)g_x(x(t))) \quad \text{a.e.}$$

The proof of [5, Thm. 3.1.7] shows that this allows us to pass to the limit in Proposition 4.4(b) to obtain

$$(-p'(t) + \zeta'(t)g_{xx}(x(t))x'(t), x'(t)) \in \partial H(t, x(t), p(t) - \zeta'(t)g_x(x(t))) \quad \text{a.e.}$$

Finally, Proposition 4.4(c) states that for some $\delta_i \in \partial f(x_i(0), x_i(T))$ one has

$$(p_i(0) - \zeta_i(0)g_x(x_i(0)), -p_i(T)) - \varepsilon_i \delta_i \in N_C(x_i(0)) \times \{0\}.$$

By passing to a further subsequence if necessary, we may assume that $\delta_i \rightarrow \delta \in \partial f(x(0), x(T))$; it follows that

$$(p(0) - \zeta(0)g_x(x(0)), p(T)) - \varepsilon\delta \in \overline{N_C}(x(0)) \times \{0\}.$$

This completes the proof. \square

COROLLARY 4.7. *If $V(\alpha) < +\infty$ and $M^0(\Sigma(\alpha)) = \{0\}$, then $M^1(\Sigma(\alpha))$ is bounded.*

Proof. If the conclusion were false, then $M^1(\Sigma(\alpha))$ would contain a sequence ζ_i with $\|\zeta_i\| \rightarrow +\infty$. Consider the vectors $z_i = \zeta_i / \|\zeta_i\|$. They form a bounded sequence, so we can extract a subsequence converging weakly to some z . By (4.2), we have

$$z_i \in M^{\varepsilon_i}(\Sigma(\alpha)), \quad \varepsilon_i = \|\zeta_i\|^{-1} \rightarrow 0.$$

It follows from Proposition 4.6 that $z_i \rightarrow z$ strongly, so $\|z\| = 1$, and that $z \in M^0(\Sigma(\alpha))$. This is a contradiction. \square

The following result, using the notation $\Sigma = \Sigma(0)$, establishes the relationship between the multiplier sets $M^e(\Sigma)$ and $N_{\text{epi } V}(0, V(0))$.

PROPOSITION 4.8. *Assume $V(0) < +\infty$, and define N, N^∞ as in (3.2). Then*

$$(4.6) \quad N_{\text{epi } V}(0, V(0)) = \text{cl co } [N \cup N^\infty].$$

In particular, $N \cup N^\infty$ contains nonzero points and therefore

$$(4.7) \quad M^1(\Sigma) \cup [M^0(\Sigma) \setminus \{0\}] \neq \emptyset.$$

Proof. Propositions 1.2 and 1.3 imply that $N_{\text{epi } V}(0, V(0))$ contains nonzero points, so (4.7) will follow immediately once we prove (4.6). Now the definitions given in (1.1) show that the inclusion \supseteq holds in (4.6). To prove the reverse inclusion, it suffices to show that any weak limit of proximal normals $(\zeta, -\varepsilon)$ as described in Proposition 4.6 lies in $N \cup N^\infty$. This is clear if $\varepsilon = 0$, for then $\zeta \in \partial^\infty V(0)$ by construction, while $\zeta \in M^0(\Sigma)$ by Proposition 4.6. So $(\zeta, 0) \in N^\infty$. Likewise, if $\varepsilon > 0$ then the vector $(\zeta, -\varepsilon)/\varepsilon = (\zeta/\varepsilon, -1)$ lies in $N_{\text{epi } V}(0, V(0))$ by construction, so $\zeta/\varepsilon \in \partial V(0)$. But we also have $\zeta/\varepsilon \in M^1(\Sigma)$ by Proposition 4.6 and the scaling property (4.2). Hence $(\zeta, -\varepsilon) \in N$. \square

Conclusion (4.7) is a succinct statement of the necessary conditions for optimality in $P(0)$. It says that some $x \in \Sigma$ either gives rise to a multiplier $\zeta \in M^1(x)$, or else yields a nontrivial multiplier $\zeta \in M^0(x)$. This is easily extended to treat each x in Σ separately: indeed, the same proof as in [5, Thm. 3.5.2] establishes the following.

THEOREM 4.9 (Necessary Conditions). *Let $V(0) < +\infty$. Then $\Sigma \neq \emptyset$, and*

$$(4.8) \quad M^1(x) \cup [M^0(x) \setminus \{0\}] \neq \emptyset \quad \forall x \in \Sigma.$$

As advertised, however, proximal normal analysis does more than offer a new approach to necessary conditions for optimality. It also enables us to write down a formula for $\partial V(0)$.

THEOREM 4.10 (Sensitivity). *Let $V(0) < +\infty$ and assume (H1)–(H6). Then*

$$(4.9) \quad \partial V(0) = \text{cl co } [\partial V(0) \cap M^1(\Sigma) + \partial^\infty V(0) \cap M^0(\Sigma)].$$

If $M^0(\Sigma) = \{0\}$ then $\partial^\infty V(0) = \{0\}$. In this case $\partial V(0)$ is a nonempty norm-compact convex subset of $AC^2[0, T]$ obeying

$$\text{ext } \partial V(0) \subseteq M^1(\Sigma).$$

Proof. Conclusion (4.9) follows straight from Propositions 4.8 and [13, Prop. 4.2]. To deduce the additional conclusions, note first that $M^0(\Sigma) = \{0\}$ implies $N^\infty = \{0\}$, so

N contains nonzero points by Proposition 4.8. In particular, $\partial V(0)$ is a nonempty set which, by (4.9), obeys

$$(4.10) \quad \partial V(0) = \text{cl co } [\partial V(0) \cap M^1(\Sigma)] \subseteq \text{cl co } M^1(\Sigma).$$

Now $M^1(\Sigma)$ is a bounded set by Corollary 4.7, so the set $\partial V(0)$ is a nonempty, closed, convex, and bounded subset of $-N_P(0)$. It follows that every sequence ζ_i chosen from $\partial V(0)$ has a weakly convergent subsequence, and that this subsequence actually converges strongly by Proposition 4.2. So $\partial V(0)$ is compact, as claimed. The conclusion that $\text{ext } \partial V(0) \subseteq M^1(\Sigma)$ follows from (4.10) and the fact that $M^1(\Sigma)$ is closed is proven in Proposition 4.6. \square

Controllability and penalization. Problem $P(0)$ is called *normal* if $M^0(\Sigma) = \{0\}$. From Theorem 4.10 we deduce that the value function for a normal problem is Lipschitz near 0, a fact which has attractive consequences both for controllability and for exact penalization. For instance, a value function which is Lipschitz near 0 must certainly be finite near 0, from which it follows that a feasible trajectory for the system

$$x'(t) \in F(t, x(t)) \quad \text{a.e.,} \quad x(0) \in C, \quad g(x(t)) \leq -\alpha(t) \quad \forall t$$

exists whenever α is chosen sufficiently near to 0 in AC^2 . Of course, a Lipschitz condition on V implies more than the local finiteness of V , so stronger controllability conclusions than this are readily available. We will not present these in detail: the reader may consult [5, Thm. 3.5.3] to discover the spirit of such results.

A related issue, with applications in numerical analysis, is exact penalization. If the original problem is normal, then all of its solutions will be found among the solutions to a modified problem in which there are no state constraints. The AC^2 -norm is used in the following result.

COROLLARY 4.11. *Suppose $P(0)$ is normal, and $x^* \in \Sigma$. Then there is a constant $M > 0$ so large that x^* provides a local solution for the problem*

$$\min \{f(x(0), x(T)) + M \|g(x(t))^+\| : x'(t) \in F(t, x(t)) \text{ a.e., } x(0) \in C\}.$$

Here $g^+ = \max\{0, g\}$, and “local solution” refers to a neighbourhood of x^* in the norm topology of AC^2 .

Proof. Since $P(0)$ is normal, $V(\cdot)$ is Lipschitz near 0 by Theorem 4.10. In particular, there exist positive constants ε and M such that

$$\|\alpha\| < \varepsilon \Rightarrow V(0) \leq V(\alpha) + M \|\alpha\|.$$

Now (H5) ensures that the mapping $x \rightarrow g(x)^+$ from AC^2 into AC^2 is continuous near x^* , so there exists $\delta > 0$ such that $\|x - x^*\| < \delta$ implies $\|g(x(\cdot))^+\| < \varepsilon$. But for any F -trajectory x with $x(0) \in C$, one has $V(-g(x(\cdot))^+) \leq f(x(0), x(T))$. So if x also obeys $\|x - x^*\| < \delta$, we obtain

$$\begin{aligned} f(x^*(0), x^*(T)) &= V(0) \\ &\leq V(-g(x)^+) + M \|g(x)^+\| \\ &\leq f(x(0), x(T)) + M \|g(x)^+\|. \end{aligned} \quad \square$$

REFERENCES

- [1] J.-P. AUBIN AND I. EKELAND, *Applied Nonlinear Analysis*, John Wiley, New York, 1984.
- [2] J. M. BORWEIN, *Epi-Lipschitz-like sets in Banach space: Theorems and examples*, to appear.
- [3] J. M. BORWEIN AND H. M. STROJWAS, *Proximal analysis and boundaries of closed sets in Banach space, part I: Theory*, *Canad. J. Math.*, 38 (1986), pp. 431–452.

- [4] J. M. BORWEIN AND H. M. STROJWAS, *Proximal analysis and boundaries of closed sets in Banach space, part II: Theory*, Canad. J. Math., to appear.
- [5] F. H. CLARKE, *Optimization and Nonsmooth Analysis*, John Wiley, New York, 1983.
- [6] ———, *Necessary conditions for nonsmooth problems in optimal control and the calculus of variations*, Ph.D. thesis, Univ. of Washington, Seattle, WA, 1973.
- [7] ———, *Perturbed optimal control problems*, IEEE Trans. Automat. Control, AC-31 (1986), pp. 535–542.
- [8] F. H. CLARKE AND P. D. LOEWEN, *The value function in optimal control: sensitivity, controllability, and time-optimality*, this Journal, 24 (1986), pp. 243–263.
- [9] J. GAUVIN, *The generalized gradient of a marginal function in mathematical programming*, Math. Oper. Res., 4 (1979), pp. 458–463.
- [10] M. R. HESTENES, *Calculus of Variations and Optimal Control Theory*, John Wiley, New York, 1966.
- [11] A. D. IOFFE, *Necessary conditions in nonsmooth optimization*, Math. Oper. Res., 9 (1984), pp. 159–189.
- [12] P. D. LOEWEN, *Proximal normal analysis in dynamic optimization*, Ph.D. thesis, Univ. of British Columbia, Vancouver, Canada, 1985.
- [13] ———, *The proximal normal formula in Hilbert space*, Nonlinear Anal., to appear.
- [14] L. W. NEUSTADT, *An abstract variational theory with applications to a broad class of optimization problems, II: Applications*, this Journal, 5 (1967), pp. 90–137.
- [15] R. T. ROCKAFELLAR, *Extensions of subgradient calculus with applications to optimization*, Nonlinear Anal., 9 (1985), pp. 665–698.
- [16] ———, *State constraints in convex problems of Bolza*, this Journal, 10 (1972), pp. 691–715.
- [17] R. B. VINTER AND G. PAPPAS, *A maximum principle for nonsmooth optimal-control problems with state constraints*, J. Math. Anal. Appl., 89 (1982), pp. 212–232.
- [18] J. WARGA, *Controllability and necessary conditions in unilateral problems without differentiability assumptions*, this Journal, 14 (1976), pp. 546–573.

AN ANALYSIS OF THE POLE-ZERO CANCELLATIONS IN H^∞ -OPTIMAL CONTROL PROBLEMS OF THE FIRST KIND*

D. J. N. LIMEBEER† AND Y. S. HUNG‡

Abstract. The aim of this paper is to study the pole-zero cancellations which occur in a class of H^∞ -optimal control problems which may be embedded in the configuration of Fig. 1. H^∞ control problems are said to be of the first kind if both $P_{12}(s)$ and $P_{21}(s)$ are square but not necessarily of the same size. It is primarily this class of problems which will concern us here. A general bound on the McMillan degree of all controllers which are stabilizing and lead to a closed loop which satisfies $\|\mathcal{R}(s)\|_\infty \leq \rho$ (ρ need not be optimal in the L^∞ -norm sense) is derived. As illustrated in Fig. 1, $\mathcal{R}(s)$ is the transfer function relating $y_1(s)$ to $u_1(s)$. If the McMillan degree of $P(s)$ in Fig. 1 is n , we show that in the single-loop (SISO) case the corresponding (unique) H^∞ -optimal controller never requires more than $n-1$ states. In the multivariable case, there is a continuum of optimal controllers whose McMillan degree satisfies this same bound, although other controllers with higher McMillan degree also exist. The derivation of these bounds require several steps, each of which is of independent system theoretic interest.

Key words. pole-zero cancellations, H^∞ -optimal control, approximation theory, Nehari's Theorem, degree bounds

AMS(MOS) subject classification. 93C35

1. Introduction. Figure 1 represents a generalized, or abstract, regulator configuration in which a large class of H^∞ -optimal control problems may be embedded. If $P_{12}(s)$ and $P_{21}(s)$ are square, we call the associated problem a problem of the first kind. (Problems of the second kind are characterized by having either $P_{12}(s)$ or $P_{21}(s)$ nonsquare. If both off-diagonal blocks are nonsquare, the problem is said to be of the third kind.) Depending on the specific design situation, the inputs $u_1(s)$ could be references, external disturbances, sensor noise signals or the outputs of models representing unknown plant dynamics. The outputs $y_1(s)$, on the other hand, may be plant outputs, plant inputs or the signals driving plant perturbation models. The H^∞ control problem for Fig. 1 is to minimize the L^∞ -norm of $\mathcal{R}(s)$ as $K(s)$ is allowed to range over the set of all stabilizing controllers. The general theory of these problems is now well developed and we refer the reader to the expository articles of Francis and Doyle [10], Doyle et al. [6], Safonov et al. [22] and the numerous references therein for details.

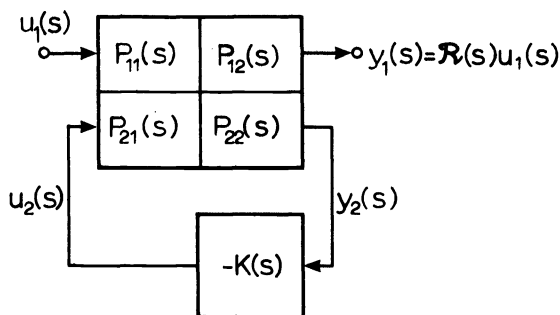


FIG. 1. Generalized regulator configuration.

* Received by the editors May 19, 1986; accepted for publication (in revised form) December 3, 1986.

† Department of Electrical Engineering, Imperial College, London, England.

‡ Electronic and Electrical Engineering Department, University of Surrey, Guildford, Surrey, England.

The purpose of this paper is to carry out a detailed analysis of the cancellation phenomena which occur as a result of H^∞ -optimality in the standard regulator configuration mentioned above. Although algorithms for computing these controllers already exist [6], [22], the procedure is so involved that issues such as McMillan degree propagation and the final controller order are obscure. A naive inspection of the procedure may lead one to suspect that the controller degree is several times higher than that of $P(s)$. Since high order controllers are inevitably preceded by computations in high dimensional state-space, expensive and unreliable computations are likely to cause difficulty in complicated design situations. For these reasons, it is our opinion that the complete structural analysis of the computational framework which is presented here will lead to improved computational methods and will also shed light on several aspects of the theory. Our approach is to analyse the entire calculation process in the state-space. This has the advantage of establishing clear links between the theoretical development and existing computer algorithms [6], [22] and also allows one to use Glover's explicit parametrization of all solutions to the Nehari extension problem [12]. This methodology has also been successfully employed in the analysis of cancellation phenomena in H^∞ problems of the second kind. This will be reported on elsewhere [16].

The paper is organized as follows: In § 2 we define notation, describe the problem in specific terms and briefly review the relevant parametrization and optimization theory. Theorem 2.1 is a reformulation of an existing result and it gives an explicit state-space characterization of all the solutions of the L^∞ -norm optimization problem in terms of bounded real type equations. In § 3, we establish by way of balancing two Riccati equations associated with the parametrization, the role of the right and left half plane zeros of $P_{12}(s)$ and $P_{21}(s)$. In the case of some problems of the first kind, the lowest achievable infinity norm of the closed loop can be expressed in terms of the solutions to these Riccati equations. Lemma 3.1 and Theorem 3.2 are new results. In § 4, we study the pole-zero cancellations which occur in the closed loop of Fig. 1 when $K(s)$ is chosen to be H^∞ -optimal (or suboptimal in a sense to be defined later). This will lead to a general McMillan degree bound for all these controllers. An illustrative example is presented which shows that midcalculation model reduction can produce undesirable effects if done in an ill-advised way. An extension of the McMillan degree bound to the case of minimum entropy controllers is also given in this section. The five results given in § 4 are all believed to be new. Section 5 contains the conclusions. All the proofs have been placed in a series of Appendices.

Some of the proofs involve long calculations. For the reader's convenience, we have written the paper so that no loss of continuity is experienced if these proofs are not studied in the first instance.

2. Notation and background theory.

2.1. Notation.

\mathbb{R}, \mathbb{C}	fields of real and complex numbers,
$\mathbb{R}(s)$	field of rational functions in s with real coefficients,
$\mathbb{F}^{m \times l}$	set of $m \times l$ matrices with elements in \mathbb{F} ($=\mathbb{R}, \mathbb{C}, \mathbb{R}(s)$ etc.),
$\mathbb{C}_+, \bar{\mathbb{C}}_+$	open (resp. closed) right half plane,
$\mathbb{C}_-, \bar{\mathbb{C}}_-$	open (resp. closed) left half plane,
$\lambda(A), \lambda_{\max}(A)$	eigenvalues of a square matrix A , largest eigenvalue of A ,
A^*	complex conjugate transpose of $A \in \mathbb{C}^{m \times l}$ (transpose if $A \in \mathbb{R}^{m \times l}$),
$\text{In}(A)$	$=(\pi, \nu, \delta)$, the inertia of A , where π, ν and δ are the number of eigenvalues of A in $\mathbb{C}_+, \mathbb{C}_-$ and the $j\omega$ (imaginary) axis,
$A \geq 0, A > 0$	A is positive semidefinite (resp. positive definite),
$A \leq 0, A < 0$	A is negative semidefinite (resp. negative definite),

RL^∞	space of matrices in $\mathbb{R}(s)^{m \times l}$ which have no poles on the $j\omega$ axis (including the point at ∞),
$\ \cdot\ _\infty$	L^∞ -norm of matrices in RL^∞ ,
RH_+^∞, RH_-^∞	subspaces of RL^∞ of matrices which have no poles in \mathbb{C}_+ (resp. \mathbb{C}_-),
Γ_G	Hankel operator associated with $G(s) \in RH_+^\infty$,
$\sigma_i(G(s))$	i th Hankel singular value of $G(s)$ (i.e. of Γ_G) in decreasing order of magnitude,
$\ G(s)\ _H$	$=\sigma_1(G(s))$, the Hankel norm of $G(s)$,
$\operatorname{Re}(s), \bar{s}, s $	the real part, complex conjugate and modulus of $s \in \mathbb{C}$,
$G^*(s)$	$=G(-\bar{s})^*$, the parahermitian conjugate of $G(s)$,
$\Rightarrow, \Leftarrow, \Leftrightarrow$	implies, is implied by, if and only if.

Associated with a transfer function matrix $G(s) \in \mathbb{R}(s)^{m \times l}$ of McMillan degree n is a state-space realization

$$(2.1a) \quad G(s) = D + C(sI - A)^{-1}B$$

where $A \in \mathbb{R}^{n \times n}$, $B \in \mathbb{R}^{n \times l}$, $C \in \mathbb{R}^{m \times n}$, and $D \in \mathbb{R}^{m \times l}$. We will use the alternative notation $G(s) = (A, B, C, D)$ or

$$(2.1b) \quad G(s) = \left[\begin{array}{c|c} A & B \\ \hline C & D \end{array} \right].$$

In the above notation, we have $G^*(s) = (-A^*, C^*, -B^*, D^*)$ and in the case that D is nonsingular, we also have $G^{-1}(s) = (A - BD^{-1}C, BD^{-1}, -D^{-1}C, D^{-1})$. If $G^{-1}(s) = G^*(s)$, then $G(s)$ is all-pass. $G(s)$ is called stable (asymptotically stable) if it has no poles in \mathbb{C}_+ (resp. $\bar{\mathbb{C}}_+$).

If $G(s) = (A, B, C, D)$ the system matrix corresponding to the given realization is defined as [19]

$$\left[\begin{array}{c|c} sI - A & -B \\ \hline C & D \end{array} \right]$$

and the system zeros are defined to be the points at which the system matrix loses normal rank. In the case when D is nonsingular, the system zeros are also given by $\lambda(A - BD^{-1}C)$. The input decoupling zeros (uncontrollable modes) are points at which $[sI - A \ B]$ loses rank. The output decoupling zeros (unobservable modes) are the points at which $[sI - A^* \ C^*]$ loses rank. In the sequel, the term “zero” refers to “system zero” unless stated otherwise. Obviously, {input decoupling zeros} \cup {output decoupling zeros} is a subset of both $\lambda(A)$ and the set of system zeros. The realization (A, B, C, D) is minimal if it has no input/output decoupling zeros. A sufficient condition for this is that all system zeros are distinct from $\lambda(A)$.

If $G_1(s) = (A_1, B_1, C_1, D_1)$ and $G_2(s) = (A_2, B_2, C_2, D_2)$ then the cascade system $G_1G_2(s)$ has a realization given by

$$(2.2) \quad \left[\begin{array}{c|c} A_1 & B_1 \\ \hline C_2 & D_2 \end{array} \right] \left[\begin{array}{c|c} A_2 & B_2 \\ \hline C_1 & D_1 \end{array} \right] = \left[\begin{array}{cc|c} A_1 & B_1C_2 & B_1D_2 \\ 0 & A_2 & B_2 \\ \hline C_1 & D_1C_2 & D_1D_2 \end{array} \right],$$

where we have taken the “multiplication” of two realizations to mean cascading the two systems. This is not to be confused with ordinary matrix multiplication. The context will always make the distinction between these two possible interpretations clear.

If a basis change T is introduced into the state-space of $G(s)$, we will take this to mean $G(s) = (TAT^{-1}, TB, CT^{-1}, D)$. The McMillan degree of $G(s)$ will be written as $\deg(G)$ and the set of poles of $G(s)$ will be denoted {poles of G }.

Let $P(s)$ be a partitioned matrix with a state-space realization given by

$$(2.3) \quad P(s) = \begin{bmatrix} P_{11} & P_{12} \\ P_{21} & P_{22} \end{bmatrix} (s) = \left[\begin{array}{c|cc} A & B_1 & B_2 \\ \hline C_1 & D_{11} & D_{12} \\ C_2 & D_{21} & D_{22} \end{array} \right],$$

then

$$(2.4) \quad P_{ij}(s) = C_i(sI - A)^{-1}B_j + D_{ij}$$

is a state-space realization of $P_{ij}(s)$. A linear fractional transformation of the partitioned matrix P and a matrix K is defined as

$$F_l(P, K) = P_{11} + P_{12}K(I - P_{22}K)^{-1}P_{21}$$

where K is of dimension $l \times m$ if P_{22} has dimension $m \times l$.

2.2. Problem description. Consider the generalized regulator configuration illustrated in Fig. 1. From the equations governing this diagram we see that the transfer function relating y_1 to u_1 is given by

$$\begin{aligned} \mathcal{R}(s) &= F_l(P, -K) \\ &= P_{11} - P_{12}K(I + P_{22}K)^{-1}P_{21}. \end{aligned}$$

We seek to bound the McMillan degrees of all the compensators $K(s)$ which simultaneously achieve an internally stable closed loop and minimize $\|\mathcal{R}(s)\|_\infty$. Throughout this paper we will assume that $P_{12}(s)$ and $P_{21}(s)$ are square (although not necessarily of the same size). We also assume that both D_{12} and D_{21} are nonsingular and that $P_{12}(s)$ and $P_{21}(s)$ have no zeros on the imaginary axis.

It is worth noting that two particular H^∞ -optimal control problems which have already received particular attention can be posed in the above setting. The first is the optimal sensitivity problem which has been analysed by Zames and others [3], [7], [8], [21], [27], [28]. In this case we wish to minimize the L^∞ -norm of a weighted sensitivity operator given by

$$(2.5) \quad \mathcal{R}_s(s) = [W_2(I + GK)^{-1}W_1](s)$$

where $G(s)$ is the transfer function of a square plant, and $W_1(s)$ and $W_2(s)$ are weighting matrices. If we put

$$(2.6) \quad P_s(s) = \left[\begin{array}{c|c} W_2W_1 & W_2G \\ \hline W_1 & G \end{array} \right] (s),$$

then direct calculation shows that

$$\mathcal{R}_s(s) = F_l(P_s(s), -K(s)).$$

The second problem is the optimal robustness problem which has been studied by Glover [13] and Kimura [15]. In the case of optimal robustness with respect to additive perturbations to the plant transfer function, we wish to minimize the L^∞ -norm of

$$(2.7) \quad \mathcal{R}_a(s) = [W_2K(I + GK)^{-1}W_1](s).$$

It can be readily shown that if we set

$$(2.8) \quad P_a(s) = \left[\begin{array}{c|c} 0 & W_2 \\ \hline -W_1 & G \end{array} \right] (s)$$

then

$$\mathcal{R}_a(s) = F_l(P_a(s), -K(s)).$$

In the sequel, we will study the general class of problems of the first kind and establish common pole-zero cancellation properties which are shared by the specific problems just mentioned.

2.3. Review of H^∞ -optimization theory. The solution of H^∞ -optimal control problems may be subdivided into two distinct steps. In the first, all the compensators which lead to an internally stable closed loop in Fig. 1 are parametrized. The second step then identifies a subclass of stabilizing compensators which minimize $\|\mathcal{R}(s)\|_\infty$ or else satisfy $\|\mathcal{R}(s)\|_\infty \leq \rho \in \mathbb{R}$. In the following sections, we will briefly describe the calculations involved in these two steps.

2.3.1. Parametrization of all stabilizing controllers. Let $P(s)$ in Fig. 1 be given by (2.3) and suppose that (A, B_2, C_2) is stabilizable and detectable. Under these conditions $K(s)$ stabilizes the feedback system in Fig. 1 if and only if it stabilizes $P_{22}(s)$. Further, such stabilizing compensators always exist [6], [22]. Let

$$(2.9) \quad P_{22}(s) = N_r(s)D_r^{-1}(s) = D_l^{-1}(s)N_l(s)$$

be right and left rational coprime fractional factorizations of $P_{22}(s)$ and

$$(2.10) \quad \begin{bmatrix} V_r & U_r \\ -N_l & D_l \end{bmatrix}(s) \begin{bmatrix} D_r & -U_l \\ N_r & V_l \end{bmatrix}(s) = \begin{bmatrix} I & 0 \\ 0 & I \end{bmatrix}$$

the corresponding Bezout identities. All the matrices in (2.10) belong to RH_+^∞ and the set of all compensators which stabilize $P_{22}(s)$, and thus also $P(s)$, are given by [4], [26]

$$(2.11) \quad K(s) = (U_l + D_r Q)(V_l - N_r Q)^{-1}(s)$$

$$(2.12) \quad = (V_r - Q N_l)^{-1}(U_r + Q D_l)(s)$$

in which the indicated inverses are assumed to exist and $Q(s) \in RH_+^\infty$. It is easy to verify that

$$(2.13) \quad K(I + P_{22}K)^{-1}(s) = (U_l + D_r Q)D_l(s).$$

Hence

$$(2.14) \quad \begin{aligned} \mathcal{R}(s) &= [P_{11} - P_{12}K(I + P_{22}K)^{-1}P_{21}](s) \\ &= [(P_{11} - P_{12}U_l D_l P_{21}) - (P_{12}D_r)Q(D_l P_{21})](s) \\ &= [T_{11} - T_{12}QT_{21}](s) \end{aligned}$$

where the $T_{ij}(s)$ are defined in an obvious way. Equation (2.14) shows that $\mathcal{R}(s)$ is parametrized linearly in $Q(s)$. Since $\mathcal{R}(s) \in RH_+^\infty$ if and only if $Q(s) \in RH_+^\infty$, we would expect that T_{11} , T_{12} and T_{21} all belong to RH_+^∞ .

Since (A, B_2) is stabilizable, there exists a state feedback matrix F such that $A - B_2 F$ is stable. Similarly, since (A, C_2) is detectable there exists an output injection matrix H such that $A - H C_2$ is stable. Given any such pair of stabilizing matrices F and H , the right and left coprime factorizations of P_{22} together with the solutions of the Bezout identities are given by [6], [18]

$$(2.15) \quad \begin{bmatrix} D_r & -U_l \\ N_r & V_l \end{bmatrix}(s) = \left[\begin{array}{c|cc} A - B_2 F & B_2 & H \\ -F & I & 0 \\ \hline C_2 - D_{22} F & D_{22} & I \end{array} \right]$$

and

$$(2.16) \quad \begin{bmatrix} V_r & U_r \\ -N_l & D_l \end{bmatrix} (s) = \left[\begin{array}{c|cc} A - HC_2 & B_2 - HD_{22} & H \\ \hline F & I & 0 \\ -C_2 & -D_{22} & I \end{array} \right].$$

Using (2.11) and (2.12), it can be verified by direct calculation that the family of all stabilizing compensators can be parametrized in terms of the linear fractional transformation [6]

$$(2.17) \quad K(s) = F_l(K_0(s), Q(s))$$

where

$$(2.18) \quad \begin{aligned} K_0(s) &= \begin{bmatrix} K_{11} & K_{12} \\ K_{21} & K_{22} \end{bmatrix} (s) = \begin{bmatrix} V_r^{-1} U_r & V_r^{-1} \\ V_l^{-1} & V_l^{-1} N_r \end{bmatrix} (s) \\ &= \left[\begin{array}{c|cc} A - B_2 F - HC_2 + HD_{22} F & -H & B_2 - HD_{22} \\ \hline -F & 0 & I \\ C_2 - D_{22} F & I & D_{22} \end{array} \right]. \end{aligned}$$

A routine computation will show that the realization of $K_0(s)$ in (2.18) is minimal if (A, B_2, C_2) is minimal.

In order to simplify later calculations, it is helpful at this point to scale (2.3) by replacing it with

$$(2.19) \quad P(s) = \left[\begin{array}{c|cc} A & B_1 & B_2 S_1 \\ \hline C_1 & D_{11} & D_{12} S_1 \\ S_2 C_2 & S_2 D_{21} & S_2 D_{22} S_1 \end{array} \right],$$

in which $S_1 = D_{12}^{-1}$ and $S_2 = D_{21}^{-1}$. From now on we will assume that $P(s)$ has been scaled so that both the (1, 2) and (2, 1) blocks of the D -matrix are identities; we therefore assume that the S_i 's have already been absorbed into B_2 , C_2 and D_{22} . Such an assumption does not incur any loss of generality in our development because the effect of any prescaling on the compensator to bring $P(s)$ into the form of (2.19) may be reversed by replacing $K(s)$ with $S_1 K(s) S_2$ at the end of the design process [22]. We now make the following specific choices of the pair of stabilizing matrices F and H , as was suggested by Doyle et al. [6]:

$$(2.20) \quad F = C_1 + B_2^* X$$

where X is the unique positive semidefinite stabilizing solution to the algebraic Riccati equation

$$(2.21) \quad X(A - B_2 C_1) + (A - B_2 C_1)^* X - X B_2 B_2^* X = 0$$

and

$$(2.22) \quad H = B_1 + Y C_2^*$$

where Y is the unique positive semidefinite stabilizing solution to the algebraic Riccati equation

$$(2.23) \quad Y(A - B_1 C_2)^* + (A - B_1 C_2) Y - Y C_2^* C_2 Y = 0.$$

It can be shown that the $T_{ij}(s)$ of (2.14) are then given by

$$(2.24) \quad T(s) = \begin{bmatrix} T_{11} & T_{12} \\ T_{21} & 0 \end{bmatrix} (s) = \left[\begin{array}{cc|cc} A - B_2 F & B_2 F & B_1 & B_2 \\ 0 & A - H C_2 & -Y C_2^* & 0 \\ \hline -B_2^* X & F & D_{11} & I \\ 0 & C_2 & I & 0 \end{array} \right].$$

Note that $T_{12}(s)$, $T_{21}(s)$ and $T_{11}(s)$ belong to RH_+^∞ as expected. Further, F and H have been chosen to make $T_{12}(s)$ and $T_{21}(s)$ inner [6].

2.3.2. Parametrization of all H^∞ -optimal controllers. In this section we briefly review the parametrization of all optimal $Q(s) \in RH_+^\infty$ which solve the minimization problem

$$(2.25) \quad \min \| (T_{11} - T_{12} Q T_{21})(s) \|_\infty = \| T_{12}^* T_{11} T_{21}^*(-s) \|_H, \quad Q \in RH_+^\infty$$

or suboptimal $Q(s) \in RH_+^\infty$ which satisfy

$$(2.26) \quad \| (T_{11} - T_{12} Q T_{21})(s) \|_\infty \leq \rho \in \mathbb{R}$$

for some given $\rho > \| T_{12}^* T_{11} T_{21}^*(s) \|_H$.

Due to the norm-preserving properties of the inner matrices $T_{12}(s)$ and $T_{21}(s)$, we may write [21]

$$(2.27) \quad \| (T_{11} - T_{12} Q T_{21})(s) \|_\infty = \| (T_{12}^* T_{11} T_{21}^* - Q)(s) \|_\infty.$$

From (2.24), we obtain

$$(2.28) \quad T_{12}^* T_{11} T_{21}^*(s) = \left[\begin{array}{cc|c} -(A - B_2 F)^* & X(B_2 D_{11} - B_1) C_2 Y & X(B_1 - B_2 D_{11}) \\ 0 & -(A - H C_2)^* & -C_2^* \\ \hline -B_2^* & (F - D_{11} C_2) Y & D_{11} \end{array} \right]$$

which shows that $T_{12}^* T_{11} T_{21}^*(s) \in RH_-^\infty$. Setting $T_{12}^* T_{11} T_{21}^*(s) = X^*(s)$ we get

$$(2.29) \quad \| (T_{11} - T_{12} Q T_{21})(s) \|_\infty = \| X(s) - Q^*(s) \|_\infty$$

which turns (2.25) into a multivariable version of the Nehari extension problem [17]. We will call any $Q(s) \in RH_+^\infty$ which satisfies (2.26) a ρ -suboptimal extension. Glover has shown that all $Q^*(s)$ which satisfy (2.25) or (2.26) may be generated by means of a balanced realization of $X(s)$. In [12], the characterization of all $Q^*(s)$ in the general nonsquare case is given in terms of a linear fractional transformation of transfer function matrices (see [12, Thm. 8.7]). We will however need a state-space version of this characterization in order to derive the main results in § 4. This is stated in the next theorem and a proof which makes use of [12, Thm. 8.7] is given in Appendix A.

THEOREM 2.1. *Let $X(s) = (A, B, C, D)$ be a stable, minimal and balanced realization with Hankel singular values*

$$\sigma_1 \geq \sigma_2 \geq \dots \geq \sigma_k > \sigma_{k+1} = \sigma_{k+2} = \dots = \sigma_{k+r} > \sigma_{k+r+1} \geq \dots \geq \sigma_n > 0.$$

Assume that the Hankel singular values have been arranged so that the gramians are given by

$$(2.30) \quad \text{diag}(\Sigma, \sigma_{k+1} I_r)$$

where

$$\Sigma = \text{diag}(\sigma_1, \dots, \sigma_k, \sigma_{k+r+1}, \sigma_n)$$

and let (A, B, C) be partitioned conformally with (2.30)

$$A = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix}, \quad B = \begin{bmatrix} B_1 \\ B_2 \end{bmatrix}, \quad C = [C_1 \quad C_2].$$

Also let

$$(2.31) \quad \Gamma = \Sigma^2 - \sigma_{k+1}^2 I.$$

Then for any error system $E(s) = X(s) - \hat{X}(s) - Q^*(s)$ with

(a) $\|E(j\omega)\|_\infty \leq \sigma_{k+1}$;

(b) $\hat{X}(s)$ is stable of McMillan degree k and $Q^*(s)$ is totally unstable there exists $\hat{A}, \hat{B}, \hat{C}$ and \hat{D} such that $E(s)$ has a realization

$$(2.32) \quad E(s) = \left[\begin{array}{cc|cc|c} A_{11} & A_{12} & 0 & 0 & B_1 \\ A_{21} & A_{22} & 0 & 0 & B_2 \\ 0 & 0 & \Gamma^{-1}(\sigma_{k+1}^2 A_{11}^* + \Sigma A_{11} \Sigma & \Gamma^{-1} C_1^* \hat{C} & \Gamma^{-1}(\Sigma B_1 + \sigma_{k+1} C_1^* \hat{D}) \\ & & -\sigma_{k+1} C_1^* \hat{D} B_1^* & & \\ 0 & 0 & -\sigma_{k+1} \hat{B} B_1^* & \hat{A} & \sigma_{k+1} \hat{B} \\ \hline C_1 & C_2 & -(C_1 \Sigma + \sigma_{k+1} \hat{D} B_1^*) & \hat{C} & \sigma_{k+1} \hat{D} \end{array} \right] = \begin{bmatrix} A_e & B_e \\ C_e & D_e \end{bmatrix}$$

which satisfies both

$$(2.33) \quad \begin{bmatrix} -(A_e P_e + P_e A_e^* + B_e B_e^*) & -(B_e D_e^* + P_e C_e^*) \\ -(D_e B_e^* + C_e P_e) & \sigma_{k+1}^2 I - D_e D_e^* \end{bmatrix} = \begin{bmatrix} L_e \\ W_e \end{bmatrix} [L_e^* | W_e^*],$$

and

$$(2.34) \quad \begin{bmatrix} -(A_e^* Q_e + Q_e A_e + C_e^* C_e) & -(C_e^* D_e + Q_e B_e) \\ -(D_e^* C_e + B_e^* Q_e) & \sigma_{k+1}^2 I - D_e^* D_e \end{bmatrix} = \begin{bmatrix} L_{ed}^* \\ W_{ed}^* \end{bmatrix} [L_{ed} | W_{ed}]$$

where in the above equations

$$P_e = \begin{bmatrix} \Sigma & 0 & I & 0 \\ 0 & \sigma_{k+1} I_r & 0 & 0 \\ I & 0 & \Sigma \Gamma^{-1} & 0 \\ 0 & 0 & 0 & \sigma_{k+1}^2 \hat{P} \end{bmatrix}, \quad Q_e = \begin{bmatrix} \Sigma & 0 & -\Gamma & 0 \\ 0 & \sigma_{k+1} I_r & 0 & 0 \\ -\Gamma & 0 & \Sigma \Gamma & 0 \\ 0 & 0 & 0 & \hat{Q} \end{bmatrix}$$

and

$$(2.35a) \quad L_e^* = [0 \ 0 \ \sigma_{k+1} \hat{W}^* C_1 \Gamma^{-1} \ \sigma_{k+1} \hat{L}^*], \quad L_{ed} = [0 \ 0 \ -\sigma_{k+1} \hat{W}_d B_e^* \ \hat{L}_d]$$

for some $\hat{L}, \hat{L}_d, \hat{W}, \hat{W}_d, \hat{P} = \hat{P}^* < 0, \hat{Q} = \hat{Q}^* < 0$ and further

$$(2.35b) \quad W_e = \sigma_{k+1} \hat{W}, \quad W_{ed} = \sigma_{k+1} \hat{W}_d \quad \text{and} \quad P_e Q_e = \sigma_{k+1}^2 I. \quad \square$$

Remark 2.1. Note that (2.33) and (2.34) are reminiscent of the state-space characterization of bounded real matrices [1], as one would expect because of the condition (a) on $E(s)$. We note however that $E(s)$ is in general not bounded real in the strict sense since it may contain unstable poles.

Remark 2.2. When giving a bounded real type state-space characterization of ρ -suboptimal extensions we make use of the idea given in [12, Remark 8.4]. In this case (2.33) and (2.34) remain in force, $\Gamma = \Sigma^2 - \rho^2 I$ replaces (2.31) (with $\sigma_k > \rho > \sigma_{k+1}$),

$$(2.36) \quad E(s) = \left[\begin{array}{cc|cc|c} A & 0 & 0 & 0 & B \\ 0 & \Gamma^{-1}(\rho^2 A^* + \Sigma A \Sigma - \rho C^* \hat{D} B^*) & \Gamma^{-1} C^* \hat{C} & \Gamma^{-1}(\Sigma B + \rho C^* \hat{D}) & \\ 0 & -\rho \hat{B} B^* & \hat{A} & \rho \hat{B} & \\ \hline C & -(C \Sigma + \rho \hat{D} B^*) & \hat{C} & \rho \hat{D} & \end{array} \right]$$

replaces (2.32),

$$(2.37) \quad P_e = \begin{bmatrix} \Sigma & I & 0 \\ I & \Sigma\Gamma^{-1} & 0 \\ 0 & 0 & \rho^2\hat{P} \end{bmatrix}, \quad Q_e = \begin{bmatrix} \Sigma & -\Gamma & 0 \\ -\Gamma & \Sigma\Gamma & 0 \\ 0 & 0 & \hat{Q} \end{bmatrix}$$

and

$$L_e^* = [0 \quad \rho\hat{W}^*C\Gamma^{-1} \quad \rho\hat{L}^*], \quad L_{ed} = [0 \quad -\rho\hat{W}_dB^* \quad \hat{L}_d],$$

replace (2.35a) and $W_e = \rho\hat{W}$, $W_{ed} = \rho\hat{W}_d$ and $P_eQ_e = \rho^2I$ replace (2.35b).

For readers who are interested in following through the proof for the ρ -suboptimal case, (A.4) in Appendix A should be replaced by

$$(2.38) \quad \begin{bmatrix} H_{11} & H_{12} \\ H_{21} & H_{22} \end{bmatrix}(s) = \left[\begin{array}{c|cc} \Gamma^{-1}(\rho^2A^* + \Sigma A \Sigma) & \Gamma^{-1}\Sigma B & -\Gamma^{-1}C^* \\ \hline C\Sigma & D & I \\ -\rho B^* & \rho I & 0 \end{array} \right].$$

Conditions (A.5) and (A.6) remain valid but (A.7) no longer applies.

Remark 2.3. Theorem 2.1 and Remark 2.2 are more general than we need in H^∞ -optimal control problems since they give a state-space characterization of all error systems associated with Hankel norm approximation problems whereas we are only interested in optimal anticausal (or Nehari) type approximations. Specifically, we will make use of Theorem 2.1 (or Remark 2.2) with $k=0$ (zeroth order Hankel approximation) so that $\hat{X}(s)=0$ and $Q(s)$ becomes a Nehari extension (or ρ -suboptimal extension) of $X(s)$. Also $\|E(s)\|_\infty = \sigma_1$ or $\|E(s)\|_\infty \leq \rho$ (where $\rho > \sigma_1$) for the optimal or ρ -suboptimal case respectively. \square

COROLLARY 2.2. In the notation of Theorem 2.1 and Remark 2.2 let $k=0$ and $U(s)=[\hat{A}, \hat{B}, \hat{C}, \hat{D}]$ (see (2.32) and the proof of Theorem 2.1). Then

(i) If $Q(s)$ is a Nehari extension of $X(s)$,

$$(2.39) \quad \deg(Q) \leq \deg(X) - r + \deg(U)$$

where r is the multiplicity of the largest Hankel singular value of X ;

(ii) If $Q(s)$ is a ρ -suboptimal extension of $X(s)$,

$$(2.40) \quad \deg(Q) \leq \deg(X) + \deg(U). \quad \square$$

This corollary follows immediately from an inspection of (2.32) and (2.36).

Remark 2.4. For $k=0$, Theorem 2.1 characterizes all

$$(2.41) \quad E(s) = X(s) - Q^*(s)$$

satisfying conditions (a) and (b). At certain points in the sequel, it is more convenient to work with a realization for $E^*(s) = X^*(s) - Q(s)$ instead of (2.41). For this purpose, we remark here that the form of the bounded real type equations (2.33) and (2.34) is invariant under parahermitian conjugation. It is easy to see that if $E(s) \rightarrow E^*(s)$, then we only need to perform the following substitutions in (2.33) and (2.34)

$$\begin{aligned} A_e &\rightarrow -A_e^*, & B_e &\rightarrow C_e^*, & C_e &\rightarrow -B_e^*, \\ D_e &\rightarrow D_e^*, & P_e &\rightarrow -Q_e, & Q_e &\rightarrow -P_e. \end{aligned}$$

3. Balancing Riccati equations. In this section we will establish some preliminary results which will be needed in the later analysis.

It is well known that the H^∞ -optimization problem given by (2.25), or equivalently (2.29), is equivalent to a matrix version of the classical Nevanlinna–Pick interpolation problem, the set of interpolation points being the right half plane zeros of $T_{12}(s)$ and $T_{21}(s)$. Since $T_{12}(s)$ and $P_{12}(s)$, and $T_{21}(s)$ and $P_{21}(s)$ have the same zeros, it is clear that the set of interpolation points is

$$\{\text{zeros of } P_{12}(s) \text{ in } \mathbb{C}_+\} \cup \{\text{zeros of } P_{21}(s) \text{ in } \mathbb{C}_+\}.$$

In this section we will bring this issue into sharp focus by balancing the two Riccati equations (2.21) and (2.23). Furthermore, we will show that the number of right half plane zeros of $P_{12}(s)$ and $P_{21}(s)$ are given, respectively, by the ranks of the solutions X and Y to these two equations.

Consider a change of basis T in the state-space of $P(s)$ in (2.19). In this new basis, $P(s)$ becomes

$$P(s) = \left[\begin{array}{c|cc} TAT^{-1} & TB_1 & TB_2 \\ \hline C_1 T^{-1} & D_{11} & I \\ C_2 T^{-1} & I & D_{22} \end{array} \right]$$

and the algebraic Riccati equation (2.21) becomes

$$(3.1) \quad X(TAT^{-1} - TB_2 C_1 T^{-1}) + (TAT^{-1} - TB_2 C_1 T^{-1})^* X - XTB_2 B_2^* T^* X = 0$$

or equivalently

$$(3.2) \quad T^* XT(A - B_2 C_1) + (A - B_2 C_1)^* T^* XT - T^* XTB_2 B_2^* T^* XT = 0.$$

This shows that the effect of the basis change on X is the congruence transformation

$$(3.3) \quad X \rightarrow T^{-*} XT^{-1}$$

where T^{-*} denotes $(T^*)^{-1}$. Similarly, in the new basis, equation (2.23) becomes

$$(3.4) \quad T^{-1} YT^{-*}(A - B_1 C_2)^* + (A - B_1 C_2)T^{-1} YT^{-*} - T^{-1} YT^{-*} C_2^* C_2 T^{-1} YT^{-*} = 0$$

and hence the effect of the basis change is

$$(3.5) \quad Y \rightarrow TYT^*.$$

(Incidentally, in problems of the second and third kind, (3.1) and/or (3.4) have an additional constant term. This term makes no difference to the balancing arguments.)

Combining (3.3) and (3.5) we obtain

$$(3.6) \quad YX \rightarrow TYXT^{-1}$$

and it is immediate from (2.20) and (2.22) that

$$(3.7) \quad F \rightarrow FT^{-1},$$

$$(3.8) \quad H \rightarrow TH.$$

Condition (3.6) shows that $\lambda(YX)$ are invariant under basis changes in the state-space of $P(s)$. Conditions (3.3) and (3.5), together with $X = X^* \geq 0$ and $Y = Y^* \geq 0$, suggest that we may use the construction in [12, Appendix B] to find a basis change T so that in the new basis

$$(3.9) \quad Y = \begin{bmatrix} \tilde{\Sigma}_1 & & & \\ & \tilde{\Sigma}_2 & & \\ & & 0 & \\ & & & 0 \end{bmatrix} = \begin{bmatrix} \Sigma_1 & 0 \\ 0 & 0 \end{bmatrix}$$

and

$$(3.10) \quad X = \begin{bmatrix} \tilde{\Sigma}_1 & & & \\ & 0 & & \\ & & \tilde{\Sigma}_3 & \\ & & & 0 \end{bmatrix}$$

are balanced diagonal matrices. The balancing of the positive definite solutions of standard LQG type Riccati equations has been studied by Joncheere and Silverman [14]. For convenience of analysis we introduce a permutation matrix J such that

$$(3.11) \quad JXJ^* = \begin{bmatrix} \tilde{\Sigma}_1 & & & \\ & \tilde{\Sigma}_3 & & \\ & & 0 & \\ & & & 0 \end{bmatrix} = \begin{bmatrix} \Sigma_2 & 0 \\ 0 & 0 \end{bmatrix}.$$

Clearly $\Sigma_1 > 0$ and $\Sigma_2 > 0$.

For notational simplicity, we will absorb the coordinate transformation matrix T into the state-space matrices and rewrite TAT^{-1} , TB_2, \dots of (3.1) and (3.4) as A, B_2, \dots . If we set $M = A - B_1 C_2$ and $C_2 = (C_{21} | C_{22})$, where the partitioning is consistent with that in (3.9), we obtain from (2.23)

$$(3.12) \quad \begin{bmatrix} \Sigma_1 & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} M_{11}^* & M_{21}^* \\ M_{12}^* & M_{22}^* \end{bmatrix} + \begin{bmatrix} M_{11} & M_{12} \\ M_{21} & M_{22} \end{bmatrix} \begin{bmatrix} \Sigma_1 & 0 \\ 0 & 0 \end{bmatrix} - \begin{bmatrix} \Sigma_1 & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} C_{21}^* \\ C_{22}^* \end{bmatrix} [C_{21} \ C_{22}] \begin{bmatrix} \Sigma_1 & 0 \\ 0 & 0 \end{bmatrix} = 0.$$

The (1, 1) block of (3.12) yields

$$(3.13) \quad \Sigma_1 M_{11}^* + M_{11} \Sigma_1 - \Sigma_1 C_{21}^* C_{21} \Sigma_1 = 0,$$

that is,

$$(3.14) \quad M_{11} - \Sigma_1 C_{21}^* C_{21} = -\Sigma_1 M_{11}^* \Sigma_1^{-1}.$$

From the (1, 2) block of (3.12) one obtains

$$(3.15) \quad \Sigma_1 M_{21}^* = 0 \Rightarrow M_{21} = 0.$$

Making use of (2.22) and what has been deduced above,

$$(3.16) \quad \begin{aligned} A - HC_2 &= M - YC_2^* C_2 \\ &= \begin{bmatrix} M_{11} - \Sigma_1 C_{21}^* C_{21} & M_{12} - \Sigma_1 C_{21}^* C_{22} \\ 0 & M_{22} \end{bmatrix} \\ &= \begin{bmatrix} -\Sigma_1 M_{11}^* \Sigma_1^{-1} & M_{12} - \Sigma_1 C_{21}^* C_{22} \\ 0 & M_{22} \end{bmatrix}. \end{aligned}$$

Applying [12, Thm. 3.3(2)] to (3.13) establishes the implication

$$(3.17) \quad \delta(\Sigma_1^{-1}) = 0 \Rightarrow 0 = \pi(-\Sigma_1^{-1}) \geq \nu(M_{11}).$$

Since we have assumed that $P_{21}(s)$ has no zeros on the imaginary axis, $\delta(M_{11}) = 0$ and therefore

$$(3.18) \quad \text{In}(M_{11}) = (\text{rank}(Y), 0, 0).$$

Note that $\{\lambda(M_{11})\} \equiv \{\text{right half plane zeros of } P_{21}(s)\}$. Since $A - HC_2$ is asymptotically stable so too is M_{22} .

Defining $Z = J(A - B_2C_1)J^*$ and $(JB_2)^* = \hat{B}_2^* = [\hat{B}_{12}^* | \hat{B}_{22}^*]$ we obtain from (2.21)

$$(3.19) \quad \begin{bmatrix} \Sigma_2 & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} Z_{11} & Z_{12} \\ Z_{21} & Z_{22} \end{bmatrix} + \begin{bmatrix} Z_{11}^* & Z_{21}^* \\ Z_{12}^* & Z_{22}^* \end{bmatrix} \begin{bmatrix} \Sigma_2 & 0 \\ 0 & 0 \end{bmatrix} - \begin{bmatrix} \Sigma_2 & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} \hat{B}_{12}^* \\ \hat{B}_{22}^* \end{bmatrix} [\hat{B}_{12}^* | \hat{B}_{22}^*] \begin{bmatrix} \Sigma_2 & 0 \\ 0 & 0 \end{bmatrix} = 0.$$

By an argument similar to the one given for (3.12), we have that

$$(3.20) \quad Z_{11} - \hat{B}_{12} \hat{B}_{12}^* \Sigma_2 = -\Sigma_2^{-1} Z_{11}^* \Sigma_2,$$

$$(3.21) \quad Z_{12} = 0,$$

$$(3.22) \quad \text{In}(Z_{11}) = (\text{rank}(X), 0, 0),$$

$$(3.23) \quad J(A - B_2F)J^* = \begin{bmatrix} -\Sigma_2^{-1} Z_{11}^* \Sigma_2 & 0 \\ Z_{21} - \hat{B}_{22} \hat{B}_{12}^* \Sigma_2 & Z_{22} \end{bmatrix},$$

and that Z_{22} is asymptotically stable. The eigenvalues of Z_{11} are the right half plane zeros of $P_{12}(s)$. Next, we partition the matrices

$$(3.24) \quad J[B_1 | B_2] = [\hat{B}_1 | \hat{B}_2] = \begin{bmatrix} \hat{B}_{11} & \hat{B}_{12} \\ \hat{B}_{21} & \hat{B}_{22} \end{bmatrix},$$

$$(3.25) \quad F = [F_1 | F_2]$$

where \hat{B}_1 and \hat{B}_2 are partitioned conformally with (3.11) and F is partitioned conformally with (3.9). Making use of (3.16), (3.23), (3.24) and (3.25), we can rewrite (2.24) as

$$(3.26) \quad \begin{bmatrix} T_{11} & T_{12} \\ T_{21} & 0 \end{bmatrix} (s) = \left[\begin{array}{cc|cc} J(A - B_2F)J^* & JB_2F & JB_1 & JB_2 \\ 0 & A - HC_2 & -YC_2^* & 0 \\ \hline -B_2^* J^* X J^* & F & D_{11} & I \\ 0 & C_2 & I & 0 \end{array} \right] \\ = \left[\begin{array}{cc|cc|cc} -\Sigma_2^{-1} Z_{11}^* \Sigma_2 & 0 & \hat{B}_{12} F_1 & \hat{B}_{12} F_2 & \hat{B}_{11} & \hat{B}_{12} \\ Z_{21} - \hat{B}_{22} \hat{B}_{12}^* \Sigma_2 & Z_{22} & \hat{B}_{22} F_1 & \hat{B}_{22} F_2 & \hat{B}_{21} & \hat{B}_{22} \\ \hline 0 & 0 & -\Sigma_1 M_{11}^* \Sigma_1^{-1} & M_{12} - \Sigma_1 C_{21}^* C_{22} & -\Sigma_1 C_{21}^* & 0 \\ 0 & 0 & 0 & M_{22} & 0 & 0 \\ \hline -\hat{B}_{12}^* \Sigma_2 & 0 & F_1 & F_2 & D_{11} & I \\ 0 & 0 & C_{21} & C_{22} & I & 0 \end{array} \right] \\ = \left[\begin{array}{cc|cc} -Z_{11}^* & \Sigma_2 \hat{B}_{12} F_1 \Sigma_1 & \Sigma_2 \hat{B}_{11} & \Sigma_2 \hat{B}_{12} \\ 0 & -M_{11}^* & -C_{21}^* & 0 \\ \hline -\hat{B}_{12}^* & F_1 \Sigma_1 & D_{11} & I \\ 0 & C_{21} \Sigma_1 & I & 0 \end{array} \right].$$

Thus

$$(3.27) \quad T_{12} = \left[\begin{array}{c|c} -Z_{11}^* & \Sigma_2 \hat{B}_{12} \\ \hline -\hat{B}_{12}^* & I \end{array} \right]; \quad T_{21} = \left[\begin{array}{c|c} -M_{11}^* & -C_{21}^* \\ \hline C_{21} \Sigma_1 & I \end{array} \right].$$

It follows from (3.13) that $T_{21}(s)$ has observability gramian Σ_1 and controllability gramian Σ_1^{-1} . Similarly, $T_{12}(s)$ has observability gramian Σ_2^{-1} and controllability gramian Σ_2 . Equations (3.18) and (3.22), $\Sigma_1 > 0$ and $\Sigma_2 > 0$ together with [12, Thm. 3.3(5)] establishes the minimality of the realizations in (3.27). To deduce that the realization (3.26) is also minimal, we note that the A -matrix of the inverse system of (3.26) is similar to

$$(3.28) \quad \begin{bmatrix} Z_{11} & (\hat{B}_{12}D_{11} - \hat{B}_{11})C_{21} \\ 0 & M_{11} \end{bmatrix}.$$

It follows from this that the realization (3.26) has no left half plane zeros and since this realization is asymptotically stable, it must be minimal as well because no pole-zero cancellations can occur. We remark however that the realization for $T_{11}(s)$ need not be minimal. Replacing the realization (2.24) by (3.26) allows the realization (2.28) for $T_{12}^* T_{11} T_{21}^*(s)$ to be replaced by

$$(3.29) \quad T_{12}^* T_{11} T_{21}^*(s) = \left[\begin{array}{cc|c} Z_{11} & (\hat{B}_{12}D_{11} - \hat{B}_{11})C_{21} & \hat{B}_{11} - \hat{B}_{12}D_{11} \\ 0 & M_{11} & -\Sigma_1 C_{21}^* \\ \hline -\hat{B}_{12}^* \Sigma_2 & F_1 - D_{11}C_{21} & D_{11} \end{array} \right]$$

which need not be minimal. The results of our analysis up to this point are now summarized in the next lemma for easy reference.

LEMMA 3.1.

- (i) The number of zeros of $P_{12}(s)$ in $\mathbb{C}_+ = \text{rank}(X)$.
- (ii) The number of zeros of $P_{21}(s)$ in $\mathbb{C}_+ = \text{rank}(Y)$.
- (iii) The realization in (3.26) is minimal with $\text{degree} = (\text{rank}(X) + \text{rank}(Y))$.
- (iv) The realizations for $T_{12}(s)$ and $T_{21}(s)$ in (3.27) are minimal with $\text{deg}(T_{12}) = \text{rank}(X)$ and $\text{deg}(T_{21}) = \text{rank}(Y)$.
- (v) $\text{deg}(T_{12}^* T_{11} T_{21}^*) \leq \text{rank}(X) + \text{rank}(Y)$.

Early in this section we showed that $\lambda(YX)$ are invariant with respect to an arbitrary similarity transformation in the state-space of $P(s)$. It is natural to ask whether or not these invariants contain any fundamental information pertaining to the optimal solutions of H^∞ control problems. We conclude this section with a result which shows that the lowest achievable L^∞ -norm for the closed loop may be expressed in terms of $\lambda_{\max}(YX)$ in the case of certain specific problems of the first kind. These problems are: (i) The unweighted optimal sensitivity problem; (ii) the unweighted optimal complementary sensitivity problem; (iii) the unweighted problem associated with optimal robustness towards multiplicative perturbations at the plant input, and (iv) the weighted optimal robustness problem.

THEOREM 3.2. If Ξ is the set of all stabilizing compensators, then

$$(3.30) \quad (i) \quad \inf_{K \in \Xi} \|(I + GK)^{-1}(s)\|_\infty = (1 + \lambda_{\max}(YX))^{1/2};$$

$$(3.31) \quad (ii) \quad \inf_{K \in \Xi} \|GK(I + GK)^{-1}(s)\|_\infty = (1 + \lambda_{\max}(YX))^{1/2};$$

$$(3.32) \quad (iii) \quad \inf_{K \in \Xi} \|KG(I + KG)^{-1}(s)\|_\infty = (1 + \lambda_{\max}(YX))^{1/2};$$

- (iv) If $W_1(s)$ and $W_2(s)$ are stable and minimum phase frequency dependent weights with proper inverses, then

$$(3.33) \quad \inf_{K \in \Xi} \|W_1 K (I + GK)^{-1} W_2(s)\|_\infty = \lambda_{\max}(YX)^{1/2}. \quad \square$$

Remark 3.1. A detailed analysis will reveal that the Riccati equations defining X and Y in the case of problems (i) and (ii) are the same and therefore that

$$(3.34) \quad \inf_{K \in \Xi} \|(I + GK)^{-1}\|_{\infty} = \inf_{K \in \Xi} \|GK(I + GK)^{-1}\|_{\infty}.$$

This result was originally proved by Kwakernaak in the SISO case and Glover in the MIMO case (private communication). In general, the (X, Y) pairs associated with the other problems given in Theorem 3.2 are different leading to different achievable L^{∞} -norm infima. It may be shown by counterexample that the results for problems (i), (ii) and (iii) do not carry over to the weighted case. \square

4. Main results. In this section we consider the pole-zero cancellation properties of the H^{∞} -optimal (or suboptimal) system of Fig. 1 and we will derive general McMillan degree bounds for all H^{∞} -optimal controllers (denoted K_{opt}) or suboptimal controllers (denoted K_{sopt}) for problems of the first kind. An outline of our development is as follows.

Let $n = \deg(P)$, $t = \deg(\mathcal{R})$ and let $m =$ (number of cancellations which occur between $P(s)$ and $K(s)$ as a result of closing the feedback loop in Fig. 1). Then

$$t = n + \deg(K) - m,$$

that is

$$\deg(K) = t + m - n.$$

To obtain an upper bound for $\deg(K)$, we proceed in two steps:

(1) Theorems 4.1, 4.2 and 2.1 establish an upper bound t_b for the McMillan degree t of all optimal closed-loop transfer functions $\mathcal{R}(s)$, and

(2) Theorem 4.3 establishes an upper bound m_b for the number of pole-zero cancellations between $P(s)$ and $K(s)$. Given such bounds, we then have

$$(4.1) \quad \deg(K) \leq t_b + m_b - n.$$

In the case of single-input-single-output (SISO) problems we will show that

$$(4.2) \quad \deg(K_{\text{opt}}) \leq n - 1,$$

$$(4.3) \quad \deg(K_{\text{sopt}}) \leq n.$$

In the case of multivariable (MIMO) problems, we will show that there is a continuum of controllers which satisfy the bounds given in (4.2) and (4.3). These results are stated in Theorem 4.4. We remark that a bound of this type has already been discovered by Glover in the special case of the optimal robustness problem [13].

In an earlier paper, Zames and Francis [28] established that there are interpolation constraints associated with both the right half plane poles and zeros of the plant in the single loop optimal sensitivity problem. If the weighted sensitivity is given by

$$(4.4) \quad s(s) = w(s)/(1 + g(s)k(s)),$$

then they have shown that

$$(4.5) \quad s(z_i) = w(z_i)$$

at each right half plane zero z_i of $g(s)$, and that

$$(4.6) \quad s(p_i) = 0$$

at each right half plane pole. These observations lead us to an interesting factorization phenomenon which may occur in H^{∞} control problems of the first kind. We will

motivate the main idea by way of the unweighted sensitivity minimization problem (weights are neglected for ease of exposition).

From (2.6), we have after prescaling that

$$(4.7) \quad P(s) = \begin{bmatrix} I & G(s)D^{-1} \\ I & G(s)D^{-1} \end{bmatrix} = \left[\begin{array}{c|cc} A & 0 & BD^{-1} \\ C & I & I \\ C & I & I \end{array} \right].$$

Substituting into (2.20) to (2.24) we get

$$(4.8) \quad \begin{bmatrix} T_{11}(s) & T_{12}(s) \\ T_{21}(s) & 0 \end{bmatrix} = \left[\begin{array}{cc|cc} A - BD^{-1}F & YC^*C & YC^* & BD^{-1} \\ 0 & A - HC & -YC^* & 0 \\ -D^{-*}B^*X & C & I & I \\ 0 & C & I & 0 \end{array} \right]$$

after the change of basis

$$(4.9) \quad T = \begin{bmatrix} I & -I \\ 0 & I \end{bmatrix}.$$

An inspection of (4.8) shows that it can be factorized as

$$(4.10) \quad \begin{bmatrix} T_{11}(s) & T_{12}(s) \\ T_{21}(s) & 0 \end{bmatrix} = \begin{bmatrix} \tilde{T}_{11}(s) & T_{12}(s) \\ I & 0 \end{bmatrix} \begin{bmatrix} T_{21}(s) & 0 \\ 0 & I \end{bmatrix}$$

where

$$(4.11) \quad \tilde{T}_{11}(s) = \left[\begin{array}{c|c} A - BD^{-1}F & YC^* \\ -D^{-*}B^*X & I \end{array} \right].$$

Consequently, $\mathcal{R}(s)$ can be written as

$$(4.12) \quad \mathcal{R}(s) = (\tilde{T}_{11}(s) - T_{12}Q(s))T_{21}(s)$$

in which we note also that $\{\text{zeros of } T_{21}\} \equiv \{\text{poles of } G\}$ and $\{\text{zeros of } T_{12}\} \equiv \{\text{zeros of } G\}$. At each pole of $G(s)$ there exists a vector x_i such that

$$(4.13) \quad T_{21}(p_i)x_i = 0 \quad \text{for all } Q(s) \text{ in } RH_+^\infty$$

which implies

$$(4.14) \quad \mathcal{R}(p_i)x_i = 0.$$

This is a generalization of (4.6) to the MIMO case. The point we want to emphasize, however, is that in certain H^∞ control problems, T_{11} may have natural all-pass common factors with T_{12} and/or T_{21} , as illustrated in (4.12). Theorem 4.1 gives a general treatment of the properties of this type of all-pass common factor.

THEOREM 4.1. *Let*

$$(4.15) \quad \begin{bmatrix} G & A_1 \\ A_2 & 0 \end{bmatrix}(s) = \left[\begin{array}{cc|cc} A_{11} & A_{12} & B_{11} & B_{12} \\ 0 & A_{22} & B_{21} & 0 \\ C_{11} & C_{12} & D & I \\ 0 & C_{22} & I & 0 \end{array} \right]$$

in which $A_1(s)$ and $A_2(s)$ are assumed inner. Suppose also that the realizations for $A_1(s)$ and $A_2(s)$ given in (4.15) are minimal. Then there exists a change of basis such that

(4.15) can be put into the form

$$(4.16) \quad \begin{bmatrix} G & A_1 \\ A_2 & 0 \end{bmatrix} (s) = \left[\begin{array}{cccc|cc} \tilde{A}_{00} & \tilde{A}_{01} & \tilde{A}_{02} & \tilde{A}_{03} & \tilde{B}_{01} & \tilde{B}_{02} \\ 0 & \tilde{A}_{11} & \tilde{A}_{12} & \tilde{A}_{13} & \tilde{B}_{11} & \tilde{B}_{12} \\ 0 & 0 & \tilde{A}_{22} & \tilde{A}_{23} & \tilde{B}_{21} & 0 \\ 0 & 0 & 0 & \tilde{A}_{33} & \tilde{B}_{31} & 0 \\ \hline \tilde{C}_{10} & \tilde{C}_{11} & \tilde{C}_{12} & \tilde{C}_{13} & D & I \\ 0 & 0 & \tilde{C}_{22} & \tilde{C}_{23} & I & 0 \end{array} \right],$$

which admits the following factorizations

$$(4.17) \quad A_1(s) = A_l(s) \tilde{A}_1(s) = \left[\begin{array}{c|c} \tilde{A}_{00} & \tilde{B}_{02} \\ \hline \tilde{C}_{10} & I \end{array} \right] \left[\begin{array}{c|c} \tilde{A}_{11} & \tilde{B}_{12} \\ \hline \tilde{C}_{11} & I \end{array} \right],$$

$$(4.18) \quad A_2(s) = \tilde{A}_2(s) A_r(s) = \left[\begin{array}{c|c} \tilde{A}_{22} & \tilde{B}_{21} \\ \hline \tilde{C}_{22} & I \end{array} \right] \left[\begin{array}{c|c} \tilde{A}_{33} & \tilde{B}_{31} \\ \hline \tilde{C}_{23} & I \end{array} \right],$$

$$(4.19) \quad G(s) = A_l(s) \tilde{G}(s) A_r(s) = \left[\begin{array}{c|c} \tilde{A}_{00} & \tilde{B}_{02} \\ \hline \tilde{C}_{10} & I \end{array} \right] \left[\begin{array}{ccc|c} \tilde{A}_{11} & \tilde{A}_{12} & \tilde{B}_{11} & \\ 0 & \tilde{A}_{22} & \tilde{B}_{21} & \\ \hline \tilde{C}_{11} & \tilde{C}_{12} & D & \end{array} \right] \left[\begin{array}{c|c} \tilde{A}_{33} & \tilde{B}_{31} \\ \hline \tilde{C}_{23} & I \end{array} \right].$$

Further

(a) $\tilde{A}_1(s)$, $\tilde{A}_2(s)$, $A_l(s)$, $A_r(s)$, are inner.

(b) The factorizations in (4.17), (4.18), and (4.19) are minimal in the sense that

$$(4.20) \quad \deg(A_1) = \deg(A_l) + \deg(\tilde{A}_1),$$

$$(4.21) \quad \deg(A_2) = \deg(A_r) + \deg(\tilde{A}_2),$$

$$(4.22) \quad \deg(G) = \deg(A_l) + \deg(\tilde{G}) + \deg(A_r).$$

(c) $A_1^* G A_2^*(s) = \tilde{A}_1^* \tilde{G} \tilde{A}_2^*(s)$ has a minimal realization given by

$$(4.23) \quad \left[\begin{array}{ccc|c} -\tilde{A}_{11}^* & \tilde{A}_{12} + \tilde{B}_{11} \tilde{B}_{21}^* + \tilde{C}_{11}^* (\tilde{C}_{12} + D \tilde{B}_{21}^*) & \tilde{C}_{11}^* D + \tilde{B}_{11} & \\ 0 & -\tilde{A}_{22}^* & -\tilde{C}_{22}^* & \\ \hline -\tilde{B}_{12}^* & \tilde{C}_{12} + D \tilde{B}_{21} & D & \end{array} \right]. \quad \square$$

The proof of this result, which is inspired by the work of Van Dooren and DeWilde [23], is given in Appendix C.

Theorem 2.1 shows that the realization (3.29) for $T_{12}^* T_{11} T_{21}^*(s)$ is controllable if and only if $T_{11}(s)$ and $T_{12}(s)$ have no common inner left divisors and that it is observable if and only if $T_{11}(s)$ and $T_{21}(s)$ have no common inner right divisors. The realization is thus minimal if and only if neither type of factor exists. However, if such inner common factors do exist, they may be extracted to form the cascade factorization

$$(4.24) \quad \begin{bmatrix} T_{11} & T_{12} \\ T_{21} & 0 \end{bmatrix} (s) = \begin{bmatrix} A_l(s) & 0 \\ 0 & I \end{bmatrix} \begin{bmatrix} \tilde{T}_{11} & \tilde{T}_{12} \\ \tilde{T}_{21} & 0 \end{bmatrix} (s) \begin{bmatrix} A_r(s) & 0 \\ 0 & I \end{bmatrix}$$

in which $A_l(s)$ and $A_r(s)$ are inner. Consequently, we have

$$(4.25) \quad \mathcal{R}(s) = A_l(s) [\tilde{T}_{11} - \tilde{T}_{12} Q \tilde{T}_{21}] (s) A_r(s).$$

Furthermore, a minimal realization for

$$(4.26) \quad \begin{bmatrix} \tilde{T}_{11} & \tilde{T}_{12} \\ \tilde{T}_{21} & 0 \end{bmatrix} (s)$$

will lead directly to a minimal realization for $\tilde{T}_{12}^* \tilde{T}_{11} \tilde{T}_{21}^*(s)$. Hence the function to be minimized in (2.25) or (2.26) can be written

$$(4.27) \quad \|(T_{11} - T_{12} Q T_{21})(s)\|_\infty = \|(\tilde{T}_{12}^* \tilde{T}_{11} \tilde{T}_{21}^* - Q)(s)\|_\infty.$$

Using Theorem 2.1 or Remark 2.2, $Q(s)$ can therefore be obtained as a Nehari extension or ρ -suboptimal extension of $\tilde{T}_{12}^* \tilde{T}_{11} \tilde{T}_{21}^*(s)$. It then follows from Theorem 2.1 and Remarks 2.2 and 2.4 that the corresponding "error system"

$$(4.28) \quad E(s) = (\tilde{T}_{12}^* \tilde{T}_{11} \tilde{T}_{21}^* - Q)(s)$$

satisfies the bounded real type equations given in (2.33) and (2.34). These equations form the basis of the hypothesis of the next theorem which enables us to deduce that the set of poles of $(\tilde{T}_{11} - \tilde{T}_{12} Q \tilde{T}_{21})(s)$ reduce to a subset of the poles of Q .

THEOREM 4.2. *Let*

$$(4.29) \quad \begin{bmatrix} G & A_1 \\ A_2 & 0 \end{bmatrix} (s) = \left[\begin{array}{cc|cc} A_{11} & A_{12} & B_{11} & B_{12} \\ 0 & A_{22} & B_{21} & 0 \\ \hline C_{11} & C_{12} & D & I \\ 0 & C_{22} & I & 0 \end{array} \right]$$

in which $A_1(s)$ and $A_2(s)$ are inner and their realizations

$$(4.30) \quad A_1(s) = \left[\begin{array}{c|c} A_{11} & B_{12} \\ \hline C_{11} & I \end{array} \right] \quad \text{and} \quad A_2(s) = \left[\begin{array}{c|c} A_{22} & B_{21} \\ \hline C_{22} & I \end{array} \right]$$

are also minimal and balanced. Then

(a)

$$(4.31) \quad \begin{aligned} A_1^* G A_2^* (s) &= \left[\begin{array}{cc|c} -A_{11}^* & -C_{11}^*(C_{12} + DB_{21}^*) - A_{12} - B_{11}B_{21}^* & C_{11}^*D + B_{11} \\ 0 & -A_{22}^* & C_{22}^* \\ \hline -B_{12}^* & -(C_{12} + DB_{21}^*) & D \end{array} \right] \\ &= \left[\begin{array}{c|c} A & B \\ \hline C & D \end{array} \right]; \end{aligned}$$

(b) for any

$$Q(s) = \left[\begin{array}{c|c} \hat{A} & \hat{B} \\ \hline \hat{C} & \hat{D} \end{array} \right]$$

such that

$$(4.32) \quad (A_1^* G A_2^* - Q)(s) = \left[\begin{array}{cc|c} A & 0 & B \\ 0 & \hat{A} & \hat{B} \\ \hline \underline{C} & -\hat{C} & \underline{D} - \hat{D} \end{array} \right] := \left[\begin{array}{c|c} \tilde{A} & \tilde{B} \\ \hline \tilde{C} & \tilde{D} \end{array} \right]$$

satisfies the bounded real type equations

$$(4.33) \quad \begin{bmatrix} -(\tilde{A}\tilde{P} + \tilde{P}\tilde{A}^* + \tilde{B}\tilde{B}^*) & -(\tilde{B}\tilde{D}^* + \tilde{P}\tilde{C}^*) \\ -(\tilde{D}\tilde{B}^* + \tilde{C}\tilde{P}) & \sigma^2 I - \tilde{D}\tilde{D}^* \end{bmatrix} = \begin{bmatrix} L \\ W \end{bmatrix} [L^* | W^*]$$

and

$$(4.34) \quad \begin{bmatrix} -(\tilde{A}^*\tilde{Q} + \tilde{Q}\tilde{A} + \tilde{C}^*\tilde{C}) & -(\tilde{C}^*\tilde{D} + \tilde{Q}\tilde{B}) \\ -(\tilde{D}^*\tilde{C} + \tilde{B}^*\tilde{Q}) & \sigma^2 I - \tilde{D}^*\tilde{D} \end{bmatrix} = \begin{bmatrix} L_d^* \\ W_d^* \end{bmatrix} [L_d | W_d]$$

in which

$$(4.35) \quad (i) \quad \tilde{P}\tilde{Q} = \sigma^2 I \text{ for some } \sigma \in \mathbb{R}, \text{ and}$$

$$(4.36) \quad (ii) \quad L^* = [0 | L_{21}^*], \quad L_d = [0 | L_{d21}]$$

where the partitioning of L and L_d is conformable with that of \tilde{B} and \tilde{C} in (4.32), we have

$$(4.37) \quad (i) \quad (G - A_1 Q A_2)(s) = \left[\frac{\hat{A}}{C_{11} Q_{13} - \hat{C}} \middle| \frac{\hat{B} - P_{23}^* B_{21}}{\hat{D}} \right]$$

where P_{23} and Q_{13} are the $(2, 3)$ and $(1, 3)$ partitions of \tilde{P} and \tilde{Q} (see (D.1) in Appendix D).

$$(4.38) \quad (ii) \quad \{\text{poles of } Q\} \supseteq \{\text{poles of } (G - A_1 Q A_2)\}.$$

Proof. See Appendix D.

Clearly, if we substitute $\tilde{T}_{11}(s)$, $\tilde{T}_{12}(s)$ and $\tilde{T}_{21}(s)$ into $G(s)$, $A_1(s)$, $A_2(s)$ of the last theorem, it follows immediately from part (b)(ii) that

$$(4.39) \quad \{\text{poles of } Q\} \supseteq \{\text{poles of } (\tilde{T}_{11} - \tilde{T}_{12} Q \tilde{T}_{21})\}$$

and this together with Theorem 4.1 and (4.25) yields

$$(4.40) \quad \{\text{poles of } Q\} \cup \{\text{poles of } A_r\} \cup \{\text{poles of } A_l\} \supseteq \{\text{poles of } \mathcal{R}\}.$$

Given (4.40), it follows that an upper bound for the McMillan degree of $\mathcal{R}(s)$ is

$$(4.41) \quad t \leq \deg(A_r) + \deg(A_l) + \deg(Q) = t_b.$$

Further, we have by Corollary 2.2 that

$$(4.42) \quad \deg(Q_{\text{opt}}) \leq \deg(\tilde{T}_{12}^* \tilde{T}_{11} \tilde{T}_{21}^*) + \deg(U) - r$$

in the case of optimal extensions (r is the multiplicity of the largest Hankel singular value of $\tilde{T}_{12}^* \tilde{T}_{11} \tilde{T}_{21}^*(s)$); and

$$(4.43) \quad \deg(Q_{\text{sopt}}) \leq \deg(\tilde{T}_{12}^* \tilde{T}_{11} \tilde{T}_{21}^*) + \deg(U)$$

in the case of ρ -suboptimal extensions. Now (3.26) and Lemma 3.1 in combination with (4.24) and Theorem 4.1 imply that

$$(4.44) \quad \deg(\tilde{T}_{12}^* \tilde{T}_{11} \tilde{T}_{21}^*) = \text{rank}(X) + \text{rank}(Y) - \deg(A_r) - \deg(A_l).$$

Direct substitution of this into the previous two inequalities and then into (4.41) yields

$$(4.45a) \quad t_{\text{opt}} \leq t_b = \text{rank}(X) + \text{rank}(Y) + \deg(U) - r$$

and

$$(4.45b) \quad t_{\text{sopt}} \leq t_b = \text{rank}(X) + \text{rank}(Y) + \deg(U)$$

which provides a McMillan degree bound for the closed-loop system and completes step 1 of our analysis.

We will now begin the second step. In order to establish a McMillan degree bound on all $K_{\text{opt}}(s)$ and $K_{\text{sopt}}(s)$ controllers, we need to bound the number of cancellations between $P(s)$ and $K(s)$ in Fig. 1; we call this bound m_b as was given in (4.1). In Theorem 4.3, we will show that every unobservable mode of the system in Fig. 1 is due to a cancellation with a zero of $P_{12}(s)$ and every uncontrollable mode is due to a cancellation with a zero of $P_{21}(s)$. (After this paper had been submitted for publication,

it came to our attention that a result similar to Theorem 4.3 had been discovered independently by Anderson and Linnemann [29].)

THEOREM 4.3. *Let*

$$(4.46) \quad \begin{bmatrix} P_{11} & P_{12} \\ P_{21} & P_{22} \end{bmatrix} (s) = \left[\begin{array}{c|cc} A & B_1 & B_2 \\ \hline C_1 & D_{11} & D_{12} \\ C_2 & D_{21} & D_{22} \end{array} \right]$$

in which $P_{12}(s) \in \mathbb{R}^{p_1 \times m_2}(s)$ with $m_2 \leq p_1$ and $P_{21}(s) \in \mathbb{R}^{p_2 \times m_1}(s)$ with $p_2 \leq m_1$. Suppose also that

$$(4.47) \quad K(s) = \left[\begin{array}{c|c} \hat{A} & \hat{B} \\ \hline \hat{C} & \hat{D} \end{array} \right]$$

is a minimal realization and that the well posedness condition $\det(I + D_{22}\hat{D}) \neq 0$ is satisfied. Then in the closed loop of Fig. 1

- (a) every unobservable mode is a zero of $P_{12}(s)$ and
- (b) every uncontrollable mode is a zero of $P_{21}(s)$.

Proof. See Appendix E.

Using this theorem, we see that the number of cancellations m between $P(s)$ and $K(s)$ is bounded above by

$$(4.48) \quad \begin{aligned} m \leq & \{\text{number of zeros of } P_{12}(s) \text{ in } \mathbb{C}_-\} \\ & + \{\text{number of zeros of } P_{21}(s) \text{ in } \mathbb{C}_-\} = m_b. \end{aligned}$$

This follows from the fact that any other cancellation (i.e. one corresponding to a right half plane zero of $P_{12}(s)$ or $P_{21}(s)$) violates the proven internal stability of the closed loop.

From (4.48) and Lemma 3.1, we have

$$(4.49) \quad \begin{aligned} m_b = & \{\text{number of zeros of } P_{12}(s) \text{ in } \mathbb{C}_-\} \\ & + \{\text{number of zeros of } P_{21}(s) \text{ in } \mathbb{C}_-\} \\ = & \{n - \text{rank}(X)\} + \{n - \text{rank}(Y)\} \\ = & 2n - \text{rank}(X) - \text{rank}(Y). \end{aligned}$$

We are now ready to prove the main theorem by combining together the results which have been established. Substitution of (4.45) and (4.49) into (4.1) proves the following.

THEOREM 4.4. *For any H^∞ -optimal control problem of the first kind, every H^∞ -optimal controller satisfies*

$$(i) \quad \deg(K_{\text{opt}}) \leq n - r + \deg(U)$$

where r is the multiplicity of the largest Hankel singular value of $\tilde{T}_{12}^* \tilde{T}_{11} \tilde{T}_{21}^*(s)$, and every ρ -suboptimal controller satisfies

$$(ii) \quad \deg(K_{\text{sopt}}) \leq n + \deg(U). \quad \square$$

Furthermore, (4.2) and (4.3) follow from the fact that $\deg(U) = 0$ if $U(s)$ is chosen constant. In the SISO case only $U = 1$ is allowed.

4.1. Computations. In this section we assemble together the ideas presented so far into an algorithm style procedure for solving H^∞ control problems of the first kind.

ALGORITHM.

- (1) Given $P(s)$ as in (2.3), do a prescaling to get $P(s)$ into the form (2.19).
- (2) Solve the Riccati equations (2.21) and (2.23) and evaluate the stabilizing matrices F and H using (2.20) and (2.22).
- (3) Assemble $T_{12}^* T_{11} T_{21}^*(s)$ as in (2.28).
- (4) Remove the hidden modes of $T_{12}^* T_{11} T_{21}^*(s)$. This can be done in a numerically reliable way by balanced truncation methods. The result is a minimal realization for $\tilde{T}_{12}^* \tilde{T}_{11} \tilde{T}_{21}^*(s)$. Note the balancing process for model reduction at this step forms the bulk of the computation required at the next step.
- (5) Determine a Nehari or ρ -suboptimal extension $Q(s)$ of $\tilde{T}_{12}^* \tilde{T}_{11} \tilde{T}_{21}^*(s)$ using Theorem 2.1 or Remark 2.2.
- (6) Back substitute $Q(s)$ into (2.17).
- (7) The previous step will typically produce a nonminimal realization for the (sub-)optimal controller $K(s)$. Again, a minimal realization may be obtained by balanced truncation methods; the bounds given in Theorem 4.4 must apply.

It should be noted that the model reduction performed in step 4 will simultaneously remove all the nonminimal states in (2.28) introduced by the left half plane zeros of $P_{12}(s)$ and $P_{21}(s)$, and the all-pass common factors shared by $T_{11}(s)$ and $T_{12}(s)$, and $T_{11}(s)$ and $T_{21}(s)$. Although Theorem 4.1 is an essential component of the theory, it does not need to be implemented in software. The second model reduction (step 7) is used to remove any nonminimal states introduced by cancellations predicted by Theorem 4.3.

4.2. Model reduction considerations. It is natural to consider reducing the number of controller states by model reduction methods such as those discussed in [12]. If we suppose that $\Delta\mathcal{R}(s)$ is the change in $\mathcal{R}(s)$ produced by the model reduction error $\Delta K(s)$, then the difficulty with this approach is that any general bound on $\|\Delta\mathcal{R}(s)\|_\infty$ in terms of a bound on $\|\Delta K(s)\|_\infty$ tends to be weak. An alternative and less direct approach is to consider the possibility of model reducing $Q(s)$ before obtaining $K(s)$ by back substitution. An argument might be that if $\Delta Q(s)$ is the perturbation produced by the model reduction of $Q(s)$, then by (2.14)

$$(4.50) \quad \mathcal{R}(s) + \Delta\mathcal{R}(s) = [T_{11} - T_{12}(Q + \Delta Q)T_{21}](s)$$

leads to

$$(4.51) \quad \|\Delta\mathcal{R}(s)\|_\infty = \|\Delta Q(s)\|_\infty,$$

since $T_{12}(s)$ and $T_{21}(s)$ are inner. Further, if the reduced order model of $Q(s)$ is obtained by retaining the first k states of a truncated balanced realization, then

$$(4.52) \quad \|\Delta\mathcal{R}(s)\|_\infty \leq 2 \sum_{i=k+1}^n \sigma_i(Q).$$

This inequality follows from [12, Thm. 9.6] and shows that it is possible to reduce the number of states of $Q(s)$ while simultaneously keeping track of the resulting maximum possible increase in $\|\mathcal{R}(s)\|_\infty$. However, contrary to the objective, this approach will tend to increase the number of controller states rather than decrease it. This is because replacing $Q(s)$ with a lower order approximation will destroy the ‘‘built in’’ cancellations predicted by our previous results. Since $K(s) = F_r(K_0(s), Q(s))$ and $\deg(K_0) = n$

we see that

$$(4.53) \quad \begin{aligned} \deg(K_{\text{opt}}) &= \deg(K_0) + \deg(Q) \\ &\quad - (\text{no. of cancellations between } K_0 \text{ and } Q_{\text{opt}}) \\ &\leq n-1 \text{ (in the case } \deg(U) = 0 \text{ in Theorem 4.4).} \end{aligned}$$

If Q_{opt} is replaced by an approximation $Q_a(s)$, then in general the cancellations no longer occur and the corresponding controller $K_a(s)$ has higher degree than that of $K_{\text{opt}}(s)$, specifically

$$\deg(K_a) = \deg(K_0) + \deg(Q_a) \geq n.$$

This point is now illustrated with an example.

Example 4.1. Consider the unweighted robust stabilization problem in which we seek

$$\inf_{K \in \Xi} \|K(I + GK)^{-1}\|_\infty \quad (\Xi \text{ is the set of stabilizing compensators}).$$

Referring back to (2.8) we recall that the corresponding $P(s)$ matrix is

$$P(s) = \begin{bmatrix} 0 & I \\ -I & G(s) \end{bmatrix}$$

and after scaling ($S_2 = -I$ and $S_1 = I$) we get

$$P(s) = \left[\begin{array}{c|cc} A & 0 & B \\ \hline 0 & 0 & I \\ -C & I & -D \end{array} \right].$$

If

$$G(s) = \{(s+3)\}/\{(s-1)(s-2)(s-3)\},$$

we get (by computer) a calculation which is based on § 4.1 that

$$Q_{\text{opt}} = \left[\begin{array}{c|c} A & B \\ \hline C & D \end{array} \right] = \left[\begin{array}{cc|c} -1.54049 & 0.89744 & 307.989 \\ -0.912165 & -0.17001 & 60.1637 \\ -0.083009 & 0.0159534 & 61.4750 \end{array} \right]$$

which has Hankel singular values 8.2979 and 2.8229. The corresponding optimal controller $K_{\text{opt}}(s)$ has McMillan degree two and has Hankel singular values 38.084 and 8.3797.

In a second calculation, we replaced $Q_{\text{opt}}(s)$ with a 1-state truncated balanced realization. In this case $K(s)$ had McMillan degree four with Hankel singular values: 38.065, 8.3596, 0.0016269 and 0.00010428. In this example therefore, removing a state from $Q_{\text{opt}}(s)$ leads to an increase of two in the McMillan degree of the controller. \square

4.3. Minimum entropy controllers. In a private communication N. J. Young pointed out to us that Arov and Krein [2] had studied a class of ρ -suboptimal extensions, which they called minimum entropy extensions. Assuming that $\|\mathcal{R}(s)\|_\infty \leq \rho = 1$ (a convenient normalization), we define the entropy of the closed loop system at some point $s_0 \in \mathbb{C}_-$ by

$$(4.54) \quad I(\mathcal{R}; s_0) = -\frac{1}{2\pi} \int_{-\infty}^{\infty} \ln |\det(I - \mathcal{R}^*(j\omega)\mathcal{R}(j\omega))| \frac{\text{Re}(s_0)}{|j\omega - s_0|^2} d\omega.$$

Using the fact that $\det(I + AB) = \det(I + BA)$, it is easy to see that inner matrices are entropy preserving. Since $\mathcal{R}(s) = T_{12}ET_{21}(s)$, where $E(s)$ is given by (4.28), it is clear that $I(\mathcal{R}; s_0) = I(E; s_0)$. From now on we will work with $E(s)$ knowing that it has the same entropy as the closed loop transfer function matrix $\mathcal{R}(s)$.

After some introductory comments we will state the result of Arov and Krein and thus show that (4.54) may be minimized while ensuring $\|\mathcal{R}(s)\|_\infty \leq 1$ with the aid of an n -state controller. It will be shown that the controller which minimizes the closed loop entropy is generated by setting $U = -H_{22}^*(s_0)$ in the general parametrization of all Nehari extensions.

From [12] we recall that

$$(4.55) \quad \begin{aligned} E^*(s) &= (\tilde{T}_{12}^* \tilde{T}_{11} \tilde{T}_{21}^*)^*(s) - H_{11}(s) \\ &\quad + H_{12}(s)U(s)(I + H_{22}(s)U(s))^{-1}H_{21}(s), \end{aligned}$$

or alternatively,

$$(4.56) \quad E^*(s) = \{A(s)U(s) + B(s)\}\{C(s)U(s) + D(s)\}^{-1}$$

in which

$$(4.57) \quad A(s) := \{(\tilde{T}_{12}^* \tilde{T}_{11} \tilde{T}_{21}^*)^*(s) - H_{11}\}H_{21}^{-1}(s)H_{22}(s) + H_{12}(s),$$

$$(4.58) \quad B(s) := \{(\tilde{T}_{12}^* \tilde{T}_{11} \tilde{T}_{21}^*)^*(s) - H_{11}\}H_{21}^{-1}(s),$$

$$(4.59) \quad C(s) := H_{21}^{-1}(s)H_{22}(s),$$

$$(4.60) \quad D(s) := H_{21}^{-1}(s).$$

Substituting from (2.38), using $\rho = 1$, we get

$$(4.61) \quad \begin{bmatrix} A(s) & B(s) \\ C(s) & D(s) \end{bmatrix} = \left[\begin{array}{cc|cc} A & 0 & -\Sigma\Gamma^{-1}C^* & -\Gamma^{-1}B \\ 0 & -A^* & \Gamma^{-1}C^* & \Gamma^{-1}\Sigma B \\ \hline C & 0 & I & 0 \\ 0 & B^* & 0 & I \end{array} \right]$$

in which we have denoted

$$(\tilde{T}_{12}^* \tilde{T}_{11} \tilde{T}_{21}^*)^*(s) = \left[\begin{array}{c|c} A & B \\ \hline C & D \end{array} \right].$$

Also, (4.61) is easily shown to be J -unitary, that is

$$(4.62) \quad \begin{bmatrix} A(s) & B(s) \\ C(s) & D(s) \end{bmatrix} \begin{bmatrix} I & 0 \\ 0 & -I \end{bmatrix} \begin{bmatrix} A^*(s) & C^*(s) \\ B^*(s) & D^*(s) \end{bmatrix} = \begin{bmatrix} I & 0 \\ 0 & -I \end{bmatrix}.$$

From the J -unitary equations we get

$$(4.63) \quad A(s)A^*(s) - B(s)B^*(s) = I,$$

$$(4.64) \quad C(s)C^*(s) - D(s)D^*(s) = -I,$$

$$(4.65) \quad A(s)C^*(s) - B(s)D^*(s) = 0.$$

Also

$$(4.66) \quad (4.64) \Rightarrow D^{-1}(s)D^{-*}(s) = I - H_{22}(s)H_{22}^*(s)$$

$$(4.67) \quad \Rightarrow \|H_{22}(s)\|_\infty \leq 1.$$

Glover [12] has shown that $H_{22}(s) \in RH_-^\infty$; a fact which is required in the proof of Theorem 4.5.

THEOREM 4.5 (Arov and Krein [2]). *If $I(H_{22}; s_0) < \infty$, $s_0 \in \mathbb{C}_-$ and $\|U(s)\|_\infty \leq 1$, then*

$$I(\mathcal{R}; s_0) = I(H_{22}; s_0) + I(U; s_0) + \ln |\det (I + H_{22}(s_0)U(s_0))|.$$

Also, there exists a unique U_0 such that $I(\mathcal{R}; s_0)$ attains a minimum value of

$$I(\mathcal{R}; s_0) = I(H_{22}; s_0) + \ln |\det (I - H_{22}(s_0)H_{22}^*(s_0))|/2.$$

The optimizing U_0 is given by

$$U_0 = -H_{22}^*(s_0).$$

A proof which mimics the discrete time proof in [2] is given in Appendix F.

5. Conclusions. Our purpose was to carry out a detailed analysis of the pole-zero cancellations which occur in the class of H^∞ -optimal control problems described in § 2.2. If $\deg(P) = n$, we have shown that SISO H^∞ controllers never require more than $n-1$ states and that MIMO problems have a continuum of controllers whose McMillan degree satisfy this same bound. A general bound on $\deg(K)$ has been derived for all Nehari and ρ -suboptimal extensions and is given in Theorem 4.4. The bounds in Theorem 4.4 are tight in the sense that there exist problems for which they are met with equality. We have found in numerous examples that these bounds typically give the actual McMillan degree of the controller.

It is our belief that state-space dimension inflation is an important consideration in practical H^∞ design problems. Apart from being interesting in its own right, a complete cancellation theory is a prerequisite for the development of reliable computational software. Example 4.1 is an illustration of how a seemingly sensible, but ill-advised intermediate model reduction step may aggravate the problem of degree inflation rather than alleviate it.

H^∞ design problems which may be embedded in Fig. 1 but with either $P_{12}(s)$ or $P_{21}(s)$, or both, nonsquare have been studied by several researchers [6], [9], [10], [11], [22], [25]. In this class of problems cancellation phenomena are more difficult to analyse [16]. However, the added complexity and iterative nature of their solution makes a cancellation theory even more essential. An additional layer of difficulty is introduced by the various scaling strategies which are introduced in the μ -synthesis work of Doyle [5], Doyle et al. [6], Safonov [20] and others.

Appendix A.

Proof of Theorem 2.1. By assumption, the Hankel singular values have been ordered so that

$$(A.1) \quad \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix} \begin{bmatrix} \Sigma & 0 \\ 0 & \sigma_{k+1}I \end{bmatrix} + \begin{bmatrix} \Sigma & 0 \\ 0 & \sigma_{k+1}I \end{bmatrix} \begin{bmatrix} A_{11}^* & A_{21}^* \\ A_{12}^* & A_{22}^* \end{bmatrix} + \begin{bmatrix} B_1 \\ B_2 \end{bmatrix} [B_1^* | B_2^*] = 0$$

and

$$(A.2) \quad \begin{bmatrix} A_{11}^* & A_{21}^* \\ A_{12}^* & A_{22}^* \end{bmatrix} \begin{bmatrix} \Sigma & 0 \\ 0 & \sigma_{k+1}I \end{bmatrix} + \begin{bmatrix} \Sigma & 0 \\ 0 & \sigma_{k+1}I \end{bmatrix} \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix} + \begin{bmatrix} C_1^* \\ C_2^* \end{bmatrix} [C_1 | C_2] = 0$$

are satisfied.

Next, we invoke [12, Thm. 8.7] which states that all error systems with the desired properties may be generated by

$$(A.3) \quad \hat{X}(s) + Q^*(s) = H_{11}(s) - H_{12}(s)U(s)(I + H_{22}(s)U(s))^{-1}H_{21}(s)$$

in which

$$(A.4) \quad H(s) = \begin{bmatrix} H_{11}(s) & H_{12}(s) \\ H_{21}(s) & H_{22}(s) \end{bmatrix} = \left[\begin{array}{c|cc} \Gamma^{-1}(\sigma_{k+1}^2 A_{11}^* + \Sigma A_{11} \Sigma) & \Gamma^{-1} \Sigma B_1 & -\Gamma^{-1} C_1^* \\ \hline C_1 \Sigma & D & I \\ -\sigma_{k+1} B_1^* & \sigma_{k+1} I & 0 \end{array} \right]$$

and

$$(A.5) \quad U(s) \in RH_-^\infty$$

satisfies

$$(A.6) \quad \|U(j\omega)\|_\infty \leq 1$$

and

$$(A.7) \quad C_2 + U(s)B_2^* = 0.$$

In order that we may obtain the required state-space characterization of all error systems we assume that an arbitrary $U(s)$ with the desired properties has a minimal state-space realization $U(s) = (\hat{A}, \hat{B}, \hat{C}, \hat{D})$. Since $\|U(j\omega)\|_\infty \leq 1 \Leftrightarrow I - U(j\omega)U^*(j\omega) \geq 0$, the bounded real lemma [1, p. 308] ensures the existence of $\hat{P} = \hat{P}^* < 0$, $\hat{Q} = \hat{Q}^* < 0$, \hat{L} , \hat{W} , \hat{L}_d and \hat{W}_d such that

$$(A.8) \quad \begin{bmatrix} -(\hat{A}\hat{P} + \hat{P}\hat{A}^* + \hat{B}\hat{B}^*) & -(\hat{B}\hat{D}^* + \hat{P}\hat{C}^*) \\ -(\hat{D}\hat{B}^* + \hat{C}\hat{P}) & I - \hat{D}\hat{D}^* \end{bmatrix} = \begin{bmatrix} \hat{L} \\ \hat{W} \end{bmatrix} [\hat{L}^* | \hat{W}^*]$$

and

$$(A.9) \quad \begin{bmatrix} -(\hat{A}^*\hat{Q} + \hat{Q}\hat{A} + \hat{C}^*\hat{C}) & -(\hat{C}^*\hat{D} + \hat{Q}\hat{B}) \\ -(\hat{D}^*\hat{C} + \hat{B}^*\hat{Q}) & I - \hat{D}^*\hat{D} \end{bmatrix} = \begin{bmatrix} \hat{L}_d^* \\ \hat{W}_d^* \end{bmatrix} [\hat{L}_d | \hat{W}_d]$$

are satisfied.

$$(A.10) \quad (A.7) \Rightarrow C_2 + U(\infty)B_2^* = 0 \Rightarrow C_2 + \hat{D}B_2^* = 0,$$

$$(A.11) \quad \begin{aligned} (A.7) \text{ and } (A.10) &\Rightarrow \hat{C}(sI - \hat{A})^{-1}\hat{B}B_2^* = 0 \\ &\Rightarrow \hat{B}B_2^* = 0 \quad (\text{since } [\hat{A}, \hat{C}] \text{ is observable}). \end{aligned}$$

The condition $\|U(j\omega)\|_\infty \leq 1$ implies that there exists a spectral factor $\Delta(s)$ such that

$$(A.12) \quad I - U^*(s)U(s) = \Delta^*(s)\Delta(s),$$

$$(A.7) \Rightarrow B_2 U^*(s)U(s)B_2^* = C_2^* C_2,$$

$$(A.13) \quad (A.12) \Rightarrow B_2 B_2^* - B_2 \Delta^*(s)\Delta(s)B_2^* = C_2^* C_2.$$

The (2, 2) blocks of (A.1) and (A.2) $\Rightarrow B_2 B_2^* = C_2^* C_2$ and this together with (A.13) means

$$(A.14) \quad \Delta(s)B_2^* = 0.$$

Multiplying (A.7) on the left by $U^*(s)$ we get

$$U^*(s)C_2 + U^*(s)U(s)B_2^* = 0$$

$$\Leftrightarrow U^*(s)C_2 + B_2^* - \Delta^*(s)\Delta(s)B_2^* = 0 \quad (\text{by (A.12)})$$

$$(A.15) \quad \Rightarrow U^*(s)C_2 + B_2^* = 0 \quad (\text{by (A.14)}).$$

Hence

$$(A.16) \quad (A.15) \Rightarrow U^*(\infty)C_2 + B_2^* = \hat{D}^*C_2 + B_2^* = 0,$$

$$(A.17) \quad (A.15) \text{ and } (A.16) \Rightarrow \hat{B}^*(sI + \hat{A}^*)^{-1}\hat{C}^*C_2 = 0$$

$$(A.18) \quad \Rightarrow \hat{C}^*C_2 = 0 \quad (\text{by the controllability of } [\hat{A}, \hat{B}]).$$

The state-space model (2.32) for $E(s)$ can be obtained by deriving a state-space realization for the linear fractional transformation (A.3) from the realizations of $H(s)$ and $U(s)$.

Equations (2.33) and (2.34) are proved by simple calculations which are reminiscent of those in [12]. We will begin with the (1, 1) block of (2.33). The validity of partitions (1, 1), (1, 2), (2, 1) and (2, 2) follows directly from (A.1).

$$\begin{aligned} \text{partition (1, 3)} &= A_{11} + (\sigma_{k+1}^2 A_{11} + \Sigma A_{11}^* \Sigma - \sigma_{k+1} B_1 \hat{D}^* C_1) \Gamma^{-1} \\ &\quad + B_1 (B_1^* \Sigma + \sigma_{k+1} \hat{D}^* C_1) \Gamma^{-1} \\ &= (A_{11} \Sigma^2 - \sigma_{k+1}^2 A_{11} + \sigma_{k+1}^2 A_{11} + \Sigma A_{11}^* \Sigma \\ &\quad - (A_{11} \Sigma + \Sigma A_{11}^*) \Sigma) \Gamma^{-1} \quad (\text{by (A.1)}) \\ &= 0. \end{aligned}$$

$$\text{partition (1, 4)} = -\sigma_{k+1} B_1 \hat{B}^* + \sigma_{k+1} B_1 \hat{B}^* = 0.$$

$$\begin{aligned} \text{partition (2, 3)} &= A_{21} + B_2 (B_1^* \Sigma + \sigma_{k+1} \hat{D}^* C_1) \Gamma^{-1} \\ &= (A_{21} \Sigma^2 - \sigma_{k+1}^2 A_{21} - (\sigma_{k+1} A_{12}^* + A_{21} \Sigma) \Sigma \\ &\quad + (A_{12}^* \Sigma + \sigma_{k+1} A_{21}) \sigma_{k+1}) \Gamma^{-1} \quad (\text{by (A.1, A.2, A.10)}) \\ &= 0. \end{aligned}$$

$$\text{partition (2, 4)} = \sigma_{k+1} B_2 \hat{B}^* = 0 \quad (\text{by (A.11)}).$$

$$\begin{aligned} \text{partition (3, 1)} &= \Gamma^{-1} \{ \sigma_{k+1}^2 A_{11}^* + \Sigma A_{11} \Sigma - \sigma_{k+1} C_1^* \hat{D} B_1^* + \Sigma^2 A_{11}^* - \sigma_{k+1}^2 A_{11}^* \\ &\quad + (\Sigma B_1 + \sigma_{k+1} C_1^* \hat{D}) B_1^* \} \\ &= \Gamma^{-1} \{ \Sigma A_{11} \Sigma + \Sigma^2 A_{11}^* - \Sigma (A_{11} \Sigma + \Sigma A_{11}^*) \} \quad (\text{by (A.1)}) \\ &= 0. \end{aligned}$$

$$\begin{aligned} \text{partition (3, 2)} &= \Gamma^{-1} \{ \Sigma^2 A_{21}^* - \sigma_{k+1}^2 A_{21}^* + (\Sigma B_1 + \sigma_{k+1} C_1^* \hat{D}) B_2^* \} \\ &= \Gamma^{-1} \{ \Sigma^2 A_{21}^* - \sigma_{k+1}^2 A_{21}^* - \Sigma (\sigma_{k+1} A_{12} + \Sigma A_{21}^*) \\ &\quad + \sigma_{k+1} (\sigma_{k+1} A_{12}^* + \Sigma A_{12}) \} \\ &= 0 \quad (\text{by (A.1), (A.2), (A.10)}). \end{aligned}$$

$$\begin{aligned} \text{partition (3, 3)} &= \Gamma^{-1} \{ (\sigma_{k+1}^2 A_{11}^* + \Sigma A_{11} \Sigma - \sigma_{k+1} C_1^* \hat{D} B_1^*) \Sigma \\ &\quad + \Sigma (\sigma_{k+1}^2 A_{11} + \Sigma A_{11}^* \Sigma - \sigma_{k+1} B_1 \hat{D}^* C_1) \\ &\quad + (\Sigma B_1 + \sigma_{k+1} C_1^* \hat{D}) (B_1^* \Sigma + \sigma_{k+1} \hat{D}^* C_1) \\ &\quad + \sigma_{k+1}^2 C_1^* \hat{W} \hat{W}^* C_1 \} \Gamma^{-1} \\ &= \Gamma^{-1} \{ \sigma_{k+1}^2 A_{11}^* \Sigma + \Sigma A_{11} \Sigma^2 + \sigma_{k+1}^2 \Sigma A_{11}^* + \Sigma^2 A_{11} \Sigma \\ &\quad - \Sigma (A_{11} \Sigma + \Sigma A_{11}^*) \Sigma \\ &\quad - \sigma_{k+1}^2 (A_{11}^* \Sigma + \Sigma A_{11}) \} \Gamma^{-1} \\ &= 0 \quad (\text{by (A.1), (A.2), (A.8)}). \end{aligned}$$

$$\begin{aligned}
\text{partition (3, 4)} &= \Gamma^{-1} \{ \sigma_{k+1}^2 C_1^* \hat{C} \hat{P} - \sigma_{k+1} \Sigma B_1 \hat{B}^* \\
&\quad + (\Sigma B_1 + \sigma_{k+1} C_1^* \hat{D}) \sigma_{k+1} \hat{B}^* + \sigma_{k+1}^2 C_1^* \hat{W} \hat{L}^* \} \\
&= \sigma_{k+1}^2 \Gamma^{-1} C_1^* \{ \hat{C} \hat{P} + \hat{D} \hat{B}^* + \hat{W} \hat{L}^* \} \\
&= 0 \quad (\text{by (A.8)}). \\
\text{partition (4, 1)} &= -\sigma_{k+1} \hat{B} \hat{B}_1^* + \sigma_{k+1} \hat{B} \hat{B}_1^* = 0. \\
\text{partition (4, 2)} &= \sigma_{k+1} \hat{B} \hat{B}_2^* = 0 \quad (\text{by (A.11)}). \\
\text{partition (4, 3)} &= \{ -\sigma_{k+1} \hat{B} \hat{B}_1^* \Sigma + \sigma_{k+1}^2 \hat{P} \hat{C}^* C_1 + \sigma_{k+1} \hat{B} (B_1^* \Sigma + \sigma_{k+1} \hat{D}^* C_1) \\
&\quad + \sigma_{k+1}^2 \hat{L} \hat{W}^* C_1 \} \Gamma^{-1} \\
&= \sigma_{k+1}^2 \{ \hat{P} \hat{C}^* + \hat{B} \hat{D}^* + \hat{L} \hat{W}^* \} C_1 \Gamma^{-1} \\
&= 0 \quad (\text{by (A.8)}). \\
\text{partition (4, 4)} &= \sigma_{k+1}^2 (\hat{A} \hat{P} + \hat{P} \hat{A}^* + \hat{B} \hat{B}^* + \hat{L} \hat{L}^*) \\
&= 0 \quad (\text{by (A.8)}).
\end{aligned}$$

The (2, 1) and therefore also the (1, 2) blocks of (2.33) are verified next.

$$\begin{aligned}
\text{partition (1, 1)} &= \sigma_{k+1} \hat{D} \hat{B}_1^* + C_1 \Sigma - C_1 \Sigma - \sigma_{k+1} \hat{D} \hat{B}_1^* = 0. \\
\text{partition (1, 2)} &= \sigma_{k+1} (\hat{D} \hat{B}_2^* + C_2) = 0 \quad (\text{by (A.10)}). \\
\text{partition (1, 3)} &= \{ \sigma_{k+1} \hat{D} (B_1^* \Sigma + \sigma_{k+1} \hat{D}^* C_1) + C_1 \Sigma^2 - \sigma_{k+1}^2 C_1 \\
&\quad - (C_1 \Sigma + \sigma_{k+1} \hat{D} \hat{B}_1^*) \Sigma + \sigma_{k+1}^2 \hat{W} \hat{W}^* C_1 \} \Gamma^{-1} \\
&= \sigma_{k+1}^2 \{ \hat{D} \hat{D}^* - I + \hat{W} \hat{W}^* \} C_1 \\
&= 0 \quad (\text{by (A.8)}).
\end{aligned}$$

$$\text{partition (1, 4)} = \sigma_{k+1}^2 \{ \hat{D} \hat{B}^* + \hat{C} \hat{P} + \hat{W} \hat{L}^* \} = 0 \quad (\text{by (A.8)}).$$

The (2, 2) block of (2.33) follows immediately from (A.8); (2.33) is thus proven. The validity of (2.34) is established in the same way—in this case use is made of equations (A.1), (A.2), (A.9), (A.16) and (A.18). Since the calculations are very similar to those used to establish (2.33), these details are omitted. \square

Appendix B.

Proof of Theorem 3.2. As one would expect, the proofs associated with problems (i) to (iii) are similar. We will therefore only prove the result in the case of (i); the sensitivity proof being marginally more intricate than the others.

Equation (2.6) shows that the $P(s)$ matrix associated with the unweighted sensitivity problem is

$$(B.1) \quad P(s) = \begin{bmatrix} I & G \\ I & G \end{bmatrix} (s)$$

which has a state-space realization

$$(B.2) \quad P(s) = \left[\begin{array}{c|cc} A & 0 & B \\ \hline C & I & D \\ C & I & D \end{array} \right] (s).$$

After scaling as in (2.19), we get

$$(B.3) \quad P(s) = \left[\begin{array}{c|cc} A & 0 & BD^{-1} \\ \hline C & I & I \\ C & I & I \end{array} \right] (s).$$

We have already established that the interpolation points are the right half plane zeros of $G(s)$ and that the left half plane zeros play no part. For simplicity, we will assume that $G(s)$ has all its zeros in \mathbb{C}_+ . If this is not the case, the Riccati equation balancing theory of this section may be used to reduce the general problem to one in which $\operatorname{Re}[\lambda(A - BD^{-1}C)] > 0$. We leave the details to the reader.

If we now substitute the various partitions of (B.3) into (2.20) to (2.23) we obtain

$$(B.4) \quad X(A - BD^{-1}C) + (A - BD^{-1}C)^*X - XBD^{-1}D^{-*}B^*X = 0,$$

$$(B.5) \quad YA^* + AY - YC^*CY = 0,$$

$$(B.6) \quad F = C + D^{-*}B^*X,$$

$$(B.7) \quad H = YC^*.$$

Since $\operatorname{Re}[\lambda(A - BD^{-1}C)] > 0$ by assumption, the stabilizing solution X to (B.4) is nonsingular.

Next, we use (B.3) and (B.4) to (B.7) in (2.28) to obtain

$$(B.8) \quad T_{12}^* T_{11} T_{21}^* = \left[\begin{array}{cc|c} -\{A - BD^{-1}(C + D^{-*}B^*X)\}^* & XBD^{-1}CY & XBD^{-1} \\ 0 & -\{A - YC^*C\}^* & C^* \\ \hline D^{-*}B^* & -D^{-*}B^*XY & I \end{array} \right].$$

Introducing the basis change

$$(B.9) \quad T = \begin{bmatrix} I & -XY \\ 0 & I \end{bmatrix}$$

gives

$$(B.10) \quad T_{12}^* T_{11} T_{21}^* = \left[\begin{array}{cc|c} -\{A - BD^{-1}(C + D^{-*}B^*X)\}^* & 0 & X(BD^{-1} - YC^*) \\ 0 & -\{A - YC^*C\}^* & C^* \\ \hline D^{-*}B^* & 0 & I \end{array} \right] \\ = \left[\begin{array}{c|c} -\{A - BD^{-1}(C + D^{-*}B^*X)\}^* & X(BD^{-1} - YC^*) \\ \hline D^{-*}B^* & I \end{array} \right].$$

The fact that the (1, 2) block of the A -matrix in (B.10) is zero may be proved using (B.4) Y and X (B.5).

The equation defining the observability gramian of $T_{12}^* T_{11} T_{21}^*(s)$ is

$$(B.11) \quad -\{A - BD^{-1}(C + D^{-*}B^*X)\}Q - Q\{A - BD^{-1}(C + D^{-*}B^*X)\}^* \\ + BD^{-1}D^{-*}B = 0.$$

Comparison of (B.11) with $X^{-1}(B.4)X^{-1} = 0$ reveals that

$$(B.12) \quad Q = -X^{-1} < 0.$$

The equation defining the controllability gramian of $T_{12}^* T_{11} T_{21}^*(s)$ is

$$(B.13) \quad \{A - BD^{-1}(C + D^{-*}B^*X)\}^*P + P\{A - BD^{-1}(C + D^{-*}B^*X)\} \\ - X(BD^{-1} - YC^*)(D^{-*}B^* - CY)X = 0.$$

Substituting $PX^{-1}(B.4)$ and $(B.4)X^{-1}P$ into (B.13) we get

$$(B.14) \quad X(A - BD^{-1}C)X^{-1}P + PX^{-1}(A - BD^{-1}C)^*X \\ + X(BD^{-1} - YC^*)(D^{-*}B^* - CY)X = 0.$$

Substitution of (B.4) and (B.5) into (B.14) now shows that (B.13) is satisfied by

$$(B.15) \quad P = -X(I + YX) < 0,$$

whence

$$PQ = I + XY$$

and

$$\|(T_{12}^* T_{11} T_{21}^*(s))\|_H = \lambda_{\max}(PQ)^{1/2} = [1 + \lambda_{\max}(YX)]^{1/2}.$$

Finally, we know (from the discussion of § 2) that

$$\inf_{K \in \Xi} \|(I + GK)^{-1}(s)\|_{\infty} = \|(T_{12}^* T_{11} T_{21}^*)^*(s)\|_H,$$

and this concludes the proof of (i). Parts (ii) and (iii) and the unweighted version of (iv) can be proved using similar calculations. To prove the weighted version of (iv), one may invoke the ideas in [13] whereby a weighted optimal robustness problem can be transformed into an equivalent unweighted problem. These details are left to the interested reader. \square

Appendix C.

Proof of Theorem 4.1. We may assume without loss of generality that the realizations for $A_1(s)$ and $A_2(s)$ are balanced. Since they are minimal also, the following six all-pass equations [12] are satisfied.

$$(C.1) \quad A_{11} + A_{11}^* + B_{12} B_{12}^* = 0,$$

$$(C.2) \quad A_{11} + A_{11}^* + C_{11}^* C_{11} = 0,$$

$$(C.3) \quad C_{11} + B_{12}^* = 0,$$

$$(C.4) \quad A_{22} + A_{22}^* + B_{21} B_{21}^* = 0,$$

$$(C.5) \quad A_{22} + A_{22}^* + C_{22}^* C_{22} = 0,$$

$$(C.6) \quad C_{22} + B_{21}^* = 0.$$

The first part of the proof will be concerned with the extraction of a maximal degree all-pass left factor $A_1(s)$ from $A_1(s)$ and $G(s)$. Let us consider

$$(C.7) \quad A_1^* G(s) = \left[\begin{array}{c|c} \frac{-A_{11}^*}{-B_{12}^*} & \frac{C_{11}^*}{I} \end{array} \right] \left[\begin{array}{cc|c} A_{11} & A_{12} & B_{11} \\ 0 & A_{22} & B_{21} \\ \hline C_{11} & C_{12} & D \end{array} \right]$$

$$= \left[\begin{array}{ccc|c} -A_{11}^* & C_{11}^* C_{11} & C_{11}^* C_{12} & C_{11}^* D \\ 0 & A_{11} & A_{12} & B_{11} \\ 0 & 0 & A_{22} & B_{21} \\ \hline -B_{12}^* & C_{11} & C_{12} & D \end{array} \right].$$

Introducing the change of basis

$$T = \left[\begin{array}{ccc} I & I & 0 \\ 0 & I & 0 \\ 0 & 0 & I \end{array} \right]$$

and then making use of (C.2) and (C.3) we obtain

$$(C.8) \quad A_1^* G(s) = \left[\begin{array}{cc|c} -A_{11}^* & A_{12} + C_{11}^* C_{12} & C_{11}^* D + B_{11} \\ 0 & A_{22} & B_{21} \\ \hline -B_{12}^* & C_{12} & D \end{array} \right].$$

Our purpose now is to show that all the uncontrollable modes in $A_1^*G(s)$ are the poles of $A_1^*(s)$. Since (A_{22}, B_{21}) is controllable by assumption, every uncontrollable mode in (C.8) is an eigenvalue of $-A_{11}^*$. First, we observe that without loss of generality, we may assume that the state-space basis of the realization in (4.15) has been chosen so that in addition to (C.1)–(C.6), we have

$$(C.9) \quad A_{12} + C_{11}^* C_{12} = 0.$$

To show that this is so, consider the following change of basis in (4.15)

$$(C.10) \quad \begin{bmatrix} \bar{A}_{11} & \bar{A}_{12} \\ 0 & \bar{A}_{22} \end{bmatrix} = \begin{bmatrix} I & T \\ 0 & I \end{bmatrix} \begin{bmatrix} A_{11} & A_{12} \\ 0 & A_{22} \end{bmatrix} \begin{bmatrix} I & -T \\ 0 & I \end{bmatrix},$$

$$(C.11) \quad [\bar{C}_{11} \quad \bar{C}_{12}] = [C_{11} \quad C_{12}] \begin{bmatrix} I & -T \\ 0 & I \end{bmatrix}.$$

We now demonstrate that T may be chosen to make

$$(C.12) \quad \bar{A}_{12} + \bar{C}_{11}^* \bar{C}_{12} = 0.$$

From (C.10) and (C.11), (C.12) is equivalent to

$$(C.13) \quad (A_{12} + TA_{22} - A_{11}T) + C_{11}^*(C_{12} - C_{11}T) = 0.$$

Using (C.2), (C.13) becomes

$$(C.14) \quad A_{11}^*T + TA_{22} + (A_{12} + C_{11}^*C_{12}) = 0.$$

Since A_{11} and A_{22} are stable, (C.14) always admits a unique solution in T . Such a T ensures that the transformed state-space matrices satisfy (C.12). For notational simplicity, we assume that this basis change has been carried out initially and we revert to the original notation. In view of (C.9), (C.8) becomes

$$(C.15) \quad A_1^*G(s) = \left[\begin{array}{cc|c} -A_{11}^* & 0 & C_{11}^*D + B_{11} \\ 0 & A_{22} & B_{21} \\ \hline -B_{12}^* & C_{12} & D \end{array} \right].$$

If the realization (C.15) has any uncontrollable modes (which must be eigenvalues of $-A_{11}^*$) we may introduce a basis change

$$(C.16) \quad T = \begin{bmatrix} U_1 & 0 \\ 0 & I \end{bmatrix}$$

in which U_1 is orthogonal, to transform (C.15) to [24]

$$A_1^*G(s) = \left[\begin{array}{ccc|c} -\tilde{A}_{00}^* & 0 & 0 & \tilde{C}_{10}^*D + \tilde{B}_{01} \\ -\tilde{A}_{01}^* & -\tilde{A}_{11}^* & 0 & \tilde{C}_{11}^*D + \tilde{B}_{11} \\ 0 & 0 & A_{22} & B_{21} \\ \hline -\tilde{B}_{02}^* & -\tilde{B}_{12}^* & C_{12} & D \end{array} \right],$$

in which all the uncontrollable modes are eigenvalues of $-\tilde{A}_{00}^*$. That is

$$(C.17) \quad \tilde{C}_{10}^*D + \tilde{B}_{01} = 0$$

and thus

$$(C.18) \quad A_1^*G(s) = \left[\begin{array}{cc|c} -\tilde{A}_{11}^* & 0 & \tilde{C}_{11}^*D + \tilde{B}_{11} \\ 0 & A_{22} & B_{21} \\ \hline -\tilde{B}_{12}^* & C_{12} & D \end{array} \right]$$

is a controllable realization.

Introducing the basis of (C.18) into (4.15) allows us to write

$$(C.19) \quad [G \ A_1](s) = \left[\begin{array}{ccc|cc} \tilde{A}_{00} & \tilde{A}_{01} & \tilde{A}_{02} & \tilde{B}_{01} & \tilde{B}_{02} \\ 0 & \tilde{A}_{11} & \tilde{A}_{12} & \tilde{B}_{11} & \tilde{B}_{12} \\ 0 & 0 & A_{22} & B_{21} & 0 \\ \hline \tilde{C}_{10} & \tilde{C}_{11} & C_{12} & D & I \end{array} \right]$$

in which we have by (C.9)

$$(C.20) \quad \begin{bmatrix} \tilde{A}_{02} \\ \tilde{A}_{12} \end{bmatrix} + \begin{bmatrix} \tilde{C}_{10}^* \\ \tilde{C}_{11}^* \end{bmatrix} C_{12} = 0.$$

Noting that orthogonal transformations map balanced realizations into balanced realizations in the case of minimal realizations of inner matrices gives

$$(C.21) \quad \begin{bmatrix} \tilde{A}_{00} & \tilde{A}_{01} \\ 0 & \tilde{A}_{11} \end{bmatrix} + \begin{bmatrix} \tilde{A}_{00}^* & 0 \\ \tilde{A}_{01}^* & \tilde{A}_{11}^* \end{bmatrix} + \begin{bmatrix} \tilde{C}_{10}^* \\ \tilde{C}_{11}^* \end{bmatrix} [\tilde{C}_{10} | \tilde{C}_{11}] = 0,$$

$$(C.22) \quad [\tilde{C}_{10} | \tilde{C}_{11}] + [\tilde{B}_{02}^* | \tilde{B}_{12}^*] = 0.$$

Substituting for \tilde{A}_{01} from (C.21), \tilde{A}_{02} from (C.20), \tilde{B}_{01} from (C.17) and \tilde{B}_{02}^* from (C.22) into (C.19) we get

$$(C.23) \quad [G \ A_1](s) = \left[\begin{array}{ccc|cc} \tilde{A}_{00} & -\tilde{C}_{10}^* \tilde{C}_{11} & -\tilde{C}_{10}^* C_{12} & -\tilde{C}_{10}^* D & -\tilde{C}_{10}^* \\ 0 & \tilde{A}_{11} & \tilde{A}_{12} & \tilde{B}_{11} & \tilde{B}_{12} \\ 0 & 0 & A_{22} & B_{21} & 0 \\ \hline \tilde{C}_{10} & \tilde{C}_{11} & C_{12} & D & I \end{array} \right]$$

$$(C.24) \quad = \left[\begin{array}{c|c} \tilde{A}_{00} & -\tilde{C}_{10}^* \\ \hline \tilde{C}_{10} & I \end{array} \right] \left[\begin{array}{cc|cc} \tilde{A}_{11} & \tilde{A}_{12} & \tilde{B}_{11} & \tilde{B}_{12} \\ 0 & A_{22} & B_{21} & 0 \\ \hline \tilde{C}_{11} & C_{12} & D & I \end{array} \right]$$

$$(C.25) \quad = A_l(s) [\bar{G} \ \tilde{A}_1](s)$$

where $A_l(s)$ and $\tilde{A}_1(s)$ are as given in (4.17) (note that $\tilde{B}_{02}^* = -\tilde{C}_{10}^*$). It follows immediately from (C.21) and (C.22) that $A_l(s)$ and hence also $\tilde{A}_1(s)$ are inner.

By using dual arguments, we can extract a maximal degree all-pass right factor $A_r(s)$ from $\bar{G}(s)$ and $A_2(s)$. We begin this calculation with the change of basis

$$(C.26) \quad T = \begin{bmatrix} I & S \\ 0 & I \end{bmatrix}$$

in the state-space of

$$(C.27) \quad \begin{bmatrix} \bar{G} \\ A_2 \end{bmatrix}(s) = \left[\begin{array}{cc|c} \tilde{A}_{11} & \tilde{A}_{12} & \tilde{B}_{11} \\ 0 & A_{22} & B_{21} \\ \hline \tilde{C}_{11} & C_{12} & D \\ 0 & C_{22} & I \end{array} \right]$$

where S is the unique solution of

$$(C.28) \quad \tilde{A}_{11} S + S A_{22}^* - \tilde{A}_{12} - \tilde{B}_{11} B_{21}^* = 0.$$

The purpose of this basis change is to transform the realization (C.27) to

$$(C.29) \quad \begin{bmatrix} \bar{G} \\ A_2 \end{bmatrix}(s) = \left[\begin{array}{cc|c} \tilde{A}_{11} & \hat{A}_{12} & \tilde{B}_{11} \\ 0 & A_{22} & B_{21} \\ \hline \tilde{C}_{11} & \hat{C}_{12} & D \\ 0 & C_{22} & I \end{array} \right]$$

in which

$$(C.30) \quad \hat{A}_{12} + \tilde{B}_{11} B_{21}^* = 0.$$

Next, a second orthogonal transformation

$$T = \begin{bmatrix} I & 0 \\ 0 & U_2 \end{bmatrix}$$

together with arguments which are duals of those invoked previously (see equations (C.15) to (C.25)) allow us to write

$$(C.31) \quad \begin{bmatrix} \bar{G} \\ A_2 \end{bmatrix} (s) = \left[\begin{array}{cc|c} \tilde{A}_{11} & \tilde{A}_{12} & \tilde{B}_{11} \\ 0 & \tilde{A}_{22} & \tilde{B}_{21} \\ \hline \tilde{C}_{11} & \tilde{C}_{12} & D \\ 0 & \tilde{C}_{22} & I \end{array} \right] \left[\begin{array}{c|c} \tilde{A}_{33} & \tilde{B}_{31} \\ \hline -\tilde{B}_{31}^* & I \end{array} \right]$$

$$(C.32) \quad = \begin{bmatrix} \tilde{G} \\ \tilde{A}_2 \end{bmatrix} (s) A_r(s)$$

in which both $\tilde{A}_2(s)$ and $A_r(s)$ are inner. Further,

$$(C.33) \quad \begin{aligned} \bar{G} A_2^*(s) &= \tilde{G} \tilde{A}_2^*(s) \\ &= \left[\begin{array}{cc|c} \tilde{A}_{11} & \tilde{A}_{12} + \tilde{B}_{11} \tilde{B}_{21}^* & \tilde{B}_{11} \\ 0 & -\tilde{A}_{22}^* & -\tilde{C}_{22}^* \\ \hline \tilde{C}_{11} & \tilde{C}_{12} + D \tilde{B}_{21}^* & D \end{array} \right] \end{aligned}$$

is an observable realization.

Equations (4.17), (4.18), (4.19) and parts (a) and (b) of the theorem have now been established and it remains for us to prove (c). Multiplying (C.33) on the left by $\tilde{A}_1^*(s)$ and using (C.21) and (C.22) we get

$$(C.34) \quad \tilde{A}_1^* \tilde{G} \tilde{A}_2^*(s) = \left[\begin{array}{cc|c} -\tilde{A}_{11}^* & \tilde{A}_{12} + \tilde{B}_{11} \tilde{B}_{21}^* + \tilde{C}_{11}^* (\tilde{C}_{12} + D \tilde{B}_{21}^*) & \tilde{C}_{11}^* D + \tilde{B}_{11} \\ 0 & -\tilde{A}_{22}^* & -\tilde{C}_{22}^* \\ \hline -\tilde{B}_{12}^* & \tilde{C}_{12} + D \tilde{B}_{21}^* & D \end{array} \right].$$

The minimality of the realization in (C.34) is established by first showing that it is observable. Using (C.21) and (C.22) we have that

$$(C.35) \quad \begin{aligned} &\left[\begin{array}{cc|c} sI + \tilde{A}_{11}^* & -\tilde{A}_{12} - \tilde{B}_{11} \tilde{B}_{21}^* - \tilde{C}_{11}^* (\tilde{C}_{12} + D \tilde{B}_{21}^*) & \\ 0 & sI + \tilde{A}_{22}^* & \\ \hline -\tilde{B}_{12}^* & \tilde{C}_{12} + D \tilde{B}_{21}^* & \end{array} \right] \\ &= \begin{bmatrix} I & 0 & -\tilde{C}_{11}^* \\ 0 & I & 0 \\ 0 & 0 & I \end{bmatrix} \left[\begin{array}{cc|c} sI - \tilde{A}_{11} & -\tilde{A}_{12} - \tilde{B}_{11} \tilde{B}_{21}^* & \\ 0 & sI + \tilde{A}_{22}^* & \\ \hline \tilde{C}_{11} & \tilde{C}_{12} + D \tilde{B}_{21}^* & \end{array} \right]. \end{aligned}$$

Since the $[A, C]$ pair in (C.33) is observable, and because the polynomial matrices on both sides of (C.35) have the same Smith form, the realization in (C.34) is also observable [19].

The minimality of the realization in (C.34) can now be established by showing that it is controllable. We note that

$$(C.36) \quad \left[\begin{array}{cc|c} sI + \tilde{A}_{11}^* & -\tilde{A}_{12} - \tilde{B}_{11}\tilde{B}_{21}^* - \tilde{C}_{11}^*(\tilde{C}_{12} + D\tilde{B}_{21}^*) & \tilde{C}_{11}^*D + \tilde{B}_{11} \\ 0 & sI + \tilde{A}_{22}^* & -\tilde{C}_{22}^* \end{array} \right] \\ = \left[\begin{array}{cc|c} sI + \tilde{A}_{11}^* & -\tilde{A}_{12} - \tilde{C}_{11}^*\tilde{C}_{12} & \tilde{C}_{11}^*D + \tilde{B}_{11} \\ 0 & sI - \tilde{A}_{22} & \tilde{B}_{21} \end{array} \right] \left[\begin{array}{ccc} I & 0 & 0 \\ 0 & I & 0 \\ 0 & -\tilde{B}_{21}^* & I \end{array} \right].$$

The first matrix on the right of (C.36) is the $[sI - A, B]$ pair of a controllable realization of $\tilde{A}_1^* \tilde{G}(s)$. To show this we assemble from (C.34), (C.31) and (C.32)

$$(C.37) \quad [\tilde{G} \quad \tilde{A}_1(s)] = \left[\begin{array}{cc|cc} \tilde{A}_{11} & \tilde{A}_{12} & \tilde{B}_{11} & \tilde{B}_{12} \\ 0 & \tilde{A}_{22} & \tilde{B}_{21} & 0 \\ \hline \tilde{C}_{11} & \tilde{C}_{12} & D & I \end{array} \right]$$

and thus

$$(C.38) \quad \tilde{A}_1^* \tilde{G}(s) = \left[\begin{array}{cc|c} -\tilde{A}_{11}^* & \tilde{A}_{12} + \tilde{C}_{11}^*\tilde{C}_{12} & \tilde{C}_{11}^*D + \tilde{B}_{11} \\ 0 & \tilde{A}_{22} & \tilde{B}_{21} \\ \hline -\tilde{B}_{12}^* & \tilde{C}_{12} & D \end{array} \right].$$

Hence

$$(C.39) \quad \tilde{A}_1^* \tilde{G} A_r(s) = \tilde{A}_1^*(s) \tilde{G}(s) \\ = \left[\begin{array}{cc|c} -\tilde{A}_{11}^* & \tilde{A}_{12} + \tilde{C}_{11}^*\tilde{C}_{12} & \tilde{C}_{11}^*D + \tilde{B}_{11} \\ 0 & \tilde{A}_{22} & \tilde{B}_{21} \\ \hline -\tilde{B}_{12}^* & \tilde{C}_{12} & D \end{array} \right] \left[\begin{array}{cc} \tilde{A}_{33} & \tilde{B}_{31} \\ -\tilde{B}_{31}^* & I \end{array} \right].$$

This cascade realization is system similar (in the sense defined in [19]) to the controllable realization in (C.18) and thus both realizations in (C.39) are controllable. This shows that the realization of (C.38) is controllable as required. The controllability and thus minimality of (C.34) now follows. This completes the proof of the minimality of (4.23). \square

Appendix D.

Proof of Theorem 4.2. Since the realizations for $A_1(s)$ and $A_2(s)$ in (4.30) are assumed to be minimal and balanced, equations (C.1) to (C.6) are again satisfied. The proof of part (a) follows by a direct calculation which is similar to the analysis contained in the proof of Theorem 4.1 and is consequently omitted.

To prove part (b), we will need a number of equations which can be deduced from various partitions of (4.33) to (4.36). Since \tilde{P} and \tilde{Q} in (4.33) and (4.34) are symmetric, we may introduce the notation

$$(D.1) \quad \tilde{P} = \begin{bmatrix} P_{11} & P_{12} & P_{13} \\ P_{12}^* & P_{22} & P_{23} \\ P_{13}^* & P_{23}^* & P_{33} \end{bmatrix}, \quad \tilde{Q} = \begin{bmatrix} Q_{11} & Q_{12} & Q_{13} \\ Q_{12}^* & Q_{22} & Q_{23} \\ Q_{13}^* & Q_{23}^* & Q_{33} \end{bmatrix}$$

where the partitioning is conformable with \tilde{A} of (4.32) in which \underline{A} is partitioned as in (4.31). The (2, 2) partition of the (1, 1) block of (4.33) gives

$$(D.2) \quad -A_{22}^* P_{22} - P_{22} A_{22} + C_{22}^* C_{22} = 0$$

and this together with (C.5) $\Rightarrow P_{22} = -I$. The (3, 2) partition of the (1, 1) block of (4.33) gives

$$(D.3) \quad \hat{A} P_{23}^* - P_{23}^* A_{22} + \hat{B} C_{22} = 0.$$

The (1, 1) partition of the (1, 1) block of (4.34) gives

$$(D.4) \quad -A_{11}Q_{11} - Q_{11}A_{11}^* + B_{12}B_{12}^* = 0$$

and this together with (C.1) $\Rightarrow Q_{11} = -I$. The (1, 3) partition of the (1, 1) block of (4.34) gives

$$(D.5) \quad -A_{11}Q_{13} + Q_{13}\hat{A} + B_{12}\hat{C} = 0.$$

Making use of (C.3), the (1, 1) partition of the (1, 2) block of (4.34) gives

$$(D.6) \quad B_{12}\hat{D} - B_{11} + Q_{12}C_{22}^* + Q_{13}\hat{B} = 0.$$

The (2, 1) partition of (4.35) gives

$$(D.7) \quad Q_{13}P_{23}^* = P_{12} + Q_{12}.$$

If we make use of (C.6), the (1, 2) partition of the (2, 1) block of (4.33) yields

$$(D.8) \quad -\hat{D}C_{22} - B_{12}^*P_{12} + C_{12} - \hat{C}P_{23}^* = 0.$$

Finally, the (1, 2) partition of the (1, 1) block of (4.34) together with (C.3) gives

$$(D.9) \quad -A_{11}Q_{12} + A_{12} + B_{11}B_{21}^* - Q_{12}A_{22}^* = 0.$$

By direct calculation we obtain

$$(D.10) \quad (G - A_1QA_2)(s) = \left[\begin{array}{cc|c} A_{11} & A_{12} & B_{11} \\ 0 & A_{22} & B_{21} \\ \hline C_{11} & C_{12} & D \end{array} \right] - \left[\begin{array}{ccc|c} A_{11} & B_{12}\hat{D}C_{22} & B_{12}\hat{C} & B_{12}\hat{D} \\ 0 & A_{22} & 0 & B_{21} \\ 0 & \hat{B}C_{22} & \hat{A} & \hat{B} \\ \hline C_{11} & \hat{D}C_{22} & \hat{C} & \hat{D} \end{array} \right].$$

The rest of the proof is based on detailed manipulations of the state-space realizations of $A_1QA_2(s)$ in (D.10). First, we introduce the change of basis

$$(D.11) \quad T = \left[\begin{array}{ccc} I & 0 & Q_{13} \\ 0 & I & 0 \\ 0 & 0 & I \end{array} \right]$$

and this together with (D.5) yields

$$(D.12) \quad A_1QA_2(s) = \left[\begin{array}{ccc|c} A_{11} & (B_{12}\hat{D} + Q_{13}\hat{B})C_{22} & 0 & B_{12}\hat{D} + Q_{13}\hat{B} \\ 0 & A_{22} & 0 & B_{21} \\ 0 & \hat{B}C_{22} & \hat{A} & \hat{B} \\ \hline C_{11} & \hat{D}C_{22} & \hat{C} - C_{11}Q_{13} & \hat{D} \end{array} \right].$$

Next, the coordinate transformation

$$(D.13) \quad T = \left[\begin{array}{ccc} I & 0 & 0 \\ 0 & I & 0 \\ 0 & -P_{23}^* & I \end{array} \right]$$

together with (D.3) gives

$$(D.14) \quad A_1QA_2(s) = \left[\begin{array}{ccc|c} A_{11} & (B_{12}\hat{D} + Q_{13}\hat{B})C_{22} & 0 & B_{12}\hat{D} + Q_{13}\hat{B} \\ 0 & A_{22} & 0 & B_{21} \\ 0 & 0 & \hat{A} & \hat{B} - P_{23}^*B_{21} \\ \hline C_{11} & \hat{D}C_{22} + \hat{C}P_{23}^* - C_{11}Q_{13}P_{23}^* & \hat{C} - C_{11}Q_{13} & \hat{D} \end{array} \right].$$

A third change of basis

$$(D.15) \quad T = \begin{bmatrix} I & -Q_{12} & 0 \\ 0 & I & 0 \\ 0 & 0 & I \end{bmatrix}$$

leads to

(D.16)

$$A_1 Q A_2(s)$$

$$= \left[\begin{array}{ccc|c} A_{11} & (B_{12}\hat{D} + Q_{13}\hat{B})C_{22} - Q_{12}A_{22} + A_{11}Q_{12} & 0 & B_{12}\hat{D} + Q_{13}\hat{B} - Q_{12}B_{21} \\ 0 & A_{22} & 0 & B_{21} \\ 0 & 0 & \hat{A} & \hat{B} - P_{23}^*B_{21} \\ \hline C_{11} & \hat{D}C_{22} + \hat{C}P_{23}^* - C_{11}Q_{13}P_{23}^* + C_{11}Q_{12} & \hat{C} - C_{11}Q_{13} & \hat{D} \end{array} \right].$$

Substituting (D.6) C_{22} and (D.9) into the (1, 2) block of the A -matrix in the above realization together with (C.5) and (C.6) gives

$$(D.17) \quad (B_{12}\hat{D} + Q_{13}\hat{B})C_{22} - Q_{12}A_{22} + A_{11}Q_{12} = A_{12}.$$

Substituting (D.6) and (C.6) into the (1, 1) block of the B -matrix in (D.16), we obtain

$$(D.18) \quad B_{12}\hat{D} + Q_{13}\hat{B} - Q_{12}B_{21} = B_{11}.$$

Finally, (D.7), (D.8) and (C.3) will verify that

$$(D.19) \quad \hat{D}C_{22} + \hat{C}P_{23}^* - C_{11}Q_{13}P_{23}^* + C_{11}Q_{12} = C_{12}.$$

Thus

$$(D.20) \quad A_1 Q A_2(s) = \left[\begin{array}{ccc|c} A_{11} & A_{12} & 0 & B_{11} \\ 0 & A_{22} & 0 & B_{21} \\ 0 & 0 & \hat{A} & \hat{B} - P_{23}^*B_{21} \\ \hline C_{11} & C_{12} & \hat{C} - C_{11}Q_{13} & \hat{D} \end{array} \right].$$

Consequently

$$(D.21) \quad (G - A_1 Q A_2)(s) = \left[\begin{array}{c|c} \hat{A} & \hat{B} - P_{23}^*B_{21} \\ \hline -\hat{C} - C_{11}Q_{13} & \hat{D} - \hat{D} \end{array} \right]$$

and this proves (b)(i). Part (b)(ii) is obvious. \square

Appendix E.

Proof of Theorem 4.3. The equations describing the closed loop of Fig. 1 are

$$\dot{x} = Ax + B_1u_1 + B_2u_2,$$

$$y_1 = C_1x + D_{11}u_1 + D_{12}u_2,$$

$$y_2 = C_2x + D_{21}u_1 + D_{22}u_2,$$

$$\dot{\hat{x}} = \hat{A}\hat{x} + \hat{B}y_2,$$

$$u_2 = -(\hat{C}\hat{x} + \hat{D}y_2).$$

Eliminating the variables u_2 and y_2 leads to the following state-space model for the closed loop

$$(E.1) \quad \begin{bmatrix} \dot{x} \\ \dot{\hat{x}} \end{bmatrix} = \begin{bmatrix} A - B_2 \hat{D} M C_2 & -B_2 [I - \hat{D} M D_{22}] \hat{C} \\ \hat{B} M C_2 & \hat{A} - \hat{B} M D_{22} \hat{C} \end{bmatrix} \begin{bmatrix} x \\ \hat{x} \end{bmatrix} + \begin{bmatrix} B_1 - B_2 \hat{D} M D_{21} \\ \hat{B} M D_{22} \end{bmatrix} [u_1],$$

$$(E.2) \quad [y_1] = [C_1 - D_{12} \hat{D} M C_2 \quad -D_{12} [I - \hat{D} M D_{22}] \hat{C}] \begin{bmatrix} x \\ \hat{x} \end{bmatrix} + [D_{11} - D_{12} \hat{D} M D_{21}] [u_1]$$

in which

$$(E.3) \quad M := (I + D_{22} \hat{D})^{-1}.$$

If s_0 is an unobservable mode of the closed loop state-space model (E.1)–(E.3), then there exists a vector $[w_1^* w_2^*]^* \neq 0$ such that

$$(E.4) \quad \begin{bmatrix} s_0 I - A + B_2 \hat{D} M C_2 & B_2 [I - \hat{D} M D_{22}] \hat{C} \\ -\hat{B} M C_2 & s_0 I - \hat{A} + \hat{B} M D_{22} \hat{C} \\ C_1 - D_{12} \hat{D} M C_2 & -D_{12} [I - \hat{D} M D_{22}] \hat{C} \end{bmatrix} \begin{bmatrix} w_1 \\ w_2 \end{bmatrix} = 0.$$

Defining

$$z_2 := -\hat{D} M C_2 w_1 - [I - \hat{D} M D_{22}] \hat{C} w_2$$

we have from (E.4) that

$$(E.5) \quad \left[\begin{array}{c|c} s_0 I - A & -B_2 \\ \hline C_1 & D_{12} \end{array} \right] \begin{bmatrix} w_1 \\ z_2 \end{bmatrix} = 0.$$

The proof of the (a) part is concluded by establishing that $[w_1^* w_2^*]^* \neq 0 \Rightarrow [w_1^* z_2^*]^* \neq 0$. Suppose for contradiction that $[w_1^* z_2^*]^* = 0$. This implies that

$$(E.6) \quad \begin{aligned} [I - \hat{D} M D_{22}] \hat{C} w_2 &= 0 \\ \Leftrightarrow (I + \hat{D} D_{12})^{-1} \hat{C} w_2 &= 0 \\ \Leftrightarrow \hat{C} w_2 &= 0. \end{aligned}$$

We also have from (E.4) that

$$(E.7) \quad (s_0 I - \hat{A}) w_2 = 0.$$

Equations (E.6) and (E.7) taken together contradict the assumed minimality of the realization in (4.47) which proves the (a) condition. The (b) part may be established by a parallel sequence of arguments. \square

Appendix F.

Proof of Theorem 4.5. We begin by pointing out that

$$(F.1) \quad (4.66) \Rightarrow I(H_{22}; s_0) = -\frac{1}{\pi} \int_{-\infty}^{\infty} [\ln |\det D(j\omega)| \operatorname{Re}(s_0) / \{|j\omega - s_0|^2\}] d\omega.$$

Next, by invoking the system of J -unitary equations (4.63) to (4.65), one may verify that (note that if $G(s)$ is J -unitary, so also is $G^*(s)$)

$$(F.2) \quad I - E(s)E^*(s) = (D + CU)^{-*} \{I - U^* U\} (D + CU)^{-1}(s)$$

so that

$$(F.3) \quad \begin{aligned} \det(I - E^* E)(s) &= \{\det(D + CU)\}^{-2} \det(I - U^* U) \\ &= \{\det(D(I + D^{-1}CU))\}^{-2} \det(I - U^* U) \\ &= \{\det(D)\}^{-2} \{\det(I + H_{22}U)\}^{-2} \det(I - U^* U). \end{aligned}$$

Therefore

$$(F.4) \quad \begin{aligned} I(E; s_0) &= -\frac{1}{2\pi} \int_{-\infty}^{\infty} \{-2 \ln |\det(D)| - 2 \ln |\det(I + H_{22}U)| \\ &\quad + \ln |\det(I + U^*U)|\} \operatorname{Re}(s_0)/\{|j\omega - s_0|^2\} d\omega \\ &= I(H_{22}; s_0) + I(U; s_0) \end{aligned}$$

$$(F.5) \quad + \frac{1}{\pi} \int_{-\infty}^{\infty} [\ln |\det(I + H_{22}U)| \operatorname{Re}(s_0)/\{|j\omega - s_0|^2\}] d\omega.$$

Since $s_0 \in \mathbb{C}_-$, $H_{22}(s) \in RH_-^\infty$ and $U(s) \in RH_-^\infty$ we have by Poisson's integral formula

$$(F.6) \quad \frac{1}{\pi} \int_{-\infty}^{\infty} [\ln |\det(I + H_{22}U)| \operatorname{Re}(s_0)/\{|j\omega - s_0|^2\}] d\omega = \ln |\det[I + H_{22}(s_0)U(s_0)]|.$$

Hence

$$(F.7) \quad I(E; s_0) = I(H_{22}; s_0) + I(U; s_0) + \ln |\det[I + H_{22}(s_0)U(s_0)]|$$

which proves the first part.

We now need to prove that

$$(F.8) \quad I(U; s_0) + \ln |\det(I + H_{22}(s_0)U(s_0))|$$

attains a minimum at $U_0 = -H_{22}^*(s_0)$. This is obvious when $H_{22}(s_0) = 0$ since $I(U; s_0) \geq 0$ and $I(0; s_0) = 0$. Let us now suppose that $H_{22}(s_0) \neq 0$ and consider the *constant* linear fractional map:

$$(F.9) \quad \Theta_0(U(s)) = (\Theta_{11}U(s) + \Theta_{12})(\Theta_{21}U(s) + \Theta_{22})^{-1}$$

where the Θ_{ij} are sub-blocks of the J -unitary matrix Θ given by

$$(F.10) \quad \Theta = \begin{bmatrix} \Theta_{11} & \Theta_{12} \\ \Theta_{21} & \Theta_{22} \end{bmatrix} = \begin{bmatrix} (I_m - X_0^* X_0)^{-1/2} & -(I_m - X_0^* X_0)^{-1/2} X_0^* \\ -(I_n - X_0 X_0^*)^{1/2} X_0 & (I_n - X_0 X_0^*)^{1/2} \end{bmatrix}$$

where

$$X_0 := -H_{22}(s_0) \in \mathbb{C}^{n \times m}.$$

By applying (F.7) to (F.9) we get

$$(F.11) \quad I(\Theta_0; s_0) = I(X_0; s_0) + I(U; s_0) + \ln |\det(I + H_{22}(s_0)U(s_0))|$$

and substituting (F.11) into (F.7) we get

$$(F.12) \quad I(E; s_0) = I(H_{22}; s_0) + I(\Theta_0; s_0) - I(X_0; s_0).$$

Since $I(\Theta_0(U(s)); s_0) \geq 0$ and $I(\Theta_0(X_0^*); s_0) = 0$ we have that the minimum value of $I(E; s_0)$ is given by

$$\begin{aligned} I(E; s_0) &= I(H_{22}; s_0) - I(X_0; s_0) \\ &= I(H_{22}; s_0) + \frac{1}{2} \ln |\det(I - H_{22}(s_0)H_{22}^*(s_0))| \quad (\text{by (F.1)}) \end{aligned}$$

which completes the proof. \square

Acknowledgment. The authors would like to thank Keith Glover for several helpful suggestions and discussions.

REFERENCES

- [1] B. D. O. ANDERSON AND S. VONGPANITLERD, *Network Analysis and Synthesis. A Modern Approach*, Prentice-Hall, Englewood Cliffs, NJ, 1972.
- [2] D. Z. AROV AND M. G. KREIN, *On the evaluation of entropy functionals and their minima in generalized extension problems*, Acta Sci. Math., 45 (1983), pp. 33–50 (in Russian).
- [3] B. C. CHANG AND B. PEARSON, *Optimal disturbance rejection in linear multivariable systems*, IEEE Trans. Automat. Control, AC-29 (1984), pp. 880–887.
- [4] C. A. DESOER, R. W. LIU, J. MURRAY AND R. SAEKS, *Feedback system design: The fractional representation approach to analysis and synthesis*, IEEE Trans. Automat. Control, AC-25 (1980), pp. 399–412.
- [5] J. C. DOYLE, *Analysis of feedback systems with structured uncertainty*, Proc. IEE, 12a (1982), pp. 242–250.
- [6] J. DOYLE ET AL., *Advances in multivariable control*, ONR/Honeywell Workshop, Office of Naval Research, Washington, DC, 1984.
- [7] B. A. FRANCIS AND G. ZAMES, *On optimal sensitivity theory for SISO feedback systems*, IEEE Trans. Automat. Control, AC-29 (1984), pp. 9–16.
- [8] B. A. FRANCIS, J. W. HELTON AND G. ZAMES, *H^∞ -optimal feedback controllers for linear multivariable systems*, IEEE Trans. Automat. Control, AC-29 (1984), pp. 888–900.
- [9] B. A. FRANCIS, *Notes on H^∞ -optimal linear feedback systems*, Lecture notes given in Linköping University, 1983.
- [10] B. A. FRANCIS AND J. C. DOYLE, *Linear control theory with an H^∞ optimality criterion*, this Journal, 25 (1987), pp. 815–844.
- [11] Y. K. FOO AND I. POSTLETHWAITE, *An H^∞ -minimax approach to the design of robust control systems*, Systems Control Lett., 5 (1984), pp. 81–82.
- [12] K. GLOVER, *All optimal Hankel-norm approximations of linear multivariable systems and their L^∞ -error bounds*, Internat. J. Control, 39 (1984), pp. 1115–1193.
- [13] ———, *Robust stabilization of linear multivariable systems: Relations to approximation*, Internat. J. Control, 43 (1986), pp. 741–766.
- [14] E. A. JONCHEERE AND L. M. SILVERMAN, *A new set of invariants for linear systems—Application to reduced order compensation design*, IEEE Trans. Automat. Control, AC-28 (1983), pp. 953–964.
- [15] H. KIMURA, *Robust stabilization for a class of transfer functions*, IEEE Trans. Automat. Control, AC-29 (1984), pp. 788–793.
- [16] D. J. N. LIMEBEER AND G. HALIKIAS, *An analysis of the pole-zero cancellations in H^∞ -optimal control problems of the second kind*, Imperial College report; this Journal, to appear.
- [17] Z. NEHARI, *On bounded bilinear forms*, Ann. of Math., 65 (1957), pp. 153–162.
- [18] C. N. NETT, C. A. JACOBSON AND M. J. BALAS, *A connection between state-space and doubly coprime fractional representations*, IEEE Trans. Automat. Control, AC-29 (1984), pp. 831–832.
- [19] H. H. ROSENBRACK, *State-space and Multivariable Theory*, John Wiley, New York, 1970.
- [20] M. G. SAFONOV, *Exact calculation of the multivariable structured singular value stability margin*, IEEE Conference on Decision and Control, Las Vegas, NV (1984), pp. 1224–1225.
- [21] M. G. SAFONOV AND M. S. VERMA, *L^∞ -sensitivity optimization and Hankel approximation*, IEEE Trans. Automat. Control, AC-30 (1985), pp. 279–280.
- [22] M. G. SAFONOV, E. A. JONCHEERE, M. VERMA AND D. J. N. LIMEBEER, *Synthesis of positive real multivariable feedback systems*, Internat. J. Control, 45 (1987), pp. 817–842.
- [23] P. M. VAN DOOREN AND P. DEWILDE, *Minimal cascade factorization of real and complex rational matrices*, IEEE Trans. Circuits and Systems, CAS-28, (1981), pp. 390–400.
- [24] P. M. VAN DOOREN, *The generalized eigenstructure problem in linear system theory*, IEEE Trans. Automat. Control, AC-26 (1981), pp. 111–129.
- [25] M. VERMA AND E. JONCHEERE, *L^∞ -compensation with mixed sensitivity as a broadband matching problem*, System Control Lett. 4, (1984), pp. 125–130.
- [26] D. C. YOULA, K. JABR AND J. J. BONGIORNO, *Modern Wiener-Hopf design of optimal controllers, Part II: The multivariable case*, IEEE Trans. Automat. Control, AC-21 (1976), pp. 319–338.
- [27] G. ZAMES, *Feedback and optimal sensitivity: Model reference transformations, multiplicative seminorms, and approximate inverses*, IEEE Trans. Automat. Control, AC-26 (1981), pp. 301–320.
- [28] G. ZAMES AND B. A. FRANCIS, *Feedback, minimax sensitivity, and optimal robustness*, IEEE Trans. Automat. Control, AC-28 (1983), pp. 585–601.
- [29] B. D. O. ANDERSON AND A. LINNEMANN, *Control of decentralized systems with distributed controller complexity*, Report AN372, Australian National University, Canberra, 1986.

AN OPTIMAL STOCHASTIC PRODUCTION PLANNING PROBLEM WITH RANDOMLY FLUCTUATING DEMAND*

W. H. FLEMING†, S. P. SETHI‡ AND H. M. SONER§

Abstract. This paper considers an infinite horizon stochastic production planning problem with demand assumed to be a continuous-time Markov chain. The problems with control (production) and state (inventory) constraints are treated. It is shown that a unique optimal feedback solution exists, after first showing that convex viscosity solutions to the associated dynamic programming equation are continuously differentiable.

Key words. production planning, stochastic optimal control, viscosity solutions

AMS(MOS) subject classifications. 93E20, 35J65, 35K60, 60J60

Introduction. Thompson and Sethi [13] consider a production-inventory model which determines production rates over time to minimize an integral representing a discounted quadratic loss function. The model is solved both with and without nonnegative production constraints. It is shown that there exists a turnpike level of inventory, to which the optimal inventory levels approach monotonically over time. The model was generalized by Sethi and Thompson [11] and Bensoussan et al. [1] by incorporating an additive white noise term in the dynamics of the inventory process.

In this paper we consider an analogue of the Thompson-Sethi model, in which the demand rate $z(t)$ is a finite state Markov chain. A similar, but technically more complicated, analysis applies if $z(t)$ is a jump Markov process or a reflected diffusion subject to bounds $0 < z_0 \leq z(t) \leq z_1 < \infty$ (see [7]). We denote by $y(t)$, $p(t)$ the inventory level and the production rate. Production is the control variable, subject to the constraint $p(t) \geq 0$. In § 5 we impose the state constraint $y(t) \geq y_{\min}$ on the inventory level.

The control objective is to minimize an expected discounted cost of the form (4.1), which involves convex holding or shortage costs $h(y)$ and productions costs $c(p)$. The value (or minimum cost) $v(y, z)$ defined in (1.4) for initial data $y(0) = y$, $z(0) = z$ obeys the dynamic programming equation (1.6). Special features of the model allow us to show that $v(\cdot, z)$ is convex and that the quantity $\partial v / \partial y$ which appears in the dynamic programming equation exists and is continuous. The optimal feedback production law $p^*(y, z)$ is expressed as a function of $\partial v / \partial y$ by formula (4.3). We do not know that $p^*(\cdot, z)$ is Lipschitz continuous. However, since $p^*(\cdot, z)$ is a nonincreasing function of y , the differential equation

$$\frac{dy^*}{dt} = p^*(y^*(t), z(t)) - z(t), \quad y^*(0) = y,$$

has a unique solution for the optimal inventory level $y^*(t)$.

* Received by the editors October 28, 1985; accepted for publication (in revised form) December 12, 1986.

† Division of Applied Mathematics, Brown University, Providence, Rhode Island 02912. The work of this author was supported by the National Science Foundation through grant MCS-812940, by the Office of Naval Research under grant N00014-83-K-0542, and by the Air Force Office of Scientific Research under grant AFOSR-810116-C.

‡ Faculty of Management Studies, University of Toronto, Toronto, Ontario, Canada M5S 1V4. The work of this author was supported by the Connaught Senior Research Fellowship and by the Natural Sciences and Engineering Research Council of Canada through grant A4619.

§ Department of Mathematics, Carnegie-Mellon University, Pittsburgh, Pennsylvania 15213. The work of this author was partly supported by the Air Force Office of Scientific Research under grant AFOSR-810116-C, and partly by the Institute for Mathematics and Its Applications with funds provided by the National Science Foundation and the Office of Naval Research.

We begin in § 1 by formulating a more general class of discounted optimal control problems with Markov chain parameter $z(t)$. It is elementary that the value function is convex in the state y , provided the state dynamics are linear in state y and control p and the cost criterion is convex jointly in (y, p) ; see Lemma 1.1. In § 2 we find that the value $v(y, z)$ is the unique solution to the dynamic programming equation, and that the gradient $\nabla_y v$ is continuous. For the continuity of $\nabla_y v$, an additional assumption (2.2) on the Hamiltonian appearing in the dynamic programming equation is needed.

In § 3 we discuss optimal controls, both from the viewpoint of dynamic programming and the theory of controlled piecewise deterministic processes. Under a strict convexity condition (3.1) on the cost criterion, there is a continuous optimal control policy $p^*(y, z)$ and the corresponding optimal control process $p^*(t)$ is unique.

These results are applied to the production planning model in § 4. Finally, in § 5 the analysis is modified to deal with a state-space constraint $y(t) \geq y_{\min}$. Such a constraint imposes an inequality (5.2) on $\partial v / \partial y$ at y_{\min} .

The production planning model considered here does not impose any upper bound on the production rate. A more interesting extension of the problem involves production processes, which are bounded from above by a stochastic process representing the capacity of the production system. The capacity process over time may be modelled as a jump process or a piecewise deterministic process [5], [14]. Moreover, there may be several different products competing for a variety of scarce capacities. This is a problem faced by flexible manufacturing systems [10], upon which the methods developed in this paper have some bearing.

1. Discounted optimal control problems with Markov chain parameters. Let us begin with a model of the following rather general form, and then specialize. Let $y(t), p(t), z(t)$ denote, respectively, state, control and parameter processes for $t \geq 0$. We assume that $y(t) \in R^n, p(t) \in K, z(t) \in Z$ for each $t \geq 0$, where R^n is n -dimensional Euclidean space, K is a closed convex subset of some Euclidean space, $0 \in K$ and Z is a finite set. The parameter process $z(t)$ is a finite state continuous time Markov chain, defined on some underlying probability space (Ω, \mathcal{F}, P) with jumping rate $q_{zz'}$ from state z to state z' . The associated generator L of the Markov chain $z(t)$ has the form

$$(1.1) \quad Lg(z) = \sum_{z' \neq z} q_{zz'} [g(z') - g(z)].$$

In the general formulation, the state dynamics are $dy(t) = f(y(t), p(t), z(t))dt, t \geq 0$. Actually we shall consider only f of the special form (1.5) below. A control process $P = \{p(t, \omega), t \geq 0, \omega \in \Omega\}$ will be called *admissible* if: (i) P is adapted to $\mathcal{F}_t = \sigma(z(s): 0 \leq s \leq t)$; (ii) $\sup \{|p(t, \omega)|: t \geq 0, \omega \in \Omega\} < \infty$; (iii) $p(t, \omega) \in K$ for all $t \geq 0$ and $\omega \in \Omega$ (in whatever follows the ω -dependence will be suppressed). Let \mathcal{A} denote the set of admissible control processes.

We consider a cost criterion $I(y, p)$ about which the following assumptions are made:

- $$(1.2) \quad \begin{aligned} & \text{(a) } I(\cdot, \cdot) \text{ is convex on } R^n \times K, \\ & \text{(b) } -C \leq I(y, p) \leq C_N(1 + |y|^m) \text{ whenever } |p| \leq N, \\ & \text{(c) } \lim_{|p| \rightarrow \infty} I(y, p)|p|^{-1} = +\infty \text{ if } K \text{ is unbounded} \end{aligned}$$

where m, C are fixed constants and C_N may depend on N . Let $\alpha > 0$ denote the

discount rate. For every $P \in \mathcal{A}$ and $y = y(0)$, $z = z(0)$, let

$$(1.3) \quad J(y, z, P) = E \int_0^\infty e^{-\alpha t} l(y(t), p(t)) dt.$$

The value function is

$$(1.4) \quad v(y, z) = \inf \{J(y, z, P) : P \in \mathcal{A}\}.$$

From now on, let us assume that the dynamics have the following special form:

$$(1.5) \quad dy(t) = [B(z(t))p(t) + C(z(t))]dt, \quad t > 0.$$

(In the simple production planning model to be considered in § 4, both $y(t)$ and $p(t)$ are scalar valued and $B(z) = 1$, $c(z) = -z$.) Instead of (1.5) we could take equally well $dy(t) = [A(z(t))y(t) + B(z(t))p(t) + C(z(t))]dt$ provided that the eigenvalues of $A(z)$ have strictly negative real parts.

LEMMA 1.1. *For each $z \in Z$, $v(\cdot, z)$ is convex on R^n and $-C \leq v(y, z) \leq C(1 + |y|^m)$ for some $C > 0$.*

This lemma is easily proved, after observing that $J(\cdot, z, \cdot)$ is convex jointly in (y, P) for each $z \in Z$ and $p(t) \equiv 0$ is an admissible control process.

The dynamic programming equation associated with this optimal stochastic control problem is as follows:

$$(1.6) \quad \alpha v(y, z) = H(y, z, \nabla v(y, z)) + Lv(y, z), \quad y \in R^n, \quad z \in Z$$

where ∇v is the gradient in y

$$Lv(y, z) = [Lv(y, \cdot)](z) = \sum_{z' \neq z} q_{zz'} [v(y, z') - v(y, z)]$$

and for $y, z, r \in R^n \times Z \times R^n$

$$(1.7) \quad H(y, z, r) = \inf_{P \in K} [l(y, p) + (B(z)p + C(z)) \cdot r].$$

Since Z is a finite set, (1.6) is a system of nonlinear first order PDE's in y , coupled through the zeroth order term Lv .

We are concerned with solutions to (1.6) belonging to the following space D_0 .

DEFINITION 1.2. We say that a real-valued function v with domain $R^n \times Z$ is in D_0 if

- (i) $v(\cdot, z)$ is convex on R^n for each $z \in Z$,
- (ii) $-C \leq v(y, z) \leq C(1 + |y|^\beta)$, for suitable C and β (depending on v),
- (iii) The gradient $\nabla v(y, z)$ is continuous.

The following "verification theorem" is standard, but for completeness we indicate the proof.

THEOREM 1.3. *Let $v \in D_0$ satisfy the dynamic programming equation (1.6). Then*

- (a) $v(y, z) \leq J(y, z, P)$ for all $P \in \mathcal{A}$,
- (b) Suppose that there are $P^* \in \mathcal{A}$, $y^*(t)$ that satisfies (1.5) with $y^*(0) = y$, $r^*(t) = \nabla v(y^*(t), z(t))$, and

$$H(y^*(t), z(t), r^*(t)) = l(y^*(t), p^*(t)) + (B(z(t))p^*(t) + C(z(t)) \cdot r^*(t))$$

a.e. in t with probability 1. Then

$$v(y, z) = J(y, z, P^*).$$

Proof. For $T < \infty$, we have the usual dynamic programming relation

$$(1.8) \quad v(y, z) \leq E \int_0^T e^{-\alpha t} l(y(t), p(t)) dt + e^{-\alpha T} v(y(T), z(T)).$$

Since any admissible P is bounded, $|y(t)| \leq c_1 t + c_2$ for suitable constants c_1, c_2 . We obtain (a) as $T \uparrow \infty$, using the polynomial growth condition (ii) in Definition 1.2. In part (b), inequality (1.8) becomes an equality. \square

In the next section we shall show that, under an additional condition (2.2) on H , the value function in fact belongs to D_0 .

2. Viscosity solutions to the dynamic programming equation. The definition of viscosity solution used here is a straightforward generalization of the original definition given by M. G. Crandall and P.-L. Lions [4]. See also [3], [9] for more information.

Let v be a continuous function on $R^n \times Z$. For each (y, z) we define convex subsets $D_y^\pm v(y, z)$ of R^n , as follows:

$$D_y^+ v(y, z) = \{r \in R^n : \limsup_{h \rightarrow 0} (v(y+h, z) - v(y, z) - r \cdot h) |h|^{-1} \leq 0\},$$

$$D_y^- v(y, z) = \{r \in R^n : \liminf_{h \rightarrow 0} (v(y+h, z) - v(y, z) - r \cdot h) |h|^{-1} \geq 0\}.$$

We say that any continuous function v is a *viscosity solution* of (1.6), if for each y, z : (i) $\alpha v(y, z) \leq H(y, z, r) + Lv(y, z)$ for all $r \in D_y^+ v(y, z)$, and (ii) $\alpha v(y, z) \geq H(y, z, r) + Lv(y, z)$ for all $r \in D_y^- v(y, z)$.

Remark 2.1. v is differentiable in the y -direction at (y, z) if and only if $D_y^+ v(y, z)$ and $D_y^- v(y, z)$ are both singletons. In this case, the singleton is the gradient $\nabla v(y, z)$. Moreover, if v is convex in y , then $D_y^+ v(y, z)$ is empty unless v is differentiable there and $D_y^- v(y, z)$ coincides with the set of subdifferentials in the sense of convex analysis,

$$(2.1) \quad D_y^- v(y, z) = \overline{\text{co}} \Gamma(y, z)$$

where

$$\Gamma(y, z) = \{r = \lim_{n \rightarrow \infty} \nabla v(y_n, z) : y_n \rightarrow y \text{ as } n \rightarrow \infty \text{ and } v(\cdot, z) \text{ is differentiable at } y_n\}$$

and where $\overline{\text{co}} \Gamma$ denotes the convex closure of Γ . (See [2, Thm. 251, pp. 63].)

We now make the additional assumption that the Hamiltonian $H(y, z, \cdot)$ is constant on no nontrivial convex set:

$$(2.2) \quad \text{If } H(y, z, \lambda r_1 + (1-\lambda)r_2) = \text{constant in } \lambda \text{ for } 0 \leq \lambda \leq 1, \text{ then } r_1 = r_2.$$

THEOREM 2.2. *Let (2.2) hold and let v be a viscosity solution to the dynamic programming equation (1.6). If, in addition $v(\cdot, z)$ is convex for each z , then $\nabla v(y, z)$ exists for all (y, z) and $\nabla v(\cdot, z)$ is continuous on R^n .*

Proof. By Remark 2.1 and formula (2.1) it suffices to show that $D_y^- v(y, z)$ is a singleton. If $v(\cdot, z)$ is differentiable at y_n , then (1.7) holds at (y_n, z) :

$$\alpha v(y_n, z) - H(y_n, z, \nabla v(y_n, z)) - Lv(y_n, z) = 0.$$

We then obtain, taking $y_n \rightarrow y$ as $n \rightarrow \infty$,

$$\alpha v(y, z) - H(y, z, r) - Lv(y, z) = 0 \quad \text{for } r \in \Gamma(y, z).$$

Moreover, $H(y, z, \cdot)$ is concave, and hence by (2.1)

$$\alpha v(y, z) - H(y, z, r) - Lv(y, z) \leq 0 \quad \text{for } r \in D_y^- v(y, z).$$

However, the viscosity property implies the opposite inequality, and hence

$$\alpha v(y, z) - H(y, z, r) - Lv(y, z) = 0 \quad \text{for } r \in D_y^- v(y, z).$$

Thus, for fixed (y, z) , $H(y, z, \cdot)$ is constant on the convex set $D_y^- v(y, z)$. By (2.2), $D_y^- v(y, z)$ is a singleton, which proves Theorem 2.2. \square

The remainder of this section consists of a proof that the value function $v(y, z)$ defined by (1.4) is a viscosity solution to (1.6): the argument is rather standard.

LEMMA 2.3. *Suppose K is bounded. Then v is a viscosity solution to (1.6).*

Proof. It is a direct modification of Theorem 1.1 of [12]. \square

THEOREM 2.4. *v is a viscosity solution to (1.6).*

Proof. Let $K_m = \{p \in K : |p| \leq m\}$ and v_m be the optimal value function of the corresponding control problem. Then, v_m is a viscosity solution to (1.7) with $H(y, z, r)$ replaced with $H_m(y, z, r) = \min_{p \in K_m} \{r \cdot [B(z)p + c(z)] + l(y, p)\}$. Also, v_m converges to v uniformly on bounded subsets of R^n as m tends to infinity.

Take $r \in D_y^- v(y_0, z_0)$. For each $\varepsilon > 0$, let $\varphi^\varepsilon(y, z) = \varphi(y, z) - \varepsilon(y - y_0)^2$ where $\varphi(y, z_0) = v(y_0, z_0) + r(y - z_0)$, and $\varphi(y, z) = v(y, z)$ if $z \neq z_0$. Since v is convex, the map $y \mapsto v(y, z_0) - \varphi^\varepsilon(y, z)$ has a strict maximum at y_0 . Therefore, $y \mapsto v_m(y, z_0) - \varphi^\varepsilon(y, z_0)$ has a maximum at y_m , and y_m converges to y_0 as m tends to infinity. But this implies that $\nabla \varphi^\varepsilon(y_m, z_0) \in D_y^- v_m(y_m, z_0)$ and the viscosity property of v_m implies that

$$\alpha v_m(y_m, z_0) \geq H_m(y_m, z_0, r - 2\varepsilon(y_m - y_0)) + Lv_m(y_m, z_0).$$

Now send m to infinity, and then ε to zero in the above inequality, to obtain

$$\alpha v(y_0, z_0) \geq H(y_0, z_0, r) + Lv(y_0, z_0) \quad \text{for all } r \in D_y^- v(y_0, z_0).$$

The reversed inequality for $r \in D_y^+ v(y, z)$ is proved by a similar argument. \square

3. Optimal controls. Let us now assume the following stricter form of convexity for the cost criterion l than what was assumed in (1.2)(a)

$$(3.1) \quad l(\lambda y_1 + (1 - \lambda)y_2, \lambda p_1 + (1 - \lambda)p_2) \\ = \lambda l(y_1, p_1) + (1 - \lambda)l(y_2, p_2) \quad \text{for some } 0 < \lambda < 1 \text{ implies } p_1 = p_2.$$

For example, for the production planning problem that will be considered in § 4, $l(y, p) = c(p) + h(y)$ and (3.1) holds if h is convex on R^n and l is strictly convex on K . Assumption (3.1) also holds, if $l(\cdot, \cdot)$ is convex and the second derivative of l in p exists and is positive at each (y, p) .

Condition (3.1) implies, in particular, strict convexity of $l(y, \cdot)$, by taking $y = y_1 = y_2$. This fact together with the superlinear growth condition (1.2)(c) imply that the minimum in (1.7) is attained at a unique $P = \Phi(y, z, r)$. Moreover, Φ is continuous on $R^n \times Z \times R^n$. Consider the control policy

$$(3.2) \quad p^*(y, z) = \Phi(y, z, \nabla v(y, z)),$$

where v is the value function. By Theorems 2.2 and 2.4, p^* is continuous. The differential equation,

$$(3.3) \quad dy(t) = [B(z(t))p^*(y(t), z(t)) + C(z(t))] dt,$$

has locally a solution $y^*(t)$. Let us assume the following:

$$(3.4) \quad \text{There exists a bounded solution } y^*(t) \text{ to (3.3) for } t \geq 0.$$

In § 4 we shall verify (3.4) in the production planning example. The control process $P^* = \{p^*(t); t \geq 0\}$, where

$$(3.5) \quad p^*(t) = p^*(y^*(t), z(t))$$

is admissible, by the superlinear growth condition (1.2)(c) and is optimal, by the verification Theorem 1.3(b). Also, a straightforward application of (3.1) yields that P^* is unique. This implies, in particular, uniqueness of $y^*(t)$. We sum these results into the following proposition.

PROPOSITION 3.1. *Let (3.1) and (3.4) hold. Suppose $P \in \mathcal{A}$ and $J(y, z, P) = v(y, z)$. Then with probability 1, $P(t) = P^*(t)$ for almost all $t > 0$. In particular, there is a unique solution to (3.3).*

Proof. For $0 < \lambda \leq 1$, let $P^\lambda = \lambda P + (1 - \lambda)P^*$, $y^\lambda = \lambda y + (1 - \lambda)y^*$, where $y(t)$ is a solution to (1.5) corresponding to P , with $y(0) = y$. By convexity of l and J , $J(y, z, P^\lambda) = v(y, z)$ which implies with probability one

$$l(y^\lambda(t), p^\lambda(t)) = \lambda l(y(t), p(t)) + (1 - \lambda)l(y^*(t), p^*(t))$$

for almost all $t \geq 0$. Assumption (3.1) then implies, with probability one, $p(t) = p^*(t)$ for almost all $t \geq 0$. \square

Remark 3.2. The optimal policy P^* was obtained by the method of dynamic programming. The theory of piecewise deterministic processes [5], [14] provides an alternate approach. In the present context, the piecewise deterministic theory considers bounded, Borel measurable functions $\pi = [0, \infty) \times R^n \times Z \rightarrow K$. Given initial data $y(0) = y$, $z(0) = z$, each such π determines an admissible control process P as follows. Let $\tau_0 = 0$ and $\tau_1 < \tau_2 < \dots$ denote the successive jump times of the Markov chain $z(t)$ and let

$$p(t) = \pi(t - \tau_i, y(\tau_i), z(\tau_i^+)), \quad \tau_i < t \leq \tau_{i+1}$$

where $y(t)$ is determined by solving (1.5) successively on each interval $[\tau_i, \tau_{i+1}]$. An optimal π^* is found as follows. For fixed z , as in (3.4), assume that there is a solution to

$$(3.6) \quad d\tilde{y}(t) = [B(z)p^*(\tilde{y}(t), z) + C(z)] dt,$$

with $\tilde{y}(0) = y$. Let

$$(3.7) \quad \pi^*(t, y, z) = p^*(\tilde{y}(t), z).$$

We claim that $\pi^*(\cdot, y, z)$ is unique, almost everywhere on $[0, \infty)$, for each y, z . This can be seen by slightly modifying the uniqueness proof above. We write the dynamic programming equation (1.6) as follows. Let

$$q_z = \sum_{z' \neq z} q_{zz'}, \quad v_1(y, z) = \sum_{z' \neq z} q_{zz'} v(y, z').$$

Then (1.6) becomes

$$(3.8) \quad (\alpha + q_z)v(y, z) = H(y, z, \nabla v) + v_1(y, z).$$

For fixed z , (3.8) is the dynamic programming equation for a discounted deterministic control problem, with dynamics (3.6), discount factor $\alpha + q_z$ and cost criterion $l(y, p) + v_1(y, z)$. As before, $p^*(\cdot, z)$ is an optimal feedback control and $\pi^*(\cdot, y, z)$ determined by (3.7) is the unique optimal (open loop) control.

4. Production planning problem. Let us return to the model mentioned in the Introduction. We now have the following:

$$y(t) = \text{inventory level at time } t \quad (y(t) \in R),$$

$$p(t) = \text{production rate at time } t \quad (p(t) \geq 0),$$

$$z(t) = \text{demand rate at time } t \quad (z(t) \in Z).$$

In the notation of § 1, we now have $n = 1$, $K = [0, \infty)$. The demand process is a finite state Markov chain, with state space $Z = \{z_1, \dots, z_M\}$. The dynamics are as follows:

$$dy(t) = [p(t) - z(t)] dt.$$

Thus, in (1.5), $B(z) = 1$ and $C(z) = -z$. It is assumed that $z_i > 0$ for all $i = 1, \dots, M$. We assume that the cost criterion has the form

$$l(y, p) = h(y) + c(p)$$

and we seek to minimize

$$(4.1) \quad J(y, z, P) = E \int_0^\infty e^{-\alpha t} [h(y(t)) + c(p(t))] dt.$$

The following assumptions are made about the holding cost h and production cost c :

(A1) h is convex, nonnegative on $(-\infty, \infty)$ with $h(0) = 0$.

(A2) c is twice continuously differentiable, nonnegative on

$$[0, \infty) \text{ with } c'(0) = c(0) = 0 \text{ and } c''(p) > 0 \text{ for } p > 0.$$

(A3) $C(|y|^\beta - 1) \leq h(y) \leq C(|y|^\gamma + 1)$ for all $y \in R$.

(A4) $C(|p|^\nu - 1) \leq c(p)$ for all $p \geq 0$,

where $C > 0$ and $\gamma, \beta, \nu > 1$ are fixed constants.

The Hamiltonian H in (1.7) now takes the form $H(y, z, r) = F(r) - zr + h(y)$ where

$$(4.2) \quad F(r) = \min_{p \geq 0} [pr + c(p)].$$

The assumption (2.2) is satisfied since $z > 0$ for all $z \in Z$, $F(r)$ is strictly concave for $r < 0$ and $F(r) = 0$ for $r \geq 0$. Theorems 2.2 and 2.4 imply that the value function $v(y, z)$ belongs to the class D_0 and is the unique viscosity solution to the dynamic programming equation.

The optimal feedback production policy is now given by

$$(4.3) \quad p^*(y, z) = \begin{cases} (c')^{-1} \left(-\frac{\partial}{\partial y} v(y, z) \right) & \text{if } \frac{\partial}{\partial y} v(y, z) > 0, \\ 0 & \text{if } \frac{\partial}{\partial y} v(y, z) \leq 0. \end{cases}$$

Since v is convex in y , and $(c')^{-1}$ is an increasing function, p^* is nonincreasing in y . Therefore, the differential equation

$$(4.4) \quad dy(t) = [p^*(y(t), z(t)) - z(t)] dt$$

has a unique solution $y^*(t)$ (see [8, Thm. 6.2].)

In the rest of the section, we shall show that y^* satisfies (3.4).

LEMMA 4.1. *There is a constant C , depending only on the initial condition $y^*(0) = y$, such that $|y^*(t)| \leq C$ for all $t \geq 0$.*

Proof. Let $\bar{y} = \sup \{y \in (-\infty, \infty) : p^*(y, z) \geq z \text{ for some } z \in Z\}$. Since v is convex in y and is nonnegative, (4.3) implies that \bar{y} is finite. Similarly, let $\tilde{y} = \inf \{y \in (-\infty, \infty) : p^*(y, z) \leq z \text{ for some } z \in Z\}$. Suppose that \tilde{y} is not finite. Then, there is $z \in Z$ such that $(\partial/\partial y)v(y, z) \geq -c'(z)$ for all $y \in R$. But this contradicts with Lemma 4.2, which follows. Now one completes the proof of the lemma, by observing that $[\tilde{y}, \bar{y}]$ is an attracting set for the differential equation (4.4). We refer to this set as the *turnpike set* in [7]. \square

Let $\bar{z} = \max \{z \in Z\}$.

LEMMA 4.2. *For each y, z , $v(y, z) \geq C(|y| - 1)$ for a suitable constant $C > c'(\bar{z})$.*

Proof. Let $\bar{v}(y)$ be the value function of the following variational problem:

$$\bar{v}(y) = \inf \left\{ \int_0^\infty e^{-\alpha t} \left[h(y(t)) + C \left| \frac{d}{dt} y(t) \right|^\nu \right] dt; y(0) = y \text{ and } y(\cdot) \in W^{1,\infty} \right\}$$

with $\nu > 1$. In view of (A4), $\bar{v}(y) \leq v(y, z)$ for a suitable $C > 0$. It suffices to show that $\bar{v}(y)|y|^{-1}$ converges to $M > \sup \{c'(z): z \in Z\}$ as y tends to $-\infty$. Since $\bar{v}(\cdot)$ is convex, $M = -\lim_n (d/dy)\bar{v}(y_n)$ where $\{y_n\}$ is any sequence which converges to $-\infty$ and $\bar{v}(\cdot)$ is differentiable at y_n , invoke Theorem 2.4 to conclude

$$(4.5) \quad \alpha \bar{v}(y_n) = \bar{F}\left(\frac{d}{dy} \bar{v}(y_n)\right) + h(y_n)$$

where $\bar{F}(r) = \sup \{rp + c|p|^\nu: -\infty < p < \infty\} = -\bar{c}|r|^{\nu/(\nu-1)}$. The positivity of \bar{v} yields $M \in [0, \infty]$. Suppose that $M < \infty$. Then divide (4.5) by $|y_n|$ and pass to the limit to obtain

$$\alpha M = \alpha \lim_n \bar{v}(y_n)|y_n|^{-1} = \lim_n h(y_n)|y_n|^{-1} = \infty.$$

Hence $M = \infty$ and the proof of the lemma is complete. \square

5. Inventory constraints. In this section, in addition to the nonnegative production constraint earlier, we impose the constraint that the inventory level cannot fall below a certain prescribed level y_{\min} . For each $y, z \in [y_{\min}, \infty) \times Z$, the set of admissible production processes $\mathcal{A}(y, z)$ is given by

$$\mathcal{A}(y, z) = \left\{ P \in \mathcal{A}: y + \int_0^t [p(s) - z(s)] ds \geq y_{\min} \text{ for all } t \geq 0 \right\}.$$

Then the corresponding value function is

$$v(y, z) = \inf \{J(y, z, P): P \in \mathcal{A}(y, z)\}.$$

The following characterization of v is a straightforward analogue of Theorem 1.1 of [12].

THEOREM 5.1. *The value function v for the constrained problem is in D_0 and is the only solution to the following equation:*

$$(5.1) \quad \alpha v(y, z) = H\left(y, z, \frac{\partial}{\partial y} v(y, z)\right) + Lv(y, z), \quad y, z \in [y_{\min}, \infty) \times Z,$$

$$(5.2) \quad \frac{\partial}{\partial y} v(y_{\min}, z) \leq -c'(z), \quad z \in Z.$$

Proof. The first two conditions in the definition of D_0 are easily verified after observing that if $P \in \mathcal{A}(y, z)$ and $\hat{P} \in \mathcal{A}(\hat{y}, z)$, then $\frac{1}{2}P + \frac{1}{2}\hat{P} \in \mathcal{A}(\frac{1}{2}y + \frac{1}{2}\hat{y}, z)$ and $J(\cdot, z, \cdot)$ is convex for each $z \in Z$.

Repeating the proofs of Theorems 2.2 and 2.4 we show that v is continuously differentiable in the y -variable on (y_{\min}, ∞) and satisfies (5.1). Define $(\partial/\partial y)v(y_{\min}, z)$ as the limit of $(\partial/\partial y)v(y, z)$ as y approaches y_{\min} from above (this limit exists due to the convexity of v in y). Now proceed as in Lemma 2.3 and use the fact that for any $P \in \mathcal{A}(y, z)$ the corresponding inventory level $y(t)$ is no less than y_{\min} , to obtain:

$$(5.3) \quad \alpha v(y_{\min}, z) \geq H(y_{\min}, z, r) + Lv(y_{\min}, z) \quad \text{for } r \leq \frac{\partial}{\partial y} v(y_{\min}, z)$$

(also, see Theorem 1.1 of [12]). Equations (5.1) and (5.3) yield

$$(5.4) \quad H(y_{\min}, z, \frac{\partial}{\partial y} v(y_{\min}, z)) \geq H(y_{\min}, z, r) \quad \text{for } r \leq \frac{\partial}{\partial y} v(y_{\min}, z).$$

The inequality (5.2) follows from (5.4), after observing that the map $r \rightarrow H(y_{\min}, z, r)$ achieves its maximum only at $r = -c'(z)$.

Uniqueness follows from the verification theorem, by observing that the optimal feedback policy P^* constructed in (3.5) is admissible on account of (5.2). \square

Remark 5.2. For each $\varepsilon > 0$, define $h^\varepsilon(y) = h(y) + [(1/\varepsilon) - 1] \max\{y_{\min} - y, 0\}$. Let v^ε be the value function of the unconstrained problem with inventory cost h^ε . Then, the following estimate is proved in [7]:

$$0 \leq v(y, z) - v^\varepsilon(y, z) \leq \sqrt{\varepsilon} K_R \quad \text{for } y, z \in [y_{\min}, R] \times Z$$

where K_R is a suitable constant.

REFERENCES

- [1] A. BENSOUSSAN, S. P. SETHI, R. VICKSON AND N. DERZKO, *Stochastic production planning with production constraints*, this Journal, 22 (1984), pp. 920-935.
- [2] F. CLARKE, *Optimization and Non-smooth Analysis*, Wiley-Interscience, New York, 1983.
- [3] M. G. CRANDALL, L. C. EVANS AND P.-L. LIONS, *Some properties of viscosity solutions of Hamilton-Jacobi equations*, Trans. AMS, 282 (1984), pp. 487-502.
- [4] M. G. CRANDALL AND P.-L. LIONS, *Viscosity solutions of Hamilton-Jacobi equations*, Trans. AMS, 277 (1983), pp. 1-42.
- [5] M. H. A. DAVIS, *Piecewise deterministic Markov processes: A general class of non-diffusion stochastic model*, J. Royal Statist. Soc. Ser. B, 46 (1984), pp. 353-388.
- [6] W. H. FLEMING AND R. RISHEL, *Deterministic and Stochastic Optimal Control*, Springer, New York, 1975.
- [7] W. H. FLEMING, S. P. SETHI AND H. M. SONER, *Turnpike sets in optimal stochastic production planning problems*, LCDS Report No. 86-2, Brown University, Providence, RI, 1986.
- [8] P. HARTMAN, *Ordinary Differential Equations*, Birkhauser-Verlag, Basel, Switzerland, 1982.
- [9] P.-L. LIONS, *Generalized Solutions of Hamilton-Jacobi Equations*, Research Notes in Mathematics, 69, Pitman, Boston, 1982.
- [10] J. G. KIMENIA AND S. B. GERSHWIN, *An algorithm for computer control of production in flexible manufacturing systems*, IIE Transactions, 15 (1983), pp. 353-362.
- [11] S. P. SETHI AND G. L. THOMPSON, *Applied Optimal Control: Applications to Management Science*, Nijhoff, Boston, 1981.
- [12] H. M. SONER, *Optimal stochastic control with state-space constraint*, II, this Journal, 24 (1986), pp. 1110-1123.
- [13] G. L. THOMPSON AND S. P. SETHI, *Turnpike horizons for production planning*, Management Sci., 26 (1980).
- [14] D. VERMES, *Optimal control of piecewise deterministic processes*, Stochastics, 14 (1985), pp. 165-208.

QUASI-NEWTON METHODS AND UNCONSTRAINED OPTIMAL CONTROL PROBLEMS*

C. T. KELLEY[†] AND E. W. SACHS[‡]

Abstract. We prove a mesh-independence result for the BFGS method in Hilbert space and apply it to a class of unconstrained optimal control problems. A new fourth order discretization scheme is used in the implementation. Observations from numerical experiments are presented.

Key words. optimal control, BFGS method

AMS(MOS) subject classifications. 65H10, 49D15

1. Introduction. Quasi-Newton methods play an important role in the numerical solution of problems in unconstrained optimization. Optimal control problems in their discretized form can be viewed as optimization problems and therefore be solved by quasi-Newton methods. Since the discretized problems do not solve the original infinite-dimensional control problem but rather approximate it up to a certain accuracy, various approximations of the control problem need to be considered. It is known that an increase in the dimension of optimization problems can have a negative effect on the convergence rate of the quasi-Newton method which is used to solve the problem. The purpose of this paper is to investigate this behavior and to explain how this drawback can be avoided for a class of optimal control problems. We show how to use the infinite-dimensional original problem to predict the speed of convergence of the BFGS-method [1], [8], [11], [23] for the finite-dimensional approximations.

In several papers [7], [15], [25], [28] the DFP-method [4], [9] and its application to optimal control problems were considered but rates of convergence were given at best for quadratic problems. In [26], [27] a linear rate of convergence was proved in Hilbert spaces and applied to optimal control. All the applications to optimal control problems were carried out for finite-dimensional approximations. This fact is important, because in [24] it was shown, that contrary to the finite-dimensional case [2], the BFGS-method can converge very slowly when applied to an infinite-dimensional problem. Hence it is desirable to know whether this convergence behavior can occur also for fine discretizations of control problems.

Sufficient [18] and characteristic [13] conditions for the superlinear rate were given in other analyses. As in the linear case for Broyden's method [29] and the conjugate gradient method [3], [10], an additional assumption on the initial approximation of the Hessian, i.e., one that approximates the true Hessian up to a compact operator, is needed to guarantee superlinear convergence (see [12]). In [10] a connection to quadratic control problems is shown. Here we want to consider nonlinear control problems and their discretization.

* Received by the editors February 18, 1986; accepted for publication November 20, 1986.

[†] Department of Mathematics, North Carolina State University, Raleigh, North Carolina 27695-8205. The work of this author was supported by National Science Foundation grants DMS-850944 and DMS-8300841.

[‡] Universität Trier, Fachbereich IV-Mathematik, Postfach 3825, D-5500 Trier, West Germany. The work of this author was supported by Air Force Office of Scientific Research grant AFOSR-85-0214.

Let $L: \mathbb{R}^{n+m+1} \rightarrow \mathbb{R}$, $f: \mathbb{R}^{n+m+1} \rightarrow \mathbb{R}^n$ for some $n, m \in \mathbb{N}$ and consider the following optimal control problem:

$$(1.1) \quad \begin{aligned} &\text{Minimize} \quad \int_0^T L(x(t), u(t), t) dt \\ &\text{subject to} \quad \dot{x}(t) = f(x(t), u(t), t), x(0) = x_0. \end{aligned}$$

If we think of $x(t) = x(u, t)$ being dependent on u , then the cost function can be written in terms of u only:

$$F(u) = \int_0^T L(x(u, t), u(t), t) dt.$$

The gradient of F is given by

$$\nabla F(u) = p(\cdot)^T f_u(x(\cdot), u(\cdot), \cdot) + L_u(x(\cdot), u(\cdot), \cdot),$$

where p solves the adjoint equation

$$(1.2) \quad -\dot{p}(t) = p(t)^T f_x(x(t), u(t), t) + L_x(x(t), u(t), t)$$

with $p(T) = 0$ (see e.g. [19]). Hence each gradient evaluation involves the solution of an additional system of differential equations. If one wants to apply Newton's method to the numerical solution of the control problems in order to use a higher order method, it becomes necessary to compute the second derivative of $F(\cdot)$. Even in the one-dimensional case, $m = n = 1$, this calculation is quite tedious. $\nabla^2 F(u)$ is given as follows.

Define $H: \mathbb{R}^3 \rightarrow \mathbb{R}$ by $H(x, u, t) = p(t)f(x, u, t) + L(x, u, t)$ for fixed $p(t)$, then with $\langle u, v \rangle = \int_0^T u(t)v(t) dt$

$$\begin{aligned} \langle w, \nabla^2 F(u)v \rangle &= \langle \xi(w), H_{xx}(x, u, \cdot) \xi(v) \rangle + \langle w, H_{uu}(x, u, \cdot) v \rangle \\ &\quad + \langle w, H_{ux}(x, u, \cdot) \xi(v) \rangle + \langle \xi(w), H_{xu}(x, u, \cdot) v \rangle \end{aligned}$$

where x solves (1.1), p solves (1.2) and $\xi(w)$ is the solution of

$$\dot{\xi}(t) = \xi(t)^T f_x(x(u, t), u(t), t) + f_u(x(u, t), u(t), t)w(t), \quad \xi(0) = 0.$$

Also for the finite-dimensional approximations, the computation of the Hessian is not a simple task. Therefore, the quasi-Newton methods where the Hessian is approximated by an updating procedure provide a useful alternative because of their superlinear rate of convergence under certain assumptions.

In order to minimize a twice Fréchet-differentiable function F over a Hilbert space H one chooses in the BFGS-method an operator $B_0 \in \mathcal{L}(H)$, the space of linear continuous invertible operators on H , and some $u_0 \in H$ and defines the following sequence: Given $u_i \in H$, $B_i \in \mathcal{L}(H)$.

(i) Solve $B_i s_i = -\nabla F(u_i)$;

(ii) Let $u_{i+1} = u_i + s_i$

$$(1.3) \quad y_i = \nabla F(u_{i+1}) - \nabla F(u_i);$$

(iii) Let

$$B_{i+1} = B_i + \frac{\langle y_i, \cdot \rangle}{\langle s_i, y_i \rangle} y_i - \frac{\langle B_i s_i, \cdot \rangle}{\langle s_i, B_i s_i \rangle} B_i s_i.$$

In this paper we will investigate finite-dimensional approximations of this algorithm.

In the second section we will show how the convergence behavior of the BFGS-method for finite-dimensional approximations is influenced by the convergence

behavior of the BFGS-method applied to the infinite-dimensional problem. The statements will be phrased in terms of the number of iterates which are required to achieve a certain termination criterion. This approach, which we also used for the solution of nonlinear integral equations [16] will be taken because no uniform convergence rate estimates are available for the family of approximating problems. This is in contrast to the analyses of discretized control problems solved by the gradient projection method or conditional gradient method [6] and Newton's method [21]. The results in § 2 can be applied to any sequence of approximating problems. In the third section we show how to utilize these techniques for finite difference approximations of the optimal control problem. We point out that our goal is not to estimate the error between the solutions of the discretized and the original problem which can be found in the existing literature [14]. Since the discretized problems also need to be solved by some algorithm we deduce statements on the convergence of this algorithm, in our case the BFGS-method, for the whole family of discretized problems. The nature of the discretized problem, i.e., the fourth order Runge-Kutta scheme with Hermite interpolation at intermediate points, does not assure the symmetry of the Hessian of the approximating problem. Since these problems, however, are still close to the original one with a self-adjoint Hessian, we used an approximate version of the BFGS-method on these problems and the numerical results are quite successful. A numerical example demonstrates that the proper choice of the initial approximation B_0 with regard to the compactness requirement leads to qualitatively different convergence behavior for the finite-dimensional problems. These results for optimal control problems underscore the observation made by the authors in [16] for integral equations and in [17] for elliptic differential equations that quasi-Newton methods can be successful, also for infinite-dimensional problems when the special structure of the problem is taken into account.

2. An approximation of the infinite-dimensional BFGS-method. The BFGS-method was introduced by Broyden [1], Fletcher [8], Goldfarb [11] and Shanno [23] as a version of quasi-Newton methods which takes into account the symmetry and positive definiteness of the Hessian of the function to be minimized. In this paper, we first consider the infinite-dimensional analysis.

Let H be a Hilbert space and

$$F: H \rightarrow R$$

a nonlinear twice continuously Fréchet-differentiable functional on H . We consider the following minimization problem.

Find $u_* \in H$ with

$$(P) \quad F(u_*) \leq F(u) \quad \text{for all } u \in H.$$

Any solution u_* of (P) is also a solution of the nonlinear equation

$$(E) \quad G(u_*) = 0$$

where $G(u) = \nabla F(u)$. The BFGS-method, outlined in § 1, finds solutions of (E) iteratively by updating linear operators B_i which approximate $G'(u_i)$ and by solving a linear equation for each step. Daniel [3] pointed out in 1965 that for conjugate gradient methods a superlinear rate of convergence can be attained in Hilbert spaces if, for example, the initial approximation B_0 of the Jacobian of G is close and differs only by a compact operator. Recently, Griewank [12] showed that a similar result holds for the BFGS-method.

We will show in § 3 that for a large class of optimal control problems the compactness condition can be satisfied in a natural way. However, the numerical solution of (E) and (P) requires some sort of discretization. If we approximate the space H by finite-dimensional spaces H^N and replace F by F^N defined on H^N , then we obtain problems of the type

$$(P^N) \quad \text{Minimize } F^N(u^N), \quad u^N \in H^N.$$

With this approach we are automatically assured that the Hessian of F^N is symmetric. The positive definiteness of the Hessian $\nabla^2 F^N$ at the solution u_*^N is an additional requirement of convergence theorems for finite-dimensional quasi-Newton methods. For control problems, however, we take the different point of view in this paper that we want to solve the Pontryagin maximum principle which is a classical approach for solving control problems by gradient or Newton's method, see e.g., the review article by Polak [20]. We approximate the necessary optimality conditions $G(u^*)=0$ and solve these approximate problems by a finite-dimensional analogue of a quasi-Newton method for the infinite-dimensional problem. The approximate functions G^N need not be gradients of scalar valued functions and, in general, they are not if defined by finite difference schemes. A choice of G^N as the gradient of some F^N would result in the loss of the simple structure of the discrete version of (1.2). Instead of (E) we solve:

$$(E^N) \quad \text{Find } u_*^N \in H^N \text{ with } G^N(u_*^N) = 0.$$

Except for simple discretization schemes such as Euler's method, however, it is not reasonable to assume that G^N has a Jacobian which is symmetric. Computational results indicate that it is still possible to treat these problems successfully with an approximate version of the infinite-dimensional BFGS-method.

Let $\langle \cdot, \cdot \rangle_N$ denote the inner product on H^N and let $G^N: H^N \rightarrow H^N$ be Fréchet-differentiable. Let some initial approximations $u_0^N \in H^N$ and $B_0^N \in \mathcal{L}(H^N)$ for the solution u_*^N and the Jacobian of G^N at u_*^N be given such that B_0^N is symmetric with regard to the inner product $\langle \cdot, \cdot \rangle_N$.

Then with $i=0$,

$$(2.1a) \quad \text{solve } B_i^N s_i^N = -G^N(u_i^N), \quad s_i^N \in H^N,$$

$$(2.1b) \quad \text{set } u_{i+1}^N = u_i^N + s_i^N, \\ y_i^N = G^N(u_{i+1}^N) - G^N(u_i^N),$$

and

$$(2.1c) \quad B_{i+1}^N = B_i^N + \frac{\langle y_i^N, \cdot \rangle_N}{\langle s_i^N, y_i^N \rangle_N} y_i^N - \frac{\langle B_i^N s_i^N, \cdot \rangle_N}{\langle s_i^N, B_i^N s_i^N \rangle_N} B_i^N s_i^N.$$

The inner product $\langle \cdot, \cdot \rangle_N$ on the N -dimensional space H^N is not the Euclidean inner product on \mathbb{R}^N and therefore, algorithm (2.1), in general, is not identical with the BFGS-method on \mathbb{R}^N . In order to ensure that the convergence behavior of $\{u_i^N\}$ is similar to that of $\{u_i\}$ for large N , we impose the following conditions on H^N .

Let $\{P^N\}$ denote a sequence of linear prolongation operators

$$P^N: H^N \rightarrow Z$$

where Z is a normed subspace of H with $\|\cdot\|_Z \cong \|\cdot\|_H$. For the application to control problems, where H^N is a space of N -dimensional vectors, P^N may be interpreted as a piecewise polynomial interpolation operator.

We introduce a notion of Z -convergence as follows: A sequence $u^N \in H^N$ is Z -convergent to $u \in Z$, i.e., $u^N \rightarrow_Z u$, if

$$(2.2) \quad \lim_{N \rightarrow \infty} \|P^N u^N - u\|_Z = 0.$$

We also have to assume that the discrete inner products approximate the original one in the following sense:

(A1) If $u_j^N \rightarrow_Z u_j \in Z$ for $j = 1, 2$, then

$$\lim_{N \rightarrow \infty} \langle u_1^N, u_2^N \rangle_N = \langle u_1, u_2 \rangle.$$

Since for control problems the evaluation of G^N includes the numerical solution of differential equations, the Z -convergence of $G^N(u^N)$ to $G(u)$ may require more smoothness of u than $u \in Z$. Hence we introduce a subspace W of smooth functions in Z and because the iterates should remain in W we assume the following:

(A2) There is $u_* \in W$ with $G(u_*) = 0$ such that G is defined for all $u \in W$ sufficiently near u_* in the Z -norm, and if for a sequence $u^N \in H^N$, $u^N \rightarrow_Z u \in W$, then

$$G^N(u^N) \xrightarrow{Z} G(u) \in W.$$

These assumptions enable us to show the following theorem.

THEOREM 2.1. Assume that (A1) and (A2) hold. Let $u_0^N \in H^N$, $u_0 \in W$, $(B_0^N)^{-1} \in \mathcal{L}(H^N)$, $B_0^{-1} \in \mathcal{L}(Z)$ such that for $i = 0$

$$(2.3) \quad B_i^{-1}(W) \subset W,$$

$$(2.4) \quad u_i^N \xrightarrow{Z} u_i$$

and

$$(2.5) \quad (B_i^N)^{-1} v^N \xrightarrow{Z} B_i^{-1} v$$

for all sequences $v^N \in H^N$ which are Z -convergent to $v \in W$. If for u_1 and B_1 as given by (1.3) we have $B_1^{-1} \in \mathcal{L}(Z)$ and $G(u_1)$ well defined, then for N large enough we have

$$(2.6) \quad (B_1^N)^{-1} \in \mathcal{L}(H^N)$$

and (2.3)–(2.5) hold for $i = 1$.

Proof. Note, that in the case

$$(2.7) \quad \langle s_0^N, y_0^N \rangle_N \neq 0 \quad \text{and} \quad \langle s_0^N, B_0^N s_0^N \rangle_N \neq 0,$$

$(B_0^N)^{-1}$ can be updated directly by

$$(2.8) \quad (B_1^N)^{-1} = (B_0^N)^{-1} + \frac{\langle s_0^N, \cdot \rangle_N}{\langle s_0^N, y_0^N \rangle_N} w_0^N + \frac{\langle w_0^N, \cdot \rangle_N}{\langle s_0^N, y_0^N \rangle_N} s_0^N - \frac{\langle w_0^N, y_0^N \rangle_N}{\langle s_0^N, y_0^N \rangle_N^2} \langle s_0^N, \cdot \rangle_N s_0^N,$$

$$w_0^N = s_0^N - (B_0^N)^{-1} y_0^N.$$

In the infinite-dimensional case a similar formula holds for B_1^{-1} . By (A2) we know that $G(u_0) \in W$ and with (2.3) we obtain $s_0 = -B_0^{-1} G(u_0) \in W$. Hence

$$u_1 \in W \quad \text{and} \quad y_0 = G(u_1) - G(u_0) \in W.$$

From this and the formula for B_1^{-1} it is easy to see that B_1^{-1} leaves W invariant. Condition (2.4) and (A2) imply that

$$G^N(u_0^N) \xrightarrow{Z} G(u_0).$$

By (2.5)

$$s_0^N \xrightarrow{Z} s_0$$

and (2.4) holds as for $i = 1$. In addition,

$$y_0^N \xrightarrow{Z} y_0$$

holds. Together with (A1) this yields

$$(2.9a) \quad \lim_{N \rightarrow \infty} \langle s_0^N, y_0^N \rangle_N = \langle s_0, y_0 \rangle,$$

$$(2.9b) \quad \lim_{N \rightarrow \infty} \langle s_0^N, B_0^N s_0^N \rangle_N = \langle s_0, B_0 s_0 \rangle.$$

Since by assumption B_1^{-1} is well defined, we know that $\langle s_0, y_0 \rangle$ and $\langle s_0, B_0 s_0 \rangle$ are both nonzero. Hence (2.9) ensures that (2.7) holds for N sufficiently large. If we update $(B_1^N)^{-1}$ and B_1^{-1} according to (2.8), then (2.4), (A1) and (2.9) together yield (2.6). Then for any sequence $v^N \in H^N$, which is Z -convergent to $v \in W$, (A1) implies (2.5) for $i = 1$. This completes the proof.

If $B_i^{-1} \in \mathcal{L}(Z)$ for $1 \leq i \leq i^*$, then we can show with an induction argument based on Theorem 2.1 that

$$(2.10) \quad \lim_{N \rightarrow \infty} \max_{1 \leq i \leq i^*} \|P^N u_i^N - u_i\|_Z = 0$$

holds for any integer $i^* \geq 1$.

A reasonable choice of a termination criterion for (E^N) is $\|G^N(u^N)\|_N$ being sufficiently small. Let $i(\varepsilon)$ denote the smallest iteration index for which the norm of the gradient is less than ε :

$$i(\varepsilon) = \min \{i \in \mathbb{N}: \|G(u_i)\| < \varepsilon\},$$

$$i_N(\varepsilon) = \min \{i \in \mathbb{N}: \|G^N(u_i^N)\|_N < \varepsilon\}.$$

The following relation holds between i and i_N .

THEOREM 2.2. *Let all the assumptions in Theorem 2.1 hold, $G(u_i)$ be defined and $B_i^{-1} \in \mathcal{L}(Z)$ for all $i \geq 1$. Then for each $\varepsilon > 0$ there exists $N_\varepsilon \in \mathbb{N}$ with*

$$(2.11) \quad i(\varepsilon + \delta) \leq i_N(\varepsilon) \leq i(\varepsilon)$$

for all $N > N_\varepsilon$ and $\delta > 0$.

Proof. Equation (2.10) and (A2) imply that for each i

$$G^N(u_i^N) \xrightarrow{Z} G(u_i)$$

and by (A1)

$$(2.12) \quad \lim_{N \rightarrow \infty} \|G^N(u_i^N)\|_N = \|G(u_i)\|.$$

Clearly, if $\|G(u_i)\| < \varepsilon$ then $\|G^N(u_i^N)\|_N < \varepsilon$ for N sufficiently large, and therefore

$$i_N(\varepsilon) \leq i(\varepsilon).$$

Hence

$$(2.13) \quad \lim_{N \rightarrow \infty} \max_{1 \leq i \leq i(\varepsilon)} \|\|G^N(u_i^N)\|_N - \|G(u_i)\|\| = 0.$$

If, now, for a given $\delta > 0$, $i_N(\varepsilon) < i(\varepsilon + \delta)$ for infinitely many N then there is $j < i(\varepsilon)$ and a sequence $N_k \rightarrow \infty$ such that

$$\|G^{N_k}(u_j^{N_k})\|_{N_k} \geq \varepsilon + \delta.$$

Letting $N_k \rightarrow \infty$ we have $\|G(u_j)\| \geq \varepsilon + \delta$ which contradicts (2.13).

Theorem 2.2 is independent of any rate of convergence for the infinite-dimensional problem and would therefore also be useful to explain slow convergence behavior for the finite-dimensional approximations. For the superlinear rate we state the following theorem.

THEOREM 2.3. *Let $G: H \rightarrow H$, $G^N: H^N \rightarrow H^N$ be continuously Fréchet-differentiable and $u_* \in H$ such that*

$$G(u_*) = 0, \quad G'(u_*) \text{ is self-adjoint, positive definite.}$$

Choose $B_0 \in \mathcal{L}(H)$, $B_0^{-1} \in \mathcal{L}(Z)$, $u_0 \in Z$ such that

$$B_0 - G'(u_*) \text{ is compact}$$

and $\|B_0 - G'(u_)\|$ and $\|u_0 - u_*\|$ are sufficiently small to yield the superlinear convergence of the BFGS-iterates for (E). Select H^N , G^N , $u_0^N \in H^N$, $B_0^N \in \mathcal{L}(H^N)$ such that (A1), (A2), (2.3)–(2.5) hold. Then for the BFGS-iterates u_i^N for (E^N) the following is true:*

For each $\varepsilon > 0$ there exists N_ε with

$$(2.14) \quad i(\varepsilon) - 1 \leq i_N(\varepsilon) \leq i(\varepsilon) \quad \text{for } N \geq N_\varepsilon.$$

Proof. The superlinear rate for (E) follows from [12]. The updates B_i^N are well defined by using an induction argument identical to the proof of Theorem 2.1.

Note now that $\|G(u_i)\|$ is a strictly decreasing sequence. To see this note that boundedness and invertibility of $G'(u_*)$ imply that there are c_1 and c_2 so that for all u sufficiently near u_*

$$c_1 \|u - u_*\| \leq \|G(u)\| \leq c_2 \|u - u_*\|.$$

We let $\lambda_n = \|u_{n+1} - u_*\| / \|u_n - u_*\|$. For n sufficiently large, $\lambda_n < c_2 / c_1$ by the superlinear convergence and hence, for n large

$$\begin{aligned} \|G(u_{n+1})\| &\leq c_2 \|u_{n+1} - u_*\| \leq c_2 \lambda_n \|u_n - u_*\| \\ &\leq \frac{c_2}{c_1} \lambda_n \|G(u_n)\| < \|G(u_n)\|. \end{aligned}$$

Hence, for δ sufficiently small

$$i(\varepsilon + \delta) \geq i(\varepsilon) - 1,$$

which implies (2.14) instead of (2.11).

Let us point out that (2.14) gives an estimate for the number of iterates necessary to achieve a certain stopping criterion for the finite-dimensional problem that is independent on the dimension N and is only determined by the iterate number of the infinite-dimensional analogue. For this problem, however, the superlinear rate can be proved directly. Theorem 2.3 also does not require that B_0^N is self-adjoint or even positive definite as long as (2.4) holds. This facilitates the application to control problems in the next section.

3. Optimal control problem. We want to apply the BFGS-method to optimal control problems which are formulated in infinite-dimensional spaces. Let

$$f: \mathbb{R}^{n+m+1} \rightarrow \mathbb{R}^n, \quad L: \mathbb{R}^{n+m+1} \rightarrow \mathbb{R}$$

be given functions. Then the optimal control problem is given by:

$$(OCa) \quad \text{Minimize} \quad \int_0^T L(x(t), u(t), t) dt$$

$$(OCb) \quad \text{subject to} \quad \dot{x}(t) = f(x(t), u(t), t), \quad x(0) = x_0.$$

Suppose that for a given control u the differential equation (OCb) is uniquely solvable with solution x denoted by

$$x = Su.$$

Then we can define an objective on $L_m^2[0, T] = H$

$$(3.1) \quad F(u) = \int_0^T L(Su(t), u(t), t) dt.$$

For the convergence analysis we assume that F is twice continuously Fréchet-differentiable on an open convex set containing u_* and all iterates. Then we can express the gradient $\nabla F(u)$ of F at u through a system of adjoint equations (see e.g. [19]):

$$(3.2) \quad \nabla F(u) = p(\cdot)^T f_u(x(\cdot), u(\cdot), \cdot) + L_u(x(\cdot), u(\cdot), \cdot)$$

where x satisfies (OCb) and p solves

$$(3.3) \quad -\dot{p}(t) = p(t)^T f_x(x(t), u(t), t) + L_x(x(t), u(t), t) \quad \text{with } p(T) = 0.$$

As we have seen in the previous section, the important condition for superlinear convergence is imposed on the Hessian of F , which is of a rather complicated form for control problems. We observe in (3.2) that x and p as solutions of differential equations depend on u through a compact operator and hence the second derivative does too. This leads to the following theorem.

THEOREM 3.1. *Suppose that F is twice Fréchet-differentiable in a neighborhood of $u \in L_2^m[0, T]$. Then with p defined in (3.3) and*

$$(3.4) \quad H(x, u) = p^T f(x, u) + L(x, u),$$

the operator $R: L_2^m[0, T] \rightarrow L_2^m[0, T]$ defined by

$$(Rv)(t) = (\nabla^2 F(u)(v))(t) - H_{uu}(x(t), u(t))v(t)$$

is compact.

The proof follows from the fact that the Hessian of F is given by

$$\nabla^2 F(u) = M^* H_{xx}(x, u) M + M^* H_x(x, u) + H_{ux}(x, u) M + H_{uu}$$

where $Mv = \xi$ denotes the solution of

$$\dot{\xi}(t) = \xi(t)^T f_x(x(t), u(t)) + v(t)^T f_u(x(t), u(t)), \quad \xi(0) = 0.$$

Obviously, M and its adjoint operator M^* are compact operators, so that the multiplication operator H_{uu} is the only noncompact part in $\nabla^2 F(u)$, which must be known exactly. Since

$$H_{uu}(x, u) = p^T f_{uu}(x, u) + L_{uu}(x, u),$$

all control problems with $f_{uu} = 0$ and L_{uu} known are examples where this demand is reasonable. This includes problems where the control u enters the differential equation linearly, i.e.,

$$\dot{x}(t) = q(t, x(t))u(t) + r(t, x(t)), \quad x(0) = x_0,$$

and the objective function L depends on u , for example, in the following form:

$$(3.5) \quad \int_0^T (N(t, x(t)) + \alpha(t)u^2(t)) dt, \quad \alpha \text{ given.}$$

In order to satisfy the compactness condition on $B_0 - \nabla^2 F(u_*)$, we would have to choose in the case of (3.5) for B_0

$$(3.6) \quad (B_0 v)(t) = \alpha(t)v(t), \quad t \in [0, T].$$

Then we might expect a superlinear rate of convergence. If $\alpha(t) \neq 1$ and we choose $B_0 = I$, we do not necessarily obtain a superlinear rate. The numerical examples illustrate this point.

As an application of the results in the second section we consider a finite difference discretization of a control problem. Various authors have studied these problems in connection with the convergence of solutions or optimal values of discretized problems to the solution or optimal value of the original problem (see e.g. [14]). In this section we investigate the performance of quasi-Newton methods when applied to the finite-dimensional approximations of optimal control problems.

In the sequel we solve $G(u) = \nabla f(u) = 0$, where for given control u the state x is computed by (OCb) and the adjoint p by (3.3). Therefore, instead of discretizing (OC) and having to compute the gradient of the finite difference scheme for (OCb), we discretize the differential equations (OCb) and (3.3) and set $\nabla F(u) = 0$ in (3.2) at the grid points. Although it is known (see e.g. [19, p. 68]) that the gradient for a discretization of the optimal control problem (OC) can be expressed through a finite difference adjoint equation, we found similar to the results in [14] that this gradient differs from a discretization of the equation $\nabla F(u) = 0$. One exception to this observation is Euler's method which, however, is inadequate for our purposes because it is a low order discretization scheme. If a Runge-Kutta scheme is applied to the state equation (OCb) or the adjoint equation (3.3) this requires the evaluation of u and x at intermediate points. We introduce these points as additional variables for the controls u only and interpolate to obtain values for x at these intermediate points.

Let $N \in \mathbb{N}$ be the discretization parameter and let us restrict for our examples to the one-dimensional case: The time interval $[0, T]$ is divided into $2N$ subinterval of equal length $h/2$, where

$$(3.7) \quad h = \frac{T}{N}.$$

Let $u \in \mathbb{R}^{2N+1}$ be given. For given x_0, \dots, x_{sj-2} compute x_{2j}, x_{2j-1} by

$$(3.8a) \quad k_1 = hf(u_{2j-2}, x_{2j-2}),$$

$$(3.8b) \quad k_2 = hf(u_{2j-1}, x_{2j-2} + \frac{1}{2}k_1),$$

$$(3.8c) \quad k_3 = hf(u_{2j-1}, x_{2j-2} + \frac{1}{2}k_2),$$

$$(3.8d) \quad k_4 = hf(u_{2j}, x_{2j-2} + k_3),$$

$$(3.8e) \quad x_{2j} = \frac{1}{6}(k_1 + 2k_2 + 2k_3 + k_4),$$

$$(3.8f) \quad k_1^+ = hf(u_{2j}, x_{2j}),$$

$$(3.8g) \quad x_{2j-1} = \frac{1}{2}(x_{2j} + x_{2j-2}) + \frac{1}{8}(k_1 + k_1^+).$$

The formulas (3.8) represent a Runge-Kutta scheme of fourth order, i.e., the global truncation error for the values $\{x_{2j}\}_{j=1}^N$ is of order $O(h^4)$. For the odd indices we use (3.8g), i.e., a Hermite interpolation based on x_{2j-2} and x_{2j} , which has a local truncation error of $O(h^4)$. Hence the error for all points $x_i, i = 0, \dots, 2N$ is of order $O(h^4)$. The

equations for the adjoint equation which is backward in time are: Given $p_{2N}, p_{2N-1}, \dots, p_{2j}$, compute p_{2j-1}, p_{2j-2} as follows. Let $\varphi(x, u, p) = pf_x(x, u) + L_x(x, u)$

$$(3.9a) \quad k_1 = h\varphi(x_{2j}, u_{2j}, p_{2j}),$$

$$(3.9b) \quad k_2 = h\varphi(x_{2j-1}, x_{2j-1}, p_{2j} + \frac{1}{2}k_1),$$

$$(3.9c) \quad k_3 = h\varphi(x_{2j-1}, x_{2j-1}, p_{2j} + \frac{1}{2}k_2),$$

$$(3.9d) \quad k_4 = h\varphi(x_{2j-2}, x_{2j-2}, p_{2j} + k_3),$$

$$(3.9e) \quad p_{2j-2} = \frac{1}{6}(k_1 + 2k_2 + 2k_3 + k_4),$$

$$(3.9f) \quad k_1^+ = h\varphi(x_{2j-2}, u_{2j-2}, p_{2j-2}),$$

$$(3.9g) \quad p_{2j-1} = \frac{1}{2}(p_{2j} + p_{2j-2}) + \frac{1}{8}(k_1 + k_1^+).$$

The endpoint condition for $p \in \mathbb{R}^{2N+1}$ is $p_{2N} = 0$. In order to apply the approximate BFGS-method in § 2 we need to satisfy assumptions (A1) and (A2). Let us choose

$$(3.10) \quad H = L_2[0, T], \quad Z = C[0, T]$$

with the usual norms. The subspace W and the finite-dimensional spaces are

$$(3.11) \quad W = C^5[0, T], \quad H^N = \mathbb{R}^{2N+1}$$

with inner products which stem from the composite Simpson's rule

$$(3.12) \quad \langle u^N, v^N \rangle_N = \frac{T}{6N} \left(u_0^N v_0^N + 2 \sum_{j=1}^{N-1} u_{2j}^N v_{2j}^N + 4 \sum_{j=1}^N u_{2j-1}^N v_{2j-1}^N + u_{2N}^N v_{2N}^N \right).$$

The prolongations $P^N: H^N \rightarrow Z$ are defined as follows:

$$(3.13) \quad \text{For } u \in \mathbb{R}^{2N+1} \text{ let } P^N u \in C[0, T] \text{ be the piecewise parabolic interpolant with } (P^N u)(t_j) = u_j, t_j = jT/2N, j = 0, \dots, 2N.$$

The nonlinear mapping $G^N: \mathbb{R}^{2N+1} \rightarrow \mathbb{R}^{2N+1}$ is defined as follows:

For $u^N \in \mathbb{R}^{2N+1}$ let $x^N \in \mathbb{R}^{2N+1}$ be the solution of (3.8) and let $p^N \in \mathbb{R}^{2N+1}$ be the solution of (3.9). Then

$$(3.14) \quad (G^N u^N)_i = p_i^N f_u(x_i^N, u_i^N) + L_u(x_i^N, u_i^N).$$

LEMMA 3.1. Assume there exists an optimal control $u_* \in C^5[0, T]$ and (OCb) is uniquely solvable for all u sufficiently close to u_* in the $C[0, T]$ -norm. If $f_x, L_x, f_u, L_u \in C^5[0, T]$, then (A1) and (A2) are satisfied with the definitions (3.10)–(3.14).

Proof. If (2.2) is satisfied, then for $u_1, u_2 \in Z$

$$(3.15) \quad \lim_{N \rightarrow \infty} \max_{0 \leq j \leq 2N} |u_{i,j}^N - u_i(t_j)| = 0, \quad i = 1, 2.$$

Since u_1 and u_2 are continuous on $[0, T]$, the quadrature rule (3.12) implies that

$$\lim_{N \rightarrow \infty} \langle u_1^N, u_2^N \rangle_N = \int_0^T u_1(t) u_2(t) dt,$$

which proves (A1).

If $u \in W = C^5[0, T]$ is chosen, then, by the smoothness assumptions on f and L , any solutions x and p of differential equations (OCb) and (3.3) lie in W and Gu is an element of W . If

$$u^N \xrightarrow{Z} u \in W,$$

then for the corresponding trajectories and solutions of the adjoint equations

$$x^N \xrightarrow{Z} x, \quad p^N \xrightarrow{Z} p.$$

By (3.14) this implies

$$G^N u^N \xrightarrow{Z} Gu.$$

Obviously (3.8)–(3.14) are not the only choices which are feasible. An Euler scheme with a simple quadrature rule would yield the same results theoretically, but the numerical observations are easier to obtain for a higher order scheme. The Runge–Kutta scheme with Hermite interpolation for intermediate points approximates the true solution up to order 4, which under appropriate smoothness on f and L , means that the gradient is approximated up to that order. The same order of accuracy holds for the second approximation, the inner product, which is done by composite Simpson's rule. In fact, one can see from the formula for the updated iterate u_1 , with the inverse of B_0 , that u_1^N differs from u_1 , by an order of four; this is why set $W = C^5[0, T]$.

The numerical example represents a simple production inventory model (see [22, p. 145]) which is described as follows:

$$\text{Min} \int_0^T e^{-\rho t} \left(\frac{d}{2} (x(t) - a(t))^2 + \frac{c}{2} (u(t) - b(t))^2 \right) dt$$

such that

$$\dot{x}(t) = u(t) - s(t), \quad x(0) = x_0.$$

The parameters were chosen as follows:

$$\rho = 1, \quad d = c = 1, \quad a = 15, \quad b = 30, \quad s(t) = t^2, \quad x_0 = 10.$$

Using Pontryagin's maximum principle, it is possible to compute the optimal control u_* for the infinite-dimensional problem. It is of the form

$$u_*(t) = \alpha_1 \lambda_1 e^{\lambda_1 t} + \alpha_2 \lambda_2 e^{\lambda_2 t} + t^2 - 2t + 4$$

where $\lambda_{1/2} = (1 \pm \sqrt{5})/2$ and α_i are other constants (see [22, p. 145f] for details). The Hessian of F for this problem is given by

$$\nabla^2 F(u)(v) = d \int_t^T e^{-\rho s} \int_0^s v(\sigma) d\sigma + c e^{-\rho t} v(t).$$

The first summand of the Hessian is a compact operator from $L_2[0, T]$ into itself and the second is not. By the theory from the previous sections it is important for the superlinear convergence of the iterates that the noncompact part is approximated correctly in the initial approximation of the Hessian. We consider two choices for B_0 :

$$(3.16) \quad B_0 = \alpha I \quad \text{and} \quad B_0 = H_{uu}$$

where $(H_{uu}v)(t) = c e^{-\rho t} v(t)$. For given T we choose α in (3.16) such that $\|\alpha I - H_{uu}\|$ is small. A simple estimate yields

$$\frac{\|G'(u_*) - H_{uu}\| - \|G'(u_*) - \alpha I\|}{\|G'(u_*)\|} \leq \frac{\|\alpha I - H_{uu}\|}{\|G'(u_*)\|} \leq 0.041 [0.19]$$

for $\alpha = 0.86$ and $T = 0.3$ [$\alpha = 0.63$ and $T = 1.0$]. For each case of α and T , this indicates that aside from the compactness the two choices for B_0 , αI and H_{uu} have roughly the same relative distance from the Jacobian at the root.

Analogous to (3.16) we choose B_0^N as diagonal matrices,

$$B_0^N = \text{diag } \alpha \quad \text{and} \quad B_0^N = \text{diag } d,$$

where

$$\alpha_j = \alpha \quad \text{and} \quad d_j = c e^{-\rho t_j}$$

for $j = 0, \dots, 2N$ with t_j as in (3.13). Furthermore, we let

$$u_0 = u_* + 1 \quad \text{and} \quad (u_0^N)_j = u_*(t_j) + 1$$

for $j = 0, \dots, 2N$.

Condition (2.3) follows from the choices of B_0^{-1} , whereas (2.4) and (2.5) are a consequence of the choice P^N .

In Tables 1 and 2 we list the number of iterates $i_N(\varepsilon)$ necessary to achieve the tolerance in the norm of the gradient.

TABLE 1
 $i_N(10^{-6})$ for $T = 0.3$.

N	10	40	100	200
$B_0 = H_{uu}$	3	3	3	3
$B_0 = 0.86I$	6	6	6	6

TABLE 2
 $i_N(10^{-4})$ for $T = 1.0$.

N	20	40	100	200
$B_0 = H_{uu}$	3	3	3	3
$B_0 = 0.63I$	7	7	7	7

TABLE 3
 $T = .3, N = 200$.

$B_0 = H_{uu}$					
i	σ_i	σ_i/σ_{i+1}	γ_i	η_i	η_i/η_{i+1}
1	.1027E+01		.2732E-01	.2911E-01	
2	.2913E-01	.2836E-01	.5859E-04	.6889E-04	.2367E-02
3	.6928E-04	.2378E-02	.3312E-06	.4041E-06	.5866E-02
4	.4042E-06	.5834E-02	.1276E-09	.1472E-09	.3641E-03
5	.1481E-09	.3664E-03	.2626E-12	.1625E-11	.1104E-01
$B_0 = .86I$					
1	.1037E+01		.9753E-01	.1047E+00	
2	.1041E+00	.1003E+00	.6389E-02	.7445E-02	.7112E-01
3	.7426E-02	.7137E-01	.5431E-03	.6308E-03	.8473E-01
4	.6276E-03	.8452E-01	.3685E-04	.4268E-04	.6766E-01
5	.4280E-04	.6819E-01	.3054E-05	.3520E-05	.8247E-01
6	.3490E-05	.8156E-01	.2092E-06	.2433E-06	.6912E-01
7	.2446E-06	.7008E-01	.1714E-07	.1968E-07	.8087E-01
8	.1948E-07	.7963E-01	.1186E-08	.1383E-08	.7027E-01
9	.1391E-08	.7144E-01	.9633E-10	.1100E-09	.7955E-01
10	.1092E-09	.7847E-01	.6715E-11	.7792E-11	.7085E-01

TABLE 4
 $T = 1, N = 200.$

$B_0 = H_{uu}$					
i	σ_i	σ_i/σ_{i+1}	γ_i	η_i	η_i/η_{i+1}
1	.1239E+01		.2302E+00	.2503E+00	
2	.2503E+00	.2021E+00	.1095E-02	.1997E-02	.7978E-02
3	.2074E-02	.8286E-02	.4516E-04	.7991E-04	.4001E-01
4	.7995E-04	.3856E-01	.3111E-06	.4719E-06	.5905E-02
5	.4824E-06	.6034E-02	.6810E-08	.1102E-07	.2366E-01
6	.1102E-07	.2283E-01	.5234E-11	.8576E-10	.7779E-02
7	.9126E-11	.8285E-03	.3077E-12	.8607E-10	.1004E+01
$B_0 = .63I$					
1	.1325E+01		.4879E+00	.4876E+00	
2	.4703E+00	.3551E+00	.4691E-01	.8533E-01	.1750E+00
3	.7501E-01	.1595E+00	.1630E-01	.2900E-01	.3398E+00
4	.2576E-01	.3434E+00	.3081E-02	.5786E-02	.1996E+00
5	.5718E-02	.2220E+00	.9685E-03	.1506E-02	.2603E+00
6	.1333E-02	.2331E+00	.2020E-03	.3692E-03	.2451E+00
7	.3532E-03	.2650E+00	.5661E-04	.9058E-04	.2454E+00
8	.8372E-04	.2370E+00	.1259E-04	.2186E-04	.2413E+00
9	.2034E-04	.2429E+00	.3386E-05	.5589E-05	.2557E+00
10	.5219E-05	.2566E+00	.7557E-06	.1295E-05	.2316E+00
11	.1214E-05	.2325E+00	.2047E-06	.3360E-06	.2595E+00
12	.3085E-06	.2542E+00	.4539E-07	.7897E-07	.2350E+00
13	.7538E-07	.2443E+00	.1233E-07	.1990E-07	.2520E+00
14	.1806E-07	.2395E+00	.2772E-08	.4874E-08	.2449E+00
15	.4655E-08	.2578E+00	.7392E-09	.1194E-08	.2450E+00
16	.1086E-08	.2333E+00	.1695E-09	.3017E-09	.2527E+00
17	.2792E-09	.2571E+00	.4441E-10	.1118E-09	.3707E+00

Tables 3 and 4 show that not only is $i_N(\varepsilon)$ quite different for αI and H_{uu} but also that the convergence behavior is qualitatively different. The numbers tabulated are as follows:

$$\sigma_i = \|s_i^N\|_N, \quad \gamma_i = \|G^N(u_i^N)\|_N,$$

$$\eta_i = \|u_i^N - \bar{u}_*^N\|_N, \quad \bar{v}^N = \{v(t_i)\}_{i=0}^{2N}.$$

The sequence σ_i/σ_{i-1} illustrates the difference between linear and superlinear convergence behavior in Tables 3 and 4. The linear rate of convergence which is observed in Tables 3 and 4 even for the case where B_0 is not H_{uu} , is to be expected from known theory (see e.g. Dennis [5]). Since the optimal control u_* for the infinite-dimensional problem is known, we also list the distance between u_i^N and u_* . The numbers for η_i indicate that for $B_0 = \alpha I$ the last three iterations are redundant because the predicted and also the actual magnitude of the discretization error exceeds the distance between u_i^N and \bar{u}_*^N . This effect is delayed when the accuracy is increased. In Table 4 also observe that an error bound of 10^{-9} which is reasonable from the discretization results in a large difference in the iteration count:

$$i_{200}(10^{-9}) = 6 \quad \text{for } B_0 = H_{uu},$$

$$i_{200}(10^{-9}) = 15 \quad \text{for } B_0 = 0.63I.$$

The computations were done on the Cyber 205 at Purdue University in single precision. Single precision on the 205 is roughly 14 place decimal accuracy.

REFERENCES

- [1] C. G. BROYDEN, *A new double-rank minimization algorithm*, AMS Notices, 16 (1969).
- [2] C. G. BROYDEN, J. E. DENNIS AND J. J. MORÉ, *On the local and superlinear convergence of quasi-Newton methods*, J. Inst. Math. Appl., 12 (1973), pp. 223-246.
- [3] J. W. DANIEL, *The conjugate gradient method for linear and nonlinear operator equations*, Ph.D. dissertation, Stanford University, Stanford, CA, 1965.
- [4] W. C. DAVIDON, *Variable metric methods for minimization*, Report ANL-90, Argonne National Laboratories, Argonne, IL, 1959.
- [5] J. E. DENNIS, *Towards a unified convergence theory for Newton-like methods*, in Nonlinear Functional Analysis and Applications, L. B. Rall, ed., Academic Press, New York, 1971, pp. 425-427.
- [6] J. C. DUNN AND E. W. SACHS, *The effect of perturbations on the convergence rates of optimization algorithms*, Appl. Math. Optim., 10 (1983), pp. 143-157.
- [7] E. R. EDGE AND W. F. POWERS, *Function-space quasi-Newton algorithms for optimal control problems with bounded controls and singular arcs*, J. Optim. Theory Appl., 20 (1976), pp. 455-479.
- [8] R. FLETCHER, *A new approach to variable metric algorithms*, Comput. J., 13 (1970), pp. 317-322.
- [9] R. FLETCHER AND M. J. D. POWELL, *A rapidly convergent descent method for minimization*, Comput. J., 6 (1963), pp. 163-168.
- [10] Z. FORTUNA, *Some convergence properties of the conjugate gradient method in Hilbert space*, SIAM J. Numer. Anal., 16 (1979), pp. 380-394.
- [11] D. G. GOLDFARB, *A family of metric methods derived by variational means*, Math. Comp., 24 (1970), pp. 23-26.
- [12] A. GRIEWANK, *The local convergence of Broyden's method on Lipschitzian problems in Hilbert spaces*, SIAM J. Numer. Anal., 24 (1987), pp. 684-705.
- [13] W. A. GRUVER AND E. SACHS, *Algorithmic Methods in Optimal Control*, Pitman, Boston-London-Melbourne, 1980.
- [14] W. W. HAGER, *Rates of convergence for discrete approximations to unconstrained control problems*, SIAM J. Numer. Anal., 13 (1976), pp. 449-472.
- [15] L. B. HORWITZ AND P. E. SARACHIK, *Davidon's method in Hilbert space*, SIAM J. Appl. Math., 16 (1968), pp. 676-695.
- [16] C. T. KELLEY AND E. W. SACHS, *Broyden's method for approximate solution of nonlinear integral equations*, J. Integral Equations, 9 (1985), pp. 25-43.
- [17] ———, *A new quasi-Newton method for some differential equations*, SIAM J. Numer. Anal., 24 (1987), pp. 516-531.
- [18] R. V. MAYORGA AND V. H. QUINTANA, *A family of variable metric methods in function space without exact line searches*, J. Optim. Theory Appl., 31 (1980), pp. 303-329.
- [19] E. POLAK, *Computational Methods in Optimization*, Academic Press, New York and London, 1971.
- [20] ———, *An historical survey of computational methods in optimal control*, SIAM Rev., 15 (1973), pp. 553-584.
- [21] E. W. SACHS, *Newton's method for singular constrained optimization problems*, Appl. Math. Optim., 11 (1984), pp. 247-276.
- [22] S. P. SETHI AND G. L. THOMPSON, *Optimal Control Theory*, Martinus Nijhoff, Boston-The Hague-London, 1981.
- [23] D. F. SHANNO, *Conditioning of quasi-Newton methods for function minimization*, Math. Comp., 24 (1970), pp. 647-657.
- [24] J. STOER, *Two examples on the convergence of certain rank-2 minimization methods for quadratic functionals in Hilbert space*, Linear Algebra Appl., 28 (1979), pp. 217-222.
- [25] H. TOKUMARU, N. ADACHI AND K. GOTO, *Davidon's method for minimization problems in Hilbert space with an application to control problems*, SIAM J. Control, 8 (1970), pp. 163-178.
- [26] P. R. TURNER AND E. HUNTLEY, *Variable metric methods in Hilbert space with applications to control problems*, J. Optim. Theory Appl., 19 (1976), pp. 381-400.
- [27] ———, *Direct-prediction quasi-Newton methods in Hilbert space with applications to control problems*, J. Optim. Theory Appl., 21 (1977), pp. 199-211.
- [28] J. WERNER, *Über die Konvergenz des Davidon-Fletcher-Powell-Verfahrens für streng konvexe Minimierungsaufgaben im Hilbertraum*, Computing, 12 (1974), pp. 167-176.
- [29] R. WINTHER, *A numerical Galerkin method for a parabolic problem*, Ph.D. dissertation, Cornell University, Ithaca, NY, 1977.

EXISTENCE OF OVERTAKING SOLUTIONS TO INFINITE DIMENSIONAL CONTROL PROBLEMS ON UNBOUNDED TIME INTERVALS*

D. A. CARLSON[†], A. HAURIE[‡] AND A. JABRANE[‡]

Abstract. The optimal control of an infinite-dimensional linear control system with unbounded time interval is considered. The results obtained are an extension to a Hilbert space setting of the results reported in Brock and Haurie [Math. Oper. Res., 1 (1976), pp. 337-346]. In particular, it is shown that (under appropriate conditions) overtaking optimal trajectories asymptotically approach a unique optimal steady state. This is the so-called *turnpike* property found in the economics literature. When this condition is combined with an associated optimal control problem, results concerning the existence of overtaking optimal solutions are obtained. Both distributed and boundary controls are considered. Finally we demonstrate the applicability of our results by considering several examples arising in the context of economics and resources management.

Key words. overtaking solutions, optimal control, distributed parameter and boundary control

AMS(MOS) subject classifications. Primary 49A10; secondary 90A

Introduction and motivation. The theory of optimal control of lumped parameter systems defined on an infinite time horizon stemmed from two main fields of applications:

- (i) the regulator problem where a stabilizing feedback with constant gain is to be designed for a given linear system, and
- (ii) the optimal economic growth model where a path of optimal capital accumulation is searched for, which permits an optimal asymptotic sustainable consumption in the economy.

A classical reference for the first problem is [23] and for the second class of problems is [1].

The extension of the regulator problem to a class of systems having infinite-dimensional state and controls (e.g. distributed parameter systems) has retained considerable attention in the systems theory literature. As examples of such generalizations we cite [25] and [12].

In the present paper we are interested in obtaining an extension to a Hilbert space setting of the existence results obtained in [3], [4], [8] and [24] for overtaking solutions of infinite horizon control problems. This previous work is clearly related to the second field of application, as it is based on an intensive use of convex analysis and it implies, as a by-product, the fundamental elements of the *turnpike* property of optimal economic growth models. This property states that, under sufficient regularity and convexity conditions, a unique optimal steady state is an asymptotic attractor of all optimal trajectories. Another distinctive feature in [8] is that it establishes a link between the optimal control problem considered and an associated convex problem of Lagrange, as introduced in [30], the solution of the second problem being also the desired solution of the first problem.

* Received by the editors November 5, 1985; accepted for publication (in revised form) December 30, 1986. This research was supported by Natural Sciences and Engineering Research Council of Canada grant A-4952, Social Science and Humanities Research Council of Canada grant 410-83-1012, Formation de Chercheurs et Action de Recherche Quebec grant EQ-428, and National Science Foundation grant DMS-8521465.

[†] Department of Mathematics, Southern Illinois University, Carbondale, Illinois 62901. Part of this author's research was done while he was visiting Groupe d'Etudes et de Recherche en Analyse des Décisions, under financial support from Direction de la Recherche, Ecole des Hautes Etudes Commerciales.

[‡] Groupe d'Etudes et de Recherche en Analyse des Décisions, Ecole des Hautes Etudes Commerciales, 5255 avenue Decelles, Montréal, Québec, Canada H3T 1V6.

Some generalizations to a Hilbert space setting, of the work of Rockafellar, have already been made by Viorel Barbu [3]. However, to our knowledge, the concept of overtaking optimality, the existence of such optimal trajectories and the relationship to the turnpike property has not yet been the object of a complete study for infinite-dimensional systems.

There is a need for such a generalization, as more economic models deal with distributed parameter systems where the *space* variable is either associated with a regional distribution of the economic variables (e.g. investment and capital as in [22], advertisement and goodwill as in [13]) or associated with the age distribution of a population as in [21] and [20]). Indeed the extension of the *turnpike* property to such models, so intimately linked to dynamic economic models since the work of Dorfman, Samuelson and Solow [14], Gale [18], McKenzie [28], as well as the proof that there exists an optimal solution to the problem posed, should be an important step toward the diffusion of these models as valid economic paradigms.

The paper is organized as follows: in § 1 the optimal control problem is formulated for the case of distributed control. In § 2 the *turnpike* property is established. In § 3 the associated problem of Lagrange is defined, its solution is proved to exist, a fundamental inequality is established, and these properties are used for proving that the solution of the associated problem of Lagrange is also overtaking optimal solution for the original problem. In § 4 these results are extended to the case of mixed (boundary and distributed) control. Finally, § 5 presents two examples dealing with typical economic and resources management models.

1. The optimal control problem. We consider a system described by the following input-output relationship:

$$(1.1) \quad x(t) = S(t)x^0 + \int_0^t S(t-s)Bu(s) ds, \quad t \geq 0$$

where E and F are separable Hilbert spaces, $x^0 \in E$, $\{S(t); t \geq 0\}$ is a strongly continuous semigroup on E with generator A , $u(\cdot) \in L^2_{loc}(R^+, F)$, the space of all strongly measurable functions $u(\cdot): R^+ \rightarrow F$, which are square-integrable on every finite interval $[0, T]$, and $B: F \rightarrow E$ a bounded linear operator.

Thus $x(\cdot)$ is the *mild solution* of the state equation

$$(1.2a) \quad \dot{x}(t) = Ax(t) + Bu(t), \quad t \geq 0,$$

$$(1.2b) \quad x(0) = x^0$$

where A is a possibly unbounded, closed and densely defined operator in E .

In addition we know (see [2], [7]) that although a mild solution need not be absolutely continuous it does satisfy the following *mild* differential equation:

$$(1.3a) \quad \forall y \in \mathcal{D}(A^*), \quad \frac{d}{dt} \langle x(t), y \rangle = \langle x(t), A^*y \rangle + \langle Bu(t), y \rangle \quad \text{a.e. } t \geq 0,$$

$$(1.3b) \quad \lim_{t \rightarrow 0^+} \langle x(t), y \rangle = \langle x^0, y \rangle$$

where A^* is the adjoint operator associated with A , with domain $\mathcal{D}(A^*)$. We impose the following additional constraints on state and control:

$$(1.4) \quad \forall t \geq 0 \quad x(t) \in X, \quad \text{a convex and closed subset of } E$$

and

$$u(t) \in U(x(t)) \subset F \quad \forall t \geq 0 \quad \text{where } U(\cdot): X \rightarrow 2^F$$

is a point to set mapping which is convex valued and satisfies

$$(1.5) \quad \alpha U(x) + (1 - \alpha)U(x') \subset U(\alpha x + (1 - \alpha)x') \quad \forall x, x' \in X, \quad \forall \alpha \in [0, 1],$$

$$(1.6) \quad \{u_n \xrightarrow{w} u, x_n \xrightarrow{w} x, u_n \in U(x_n) \forall n\} \Rightarrow u \in U(x)$$

where \xrightarrow{w} stands for the weak convergence in E . Notice that (1.6) implies that $U(x)$ is weakly closed, for each x in X .

The performance of the system on any interval $[0, t]$, $t \geq 0$ is evaluated by the cost functional

$$(1.7) \quad x_0(t; u(\cdot), x(\cdot)) \triangleq \int_0^t f_0(x(s), u(s)) ds$$

where $f_0: E \times F \rightarrow R$ is a convex functional which is lower semicontinuous on $E \times F$ and satisfies the following growth condition: there exists $K_1 > 0$ and $K > 0$ such that

$$(1.8) \quad \|x\|^2 + \|u\|^2 > K_1 \Rightarrow f_0(x, u) \geq K(\|x\|^2 + \|u\|^2).$$

We call $\mathcal{S}(x^0)$ the set of pairs $(u(\cdot), x(\cdot))$ where

- (i) $u(\cdot) \in L_{loc}^2(R^+, F)$,
- (ii) $x(\cdot)$ is given by (1.1),
- (iii) $x(t) \in X$ and $u(t) \in U(x(t))$ for all $t \geq 0$.

Then $u(\cdot)$ is called an *admissible control* at x^0 , and $x(\cdot)$ is the *associated trajectory*.

Since we consider our model on an infinite horizon, we do not a priori assume the convergence of (1.7) as $t \rightarrow \infty$. Hence we need to consider the following weaker notions of optimality.

DEFINITION 1. $(\hat{u}(\cdot), \hat{x}(\cdot)) \in \mathcal{S}(x^0)$ is *overtaking* (resp. *weakly overtaking*) *optimal* at x^0 if for any other pair $(u(\cdot), x(\cdot)) \in \mathcal{S}(x^0)$

$$(1.9) \quad \liminf_{t \rightarrow \infty} \{x_0(t; u(\cdot), x(\cdot)) - x_0(t; \hat{u}(\cdot), \hat{x}(\cdot))\} \geq 0, \\ \text{(resp. } \limsup_{t \rightarrow \infty} \{x_0(t; u(\cdot), x(\cdot)) - x_0(t; \hat{u}(\cdot), \hat{x}(\cdot))\} \geq 0. \quad \square$$

These are the optimality concepts used in [8]. The weak overtaking optimality concept was also studied in [19] from the point of view of necessary conditions for optimality in lumped parameter systems. For other recent contributions to the theory of optimal control on infinite time horizon making use of these solution concepts see [9], [10], [11], [17] and [24].

2. The turnpike property. In this section we investigate the asymptotic convergence properties of optimal trajectories in the sense of Definition 1 in § 1. In the economics literature these are the so-called *turnpike* properties first introduced by Samuelson [32] in the context of economic growth theory.

Assume the following.

Assumption 1. The optimal steady state problem (OSSP) consisting of

$$\begin{aligned} & \text{Min } f_0(x, u) \\ & \text{s.t. } 0 = \langle x, A^*y \rangle + \langle Bu, y \rangle \quad \forall y \in \mathcal{D}(A^*), \\ & \quad x \in X, \\ & \quad u \in U(x) \end{aligned}$$

has a solution denoted (\bar{x}, \bar{u}) , with \bar{x} uniquely defined.

By the convexity assumptions already made on $f_0(\cdot, \cdot)$, X , and $U(\cdot)$, the OSSP is a convex programming problem in a Hilbert space. Thus there exists $\bar{p} \in \mathcal{D}(A^*)$ such that (see [31], [15])

$$(2.1) \quad f_0(\bar{x}, \bar{u}) \leq f_0(x, u) - (\langle x, A^* \bar{p} \rangle + \langle Bu, \bar{p} \rangle)$$

for every x in X and u in $U(x)$. Let $L_0: E \times F \rightarrow [0, +\infty]$ be defined by

$$(2.2) \quad L_0(x, u) = \begin{cases} f_0(x, u) - f_0(\bar{x}, \bar{u}) - \langle x, A^* \bar{p} \rangle - \langle Bu, \bar{p} \rangle & \text{if } x \in X \text{ and } u \in U(x), \\ +\infty & \text{otherwise.} \end{cases}$$

Then we have $L_0(\bar{x}, \bar{u}) = 0$. Furthermore, since L_0 differs from f_0 through an affine function of x and u , it still satisfies the growth property (1.8), with f_0 replaced by L_0 .

As a consequence of this property we have the following technical lemma, whose proof is relegated to § 7.

LEMMA 1. *If an admissible pair $(\tilde{x}(\cdot), \tilde{u}(\cdot)) \in \mathcal{S}(x_0)$ is such that*

$$(2.3) \quad \int_0^\infty L_0(\tilde{x}(t), \tilde{u}(t)) dt < \infty;$$

then necessarily $\tilde{x}(\cdot)$ is bounded and for every fixed $T > 0$ there exists a constant $C(T)$ s.t.

$$(2.4) \quad \forall t \geq 0 \quad \int_t^{t+T} \|\tilde{u}(s)\|^2 ds \leq C(T).$$

Proof. The proof is given in § 7.

We can now establish the *Weak Turnpike Theorem*.

THEOREM 1. *Under Assumption 1 if $(\tilde{u}(\cdot), \tilde{x}(\cdot)) \in \mathcal{S}(x^0)$ is such that*

$$(2.5) \quad \limsup_{T \rightarrow \infty} \left[\int_0^T (f_0(\tilde{x}(t), \tilde{u}(t)) - f_0(\bar{x}, \bar{u})) dt \right] = \alpha < \infty,$$

then necessarily

$$(2.6) \quad \frac{1}{T} \int_0^T \tilde{x}(t) dt \xrightarrow{w} \bar{x} \text{ as } t \rightarrow \infty.$$

Proof. First we show that (2.3) is satisfied. As a result of the nonnegativity of L_0 it suffices to exhibit a sequence $\{T_k\}$, $T_k \rightarrow \infty$ as $k \rightarrow \infty$, such that, for a given constant C

$$(2.7) \quad \forall k = 1, 2, \dots \quad \int_0^{T_k} L_0(\tilde{x}(t), \tilde{u}(t)) dt < C.$$

According to (2.5), (2.7) will be satisfied if for a given sequence $\{T_k\}$ the set

$$\{\langle \tilde{x}(T_k), \bar{p} \rangle : k = 1, 2, \dots\}$$

is bounded. We show now that one can obtain such a sequence.

We claim, first, that under Assumption 1 there exists a constant $\bar{C} > 0$ such that

$$(2.8) \quad \left\| \frac{1}{T} \int_0^T \tilde{x}(t) dt \right\| \leq \bar{C} \quad \forall T > 0.$$

Suppose the contrary, then there exists a sequence $\{T_k\}$ $T_k \rightarrow \infty$ as $k \rightarrow \infty$ such that

$$\left\| \frac{1}{T_k} \int_0^{T_k} \tilde{x}(t) dt \right\| \rightarrow +\infty \quad \text{as } k \rightarrow \infty.$$

Using Jensen's inequality on f_0 and relation (1.8) we obtain the following from (2.5)

$$(2.9) \quad K \left\| \frac{1}{T_k} \int_0^{T_k} \tilde{x}(t) dt \right\|^2 - f_0(\bar{x}, \bar{u}) \leq \frac{1}{T_k} \int_0^{T_k} \{f_0(\tilde{x}(t), \tilde{u}(t)) - f_0(\bar{x}, \bar{u})\} dt \leq \frac{\alpha}{T_k}.$$

As k tends to infinity the right-hand side in (2.9) goes to zero whereas the left-hand side goes to infinity. This is a contradiction, and so (2.8) holds.

Notice that a similar argument would show that there exists a positive constant \bar{C}' such that

$$(2.10) \quad \left\| \frac{1}{T} \int_0^T \tilde{u}(t) dt \right\| \leq \bar{C}' \quad \forall T > 0.$$

Suppose, now, that for every sequence $\{T_k\}$, the set $\{\langle x(T_k), \bar{p} \rangle : k = 1, \dots\}$ is unbounded or, equivalently, that

$$(2.11) \quad \forall C > 0, \exists \bar{T} > 0 \quad \text{s.t. } \forall t > \bar{T} \\ |\langle \tilde{x}(t), \bar{p} \rangle| > C.$$

Since $\tilde{x}(\cdot)$ is continuous, by (2.11), $\langle \tilde{x}(t), \bar{p} \rangle$ keeps a constant sign in $[\bar{T}, \infty)$ and therefore, for any $T > \bar{T}$

$$(2.12) \quad \left| \frac{1}{T} \int_{\bar{T}}^T \langle x(t), \bar{p} \rangle dt \right| = \frac{1}{T} \int_{\bar{T}}^T |\langle x(t), \bar{p} \rangle| dt.$$

Take C such that

$$(2.13) \quad C > \bar{C} \|\bar{p}\|.$$

Then, using (2.8) and (2.12), we obtain

$$(2.14) \quad \left| \frac{1}{T} \int_{\bar{T}}^T \langle \tilde{x}(t), \bar{p} \rangle dt \right| \leq \bar{C} \|\bar{p}\| + \frac{1}{T} \left\| \int_0^{\bar{T}} \tilde{x}(t) dt \right\| \cdot \|\bar{p}\|.$$

When T goes to infinity in (2.14) the R.H.S. tends to $\bar{C} \|\bar{p}\|$ whereas the L.H.S. remains greater than C , and this contradicts (2.13). This establishes (2.7).

Integrating the state equation in the mild form (2.3), we have

$$(2.15) \quad \frac{1}{T} \langle \tilde{x}(T) - x^0, z \rangle = \left\langle \frac{1}{T} \int_0^T \tilde{x}(t) dt, A^* z \right\rangle + \left\langle B \frac{1}{T} \int_0^T \tilde{u}(t) dt, z \right\rangle.$$

By Lemma 1, $\tilde{x}(\cdot)$ is bounded, and since X and $U(\cdot)$ are convex, any weak cluster point (x^*, u^*) of the set

$$\left\{ \left(\frac{1}{T} \int_0^T \tilde{x}(t) dt, \frac{1}{T} \int_0^T \tilde{u}(t) dt \right), T > 0 \right\}$$

is an admissible pair for the OSSP. By Jensen's inequality and the lower semicontinuity of f_0 , the inequality (2.5) yields

$$(2.16) \quad f_0(x^*, u^*) \leq f_0(\bar{x}, \bar{u}).$$

Therefore, by the uniqueness of \bar{x} (Assumption 1), we have from (2.16) that

$$x^* = \bar{x}$$

and this completes the proof. \square

Remark 1. The *weak turnpike property* relates to the weak convergence of the average state $(1/T) \int_0^T \tilde{x}(t) dt$ toward \bar{x} as soon as the inequality (2.5) is satisfied.

Remark 2. As a consequence of (2.16) the average control $(1/T) \int_0^T \tilde{u}(t) dt$ has to converge weakly toward an optimal control \bar{u} for the OSSP.

In order to obtain a stronger turnpike property we introduce the set

$$(2.17) \quad G = \{x \in E : \exists u \in F \text{ s.t. } L_0(x, u) = 0\}$$

and the following definition.

DEFINITION 2. Let \mathcal{F} be the family of all trajectories $x(\cdot) \in \mathcal{S}(X)$ such that

$$(2.18) \quad x(t) \in G \quad \text{a.e. on } [0, \infty).$$

We say that G has Property \mathcal{C} (for *Convergence*) if $x(t) \xrightarrow{w} \bar{x}$ as $t \rightarrow \infty$, uniformly in \mathcal{F} .

The following theorem is an adaptation of results obtained in [24] for lumped parameter systems.

THEOREM 2. Under Assumption 1, if G has the property \mathcal{C} and if a feasible pair $(\tilde{x}(\cdot), \tilde{u}(\cdot))$ is such that

$$(2.19) \quad \int_0^\infty L_0(\tilde{x}(t), \tilde{u}(t)) dt < \infty;$$

then, necessarily, $\tilde{x}(t)$ converges weakly to \bar{x} as $t \rightarrow \infty$.

Proof. Assume the contrary. Then, there exist $\bar{y} \in E$, $\bar{y} \neq 0$, times $t_k \rightarrow \infty$ and an $\varepsilon > 0$ such that

$$(2.20) \quad |\langle \tilde{x}(t_k) - \bar{x}, \bar{y} \rangle| \geq \varepsilon \quad \text{for all } k \geq 1.$$

Let t_0 be such that, for any $x(\cdot) \in \mathcal{F}$,

$$(2.21) \quad |\langle x(t) - \bar{x}, \bar{y} \rangle| < \frac{\varepsilon}{2} \quad \text{a.e. on } (t_0, \infty).$$

Define $u_k(t) = \tilde{u}(t_k - t_0 + t)$ and $x_k(t) = \tilde{x}(t_k - t_0 + t)$, $t \geq 0$. A simple calculation involving (1.1) shows that $(x_k(\cdot), u_k(\cdot))$ is a feasible pair such that $x_k(0) = \tilde{x}(t_k - t_0)$. Further, Lemma 1 implies that $\{u_k(\cdot)\}$ is bounded in $L^2(0, T, F)$ for any fixed $T > 0$, and $\{x(t_k - t_0)\}$ is bounded in E . Thus, extracting a subsequence if necessary, we may suppose that for each $T > 0$

$$(2.22) \quad \begin{aligned} u_k &\rightarrow u^* \quad \text{weakly in } L^2(0, T, F), \\ x_k(0) &\rightarrow x_0^* \quad \text{weakly in } E. \end{aligned}$$

Furthermore, it is easy to see that

$$(2.23) \quad x_k \rightarrow x^* \quad \text{weakly in } L^2(0, T; E)$$

and $(x^*(\cdot), u^*(\cdot))$ is a feasible pair such that

$$(2.24) \quad \lim_{t \rightarrow 0^+} \langle x^*(t), y \rangle = \langle x_0^*, y \rangle \quad \forall y \in \mathcal{D}(A^*).$$

Since $L_0 \geq 0$, then, using (2.19) we obtain

$$(2.25) \quad \int_0^T L_0(x_k(t), u_k(t)) dt = \int_{t_k - t_0}^{T + t_k - t_0} L_0(x(t), u(t)) dt \rightarrow 0 \quad \text{as } k \rightarrow \infty.$$

On the other hand the lower semicontinuity of f_0 implies that, for every $T > 0$, the convex function $\varphi: (x(\cdot), u(\cdot)) \rightarrow \int_0^T L_0(x(t), u(t)) dt$ is lower semicontinuous (l.s.c.) on $L^2(0, T, E) \times L^2(0, T, F)$. Since L_0 is convex and l.s.c., then for every $k \in \mathbb{R}$, the set

$$\phi = \left\{ (x, u) \in L^2(0, T, E) \times L^2(0, T, F): \int_0^T L_0(x(t), u(t)) dt \leq k \right\}$$

is convex and closed. It is well known that a convex set is closed iff it is weakly closed. Hence ϕ is weakly closed and convex and then the function φ is weakly l.s.c. Hence (2.25), (2.22) and (2.23) imply that

$$\int_0^T L_0(x^*(t), u^*(t)) dt = 0 \quad \forall T > 0.$$

Since $L_0 \geq 0$, this implies that $L_0(x^*(t), u^*(t)) = 0$ a.e. on $[0, +\infty)$. Hence $x^*(\cdot)$ is in \mathcal{F} , and, according to Property \mathcal{C} , $x^*(t)$ converges weakly to \bar{x} as $t \rightarrow \infty$. For k sufficiently large we have

$$\begin{aligned} |\langle \tilde{x}(t_k) - \bar{x}, \bar{y} \rangle| &= |\langle x_k(t_0) - \bar{x}, \bar{y} \rangle| \\ &\leq |\langle x_k(t_0) - x^*(t_0), \bar{y} \rangle| + |\langle x^*(t_0) - \bar{x}, \bar{y} \rangle| \\ &\leq \varepsilon, \end{aligned}$$

which contradicts (2.20). \square

Remark 3. The turnpike property relates to the weak convergence of $\tilde{x}(t)$ toward \bar{x} as $t \rightarrow \infty$, when the inequality (2.19) holds.

Remark 4. If, instead of weak convergence, strong convergence (in the norm of E) is assumed in Property \mathcal{C} , then a strong turnpike property holds where $\|\tilde{x}(t) - \bar{x}\| \rightarrow 0$ as $t \rightarrow \infty$, when (2.19) holds.

The usefulness of Theorem 2 for infinite horizon optimal control problems is illustrated by the following corollary.

COROLLARY 1. In addition to the hypotheses given in Theorem 2, let us suppose that there exists a pair $(\tilde{u}(\cdot), \tilde{x}(\cdot))$ in $\mathcal{P}(x^0)$ such that

$$(2.26) \quad \int_0^\infty L_0(\tilde{x}(t), \tilde{u}(t)) dt < \infty.$$

Then if in the class of all bounded trajectories, there exists an overtaking optimal solution, say $(\hat{u}(\cdot), \hat{x}(\cdot))$, it happens that

$$\lim_{t \rightarrow +\infty} \hat{x}(t) = \bar{x} \quad \text{in the weak sense.}$$

Proof. Since $(\hat{u}(\cdot), \hat{x}(\cdot))$ is overtaking optimal we have

$$\liminf_{T \rightarrow \infty} \left\{ \int_0^T f_0(\tilde{x}(t), \tilde{u}(t)) - f_0(\hat{x}(t), \hat{u}(t)) dt \right\} \geq 0.$$

Thus we may write,

$$\begin{aligned} 0 \leq \liminf_{T \rightarrow \infty} \left[\int_0^T L_0(\tilde{x}(t), \tilde{u}(t)) - L_0(\hat{x}(t), \hat{u}(t)) \right. \\ \left. + \int_0^T \left\{ \frac{d}{dt} \langle \tilde{x}(t) - \hat{x}(t), \bar{p} \rangle \right\} dt \right]. \end{aligned}$$

Therefore, for any $\varepsilon > 0$ there exists $T(\varepsilon) > 0$ such that for all $T \geq T(\varepsilon)$,

$$-\varepsilon \leq \int_0^T L_0(\tilde{x}(t), \tilde{u}(t)) - L_0(\hat{x}(t), \hat{u}(t)) dt + \langle \tilde{x}(T) - \hat{x}(T), \bar{p} \rangle,$$

or equivalently,

$$0 \leq \int_0^T L_0(\hat{x}(t), \hat{u}(t)) dt < \varepsilon + \int_0^T L_0(\tilde{x}(t), \tilde{u}(t)) dt + \langle \tilde{x}(T) - \hat{x}(T), \bar{p} \rangle.$$

The boundedness of \hat{x} along with the assumption made on $(\tilde{u}(\cdot), \tilde{x}(\cdot))$ ensures that the right-hand side of the above inequality is bounded as $T \rightarrow +\infty$. Thus,

$$\int_0^{+\infty} L_0(\hat{x}(t), \hat{u}(t)) dt < \infty,$$

so that by Theorem 2 we must have

$$\lim_{t \rightarrow \infty} \hat{x}(t) = \bar{x} \quad \text{in the weak sense.} \quad \square$$

The assumption made on the pair $(\tilde{u}(\cdot), \tilde{x}(\cdot))$ given by (2.26) above is guaranteed for example if there exists a pair $(u(\cdot), x(\cdot)) \in \mathcal{S}(x^0)$ defined only on $[0, T]$ for some finite $T > 0$ satisfying the terminal condition $x(T) = \bar{x}$. Having such a control allows us to define $(\tilde{u}(\cdot), \tilde{x}(\cdot))$ in $\mathcal{S}(x^0)$ by the following:

$$(\tilde{u}(t), \tilde{x}(t)) = \begin{cases} (u(t), x(t)) & \text{if } 0 \leq t < T, \\ (\bar{u}, \bar{x}) & \text{if } T \leq t. \end{cases}$$

Conditions for such a controllability property are well known; see for example [2, § 4.9].

3. Sufficient conditions for existence of overtaking optimal solutions. In this section we associate with the originally defined control problem a so-called *Associated Problem of Lagrange* (APL). For this problem we are able to adapt an approach due to V. Barbu [3], [4], which shows the existence of a solution which is optimal in the usual sense. Under Assumption 1 (resp. Property \mathcal{C}) we will show that a solution to the APL leads to a weakly overtaking (resp. overtaking) optimal solution for the original problem.

We begin by defining the APL as consisting of minimizing the functional

$$(3.1) \quad \int_0^\infty L_0(x(t), u(t)) dt$$

over all pairs $(u(\cdot), x(\cdot)) \in \mathcal{S}(x^0)$.

Assumption 2. There exists $(\tilde{u}(\cdot), \tilde{x}(\cdot)) \in \mathcal{S}(x^0)$ such that

$$(3.2) \quad \int_0^\infty L_0(\tilde{x}(t), \tilde{u}(t)) dt < \infty.$$

LEMMA 2. Under Assumptions 1 and 2 there exists $(\hat{u}(\cdot), \hat{x}(\cdot)) \in \mathcal{S}(x^0)$ such that for all $(u(\cdot), x(\cdot)) \in \mathcal{S}(x^0)$

$$(3.3) \quad \int_0^\infty L_0(\hat{x}(t), \hat{u}(t)) dt \leq \lim_{T \rightarrow \infty} \int_0^T L_0(x(t), u(t)) dt.$$

Proof. Let

$$\varphi = \inf \left\{ \int_0^\infty L_0(x(t), u(t)) dt, (x, u) \in \mathcal{S}(x^0) \right\}.$$

By Assumption 2 φ is finite. Let $(x_n(\cdot), u_n(\cdot)) \in \mathcal{S}(x^0)$ be a minimizing sequence. For any fixed $T > 0$ we claim that the sequence $\{(x_n(\cdot), u_n(\cdot))\}$ is bounded in $L^2(0, T, E) \times L^2(0, T, F)$.

Suppose the contrary, i.e., there exists a divergent subsequence $\{(x_{n_k}(\cdot), u_{n_k}(\cdot))\}$. Consider the family of sets Ω_k defined by

$$\Omega_k = \{t \in [0, T]: \|x_{n_k}(t)\|^2 + \|u_{n_k}(t)\|^2 > K_1\}.$$

Obviously we have

$$(3.4) \quad \int_{\Omega_k} \{\|x_{n_k}(t)\|^2 + \|u_{n_k}(t)\|^2\} dt \rightarrow \infty \quad \text{as } k \rightarrow \infty.$$

On the other hand, the coercivity assumption (1.8) made on f_0 shows that, since L_0 is nonnegative,

$$(3.5) \quad \begin{aligned} K \int_{\Omega_k} \{\|x_{n_k}(t)\|^2 + \|u_{n_k}(t)\|^2\} dt &\leq \int_{\Omega_k} L_0(x_{n_k}(t), u_{n_k}(t)) dt \\ &\leq \int_0^\infty L_0(x_{n_k}(t), u_{n_k}(t)) dt. \end{aligned}$$

The right-hand side is finite, and thus contradicts (3.4). Therefore, for any fixed $T > 0$, the sequence $\{u_n\}$, respectively $\{x_n\}$, is bounded in $L^2(0, T, F)$, respectively $L^2(0, T, E)$.

By extracting a subsequence if necessary, we may find $(\hat{x}(\cdot), \hat{u}(\cdot))$ such that for each $T > 0$

$$(3.6) \quad u_n \rightarrow \hat{u} \quad \text{weakly in } L^2(0, T, F),$$

$$(3.7) \quad x_n \rightarrow \hat{x} \quad \text{weakly in } L^2(0, T, E),$$

and moreover, where $\hat{u} \in L_{\text{loc}}^2(\mathbb{R}^+, F)$ and $\hat{x}(t) = S(t)x^0 + \int_0^t S(t-s)B\hat{u}(s) ds$, for $t \geq 0$. The hypotheses made on $U(\cdot)$ and X show that $(\hat{x}(\cdot), \hat{u}(\cdot)) \in \mathcal{S}(x^0)$. Relations (3.6), (3.7) combined with the weak l.s.c. of the functions $\delta_T: L^2(0, T, E) \times L^2(0, T, F) \rightarrow [0, +\infty]$ defined by

$$\delta_T(x, u) = \int_0^T L_0(x(t), u(t)) dt,$$

shows that $\int_0^\infty L_0(\hat{x}(t), \hat{u}(t)) dt \leq \varphi$. Hence the result follows. \square

The next lemma establishes a comparison property between the performance criteria associated with two admissible pairs for the original control problem.

LEMMA 3. Let $(\hat{x}(\cdot), \hat{u}(\cdot))$ and $(x(\cdot), u(\cdot))$ be in $\mathcal{S}(x^0)$ and assume that $(\hat{x}(\cdot), \hat{u}(\cdot))$ is a solution to the APL. Then the following holds:

$$(3.8) \quad \begin{aligned} &\forall \varepsilon > 0, \exists T(\varepsilon) > 0 \quad \text{s.t. } \forall T \geq T(\varepsilon) \\ &\int_0^T \{f_0(\hat{x}(t), \hat{u}(t)) - f_0(x(t), u(t))\} dt < \varepsilon + \langle x(T) - \hat{x}(T), \bar{p} \rangle. \end{aligned}$$

Proof. The proof is similar to the one of Lemma 3.3 in [8], except for the fact that a calculus of variations formalism was used in [8]. \square

The next two theorems are then direct extensions of similar results obtained in [8] for lumped parameter systems.

THEOREM 3. *Under Assumptions 1 and 2 there exists a weakly overtaking optimal pair.*

Proof. By Lemma 2 there exists a solution x^* for the APL. Assume that x^* is not weakly overtaking optimal. Then there is a pair $(x(\cdot), u(\cdot)) \in \mathcal{S}(x^0)$ such that

$$(3.9) \quad \forall \varepsilon > 0, \quad \exists T_1 > 0: \quad \forall t > T_1 \\ \int_0^t f_0(x^*(s), u^*(s)) ds > \int_0^t f_0(x(s), u(s)) ds + \varepsilon.$$

By Lemma 3 we have

$$(3.10) \quad \forall \varepsilon' > 0, \quad \exists T_2 > 0 \quad \text{s.t.} \quad \forall t > T_2 \\ \int_0^t \{f_0(x^*(s), u^*(s)) - f_0(x(s), u(s))\} ds < \varepsilon' + \langle x(t) - x^*(t), \bar{p} \rangle;$$

taking $\varepsilon' = \varepsilon/2$ and $T \geq \max(T_1, T_2)$ we have

$$(3.11) \quad \frac{\varepsilon}{2} < \langle x(T) - x^*(T), \bar{p} \rangle.$$

Remembering that $\lim_{T \rightarrow \infty} (1/T) \int_0^T x^*(t) dt = \bar{x}$ in the weak sense, then, if the same property holds for $x(\cdot)$, we have

$$\frac{\varepsilon}{2} \leq \lim_{t \rightarrow \infty} \frac{1}{T} \int_0^T \langle x(t) - x^*(t), \bar{p} \rangle dt = 0,$$

which is a contradiction.

Suppose now that $\lim_{T \rightarrow \infty} (1/T) \int_0^T x(t) dt \neq \bar{x}$ in the weak sense. The contrapositive of Theorem 1 shows that

$$(3.12) \quad \int_0^\infty \{f_0(x(t), u(t)) - f_0(\bar{x}, \bar{u})\} dt = \infty.$$

A simple calculation, using (3.7), shows that for every $\varepsilon > 0$, there exists $T' > 0$ s.t. for every $T > T'$

$$(3.13) \quad \int_0^T L_0(x^*(t), u^*(t)) dt + \langle x^*(T) - x_0, \bar{p} \rangle - \varepsilon \\ > \int_0^T \{f_0(x(t), u(t)) - f_0(\bar{x}, \bar{u})\} dt.$$

Since $x^*(\cdot)$ is bounded and is a solution of the APL, (3.13) contradicts (3.10). \square

THEOREM 4. *If, in addition to the assumptions of the previous theorem, one assumes that the set G satisfies Property \mathcal{C} , then there exists an overtaking trajectory in the subclass $\mathcal{S}_B(x_0)$ consisting of bounded trajectories emanating from x_0 .*

Proof. By Lemma 2, there exists a solution x^* for the APL and by Lemma 1 $(x^*(\cdot), u^*(\cdot)) \in \mathcal{S}_B(x_0)$. Consider any other pair $(x(\cdot), u(\cdot)) \in \mathcal{S}_B(x^0)$ and define

$$(3.14) \quad e(T) = \int_0^T \{f(x^*(t), u^*(t)) - f(x(t), u(t))\} dt.$$

Using Lemma 3 and (3.14), we obtain

$$(3.15) \quad \forall \varepsilon > 0, \quad \exists T > 0: \quad \forall t \geq T \quad e(t) < \varepsilon + \langle x(t) - x^*(t), \bar{p} \rangle;$$

by Theorem 2, $x^*(t)$ converges weakly to \bar{x} as $t \rightarrow +\infty$. Assume that the same holds for $x(\cdot)$. Then taking t sufficiently large in (3.15), we have

$$\forall \varepsilon > 0, \quad \exists T > 0 \quad \text{s.t.} \quad \forall t \geq T$$

$$\int_0^t f_0(x^*(t), u^*(t)) dt < \int_0^t f_0(x(t), u(t)) dt + \varepsilon.$$

Thus x^* is overtaking optimal.

Now if $\lim_{t \rightarrow \infty} x(t) \neq \bar{x}$ in the weak sense then, by Theorem 2,

$$(3.16) \quad \int_0^\infty L_0(x(t), u(t)) dt = +\infty.$$

A simple calculation using (3.14) shows that

$$(3.17) \quad e(T) = \int_0^T \{L_0(x^*(t), u^*(t)) - L_0(x(t), u(t))\} dt + \langle x^*(T) - x(T), \bar{p} \rangle.$$

Since $x(\cdot)$ and $x^*(\cdot)$ are bounded and x^* is a solution for the APL, then, taking the limit in (3.17) and using (3.16) we have

$$\lim_{T \rightarrow \infty} e(T) = -\infty,$$

and therefore, x^* is overtaking optimal in $\mathcal{S}_B(x_0)$. \square

4. Extension to the case of systems with distributed and boundary controls. We now extend the existence results obtained in § 3 to the case of infinite-dimensional control problems with distributed and boundary control. By appealing to the boundary control theory of Fattorini [16] and Barbu [5] we transform the boundary control system to a linear evolution equation on a Hilbert space. This allows us to exploit the results of previous sections and give sufficient conditions for the “turnpike” property to hold as well as conditions for the existence of overtaking solutions.

We consider the following control system:

$$(4.1) \quad \dot{x}(t) = \sigma x(t) + B_1 u_1(t),$$

$$(4.2) \quad \gamma x(t) = B_2 u_2(t),$$

with initial condition

$$(4.3) \quad x(0) = x^0,$$

and the following additional constraints

$$(4.4) \quad x(t) \in X \subset E_1, \quad u_i(t) \in U_i(x(t)) \subset F_i, \quad i = 1, 2, t \in [0, \infty)$$

where E_1 , E_2 , F_1 and F_2 are separable Hilbert spaces, $x^0 \in E_1$, σ is a closed, linear and densely defined operator on E_1 , γ is a linear operator (the boundary operator) with domain in E_1 and range in E_2 , and $B_i: F_i \rightarrow E_i$, $i = 1, 2$, are linear continuous operators.

The sets X and $U_i(x)$, $i = 1, 2$, $x \in X$, satisfy the assumptions (1.4)–(1.6).

A solution to the system (4.1)–(4.3) satisfies the following input–output relation

$$(4.5) \quad x(t) = S(t)x^0 + \int_0^t S(t-s)(B_1 u_1(s) + \sigma B u_2(s)) ds - A \int_0^t S(t-s) B u_2(s) ds$$

where the operators A , B and $S(\cdot)$ are defined in the assumptions below.

(A1) We assume that $\mathcal{D}(\sigma) \subset \mathcal{D}(\gamma)$ (here $\mathcal{D}(\cdot)$ denotes the domain), and that the restriction of γ to $\mathcal{D}(\sigma)$ is continuous w.r.t. the graph norm of $\mathcal{D}(\sigma)$.

(A2) The operator $A: \mathcal{D}(A) \subset E_1 \rightarrow E_1$, defined by

$$(4.6) \quad Ay = \sigma y,$$

$$(4.7) \quad y \in \mathcal{D}(A) \triangleq \{y \in \mathcal{D}(\sigma): \gamma y = 0\}$$

is the infinitesimal generator of a strongly continuous semigroup $\{S(t): t \geq 0\}$ on E_1 .

(A3) There exists a linear continuous operator $B: F_2 \rightarrow E_1$, such that

$$(4.8) \quad \sigma B \in \mathcal{L}(F_2, E_1),$$

$$(4.9) \quad \gamma(Bv) = B_2 v \quad \forall v \in F_2,$$

$$(4.10) \quad \|Bv\|_{E_1} \leq c \|B_2 v\|_{E_2} \quad \forall v \in F_2.$$

(A4) For each $t \geq 0$ and $v \in L^2_{\text{loc}}([0, \infty), F_2)$

$$(4.11) \quad \int_0^t S(t-s) B v(s) ds \in \mathcal{D}(A)$$

and there exists $\delta \in L^2_{\text{loc}}(0, \infty)$ such that

$$(4.12) \quad \|AS(t)B\| \leq \delta(t) \quad \text{a.e.},$$

$$(4.13) \quad \forall t \geq 0, \quad \forall h \in [0, 1] \quad \int_t^{t+h} \delta(s) ds \leq \mathcal{K}.$$

Under the above hypotheses, the expression (4.5) is well defined for $u_i(\cdot) \in L^2_{\text{loc}}([0, \infty), F_i)$, $i = 1, 2$, and agrees with the formulation presented in [5] by Barbu.

The performance of the system on any interval $[0, t]$, $t \geq 0$, is evaluated by the cost functional

$$(4.14) \quad x_0(t; u(\cdot), x(\cdot)) = \int_0^t f_0(x(s), u_1(s), u_2(s)) ds$$

where $f_0: E_1 \times F_1 \times F_2 \rightarrow \mathbb{R}$ is a convex lower semicontinuous functional, which satisfies the following coercivity assumption: there exist positive constants K_1 and K such that the following hold:

$$(4.15) \quad \|x\|^2 + \|u_1\|^2 + \|u_2\|^2 \geq K_1 \quad \text{implies} \quad \frac{f_0(x, u_1, u_2)}{\|x\|^2 + \|u_1\|^2 + \|u_2\|^2} \geq K.$$

The steady-state system corresponding to (4.1), (4.2) is defined by

$$(4.16) \quad 0 = \sigma x + B_1 u_1,$$

$$(4.17) \quad \gamma x = B_2 u_2.$$

Under (A2), (A3), this system can be rewritten as follows:

$$(4.18) \quad 0 = A(x - Bu_2) + B_1u_1 + \sigma Bu_2.$$

Assume the following.

Assumption 3. The optimal steady-state problem (OSSP) consisting of

$$\text{Min } f_0(x, u_1, u_2)$$

$$(4.19) \quad \text{s.t. } 0 = \langle x - Bu_2, A^*z \rangle + \langle B_1u_1 + \sigma Bu_2, z \rangle \quad \forall z \in \mathcal{D}(A^*),$$

$$(4.20) \quad x \in X, \quad u_2 \in U_1(x), \quad u_2 \in U_2(x)$$

has a unique solution $(\bar{x}, \bar{u}_1, \bar{u}_2)$.

THEOREM 5. Under Assumption 3 if $((u_1(\cdot), u_2(\cdot)), x(\cdot)) \in \mathcal{S}(x^0)$ is such that

$$(4.21) \quad \limsup_{T \rightarrow \infty} \int_0^T (f_0(x(t), u_1(t), u_2(t)) - f_0(\bar{x}, \bar{u}_1, \bar{u}_2)) dt < \infty,$$

then

$$(4.22) \quad \frac{1}{T} \int_0^T x(t) dt \xrightarrow{w} \bar{x} \quad \text{as } T \rightarrow \infty,$$

and

$$(4.23) \quad \frac{1}{T} \int_0^T u_i(t) dt \xrightarrow{w} \bar{u}_i, \quad i = 1, 2 \quad \text{as } T \rightarrow \infty.$$

Proof. Proceeding exactly as in Theorem 1, it is easy to show that any weak cluster point $(\tilde{x}, \tilde{u}_1, \tilde{u}_2)$ of the set

$$\left\{ \left(\frac{1}{T} \int_0^T x(t) dt, \frac{1}{T} \int_0^T u_1(t) dt, \frac{1}{T} \int_0^T u_2(t) dt \right) : T > 0 \right\}$$

satisfies

$$(4.24) \quad f_0(\tilde{x}, \tilde{u}_1, \tilde{u}_2) \leq f_0(\bar{x}, \bar{u}_1, \bar{u}_2).$$

It remains to show that $(\tilde{x}, \tilde{u}_1, \tilde{u}_2)$ satisfies (4.19). From (4.5) we may rewrite $x(t)$ as

$$(4.25) \quad x(t) = \alpha(t) + \beta(t) - A\eta(t)$$

where

$$(4.26) \quad \alpha(t) = S(t)x^0 + \int_0^t S(t-s)B_1u_1(s) ds,$$

$$(4.27) \quad \beta(t) = \int_0^t S(t-s)\sigma Bu_2(s) ds,$$

$$(4.28) \quad \eta(t) = \int_0^t S(t-s)Bu_2(s) ds.$$

The functions, $\alpha(\cdot)$, $\beta(\cdot)$, and $\eta(\cdot)$ are the mild solutions to the systems $\dot{\alpha}(t) = A\alpha(t) + B_1u_1(t)$, $\alpha(0) = x^0$; $\dot{\beta}(t) = A\beta(t) + \sigma Bu_2(t)$, $\beta(0) = 0$; $\dot{\eta}(t) = A\eta(t) + Bu_2(t)$, $\eta(0) = 0$, respectively. Hence for all $z \in \mathcal{D}(A^*)$

$$(4.29) \quad \langle \alpha(t), A^*z \rangle = \frac{d}{dt} \langle \alpha(t), z \rangle - \langle B_1u_1(t), z \rangle,$$

$$(4.30) \quad \langle\langle \beta(t), A^*z \rangle\rangle = \frac{d}{dt} \langle\langle \beta(t), z \rangle\rangle - \langle\langle \sigma B u_2(t), z \rangle\rangle.$$

Moreover, since $\eta(t) \in \mathcal{D}(A)$ by (A6), $\eta(\cdot)$ is a weak solution and so it satisfies

$$(4.31) \quad \langle\langle A\eta(t), y \rangle\rangle = \frac{d}{dt} \langle\langle \eta(t), y \rangle\rangle - \langle\langle B u_2(t), y \rangle\rangle \quad \forall y \in E.$$

Combining (2.14), (2.15), (2.16) gives, for all $z \in \mathcal{D}(A^*)$

$$(4.32) \quad \begin{aligned} \langle\langle x(t), A^*z \rangle\rangle &= \frac{d}{dt} \langle\langle \alpha(t) + \beta(t) - A\eta(t), z \rangle\rangle \\ &\quad - \langle\langle B_1 u_1(t), z \rangle\rangle - \langle\langle \sigma B u_2(t), z \rangle\rangle + \langle\langle B u_2(t), A^*z \rangle\rangle. \end{aligned}$$

Notice that we apply (4.31) with $y = A^*z$. Integrating (4.32) and using (4.25), we obtain

$$(4.33) \quad \begin{aligned} \frac{1}{T} \int_0^T \langle\langle x(t), A^*z \rangle\rangle dt &= \frac{1}{T} \langle\langle x(T) - x^0, z \rangle\rangle \\ &\quad - \frac{1}{T} \int_0^T \langle\langle B_1 u_1(t) + \sigma B u_2(t), z \rangle\rangle dt \\ &\quad + \frac{1}{T} \int_0^T \langle\langle B u_2(t), A^*z \rangle\rangle dt. \end{aligned}$$

Letting $T \rightarrow \infty$, we obtain (4.19) for any weak cluster point $(\tilde{x}, \tilde{u}_1, \tilde{u}_2)$, if $x(\cdot)$ is bounded. By taking account of (4.12) and (4.13) the proof of the boundedness of $x(\cdot)$ is a direct adaptation of the proof of Lemma 1. \square

Once again the convexity assumptions we have imposed allow us to define the nonnegative convex lower semicontinuous functional $L_0: E_1 \times F_1 \times F_2 \rightarrow \mathbb{R}$ by

$$(4.34) \quad L_0(x, u_1, u_2) = \begin{cases} f_0(x, u_1, u_2) - f_0(\bar{x}, \bar{u}_1, \bar{u}_2) - \langle\langle x - B u_2, A^* \bar{p} \rangle\rangle - \langle\langle B_1 u_1 + \sigma B u_2, \bar{p} \rangle\rangle \\ \quad \forall x \in X, \quad u_i \in U_i(x), \quad i = 1, 2, \\ +\infty \quad \text{otherwise} \end{cases}$$

where \bar{p} is a fixed element in $\mathcal{D}(A^*)$.

THEOREM 6. *Under Assumption 3, if G satisfies property \mathcal{C} then every pair $(u(\cdot), x(\cdot)) \in \mathcal{S}(x^0)$ which satisfies*

$$(4.35) \quad \int_0^\infty L_0(x(t), u_1(t), u_2(t)) dt < \infty$$

has the property that

$$(4.36) \quad x(t) \xrightarrow{w} \bar{x} \quad \text{as } t \rightarrow \infty.$$

Proof. The proof is identical to the proof of Theorem 2 in [2], with obvious modifications to account for the boundary control u_2 . \square

With these results it is now a routine task to obtain the desired existence results. The statements and the proofs of the existence theorems remain exactly the same as in § 3.

THEOREM 7. *If there exists $(\tilde{u}(\cdot), \tilde{x}(\cdot)) \in \mathcal{S}(x^0)$ such that (4.35) holds and if Assumption 3 holds then there exists a weakly overtaking optimal pair $((u_1^*(\cdot), u_2^*(\cdot)), x^*(\cdot))$.*

THEOREM 8. *If in addition to hypotheses of Theorem 7, the set G satisfies property \mathcal{C} , then there exists an overtaking optimal pair.*

5. Examples.

5.1. A regional economic growth model. The following model is a slight modification to a model proposed in [20, § 6.3].

$$(5.1) \quad \text{Min} \int_0^\infty \int_0^h f_0(C(t, y), K(t, y)) dy dt$$

$$(5.2) \quad \text{s.t.} \quad \frac{\partial}{\partial t} K(t, y) = I(t, y) + \alpha \frac{\partial^2}{\partial y^2} K(t, y) - \mu(y) K(t, y),$$

$$I(t, y) + C(t, y) \leq F(K(t, y)),$$

$$(5.3) \quad I(t, y) \geq 0, \quad C(t, y) \geq 0, \quad K(t, y) \geq 0,$$

$$\forall t \geq 0, \quad \forall y \in [0, h],$$

$$K(0, y) = K^0(y) \quad \text{given for } y \in [0, h],$$

$$(5.4) \quad \frac{\partial K}{\partial y}(t, 0) = \frac{\partial K}{\partial y}(t, h) = 0.$$

Here $K(t, y)$ represents the stock of capital in the economy at time t which is available at location y . The control variables $I(t, y)$ and $C(t, y)$ represent respectively the investment rate and consumption rate at time t , location y . $F(\cdot)$ is the production function, assumed to be concave and Lipschitz continuous, and $\mu(y)$ is the depreciation rate of capital at location y , assumed to be a continuous function of y . The boundary conditions represent the fact that all capital remains in the region $[0, h]$. The cost functional $f_0(C(t, y), K(t, y))$ is the negative of the utility derived from consumption and capitalization at (t, y) . We assume that f_0 is strictly convex, in C , decreasing and convex in K . The meaning of having K in the performance criterion could be related to, e.g., environmental costs associated with the presence of production equipments.

For a justification of the diffusion term in (4.14) we refer the reader to [20].

We now formulate this system in the framework considered here by introducing

$$E \triangleq L^2[0, h],$$

$$X \triangleq \{x \in E: x(y) \geq 0, x(y) \leq \tilde{K}\},$$

$$(5.5) \quad U(x) \triangleq \{u \in E \times E: u(y) = (I(y), C(y)) \geq 0, I(y) + C(y) \leq F(x(y))\},$$

$$A \triangleq \alpha \frac{\partial^2}{\partial y^2} - \mu(y) \quad \text{with} \quad \mathcal{D}(A) = \left\{ x \in E: \frac{\partial x}{\partial y}, \frac{\partial^2 x}{\partial y^2} \in E \text{ and } \frac{\partial x}{\partial y}(0) = \frac{\partial x}{\partial y}(h) = 0 \right\}.$$

The upper bound \tilde{K} for the state variable x which corresponds to the capital stock K is ensured by the classical Inada economic growth conditions (see [1]). This boundedness property is stronger than the coercivity condition (1.8). We intend to show that there exists an overtaking (or weakly overtaking) solution at any initial state $x_0 = K_0$

which is *strictly sustainable*, i.e., such that there exists a consumption function C_0 and an investment function I_0 such that, a.e. on $[0, h]$

$$\begin{aligned}
 (5.6) \quad & \frac{\partial^2 K_0(y)}{\partial y^2} - \mu K_0(y) + I_0(y) = 0, \\
 & I_0(y) + C_0(y) < F(K_0(y)), \\
 & I_0(y) > 0, \quad C_0(y) > 0, \\
 & \frac{\partial K}{\partial y}(0) = \frac{\partial K}{\partial y}(h) = 0.
 \end{aligned}$$

We shall establish the following properties: (i) the OSSP has a unique solution $(\bar{K}, \bar{C}, \bar{I})$, (ii) the steady state \bar{K} is asymptotically reachable from K_0 , and (iii) the set G has the Property \mathcal{C} if f_0 is also strictly convex in K . Properties (i) and (ii) imply the existence of a weakly overtaking solution at K_0 . Properties (i)–(iii) imply the existence of an overtaking solution of K_0 . (Theorems 3 and 4, respectively.)

(i) The OSSP is formulated as follows:

$$\begin{aligned}
 (5.7) \quad & \text{Min } \int_0^h f_0(K(y), C(y)) dy \\
 & \text{s.t. } \frac{\partial^2 K(y)}{\partial y^2} - \mu K(y) + I(y) = 0, \\
 & I(y) + C(y) \leq F(K(y)), \\
 & K(y) \geq 0, \quad I(y) \geq 0, \quad C(y) \geq 0, \\
 & \frac{\partial K}{\partial y}(0) = 0 = \frac{\partial K}{\partial y}(h).
 \end{aligned}$$

This is a standard optimal control problem for which there exists a solution (e.g. see Lee and Markus [27]). The strict convexity of f_0 w.r.t. C ensures the uniqueness of the optimal steady state consumption \bar{C} . Since f_0 is strictly decreasing in C , the output constraint is active at the optimum, and the optimal steady state solution satisfies

$$\begin{aligned}
 (5.8) \quad & \frac{\partial^2 \bar{K}(y)}{\partial y^2} - \mu \bar{K}(y) + F(\bar{K}(y)) - \bar{C}(y) = 0, \\
 & \bar{K}(y) \geq 0, \quad \bar{C}(y) \geq 0, \\
 & \frac{\partial \bar{K}}{\partial y}(0) = 0, \quad \frac{\partial \bar{K}}{\partial y}(h) = 0.
 \end{aligned}$$

Since F is Lipschitz, the uniqueness of \bar{C} implies the uniqueness of \bar{K} . This establishes that this system satisfies Assumption 1. The multiplier \bar{p} will be defined as the adjoint variable in the Pontryagin maximum principle for the control problem (5.7).

(ii) To establish the asymptotic reachability of \bar{K} from K_0 , let us consider, for a given $\alpha > 0$ the following function

$$(5.8) \quad \tilde{K}(t, y) = e^{-\alpha t} K_0(y) + (1 - e^{-\alpha t}) \bar{K}(y).$$

Let us also define the functions

$$(5.9) \quad \tilde{I}(t, y) = e^{-\alpha t} [I_0(y) - \alpha(K_0(y) - \bar{K}(y))] + (1 - e^{-\alpha t}) \bar{I}(y),$$

$$(5.10) \quad \tilde{C}(y, t) = e^{-\alpha t} [C_0(y) + \alpha(K_0(y) - \bar{K}(y))] + (1 - e^{-\alpha t}) \bar{C}(y).$$

If α is taken small enough, since K_0 is strictly sustainable (5.6), \tilde{I} and \tilde{C} remain nonnegative for all $t \geq 0$. Furthermore

$$(5.11) \quad \tilde{I}(t, y) + \tilde{C}(t, y) = e^{-\alpha t}(I_0(y) + C_0(y)) + (1 - e^{-\alpha t})(\bar{I}(y) + \bar{C}(y))$$

and, by the concavity of F ,

$$(5.12) \quad \tilde{I} + \tilde{C} \leq F(\tilde{K}).$$

It is straightforward to check that $(\tilde{K}, \tilde{I}, \tilde{C})$ constitute an admissible pair $(x, u) \in \mathcal{S}(x_0)$, which converges asymptotically toward \bar{K} .

Now let us consider the integral

$$(5.13) \quad J(T) = \int_0^T \int_0^h \{f_0(\tilde{C}(t, y), \tilde{K}(t, y)) - f_0(\bar{C}(y), \bar{K}(y))\} dy dt.$$

By the convexity of f_0 , we have

$$(5.14) \quad J(T) \leq \int_0^T \int_0^h e^{-\alpha t} [f_0(C_0(y) + \alpha(K_0(y) - \bar{K}(y)), K_0(y)) - f_0(\bar{C}(y), \bar{K}(y))] dy dt.$$

Clearly $J(T)$ remains bounded above when $T \rightarrow \infty$. This establishes Assumption 2.

Therefore Theorem 3 applies, and there exists a weakly overtaking optimal trajectory at K_0 .

(iii) Assume now that f_0 is also strictly convex in K . Then the set G reduces to the singleton $\{\bar{K}\}$ and, therefore, the Property \mathcal{C} is trivially verified. In that case, Theorem 4 guarantees the existence of an overtaking solution at K_0 .

5.2. The cattle ranching problem. This problem has been formulated in [13]. It concerns a cattle rancher who must decide the number of cattle in different age groups to be bought and sold at each instant in order to maximize his profit.

To conform to our notation we reformulate the problem as follows.

Let $x(t, y)$ denote the density of cattle of age y at time t , $y \in [0, h] \subset \mathbb{R}$. The cattle are slaughtered at age h . The control variable $u(t, y)$ is the number of y -aged cattle bought at time t ; when $u(t, y)$ is negative, that means a sale. The process is governed by the system:

$$(5.15) \quad \frac{\partial x}{\partial t} = -\frac{\partial x}{\partial y} + u,$$

$$(5.16) \quad x(0, y) = x^0(y)$$

where $x^0(\cdot): [0, h] \rightarrow \mathbb{R}$ is the given initial state which is assumed to satisfy

$$(5.17) \quad x^0(0) = 0.$$

The birth rate of the cattle at any time t is assumed to be 0, which means that the following boundary condition holds:

$$(5.18) \quad x(t, 0) = 0, \quad t \geq 0.$$

In addition, we assume the state constraints

$$(5.19) \quad 0 \leq x(t, y) \leq \alpha \quad \text{a.e. } t \geq 0 \quad \forall y \in [0, h]$$

for a given $\alpha > 0$ which represents the ranch capacity constraint. The control $u(t, y)$ is supposed to be nonnegative which means that no cattle is sold below the age of maturity.

$$(5.20) \quad 0 \leq u(t, y) \quad \text{a.e. } t \geq 0 \quad \forall y \in [0, h].$$

The objective function for the cattle rancher is given by

$$(5.21)$$

$$x_0(T; u(\cdot), x(\cdot)) = \int_0^T \left[\int_0^h \{P(y)u(t, y) + C(y)x(t, y) + D[u(t, y) - d(y)]^2\} dy - Qx(t, h) \right] dt$$

where $P(y)$ is the unit price of cattle of age y , $C(y)$ is the unit feeding cost of cattle of age y , Q is the selling price of mature cattle (of age h), and D is a positive constant associated with a quadratic cost for deviation of the control from a desired level $d(y)$.

To formulate (5.18)–(5.22) as a problem of type (1.1) we set

$$E = L^2(0, h) \quad \text{with usual norm,}$$

$$A = -\frac{\partial}{\partial y} \quad \text{with domain } \mathcal{D}(A) = H_0^1(0, h),$$

the Sobolev space of functions which are $L^2(0, h)$ with derivatives, in the sense of distributions, in $L^2(0, h)$, and vanish at $y = 0$ ([25]). The solution of (5.15) with (5.16) is given by

$$(5.22) \quad x(t, \tau) = \begin{cases} x_0(\tau - t) + \int_0^t u(s, \tau - t + s) ds & \text{if } 0 \leq t \leq \tau, \\ \int_0^\tau u(t - \tau + s, s) ds & \text{if } t \geq \tau. \end{cases}$$

Introduce

$$(5.23) \quad X \triangleq \{x: x \in L^2(0, h), x(0) = 0, 0 \leq x \leq \alpha\},$$

$$(5.24) \quad U(x) \triangleq U = \{u: u \in L^2(0, h), 0 \leq u(y)\},$$

$$(5.25) \quad g(x, u) = \int_0^h \{P(y)u(y) + C(y)x(y) + D(u(y) - d(y))^2\} dy,$$

$$(5.26) \quad f_0(x, u) = -Qx(h) + g_0(x, u).$$

It can be readily checked that X is closed and convex in E , $U(\cdot) \equiv U$ satisfies (1.4)–(1.7), and g_0 is convex and continuous if $P(\cdot)$, $C(\cdot)$ and $d(\cdot)$ are continuous. However, f_0 is not well defined over $L^2(0, h) \times L^2(0, h)$ because of the boundary term $Qx(h)$. Nevertheless, the theory still holds with minor modifications. We shall show that (i) the OSSP is well defined and admits a unique solution, (ii) there exists a functional L'_0 , slightly different from L_0 introduced in § 2, (iii) the reachability of the optimal steady state and the Property \mathcal{C} are guaranteed.

(i) The OSSP can be formulated as follows:

$$(5.27) \quad \text{Min} \left[-Qx(h) + \int_0^h \{C(\tau)x(\tau) + P(\tau)u(\tau) + D(u(\tau) - d(\tau))^2\} d\tau \right]$$

$$(5.28) \quad \text{s.t.} \quad \frac{dx(\tau)}{d\tau} = u(\tau),$$

$$(5.29) \quad x(0) = 0,$$

$$(5.30) \quad 0 \leq x(\tau) \leq \alpha,$$

$$(5.31) \quad 0 \leq u(\tau).$$

This problem is well defined since (5.28), (5.29) admits a continuous solution yielding

$$(5.32) \quad x(h) = x(0) + \int_0^h u(\tau) d\tau.$$

The convex optimal control problem (5.27)–(5.31) admits a unique solution which satisfies the maximum principle. Thus there exists an adjoint function

$$\bar{p}(\cdot) : [0, h] \rightarrow \mathbb{R}$$

such that

$$(5.33) \quad \frac{d\bar{p}(y)}{dy} = -C(y) \quad \text{a.e. on } [0, h],$$

$$(5.34) \quad \bar{p}(h) = -Q$$

and such $\bar{u}(y)$ maximizes the concave function

$$(5.35) \quad -u(\tau)[\bar{p}(\tau) + P(\tau) - 2Dd(\tau) + Du(\tau)]$$

s.t. $u(\tau) \geq 0$. Therefore the OSSP has a unique solution given by

$$\begin{aligned} \bar{p}(\tau) &= -Q + \int_{\tau}^h C(s) ds, \\ \bar{u}(\tau) &= \begin{cases} 0 & \text{if } \bar{p}(\tau) + P(\tau) - 2Dd(\tau) > 0, \\ \frac{1}{2D}(2Dd(\tau) - \bar{p}(\tau) - P(\tau)) & \text{otherwise,} \end{cases} \\ \bar{x}(\tau) &= \int_0^{\tau} \bar{u}(s) ds, \quad \tau \in [0, h]. \end{aligned}$$

(ii) Let us consider the functional

$$(5.36) \quad L'_0(x, u) = g_0(x, u) - f_0(\bar{x}, \bar{u}) + \langle \bar{p}', x \rangle + \langle \bar{p}, u \rangle$$

where $\bar{p}'(\tau)$ stands for $d\bar{p}(\tau)/d\tau$. If we assume that x is a $\mathcal{D}(A)$ then $L'_0(x, u)$ must be nonnegative, since $x(h)$ is well defined, and (5.33) reduces to the separation property for the OSSP. Since $\mathcal{D}(A)$ is dense in E , $L'_0(x, u)$ is nonnegative in $E \times F$. The existence theory of § 3 can then be developed with L'_0 replacing L_0 .

(iii) Let $\tilde{u}(t, y)$ be defined by

$$\tilde{u}(t, y) = \begin{cases} 0 & \text{if } t < h, \\ \bar{u}(y) & \text{if } t \geq h. \end{cases}$$

Then it is easy to check that the associated state trajectory satisfies $\tilde{x}(t, y) = \bar{x}(y)$ for $t > 2h$. We thus have reachability of the steady state in a finite time. Property \mathcal{C} is trivially verified for L'_0 . Therefore there exists an overtaking solution for any initial condition x_0 .

5.3. Example with boundary control. To illustrate the application of the above results concerning the boundary control problems we present the following simple example. This example is the infinite horizon analogue of the "Mixed Dirichlet Problem" considered in Barbu [5]. Specifically, we let $\Omega \subset \mathbb{R}^n$ be a bounded open set with smooth boundary Γ and we let X and U_2 denote fixed, closed bounded intervals of \mathbb{R} . The boundary-control system we consider is as follows:

$$\begin{aligned}
 (5.37) \quad & \frac{\partial x}{\partial t}(t, \tau) - \Delta_\tau x(t, \tau) = 0 \quad \text{for } (t, \tau) \in [0, \infty) \times \Omega, \\
 & x|_\Gamma = u_2 \quad \text{for } t \in [0, \infty), \\
 & x(0, \tau) = x^0(\tau) \quad \text{for } \tau \in \Omega, \\
 & x(t, \tau) \in X \quad \text{for } (t, \tau) \in [0, \infty) \times \Omega, \\
 & u_2(t, \tau) \in U_2 \quad \text{for } (t, \tau) \in [0, \infty) \times \Gamma.
 \end{aligned}$$

Following Barbu [5] we let $E_1 = F_1 = L^2(\Omega)$, $E_2 = H^{-1/2}(\Gamma)$ and $F_2 = L^2(\Gamma)$. Further, we define the operators $B_i: E_i \rightarrow F_i$, $i = 1, 2$, by $B_1 \triangleq 0$, $B_2 \triangleq I$ and we define $\sigma: \mathcal{D}(\sigma) \rightarrow E_1$ as $\sigma = \Delta_\tau$, $\mathcal{D}(\sigma) = \{x \in L^2(\Omega): \Delta x \in L^2(\Omega)\}$. We define the operator $\gamma: E_1 \rightarrow E_2$ as the "trace" operator $\gamma_0 y$ which is well defined and belongs to $H^{-1/2}(\Gamma)$ for each $y \in \mathcal{D}(\sigma)$. Finally, we take $A: \mathcal{D}(A) \rightarrow E_1$ to be $A = \Delta$, $\mathcal{D}(A) = H_0^1(\Omega) \cap H^2(\Omega)$, and we define the linear operator $B: F_2 \rightarrow F_1$ by $Bu = w_u$, where $w_u \in L^2(\Omega)$ is the unique solution (generalized) to the Dirichlet boundary-value problem

$$\Delta w_u = 0 \quad \text{in } \Omega, \quad w_u|_\Gamma = u.$$

With this notation it follows, exactly as in [5], that the hypotheses (A1)–(A4) are satisfied. Further, since X and U_2 are compact intervals the Assumptions 1 and 2 are trivially met.

To describe the performance of the system on any interval $[0, t]$ we let $g: \Omega \times \mathbb{R} \rightarrow \mathbb{R}$ and $h_0: U_2 \rightarrow \mathbb{R}$ be two given functions satisfying the following conditions:

(i) g is continuous and convex in x , measurable in τ , and satisfies the growth condition

$$|g(\tau, x)| \leq c|x|^2 + \zeta(\tau) \quad \text{a.e. in } \Omega \times \mathbb{R}$$

where $c > 0$ is constant and $\zeta \in L^2(\Omega)$.

(ii) h_0 is convex and lower semicontinuous.

With these functions we define $f_0: E_1 \times F_2 \rightarrow \mathbb{R}$ by

$$f_0(x, u_2) = \int_\Omega g(\tau, x(\tau)) \, d\tau + \int_\Gamma h(u_2(\sigma)) \, d\sigma.$$

This function is seen to be lower semicontinuous (see [5]). To summarize, the optimal control problem we consider consists of "minimizing" the integral

$$\int_0^{+\infty} f_0(x(t), u_2(t)) \, dt$$

over all pairs of functions $(u_2(\cdot), x(\cdot))$ in $L^2_{\text{loc}}([0, \infty); L_2(\Gamma)) \times C([0, \infty); L_2(\Omega))$ satisfying the relations (1.5) and (1.4) (with the obvious modifications to account for the absence of the distributed control u_1).

The optimal steady-state problem consists of minimizing

$$f_0(x, u_2)$$

over all $(u_2, x) \in L_2(\Gamma) \times L_2(\Omega)$ satisfying the Dirichlet boundary-value problem

$$\begin{aligned}\Delta_\tau x(\tau) &= 0 \quad \text{in } \Omega, \\ x|_\Gamma &= u_2\end{aligned}$$

as well as the constraints

$$x(\tau) \in X \quad \text{and} \quad u_2(\tau) \in U_2.$$

For $u_2 \in L^2(\Omega)$, the Dirichlet boundary-value problem above has a unique generalized solution $\bar{x} \in L^2(\Omega)$. Therefore, if the interval X is taken sufficiently large there exists feasible pairs (\bar{u}_2, \bar{x}) and so the optimal steady-state problem will have solutions. To assert uniqueness we need only require strict convexity of f_0 .

Consequently, the existence results of § 4 are applicable if one can show there exists a feasible pair (\tilde{u}_2, \tilde{x}) satisfying

$$\int_0^{+\infty} L_0(x(t), u_2(t)) dt < \infty$$

where L_0 is given as in (2.19). To this end, take $\tilde{u}_2(t) = \tilde{u}$ a.e. and $\tilde{x}(\cdot)$ the corresponding trajectory emanating from x_0 . Since (\bar{u}_2, \bar{x}) is a steady state pair, it satisfies (5.3) with $x^0(\tau) = \bar{x}(\tau)$ and $u_2(t, \tau) = \bar{u}_2(\tau)$. Hence the difference $(\tilde{x}(t) - \bar{x})$ satisfies (5.3) with $u_2 \equiv 0$ and x_0 replaced by $x_0 - \bar{x}$ and then

$$\tilde{x}(t) - \bar{x} = S(t)(x_0 - \bar{x}).$$

Since the semigroup generated by Δ_τ is exponentially stable we have

$$\|\tilde{x}(t) - \bar{x}\| \leq M e^{-\mu t}$$

where M at μ are positive constants.

If the cost function g is quadratic, i.e.,

$$g(\tau, x(\tau)) = c(\tau)\|x(\tau)\|^2 + g(\tau),$$

then the exponential stability implies Assumption 2, therefore there exists an overtaking solution at any x_0 .

For more formulations of the function g , checking Assumption 2 would require more specific computations.

6. Conclusions. The results presented here extend the existence theory for overtaking optimal solutions of infinite horizon optimal control problems given in Brock and Haurie [8] to the case where the state dynamics are described by a mild solution of a linear control system on a Hilbert space. These results permit us to consider both distributed and boundary control problems, and their applicability is demonstrated by the examples given in § 5. In addition our work incorporates the recent advances to the theory given in Leizarowitz [24] which eliminate the compactness of the state constraint set, X , as well as the uniqueness of the optimal steady state. The compactness of X is removed by assuming that the cost integrand satisfies the coercivity relation (1.8) and the uniqueness of the optimal steady state is removed by assuming that the

set G (given in (2.7)) enjoys the convergence property \mathcal{C} . The set G utilized in [24] differs from ours. In our notation, the set G of [24] is a subset of $\mathcal{D}(A) \times E$ and is defined by

$$G = \{(x, z): \exists u \in U(x) \text{ such that } L_0(x, u) = 0 \text{ and } z = Ax + Bu\}.$$

In our setting $G \subset E$. Further, in [24] the convergence property \mathcal{C} is shown to hold if the differential inclusion

$$\dot{x}(t) \in \{z \in E: (x(t), z) \in G\}$$

has no “elliptic solutions” (see [24, Thm. 5.2]). Moreover the nonexistence of elliptic solutions is a generic property (see [24, Thm. 7.1]). At present it is not known if a similar result holds in the setting considered here. Further, as we consider only mild solutions of the control system, the arguments given in [24] clearly cannot be extended directly. Hence the question requires further investigation.

In closing, we mention that another direction for research would be the introduction of nonlinearities into the control system. From our remarks in the Introduction it is evident that as the development of distributed parameter economic models progresses, there will be a need to consider such problems.

7. Proof of Lemma 1. Let $(\tilde{u}, \tilde{x}) \in \mathcal{S}(x_0)$ be as indicated in Lemma 1. We first show that for each $T > 0$, there exists a constant $C(T) > 0$ such that

$$(2.4) \quad \int_t^{T+t} \|\tilde{u}(s)\|^2 ds < C(T).$$

To do this we proceed by contradiction and suppose there exists $T > 0$ and a sequence $\{t_k\}_{k=1}^\infty$ with $t_k \rightarrow \infty$ as $k \rightarrow \infty$ such that

$$\lim_{k \rightarrow \infty} \int_{t_k}^{t_k+T} \|\tilde{u}(s)\|^2 ds = +\infty.$$

For each index, k , we let $\Omega_k = \{t \in [t_k, t_k + T]: \|\tilde{u}(t)\|^2 \geq K_2\}$, where K_2 is as given in the growth condition (1.8) (with f_0 replaced by L_0) and let $\Omega'_k = [t_k, t_k + T] \setminus \Omega_k$. With this notation we observe that for each $k = 1, 2, \dots$,

$$\int_{\Omega_k} \|\tilde{u}(t)\|^2 dt \leq K_2 \text{ measure } (\Omega'_k) \leq K_2 T.$$

This implies that

$$\lim_{k \rightarrow \infty} \int_{\Omega_k} \|\tilde{u}(t)\|^2 dt = +\infty,$$

and so from the growth condition (1.8) and the nonnegativeness of L_0 , we arrive at

$$\begin{aligned} +\infty &= \lim_{k \rightarrow \infty} \int_{\Omega_k} \|\tilde{u}(t)\|^2 dt \\ &\leq \lim_{k \rightarrow \infty} \int_{\Omega_k} L_0(\tilde{x}(s), \tilde{u}(s)) ds \\ &\leq \int_0^{+\infty} L_0(\tilde{x}(s), \tilde{u}(s)) ds \\ &< +\infty, \end{aligned}$$

an obvious contradiction of (2.3). Hence the conclusion (2.4) holds.

We now show that $\|\tilde{x}(t)\|$ is bounded for all $t \geq 0$. To this end, we define for each $T > 0$, the set

$$\Omega_T = \{t \geq T: \|\tilde{x}(t)\|^2 > K_2\}$$

where K_2 is as indicated above. As a result of the growth condition (1.8), the nonnegativeness of L_0 , and the integrability condition (2.3) we have

$$\begin{aligned} 0 \leq \text{measure}(\Omega_T) &\leq \frac{1}{K_2} \int_{\Omega_T} \|\tilde{x}(t)\|^2 dt \\ &\leq \frac{1}{K_2} \int_{\Omega_T} L_0(\tilde{x}(t), \tilde{u}(t)) dt \\ &\leq \frac{1}{K_2} \int_T^\infty L_0(\tilde{x}(t), \tilde{u}(t)) dt, \end{aligned}$$

which implies

$$\lim_{T \rightarrow \infty} [\text{measure}(\Omega_T)] = 0.$$

Therefore, we can choose $T > 1$ sufficiently large so that

$$\text{measure}(\Omega_T) < \frac{1}{2}.$$

We now observe that since the map $t \rightarrow \|\tilde{x}(t)\|$ is continuous on $[0, T]$, there exists a constant $A_1 > 0$ so that

$$\|\tilde{x}(t)\| \leq A_1 \quad \text{on } [0, T].$$

Further, for $t \in [T, \infty) \setminus \Omega_T$ we have

$$\|\tilde{x}(t)\| \leq \sqrt{K_2}.$$

Hence the desired result follows if we can show $\|\tilde{x}(t)\|$ is bounded on Ω_T . To see this we observe that, since $\text{measure}(\Omega_T) < \frac{1}{2}$, for each $t \in \Omega_T$, there exists $h \in [0, 1]$ so that $t+h \notin \Omega_T$. Indeed if no such h exists, the interval $[t-h, t] \subset \Omega_T$ implying $\text{measure}(\Omega_T) > 1$. By appealing to the variation of constants formula (1.1) with $\sigma = t-h$ we obtain

$$\begin{aligned} \|\tilde{x}(t)\| &= \|\tilde{x}(\sigma+h)\| \\ &= \left\| S(\sigma+h)x_0 + \int_0^{\sigma+h} S(\sigma+h-s)B\tilde{u}(s) ds \right\| \\ &= \left\| S(h) \left[S(\sigma)x_0 + \int_0^\sigma S(\sigma-s)B\tilde{u}(s) ds \right] + \int_\sigma^{\sigma+h} S(\sigma+h-s)B\tilde{u}(s) ds \right\| \\ &= \left\| S(h)\tilde{x}(\sigma) + \int_\sigma^{\sigma+h} S(\sigma+h-s)B\tilde{u}(s) ds \right\|. \end{aligned}$$

Since $\{S(t): t \geq 0\}$ is a strongly continuous semigroup, there exists $C_1 > 0$ such that $\|S(t)\| \leq C_1$ for all $t \geq 0$. Hence, by the triangle inequality,

$$\begin{aligned} \|\tilde{x}(t)\| &\leq C_1 \|\tilde{x}(\sigma)\| + C_1 \|B\| \int_\sigma^{\sigma+h} \|\tilde{u}(s)\| ds \\ &\leq C_1 \|\tilde{x}(\sigma)\| + C_1 \|B\| \int_\sigma^{\sigma+1} \|\tilde{u}(s)\| ds \\ &\leq C_1 \|\tilde{x}(\sigma)\| + C_1 \|B\| \left(\int_\sigma^{\sigma+1} \|\tilde{u}(s)\|^2 ds \right)^{1/2} \end{aligned}$$

where the last inequality follows from the Hölder Inequality.

From the first part of our proof (with $T = 1$) there exists a constant $C = C(1) > 0$ such that for all $\sigma > 0$

$$\int_{\sigma}^{\sigma+1} \|\tilde{u}(s)\|^2 ds \leq C(1).$$

Therefore we obtain, with $C_2 = C_1 \|B\| \sqrt{C(1)}$,

$$\begin{aligned} \|\tilde{x}(t)\| &\leq C_1 \|\tilde{x}(\sigma)\| + C_2 \\ &= C_1 \|\tilde{x}(t-h)\| + C_2 \\ &\leq C_1 \max \{A_1, \sqrt{K_2}\} + C_2 \end{aligned}$$

for all $t \in \Omega_T$, and so the desired result follows.

REFERENCES

- [1] K. J. ARROW AND M. KURZ, *Public Investment, the Rate of Return, and Optimal Fiscal Policy*, Johns Hopkins Press, Baltimore, MD, 1970.
- [2] A. V. BALAKRISHNAN, *Applied Functional Analysis*, Springer-Verlag, Berlin-New York, 1976.
- [3] V. BARBU, *Convex control problems and Hamiltonian systems on infinite intervals*, this Journal, 16 (1978), pp. 895-911.
- [4] ———, *On convex control problems on infinite intervals*, J. Math. Anal. Appl., 65 (1978), pp. 687-702.
- [5] ———, *Boundary control problems with convex cost criterion*, this Journal, 18 (1980), pp. 227-243.
- [6] N. BOURBAKI, *Eléments de mathématique, Livre 4, Intégration*, Hermann, Paris, 1965.
- [7] H. BREZIS, *Opérateurs maximaux monotones et semi groupe de contractions dans les espaces de Hilbert*, North-Holland, Amsterdam-New York, 1973.
- [8] W. A. BROCK AND A. HAURIE, *On existence of overtaking optimal trajectories over an infinite time horizon*, Math. Oper. Res., 1 (1976), pp. 337-346.
- [9] D. A. CARLSON, *On the existence of optimal solutions for infinite horizon optimal control problems*, Ph.D. thesis, Univ. of Delaware, Newark, DE, 1983.
- [10] ———, *A Carathéodory-Hamilton-Jacobi theory for infinite horizon optimal control problems*, J. Optim. Theory Appl., 48 (1986), pp. 265-87.
- [11] ———, *On the existence of catching up optimal solutions for Lagrange problems defined on unbounded intervals*, J. Optim. Theory Appl., 49 (1986), pp. 207-225.
- [12] M. DELFOUR, *The largest class of hereditary systems defining a C_0 semigroup on the product space*, Canad. J. Math., 32 (1980) pp. 969-978.
- [13] N. DERZKO AND S. P. SETHI, *Distributed parameter systems approach to the optimal cattle ranching problem*, Optimal Control Appl. Methods, 1 (1980), pp. 3-10.
- [14] R. DORFMAN, P. A. SAMUELSON AND R. M. SOLOW, *Linear Programming and Economic Analysis*, McGraw-Hill, New York, 1958.
- [15] I. EKELAND AND R. TEMAN, *Convex Analysis and Variational Problems*, North-Holland, Amsterdam-New York, 1976.
- [16] H. O. FATTORINI, *Boundary control systems*, this Journal, 6 (1968), pp. 349-385.
- [17] C. D. FEINSTEIN AND D. G. LUENBERGER, *Analysis of the asymptotic behavior of optimal control trajectories: the implicit programming problem*, this Journal, 19 (1981), pp. 561-585.
- [18] D. GALE, *On optimal development in a multi-sector economy*, Rev. Econom. Stud., 34 (1967), pp. 1-18.
- [19] H. HALKIN, *Necessary conditions for optimal control problems with infinite horizon*, Econometrica, 42 (1974), pp. 267-273.
- [20] A. HAURIE, *Stability and optimal exploitations over an infinite time horizon of interacting populations*, Optimal Control Appl. Methods, 3 (1982), pp. 241-256.
- [21] A. HAURIE, S. SETHI AND R. HARTL, *Optimal control of an age-structured population model with applications to social services planning*, Large Scale Systems, 6 (1984), pp. 133-158.
- [22] W. ISARD AND P. LIOSSATOS, *Spatial Dynamics and Space-Time Development*, North-Holland, Amsterdam-New York, 1979.
- [23] R. E. KALMAN, *Contribution to the theory of optimal control*, Bol. Soc. Mat. Mexicana, (2) 5 (1960), pp. 102-119.
- [24] A. LEIZAROWITZ, *Existence of overtaking optimal trajectories for problems with convex integrands*, Math. Oper. Res., 10 (1985), pp. 450-461.

- [25] J. L. LIONS, *Optimal Control of Systems Governed by Partial Differential Equations*, Springer-Verlag, Berlin-New York, 1971.
- [26] L. A. LUSTERNIK AND V. J. SOBOLEV, *Elements of Functional Analysis*, Gordon and Breach, New York, 1962.
- [27] E. B. LEE AND L. MARKUS, *Foundations of Optimal Control Theory*, John Wiley, New York, 1967.
- [28] L. W. MCKENZIE, *Turnpike theory*, *Econometrica*, 44 (1976), pp. 841-866.
- [29] B. J. PETTIS, *On integration in vector spaces*, *Trans. Amer. Math. Soc.*, 44 (1938), pp. 277-304.
- [30] T. R. ROCKAFELLAR, *Saddle points of Hamiltonian systems in convex problems of Lagrange*, *J. Optim. Theory Appl.*, 12 (1973), pp. 367-390.
- [31] ———, *Convex Analysis*, Princeton University Press, Princeton, NJ, 1969.
- [32] P. A. SAMUELSON, *The Collected Scientific Papers of Paul A. Samuelson*, J. E. Stiglitz, ed., MIT Press, Cambridge, MA, 1966.

SENSITIVITY ANALYSIS OF CONTROL CONSTRAINED OPTIMAL CONTROL PROBLEMS FOR DISTRIBUTED PARAMETER SYSTEMS*

JAN SOKOŁOWSKI†

Abstract. The differential stability of solutions to control constrained quadratic optimal control problems for distributed parameter systems is investigated in this paper. The form of the right-derivatives of optimal controls for such problems with respect to the real parameter is derived. The differential sensitivity of optimal controls with respect to the perturbations of the coefficients of the state equation as well as to the deformations of the domain of integration is considered. The right-derivative of an optimal control with respect to the parameter is obtained in the form of an optimal control to the auxiliary control constrained optimal control problem.

Key words. distributed parameter system, sensitivity analysis, right-derivative of optimal control, Euler derivative, Lagrange derivative, shape sensitivity analysis

AMS(MOS) subject classification. 49B22

1. Introduction. This paper is devoted to the sensitivity analysis of a class of optimal control problems for distributed parameter systems. Sufficient conditions for the directional differentiability of an optimal control with respect to the parameter for an abstract quadratic control constrained optimal control problem are presented. The conditions are verified for a class of optimal control problems provided the related metric projection onto the set of admissible controls is directionally differentiable. The differentiable stability of solutions to convex, constrained optimal control problems is investigated by Malanowski [5] for a system of ordinary differential equations. The same method of the proof is used by Malanowski and Sokołowski [6] for the sensitivity analysis of a control problem for a system of elliptic equations.

Our approach is based on results of the sensitivity analysis of solutions to abstract variational inequalities and on differential stability of the metric projection onto the set of admissible controls. We refer the reader to Sokołowski [10], for related results on the sensitivity analysis of the convex, constrained optimization problems.

Since in many applications the functional parameters of a model for a distributed parameter system are not known exactly, the presented results on the sensitivity analysis can be used to compute increments of the optimal control corresponding to increments of the functional parameters of the model. Some related numerical results for an optimal control problem arising from air quality control are presented in [15]. On the other hand the results on the shape sensitivity analysis of an optimal control problem can be used for the optimal design of the distributed parameter systems. We refer the reader to [16] for related results on the shape sensitivity analysis of boundary optimal control problems for parabolic systems.

In this article we investigate the local sensitivity of optimal controls for two different examples. In § 3, a boundary optimal control problem for a system described by a parabolic equation is considered. The local sensitivity of an optimal control with respect to the perturbations of the boundary conditions is examined.

* Received by the editors August 30, 1983; accepted for publication (in revised form) December 28, 1986.

† Systems Research Institute of the Polish Academy of Sciences, ul. Newelska 6, 01-447 Warszawa, Poland.

In § 4, a distributed control problem for the Laplace equation is taken under consideration. The local sensitivity of an optimal control with respect to the deformations of the domain of integration is investigated. The existence of the so-called Euler (material) derivative of an optimal control in the direction of a vector field is shown. Furthermore, the form of the so-called Lagrange derivative of an optimal control is derived.

We refer the reader to Haraux [2] and Mignot [7] for abstract results on differential stability of solutions to variational inequalities. Applications of some abstract results of this type to the shape sensitivity analysis of elliptic free boundary problems are given by Sokołowski and Zolesio [11]–[13]. A compilation of results on sensitivity analysis in nonlinear programming is presented in Fiacco [1].

The outline of this paper is as follows. Section 2 describes the results obtained for an abstract optimal control problem. In § 3 the sensitivity analysis of a boundary optimal control problem of a parabolic initial-boundary problem is presented. Section 4 describes the results obtained for the shape sensitivity of a distributed optimal control problem for the Laplace equation. Standard notation [4] is used throughout this paper.

2. Sensitivity analysis of an optimal control problem. Our object is to determine the form of right derivative of an optimal control problem to the abstract optimal control problem (P^ε) defined in § 2.1 with respect to the parameter ε at $\varepsilon = 0$. Let u^ε denote a unique optimal solution to problem (P^ε) , $\varepsilon \in [0, \delta)$, $\delta > 0$. We prove the existence and uniqueness of an element q , the so-called sensitivity coefficient for the problem (P^0) , such that for $\varepsilon > 0$, ε small enough

$$(2.1) \quad u^\varepsilon = u^0 + \varepsilon q + o(\varepsilon).$$

2.1. Optimal control problem. We shall use the following notation. Let U, Y, W, H be Hilbert spaces. We identify U with its dual U' and we denote by $\Lambda \in \mathcal{L}(H, H')$ the canonical isomorphism. Let there be given linear, continuous mappings $L^\varepsilon \in \mathcal{L}(Y, W)$, $B^\varepsilon \in \mathcal{L}(U, W)$, $C^\varepsilon \in \mathcal{L}(Y, H)$, $N^\varepsilon \in \mathcal{L}(U, U)$, $\varepsilon \in [0, \delta)$ and elements $z^\varepsilon \in H$, $v^\varepsilon \in U$, $\varepsilon \in [0, \delta)$. We assume that there exists the inverse mapping

$$(2.2) \quad (L^\varepsilon)^{-1} \in \mathcal{L}(W, Y) \quad \forall \varepsilon \in [0, \delta)$$

and denote

$$(2.3) \quad S^\varepsilon = C^\varepsilon (L^\varepsilon)^{-1} B^\varepsilon \in \mathcal{L}(U, H),$$

$$(2.4) \quad A^\varepsilon = (S^\varepsilon)^* \Lambda S^\varepsilon + N^\varepsilon \in \mathcal{L}(U, U)$$

where $(S^\varepsilon)^* \in \mathcal{L}(H', U)$ denotes the adjoint mapping.

Let $N^\varepsilon \in \mathcal{L}(U, U)$ be a selfadjoint mapping for all $\varepsilon \in [0, \delta)$ such that $N^0 = NI$. I denotes identity mapping in U . We assume that there exists a constant $\varepsilon > 0$ such that

$$(2.5) \quad (N^\varepsilon v, v)_U \geq \alpha \|v\|_U^2 \quad \forall v \in U, \quad \forall \varepsilon \in [0, \delta).$$

Let us denote by U_{ad} the set of admissible controls. We assume that $U_{\text{ad}} \subset U$ is a closed and convex subset.

We shall consider an optimal control problem for the system described by the state equation of the form

$$(2.6) \quad L^\varepsilon y^\varepsilon(v) = B^\varepsilon v + \eta^\varepsilon, \quad \forall v \in U, \quad \forall \varepsilon \in [0, \delta)$$

where $\eta^\varepsilon \in W$, $\varepsilon \in [0, \delta]$ are given elements. We assume in the sequel that $\eta^\varepsilon = 0$, for all $\varepsilon \in [0, \delta)$.

Let us consider the following optimal control problem.

Problem (P^ε). Find an element $u^\varepsilon \in U_{\text{ad}}$ which minimizes the cost functional

$$(2.7) \quad J^\varepsilon(v) = \frac{1}{2} \|C^\varepsilon y^\varepsilon(v) - z^\varepsilon\|_H^2 + \frac{1}{2} (N^\varepsilon(v - v^\varepsilon), v - v^\varepsilon)_U$$

over the set U_{ad} .

It can be verified [3] that under our assumptions for a fixed parameter $\varepsilon \in [0, \delta)$, there exists a unique element $u^\varepsilon \in U_{\text{ad}}$ such that

$$(2.8) \quad J^\varepsilon(u^\varepsilon) \leq J^\varepsilon(v) \quad \forall v \in U_{\text{ad}}, \quad \forall \varepsilon \in [0, \delta).$$

It can be shown that the element u^ε is given by a unique solution of the following variational inequality:

$$(2.9) \quad \begin{aligned} u^\varepsilon &\in U_{\text{ad}}, \\ a^\varepsilon(u^\varepsilon, v - u^\varepsilon) &\geq (f^\varepsilon, v - u^\varepsilon)_U \quad \forall v \in U_{\text{ad}} \end{aligned}$$

where the bilinear form $a^\varepsilon(\cdot, \cdot): U \times U \rightarrow \mathbb{R}$ and the element $f^\varepsilon \in U$ are defined as given below:

$$(2.10) \quad a^\varepsilon(u, v) = (A^\varepsilon u, v)_U \quad \forall u, v \in U,$$

$$(2.11) \quad f^\varepsilon = (S^\varepsilon)^* \Lambda z^\varepsilon + N^\varepsilon v^\varepsilon.$$

Let us consider the variational inequality (2.9) for $\varepsilon = 0$. We denote by $\Pi: U \rightarrow U$ the mapping defined as follows:

For any $f \in U$ the element Πf is given by a unique solution of the variational inequality

$$(2.12) \quad \begin{aligned} \Pi f &\in U_{\text{ad}}, \\ a^0(\Pi f, v - \Pi f) &\geq (f, v - \Pi f)_U \quad \forall v \in U_{\text{ad}}. \end{aligned}$$

It can be proved that the element $\Pi f \in U$ minimizes the cost functional

$$(2.13) \quad I(v) = \frac{1}{2} \|S^0 v\|_H^2 + N \|v - f\|_U^2$$

over the set U_{ad} . Furthermore, $u^0 = \Pi f^0$. In order to derive the form of the sensitivity coefficient q in (2.1) we assume that there exists a continuous and positively homogeneous mapping

$$(2.14) \quad \Pi': U \rightarrow U$$

such that for $\tau > 0$, τ small enough

$$(2.15) \quad \forall v \in U: \Pi(f^0 + \tau v) = \Pi(f^0) + \tau \Pi'(v) + o(\tau) \quad \text{in } U$$

where $\|o(\tau)\|_U / \tau \rightarrow 0$ with $\tau \downarrow 0$.

Remark 1. In some cases the assumption (2.15) can be verified [10] using the related results [2], [7], [9] on the differential stability of the metric projection in space U onto the convex set U_{ad} . The explicit form of the mapping Π' is derived in Sokołowski [10] for the particular case of the so-called polyhedral set U_{ad} . We present an example of the polyhedral set U_{ad} in § 3.

THEOREM 1. Assume that condition (2.15) is verified; furthermore assume that for $\varepsilon > 0$, ε small enough,

$$(2.16) \quad A^\varepsilon = A^0 + \varepsilon A_1 + R_\varepsilon \quad \text{in } \mathcal{L}(U, U),$$

$$(2.17) \quad f^\varepsilon = f^0 + \varepsilon f_1 + r_\varepsilon \quad \text{in } U$$

where $A_1, R_\varepsilon \in \mathcal{L}(U, U)$, $f_1, r_\varepsilon \in U$ are given elements for $\varepsilon \in [0, \delta)$ such that $\|R_\varepsilon\|_{\mathcal{L}(U, U)}/\varepsilon \rightarrow 0$, $\|r_\varepsilon\|_U/\varepsilon \rightarrow 0$ with $\varepsilon \downarrow 0$. Then for $\varepsilon > 0$, ε small enough

$$(2.18) \quad u^\varepsilon = u^0 + \varepsilon \Pi'(f_1 - A_1 u^0) + o(\varepsilon) \quad \text{in } U$$

where $\|o(\varepsilon)\|_U/\varepsilon \rightarrow 0$ with $\varepsilon \downarrow 0$.

Proof. It can be proved that by (2.9), (2.16), (2.17) there exists a constant $C < +\infty$ such that

$$(2.19) \quad \|u^\varepsilon - u^0\|_U \leq C\varepsilon \quad \forall \varepsilon \in [0, \delta).$$

From (2.9), (2.16), (2.17), taking into account (2.19) we obtain

$$(2.20) \quad u^\varepsilon = \Pi(f^0 + \varepsilon(f_1 - A_1 u^0) + r(\varepsilon))$$

where $\|r(\varepsilon)\|_U/\varepsilon \rightarrow 0$ with $\varepsilon \downarrow 0$.

On the other hand, there exists a constant C_1 such that

$$(2.21) \quad \|\Pi(f_1) - \Pi(f_2)\|_U \leq C_1 \|f_1 - f_2\|_U \quad \forall f_1, f_2 \in U.$$

Hence

$$(2.22) \quad \Pi(f^0 + \varepsilon(f_1 - A_1 u^0) + r(\varepsilon)) = \Pi(f^0 + \varepsilon(f_1 - A_1 u^0)) + o(\varepsilon)$$

where $\|o(\varepsilon)\|_U/\varepsilon \rightarrow 0$ with $\varepsilon \downarrow 0$. We can conclude from (2.15), (2.20), (2.22) that

$$(2.23) \quad \begin{aligned} u^\varepsilon &= \Pi(f^0 + \varepsilon(f_1 - A_1 u^0)) + o(\varepsilon) \\ &= \Pi(f^0) + \varepsilon \Pi(f_1 - A_1 u^0) + o(\varepsilon) \\ &= u^0 + \varepsilon \Pi'(f_1 - A_1 u^0) + o(\varepsilon) \end{aligned}$$

which completes the proof.

Remark 2. Let us assume that for $\varepsilon \in [0, \delta)$:

$$(2.24) \quad N^\varepsilon = NI + \varepsilon N_1 + r_1(\varepsilon), \quad N_1, r_1(\varepsilon) \in \mathcal{L}(U, U),$$

$$(2.25) \quad S^\varepsilon = S^0 + \varepsilon S_1 + r_2(\varepsilon), \quad S_1, r_2(s) \in \mathcal{L}(U, H),$$

$$(2.26) \quad z^\varepsilon = z^0 + \varepsilon z_1 + r_3(\varepsilon), \quad z_1, r_3(\varepsilon) \in H,$$

$$(2.27) \quad v^\varepsilon = v^0 + \varepsilon v_1 + r_4(\varepsilon), \quad v_1, r_4(\varepsilon) \in U$$

where $r_i(\varepsilon)/\varepsilon \rightarrow 0$ with $\varepsilon \downarrow 0$ for $i = 1, \dots, 4$ in norms of spaces $\mathcal{L}(U, U)$, $\mathcal{L}(U, H)$, H , U , respectively.

It can be shown that (2.24)–(2.27) imply (2.16), (2.17). In this case the element $f_1 \in U$ and the mapping $A_1 \in \mathcal{L}(U, U)$ are defined as follows:

$$(2.28) \quad f_1 = (S^0)^* \Lambda z_1 + S_1^* \Lambda z_0 + N v_1 + N_1 v^0,$$

$$(2.29) \quad A_1 = S_1^* \Lambda S^0 + (S^0)^* \Lambda S_1 + N_1.$$

Remark 3. Theorem 1 is presented in a different setting in [9], [11], [12].

3. Sensitivity analysis of boundary optimal control problems. This section is devoted to the sensitivity analysis of a boundary optimal control problem (P^ε) for a parabolic equation. The functional coefficient $c_\varepsilon(\cdot, \cdot)$ in the mixed boundary conditions depends on the parameter $\varepsilon \in [0, \delta)$. We prove that the optimal control u^ε is right differentiable with respect to the parameter ε . Furthermore, the sensitivity coefficient q for the problem (P^0) is given by a unique solution of an auxiliary optimal control problem (Q) .

3.1. Parabolic equation. Let $\Omega \subset R^n$ be a given bounded domain with smooth boundary $\Gamma = \partial\Omega$. We denote $Q = \Omega \times (0, t_1)$, $\Sigma = \Gamma \times (0, t_1)$ where $t_1 > 0$ is a given constant. We shall use the following notation [7] for functional spaces:

$$(3.1) \quad U = H = L^2(\Sigma),$$

$$(3.2) \quad W = L^2(0, t_1; (H^1(\Omega))'),$$

$$(3.3) \quad Y = W(0, t_1) = \left\{ \phi \in L^2(0, t_1; H^1(\Omega)) \left| \frac{\partial \phi}{\partial t} \in L^2(0, t_1; (H^1(\Omega))') \right. \right\}.$$

We assume that there are given elements $c_0(\cdot, \cdot)$, $c_1(\cdot, \cdot)$, $r_\varepsilon(\cdot, \cdot) \in L^\infty(\Sigma)$, $\varepsilon \in [0, \delta)$ such that $c_0(x, t) \geq c > 0$ for a.e. $(x, t) \in \Sigma$, $\|r_\varepsilon\|_{L^\infty(\Sigma)}/\varepsilon \rightarrow 0$ with $\varepsilon \downarrow 0$.

We denote

$$(3.4) \quad c_\varepsilon(x, t) = c_0(x, t) + \varepsilon c_1(x, t) + r_\varepsilon(x, t), \quad (x, t) \in \Sigma.$$

Let $v \in L^2(\Sigma)$ be a given control. The state $y^\varepsilon(v) \in W(0, t_1)$ is given by a unique weak solution of the following parabolic equation:

$$(3.5) \quad \frac{\partial y}{\partial t}(v; x, t) - \Delta y^\varepsilon(v; x, t) = 0 \quad \text{in } Q,$$

$$(3.6) \quad \frac{\partial y}{\partial n}(v; x, t) + c_\varepsilon(x, t)y^\varepsilon(v; x, t) = v(x, t) \quad \text{on } \Sigma,$$

$$(3.7) \quad y^\varepsilon(v; x, 0) = 0 \quad \text{on } \Omega.$$

It can be shown [4] that $y^\varepsilon(v; \cdot, \cdot) \in H^{3/2, 3/4}(Q)$ therefore the trace $y^\varepsilon(v)|_\Sigma \in L^2(\Sigma)$ is well defined, hence a bounded linear mapping

$$(3.8) \quad S_\varepsilon : L^2(\Sigma) \ni v \rightarrow y^\varepsilon(v)|_\Sigma \in L^2(\Sigma)$$

is defined for all $\varepsilon \in [0, \delta)$. It can be verified [4] that the mapping (3.8) is compact and selfadjoint for every $\varepsilon \in [0, \delta)$. Furthermore, for $\varepsilon > 0$, ε small enough

$$(3.9) \quad S_\varepsilon = S_0 + \varepsilon S_1 + R_\varepsilon \quad \text{in } \mathcal{L}(L^2(\Sigma), L^2(\Sigma))$$

where $\|R_\varepsilon\|_{\mathcal{L}(L^2(\Sigma), L^2(\Sigma))}/\varepsilon \rightarrow 0$ with $\varepsilon \downarrow 0$. The linear, compact, selfadjoint mapping S_1 is defined as follows:

$$(3.10) \quad S_1 v = \overset{\text{def}}{z(v)}|_\Sigma \quad \forall v \in L^2(\Sigma)$$

where the element $z(v) \in W(0, t_1)$ is given by a unique weak solution of the parabolic equation

$$(3.11) \quad \frac{\partial z}{\partial t}(v; x, t) - \Delta z(v; x, t) = 0 \quad \text{in } Q,$$

$$(3.12) \quad \frac{\partial z}{\partial n}(v; x, t) + c_0(x, t)z(v; x, t) = -c_1(x, t)y^0(v; x, t) \quad \text{on } \Sigma,$$

$$(3.13) \quad z(v; x, 0) = 0 \quad \text{on } \Omega.$$

In this example the mapping (2.4) is defined by

$$(3.14) \quad A^\varepsilon = S_\varepsilon S_\varepsilon + NI \in \mathcal{L}(L^2(\Sigma), L^2(\Sigma))$$

where $N > 0$ is a given constant.

3.2. Boundary optimal control problem. Let us consider the following optimal control problem.

Problem (P^ε). Find an element $u^\varepsilon \in L^2(\Sigma)$ which minimizes the cost functional

$$(3.15) \quad I_\varepsilon(v) = \frac{1}{2} \|S_\varepsilon v - z_d\|_{L^2(\Sigma)}^2 + \frac{N}{2} \|v\|_{L^2(\Sigma)}^2$$

over the set of admissible controls of the form

$$(3.16) \quad U_{\text{ad}} = \{v \in L^2(\Sigma) | 0 \leq v(x, t) \leq M \text{ for a.e. } (x, t) \in \Sigma\}$$

where $M > 0$ is a given constant and $z_d \in L^2(\Sigma)$ is a given element.

A unique optimal control u^ε is given [8] by a unique solution of the following optimality system.

Find elements $(u^\varepsilon, y^\varepsilon, p^\varepsilon) \in U_{\text{ad}} \times W(0, t_1) \times W(0, t_1)$ which satisfy the following system:

State equation.

$$(3.17) \quad \begin{aligned} \frac{\partial y^\varepsilon}{\partial t}(u^\varepsilon; x, t) - \Delta y^\varepsilon(u^\varepsilon; x, t) &= 0 \quad \text{in } Q, \\ \frac{\partial y^\varepsilon}{\partial n}(u^\varepsilon; x, t) + c_\varepsilon(x, t)y^\varepsilon(u^\varepsilon; x, t) &= u^\varepsilon(x, t) \quad \text{on } \Sigma, \\ y^\varepsilon(u^\varepsilon; x, 0) &= 0 \quad \text{on } \Omega. \end{aligned}$$

Adjoint-state equation.

$$(3.18) \quad \begin{aligned} -\frac{\partial p^\varepsilon}{\partial t}(x, t) - \Delta p^\varepsilon(x, t) &= 0 \quad \text{in } Q, \\ \frac{\partial p^\varepsilon}{\partial n}(x, t) + c_\varepsilon(x, t)p^\varepsilon(x, t) &= y^\varepsilon(u^\varepsilon; x, t) - z_d(x, t) \quad \text{on } \Sigma, \\ p^\varepsilon(x, t_1) &= 0 \quad \text{on } \Omega. \end{aligned}$$

Optimality condition.

$$(3.19) \quad \int_{\Sigma} (Nu^\varepsilon(x, t) + p^\varepsilon(x, t))(v(x, t) - u^\varepsilon(x, t)) \, d\Sigma \geq 0 \quad \forall v \in U_{\text{ad}}.$$

The optimality condition (3.19) can be represented in the equivalent form

$$(3.20) \quad u^\varepsilon = P_{U_{\text{ad}}} \left(-\frac{1}{N} p^\varepsilon \right)$$

where for any element $f \in L^2(\Sigma)$ its metric projection $P_{U_{\text{ad}}}f$ in space $L^2(\Sigma)$ onto the set U_{ad} is given by a unique solution of the following variational inequality:

$$P_{U_{\text{ad}}}f \in U_{\text{ad}},$$

$$(3.21) \quad \int_{\Sigma} \{((P_{U_{\text{ad}}}f)(x, t) - f(x, t))(v(x, t) - (P_{U_{\text{ad}}}f)(x, t))\} \, d\Sigma \geq 0 \quad \forall v \in U_{\text{ad}}.$$

3.3. Sensitivity analysis. We derive the form of the sensitivity coefficient $q \in L^2(\Sigma)$ for the problem (P^0).

We will use the following notation:

$$(3.22) \quad \eta = S_1 z_1 - (S_0 S_1 + S_1 S_0) u^0, \quad \eta \in L^2(\Sigma),$$

$$(3.23) \quad \chi = \frac{1}{N} S_0 (z_d - S_0 u^0), \quad \chi \in L^2(\Sigma).$$

Furthermore, by $K(\Sigma) \subset L^2(\Sigma)$ we denote the closed, convex cone of the following form:

$$(3.24) \quad K(\Sigma) = \left\{ v \in L^2(\Sigma) \mid v(x, t) \geq 0 \text{ a.e. on } \Xi_0, v(x, t) \leq 0 \text{ a.e. on } \Xi_M, \right. \\ \left. \int_{\Xi_0 \cup \Xi_M} \chi(x, t) v(x, t) d\Sigma = 0 \right\}$$

where

$$(3.25) \quad \Xi_0 = \{(x, t) \in \Sigma \mid u^0(x, t) = 0\},$$

$$(3.26) \quad \Xi_M = \{(x, t) \in \Sigma \mid u^0(x, t) = M\}.$$

THEOREM 2. *There exists a unique sensitivity coefficient $q \in L^2(\Sigma)$ for the problem (P^0) , i.e., for $\varepsilon > 0$, ε small enough*

$$(3.27) \quad u^\varepsilon = u^0 + \varepsilon q + o(\varepsilon) \quad \text{in } L^2(\Sigma)$$

where $\|o(\varepsilon)\|_{L^2(\Sigma)}/\varepsilon \rightarrow 0$ with $\varepsilon \downarrow 0$.

The element q is given by a unique solution of the following optimal control problem.

Problem (Q). Find an element $q \in L^2(\Sigma)$ which minimizes the cost functional

$$(3.28) \quad I(v) = \frac{1}{2} \|S_0 v\|_{L^2(\Sigma)}^2 + \frac{N}{2} \left\| v - \frac{1}{N} \eta \right\|_{L^2(\Sigma)}^2$$

over the cone $K(\Sigma) \subset L^2(\Sigma)$.

Proof. We shall apply Theorem 1. To do this we must verify all assumptions of this theorem.

Since in our case $A^\varepsilon = S_\varepsilon S_\varepsilon + NI$ it follows from (3.9) that (2.16) holds with $A_1 = S_0 S_1 + S_1 S_0$. Since $f^\varepsilon = S_\varepsilon z_d$, it follows that (2.17) holds with $f_1 = S_1 z_d$.

Let us verify condition (2.15). For a given function $f \in L^2(\Sigma)$, the corresponding element $\Pi f \in L^2(\Sigma)$ minimizes the cost functional

$$(3.29) \quad I(v) = \frac{1}{2} \|S_0 v\|_{L^2(\Sigma)}^2 + \frac{N}{2} \left\| v - \frac{1}{N} f \right\|_{L^2(\Sigma)}^2$$

over the set (3.16).

It can be shown that there exists a constant $C < +\infty$ such that

$$(3.30) \quad \|\Pi(f^0 + \tau h) - \Pi f^0\|_{L^2(\Sigma)} \leq C\tau \|h\|_{L^2(\Sigma)}.$$

From (3.30) it follows that there exists an element $\Pi' \in L^2(\Sigma)$ such that for a subsequence still denoted $\tau \downarrow 0$

$$(3.31) \quad \Pi(f^0 + \tau h) = \Pi f^0 + \tau \Pi' + r(\tau) \quad \text{in } L^2(\Sigma)$$

where $r(\tau)/\tau \rightarrow 0$ weakly in $L^2(\Sigma)$ with $\tau \downarrow 0$. It can be shown [3] that for $\tau > 0$, τ small enough

$$(3.32) \quad P_{U_{ad}}(\chi + \tau v) = P_{U_{ad}}(\chi) + \tau P_{K(\Sigma)}(v) + o(\tau) \quad \text{in } L^2(\Sigma) \quad \forall v \in L^2(\Sigma)$$

where $\|o(\tau)\|_{L^2(\Sigma)}/\tau \rightarrow 0$ with $\tau \downarrow 0$.

Let us denote $r_1(\tau) = S_0 S_0 r(\tau)$, then we have $\|r_1(\tau)\|_{L^2(\Sigma)}/\tau \rightarrow 0$ with $\tau \downarrow 0$ since $S_0 \in \mathcal{L}(L^2(\Sigma), L^2(\Sigma))$ is compact.

Observe that

$$\begin{aligned} \Pi(f^0 + \tau h) &= P_{U_{ad}}\left(\frac{1}{N}(f^0 + \tau h - S_0 S_0 \Pi(f^0 + \tau h))\right) \\ (3.33) \quad &= P_{U_{ad}}\left(\frac{1}{N}(f^0 - S_0 S_0 u^0) + \frac{\tau}{N}(h - S_0 S_0 \Pi') + r_1(\tau)\right) \\ &= P_{U_{ad}}(\chi) + \tau P_{K(\Sigma)}\left(\frac{1}{N}(h - S_0 S_0 \Pi')\right) + o(\tau). \end{aligned}$$

From (3.31), (3.33) it follows that

$$(3.34) \quad \Pi' = P_{K(\Sigma)}\left(\frac{1}{N}(h - S_0 S_0 \Pi')\right),$$

i.e., $\Pi' = \Pi'(h)$ minimizes the cost functional

$$(3.35) \quad I(v) = \frac{1}{2} \|S_0 v\|_{L^2(\Sigma)}^2 + \frac{N}{2} \left\| v - \frac{1}{N} h \right\|_{L^2(\Sigma)}^2$$

over the set $K(\Sigma)$.

Denote $\eta = f_1 - A_1 u^0$ then by Theorem 1 we obtain $q = \Pi'(\eta)$.

4. Shape sensitivity analysis. This section is devoted to the sensitivity analysis of an optimal control problem (P_ε) for the Laplace equation defined in a domain $\Omega_\varepsilon \subset \mathbb{R}^n$. In this case the domain of integration depends on the parameter $\varepsilon \in [0, \delta)$.

We recall briefly [14] how to construct a family of domains $\{\Omega_\varepsilon\} \subset \mathbb{R}^n$, $\varepsilon \in [0, \delta)$, depending on a given vector field $V(\cdot, \cdot)$. We prove that an optimal control u_ε to problem (P_ε) is right differentiable with respect to the parameter ε .

The sensitivity coefficient for problem (P^0) is obtained in the form of an optimal control for the auxiliary optimal control problem.

4.1. Family of domains $\{\Omega_\varepsilon\}$. We define a family of domains $\{\Omega_\varepsilon\} \subset \mathbb{R}^n$ depending on the parameter $\varepsilon \in [0, \delta)$. Let $\Omega \subset \mathbb{R}^n$ be a given domain with a smooth boundary $\Gamma = \partial\Omega$. Let us assume that there is given a vector field $V(\cdot, \cdot) \in C^1([0, \delta_0]; C^2(\mathbb{R}^n, \mathbb{R}^n))$. We denote by $T_\varepsilon = T_\varepsilon(V)$ the mapping

$$(4.1) \quad T_\varepsilon(V): \mathbb{R}^n \ni X \rightarrow x(\varepsilon, X) \in \mathbb{R}^n$$

where the vector function $x(t) = x(t, X)$, $t \in [0, \delta)$, $X \in \mathbb{R}^n$, is given by the solution of the ordinary differential equation:

$$\begin{aligned} (4.2) \quad &\frac{d}{dt} x(t) = V(t, x(t)), \quad t \in (0, \delta), \\ &x(0) = X. \end{aligned}$$

The mapping (4.1) is defined for $\varepsilon \in [0, \delta)$, $\delta < \delta_0$. We denote by $DT_\varepsilon(X)$, $DT_\varepsilon^{-1}(X)$, ${}^*DT_\varepsilon^{-1}(X)$ the Jacobian matrix of the mapping (4.1) evaluated at point $X \in \Omega$, inverse

of matrix $DT_\varepsilon(X)$ and transposed matrix $DT_\varepsilon^{-1}(X)$, respectively. We denote by Ω_ε the domain of the form

$$(4.3) \quad \Omega_\varepsilon = \{x \in R^n \mid \exists X \in \Omega \text{ such that } x = x(\varepsilon, X)\}, \text{ i.e., } \Omega_\varepsilon = T_\varepsilon(V)(\Omega).$$

We will consider the Laplace equation defined in the domain Ω_ε

$$(4.4) \quad \begin{aligned} -\Delta y_\varepsilon(v_\varepsilon; x) &= v_\varepsilon(x) \quad \text{in } \Omega_\varepsilon, \\ y_\varepsilon(x) &= 0 \quad \text{on } \partial\Omega_\varepsilon \end{aligned}$$

where $v_\varepsilon \in L^2(\Omega_\varepsilon)$ is a given element. The solution to (4.4) is uniquely determined and belongs to $H_0^1(\Omega_\varepsilon) \cap H^2(\Omega_\varepsilon)$. Hence a bounded, selfadjoint, compact linear mapping

$$(4.5) \quad F_\varepsilon : L^2(\Omega_\varepsilon) \ni v_\varepsilon \rightarrow y_\varepsilon(v_\varepsilon) \in L^2(\Omega_\varepsilon)$$

is defined for any $\varepsilon \in [0, \delta)$.

4.2. Optimal control problem. Let $z_d \in H^1(R^n)$ be a given element. We denote $z_\varepsilon = z_d|_{\Omega_\varepsilon} \in H^1(\Omega_\varepsilon)$.

Let us consider the following optimal control problem.

Problem (P_ε) . Find an element $u_\varepsilon \in L^2(\Omega_\varepsilon)$ which minimizes the cost functional

$$(4.6) \quad J_\varepsilon(v_\varepsilon) = \frac{1}{2} \|F_\varepsilon v_\varepsilon - z_\varepsilon\|_{L^2(\Omega_\varepsilon)}^2 + \frac{N}{2} \|v_\varepsilon\|_{L^2(\Omega_\varepsilon)}^2$$

over the set of admissible controls

$$(4.7) \quad U_{\text{ad}}^\varepsilon = \{v_\varepsilon \in L^2(\Omega_\varepsilon) \mid 0 \leq v_\varepsilon(x) \leq M \text{ a.e. on } \Omega_\varepsilon\}$$

where $M > 0$ is a given constant.

It can be shown that an optimal control u_ε for the problem (P_ε) is given by a unique solution of the following optimality system:

$$(4.8) \quad \begin{aligned} -\Delta y_\varepsilon &= u_\varepsilon \quad \text{in } \Omega_\varepsilon, & y_\varepsilon &= 0 \quad \text{on } \partial\Omega_\varepsilon, \\ -\Delta p_\varepsilon &= y_\varepsilon - z_\varepsilon \quad \text{in } \Omega_\varepsilon, & p_\varepsilon &= 0 \quad \text{on } \partial\Omega_\varepsilon, \\ u_\varepsilon &\in U_{\text{ad}}^\varepsilon: \int_{\Omega_\varepsilon} (Nu_\varepsilon + p_\varepsilon)(v_\varepsilon - u_\varepsilon) dx \geq 0 \quad \forall v_\varepsilon \in U_{\text{ad}}^\varepsilon, \end{aligned}$$

hence the optimal control takes the following form:

$$(4.9) \quad u_\varepsilon(x) = \max \left\{ 0, \min \left\{ -\frac{1}{N} p_\varepsilon(x), M \right\} \right\}, \quad x \in \Omega_\varepsilon$$

since $p_\varepsilon \in H_0^1(\Omega_\varepsilon)$ from (4.9), it follows that

$$(4.10) \quad u_\varepsilon \in H_0^1(\Omega_\varepsilon) \quad \forall \varepsilon \in [0, \delta).$$

We denote by $\tilde{u}_\varepsilon \in H_0^1(R^n)$ an extension of the function $u_\varepsilon \in H_0^1(\Omega_\varepsilon)$:

$$(4.11) \quad \tilde{u}_\varepsilon(x) = \begin{cases} u_\varepsilon(x), & x \in \Omega_\varepsilon, \\ 0, & x \in R^n \setminus \Omega_\varepsilon, \end{cases} \quad \begin{matrix} \varepsilon \in [0, \delta), \\ \varepsilon \in [0, \delta). \end{matrix}$$

4.3. Sensitivity analysis. We will prove that the function $\tilde{u}_\varepsilon|_\Omega \in L^2(\Omega)$ is right differentiable with respect to the parameter ε at $\varepsilon = 0$. This means that there exists the so-called Lagrange derivative $u'(\Omega)$ of the optimal control u_ε in the direction of a vector field $V(\cdot, \cdot)$.

We will use the following notation. Let us consider the system of two coupled Laplace equations:

$$(4.12) \quad \begin{aligned} -\Delta y &= 0 \quad \text{in } \Omega, & y &= v_n \frac{\partial z}{\partial n} \quad \text{on } \partial\Omega, \\ -\Delta z &= \phi \quad \text{in } \Omega, & z &= 0 \quad \text{on } \partial\Omega \end{aligned}$$

where $\phi \in L^2(\Omega)$ is a given function, $n(x)$, $x \in \partial\Omega$ denotes the unit outward normal vector on $\partial\Omega$ and $v_n(x) = \langle V(0, x), n(x) \rangle_{R^n}$, $x \in \partial\Omega$ is the normal component of the vector field on $\partial\Omega$. The solution (y, z) to (4.12) is uniquely determined and belongs to $H^1(\Omega) \times (H_0^1(\Omega) \cap H^2(\Omega))$ hence a linear, selfadjoint, compact mapping

$$(4.13) \quad F_1 : L^2(\Omega) \ni \phi \rightarrow y(\phi) \in L^2(\Omega)$$

is defined.

We denote by $\eta, \chi \in L^2(\Omega)$ the elements:

$$(4.14) \quad \chi = \frac{1}{N} F_0(z_d - F_0 u_0),$$

$$(4.15) \quad \eta = F_1 z_d - (F_1 F_0 + F_0 F_1) u_0.$$

Furthermore we denote by $K(\Omega)$ the cone

$$(4.16) \quad K(\Omega) = \left\{ \phi \in L^2(\Omega) \left| \begin{aligned} &\phi(x) \geq 0 \text{ a.e. on } \Xi_0, \phi(x) \leq 0 \text{ a.e. on } \Xi_M, \\ &\int_{\Xi_0 \cup \Xi_M} \chi(x) \phi(x) dx = 0 \end{aligned} \right. \right\}.$$

THEOREM 3. *There exists a unique element $u'(\Omega) \in L^2(\Omega)$ such that for $\varepsilon > 0$, ε small enough:*

$$(4.17) \quad \tilde{u}_\varepsilon|_\Omega = u_0 + \varepsilon u'(\Omega) + o(\varepsilon) \quad \text{in } L^2(\Omega)$$

where $\|o(\varepsilon)\|_{L^2(\Omega)}/\varepsilon \rightarrow 0$ with $\varepsilon \downarrow 0$.

The Lagrange derivative $u'(\Omega)$ is given by a unique solution of the following optimal control problem.

Problem (P'). Find an element $u'(\Omega) \in L^2(\Omega)$ which minimizes the cost functional

$$(4.18) \quad I(v) = \frac{1}{2} \|F_0 v\|_{L^2(\Omega)}^2 + \frac{N}{2} \left\| v - \frac{1}{N} \eta \right\|_{L^2(\Omega)}^2$$

over the cone $K(\Omega) \subset L^2(\Omega)$.

Proof. The proof is divided into three parts.

Step 1. We apply Theorem 1 in order to prove that there exists a unique element $\dot{u}(\Omega) \in L^2(\Omega)$ such that

$$(4.19) \quad \lim_{\varepsilon \downarrow 0} \|(u_\varepsilon \circ T_\varepsilon - u_0)/\varepsilon - \dot{u}(\Omega)\|_{L^2(\Omega)} = 0.$$

The element $\dot{u}(\Omega)$ denotes the Euler (material) derivative of an optimal control u_ε in the direction of a vector field $V(\cdot, \cdot)$.

Step 2. Taking into account the relationship between the Lagrange derivative and the Euler derivative [14]:

$$(4.20) \quad u'(\Omega) = \dot{u}(\Omega) - \langle \text{grad } u_0, V(0) \rangle_{R^n}$$

we derive the form of the Lagrange derivative $u'(\Omega)$.

Step 3. We prove that the domain derivative $u'(\Omega)$ minimizes the cost functional (4.18) over the cone $K(\Omega)$.

Step 1. Euler Derivative $\dot{u}(\Omega)$. We define an optimal control problem (P^ε) such that a unique optimal control u^ε for this problem takes the form

$$(4.21) \quad u^\varepsilon = u_\varepsilon \circ T_\varepsilon \quad \forall \varepsilon \in [0, \delta).$$

To this end we transport the state equation (4.4), the cost functional (4.6) and the set of admissible controls (4.7) from the domain Ω_ε to the fixed domain Ω using the mapping (4.1). We obtain the state equation in the form of the following elliptic boundary-value problem

$$(4.22) \quad \begin{aligned} -\text{div}(B^\varepsilon(x) \cdot \text{grad } y^\varepsilon(v; x)) &= b^\varepsilon(x)v(x) \quad \text{in } \Omega, \\ y^\varepsilon(v; x) &= 0 \quad \text{on } \partial\Omega \end{aligned}$$

where $v \in L^2(\Omega)$ is a given element, $B^\varepsilon = \det(DT_\varepsilon)^* DT_\varepsilon^{-1}$. DT_ε^{-1} is a matrix function and $b^\varepsilon = \det(DT_\varepsilon)$, $\varepsilon \in [0, \delta)$.

The solution $y^\varepsilon(v)$ of the problem (4.21) is uniquely determined and belongs to $H_0^1(\Omega) \cap H^2(\Omega)$. Hence a compact, selfadjoint linear mapping

$$(4.23) \quad F^\varepsilon : L^2(\Omega) \ni v \rightarrow y^\varepsilon(v) \in L^2(\Omega)$$

is defined.

It can be shown that [14]

$$(4.24) \quad F^\varepsilon = F_0 + \varepsilon F' + R_\varepsilon \quad \text{in } \mathcal{L}(L^2(\Omega), L^2(\Omega))$$

where the mapping F_0 is defined by (4.5) for $\varepsilon = 0$, the mapping F' is defined below and $\|R_\varepsilon\|_{\mathcal{L}(L^2(\Omega), L^2(\Omega))} / \varepsilon \rightarrow 0$ with $\varepsilon \downarrow 0$. The mapping $F' \in \mathcal{L}(L^2(\Omega), L^2(\Omega))$ is defined in the following way:

For any element $\phi \in L^2(\Omega)$, the element $w = F'\phi$ is given by a unique solution of the following boundary value problem:

$$-\Delta w(x) = \text{div } V(0, x)\phi(x) + \text{div}(B'(x) \cdot \text{grad } z(x)) \quad \text{in } \Omega, \quad w(x) = 0 \quad \text{on } \partial\Omega.$$

Here $z = F_0\phi \in H_0^1(\Omega) \cap H^2(\Omega)$, and

$$B'(x) = \lim_{\varepsilon \downarrow 0} (B^\varepsilon(x) - B^0(x)) / \varepsilon = \text{div } V(0, x)I - *DV(0, x) - DV(0, x), \quad x \in \Omega.$$

$DV(0, x)$ is the Jacobian matrix of $V(0, x)$, $x \in \Omega$.

We define the cost functional $J^\varepsilon(v)$, $v \in L^2(\Omega)$ of the form

$$(4.25) \quad J^\varepsilon(v) = \frac{1}{2}(b^\varepsilon(F^\varepsilon v - z^\varepsilon), F^\varepsilon v - z^\varepsilon)_{L^2(\Omega)} + \frac{N}{2}(b^\varepsilon v, v)_{L^2(\Omega)}.$$

Here $z^\varepsilon = z_d \circ T_\varepsilon \in H^1(\Omega)$, and the set of admissible controls

$$(4.26) \quad U_{\text{ad}} = \{v \in L^2(\Omega) \mid 0 \leq v(x) \leq M \text{ a.e. on } \Omega\}.$$

Let us consider the optimal control problem.

Problem (P^ε) . Find an element $u^\varepsilon \in L^2(\Omega)$ which minimizes the cost functional (4.25) over the set (4.26).

It can be verified that a unique optimal control u^ε for the problem (P^ε) takes the form (4.21). Let us verify the assumptions of Theorem 1 in the case of problem (P^ε) . In this case $U = H = W = L^2(\Omega)$, $Y = H_0^1(\Omega) \cap H^2(\Omega)$, the mapping $A^\varepsilon \in \mathcal{L}(L^2(\Omega), L^2(\Omega))$ and the element $f^\varepsilon \in L^2(\Omega)$ are given by

$$(4.27) \quad A^\varepsilon = F^\varepsilon b^\varepsilon F^\varepsilon + N b^\varepsilon,$$

$$(4.28) \quad f^\varepsilon = F^\varepsilon b^\varepsilon z^\varepsilon.$$

It can be shown [14] that

$$(4.29) \quad b^\varepsilon = 1 + \varepsilon \operatorname{div} (V(0)) + o(\varepsilon) \quad \text{in } L^\infty(\Omega),$$

$$(4.30) \quad z^\varepsilon = z_d + \varepsilon \langle \operatorname{grad} z_d, V(0) \rangle_{R^n} + o(\varepsilon) \quad \text{in } L^2(\Omega),$$

therefore, by (4.24) and (4.27)–(4.30) it follows that the assumptions (2.16), (2.17) are satisfied, and we have

$$(4.31) \quad A_1 = F'F_0 + F_0F' + F_0\beta F_0 + N\beta,$$

$$(4.32) \quad f_1 = F'z_d + F_0\beta z_d + F_0 \langle \operatorname{grad} z_d, V(0) \rangle_{R^n}.$$

Here $\beta = \operatorname{div} (V(0))$, $V(0) = V(0, \cdot)$.

Finally, assumption (2.15) can be verified using exactly the same argument as in the proof of Theorem 2. The element $\Pi'(h)$ is a unique minimizer of the cost functional

$$(4.33) \quad I(v) = \frac{1}{2} \|F_0 v\|_{L^2(\Omega)}^2 + \frac{N}{2} \left\| v - \frac{1}{N} h \right\|_{L^2(\Omega)}^2$$

over the cone $K(\Omega)$ defined by (4.16), for any $h \in L^2(\Omega)$. Since all assumptions of Theorem 1 are verified, from (2.18) it follows that

$$(4.34) \quad \begin{aligned} \dot{u}(\Omega) = \Pi'(f_1 - A_1 u_0) = \Pi'(F'z_d + F_0\beta z_d + F_0 \langle \operatorname{grad} z_d, V(0) \rangle_{R^n} \\ + (F'F_0 + F_0F' + F_0\beta F_0 + N\beta)u_0). \end{aligned}$$

Step 2. Lagrange Derivative $u'(\Omega)$. From (4.20), (4.34) we obtain

$$(4.35) \quad \begin{aligned} u'(\Omega) = \Pi'(F'z_d + F_0\beta z_d + F_0 \langle \operatorname{grad} z_d, V(0) \rangle_{R^n} \\ + (F'F_0 + F_0F' + F_0\beta F_0 + N\beta)u_0) - \langle \operatorname{grad} u_0, V(0) \rangle_{R^n} \end{aligned}$$

and

$$(4.36) \quad u'(\Omega) \in K_v(\Omega) = \{\phi \in L^2(\Omega) \mid \exists \psi \in K(\Omega) \text{ such that } \phi = \psi - \langle \operatorname{grad} u_0, V(0) \rangle_{R^n}\}.$$

Observe that $\operatorname{grad} u_0(x) = 0$ a.e. $\Xi_0 \cup \Xi_M$ hence $K_v(\Omega) = K(\Omega)$ for any vector field V whence

$$(4.37) \quad u'(\Omega) \in K(\Omega).$$

From (4.35), (4.37) it follows that the Lagrange derivative $u'(\Omega) \in L^2(\Omega)$ is a unique minimizer of the cost functional

$$(4.38) \quad \begin{aligned} I_1(v) = \frac{1}{2} \|F_0 v + F_0 \langle \operatorname{grad} u_0, V(0) \rangle_{R^n}\|_{L^2(\Omega)}^2 \\ + \frac{N}{2} \left\| v + \langle \operatorname{grad} u_0, V(0) \rangle_{R^n} \right. \\ \left. - \frac{1}{N} ((F'F_0 + F_0F' + F_0\beta F_0 + N\beta)u_0 \right. \\ \left. + F'z_d + F_0\beta z_d + F_0 \langle \operatorname{grad} z_d, V(0) \rangle_{R^n}) \right\|_{L^2(\Omega)} \end{aligned}$$

over the cone $K(\Omega)$.

Therefore the Lagrange derivative is given by a unique solution of the following variational inequality:

$$(4.39) \quad \begin{aligned} u'(\Omega) &\in K(\Omega), \\ a_0(u'(\Omega), \phi - u'(\Omega)) &\geq G(u_0, V(0), \phi - u'(\Omega)) \quad \forall \phi \in K(\Omega) \end{aligned}$$

where

$$(4.40) \quad a_0(u, \phi) = (F_0 u, F_0 \phi)_{L^2(\Omega)} + N(u, \phi)_{L^2(\Omega)} \quad \forall u, \phi \in L^2(\Omega),$$

$$(4.41) \quad \begin{aligned} G(u_0, V(0), \phi) &= (f_1, \phi)_{L^2(\Omega)} - a_0(\langle \text{grad } u_0, V(0) \rangle_{R^n}, \phi) \\ &\quad - (A_1 u_0, \phi)_{L^2(\Omega)} \quad \forall \phi \in \{K(\Omega) - K(\Omega)\}. \end{aligned}$$

Step 3. We must prove that

$$(4.42) \quad G(u_0, V(0), \phi) = (\eta, \phi)_{L^2(\Omega)} \quad \forall \phi \in \{K(\Omega) - K(\Omega)\}.$$

We will use the same argument as in [13]. It can be shown [14] that for any vector field $V(\cdot, \cdot) \in C^1([0, \delta]; C^2(R^n, R^n))$ such that

$$(4.43) \quad v_n \stackrel{\text{def}}{=} \langle V(0), n \rangle_{R^n} = 0 \quad \text{on } \partial\Omega$$

we have $u'(\Omega) = 0$. Since the mapping

$$C^2(R^n, R^n) \ni V(0) \rightarrow G(u_0, V(0), \phi) \in R$$

is linear for all $\phi \in \{K(\Omega) - K(\Omega)\}$ it implies that $G(u_0, V(0), \phi) = 0$ for every $\phi \in \{K(\Omega) - K(\Omega)\}$ for any vector field $V(0)$ which satisfies (4.42). Hence there exists the distribution $g_n(\phi) \in \mathcal{D}'_1(\partial\Omega)$ such that

$$(4.44) \quad G(u_0, V(0), \phi) = \langle g_n(\phi), v_n \rangle_{\mathcal{D}'_1(\partial\Omega) \times \mathcal{D}_1(\partial\Omega)} \quad \forall \phi \in \{K(\Omega) - K(\Omega)\}.$$

From (4.41) by integration by parts, taking into account (4.43) we obtain (4.42) in the following way.

In what follows in view of (4.44), we neglect terms of the form

$$\int_{\Omega} \langle V(0, x), R(x) \rangle_{R^n} dx$$

where $R(\cdot) \in L^2(\Omega; R^n)$ is a given element. For example, we write

$$\begin{aligned} \int_{\Omega} \xi(x) \operatorname{div} V(0, x) dx &= - \int_{\Omega} \langle V(0, x), \operatorname{grad} \xi(x) \rangle_{R^n} dx + \int_{\partial\Omega} v_n(x) \xi(x) dx \\ &= \int_{\partial\Omega} v_n(x) \xi(x) dx + \dots \quad \text{for any } \xi \in H^1(\Omega) \\ &= 0 + \dots \quad \text{for any } \xi \in H^1_0(\Omega). \end{aligned}$$

Let us consider subsequently all terms in (4.41).

(i) For the first term

$$(f_1, \phi)_{L^2(\Omega)} = (F' z_d, \phi)_{L^2(\Omega)} + (F_0 \beta z_d, \phi)_{L^2(\Omega)} + (F_0 \langle \operatorname{grad} z_d, V(0) \rangle_{R^n}, \phi)_{L^2(\Omega)}$$

we have

$$\begin{aligned} (F' z_d, \phi)_{L^2(\Omega)} &= (F_1 z_d, \phi)_{L^2(\Omega)} + \dots, \\ (F_0 \beta z_d, \phi)_{L^2(\Omega)} &= 0 + \dots, \\ (F_0 \langle \operatorname{grad} z_d, V(0) \rangle_{R^n}, \phi)_{L^2(\Omega)} &= 0 + \dots, \end{aligned}$$

therefore,

$$(f_1, \phi) = (F_1 z_d, \phi)_{L^2(\Omega)} + \dots$$

(ii) For the second term

$$\begin{aligned} -a_0(\langle \text{grad } u_0, V(0) \rangle_{R^n}, \phi) &= -(F_0(\langle \text{grad } u_0, V(0) \rangle_{R^n}), F_0 \phi)_{L^2(\Omega)} \\ &\quad - (\langle \text{grad } u_0, V(0) \rangle_{R^n}, \phi)_{L^2(\Omega)} \\ &= -(\langle \text{grad } u_0, V(0) \rangle_{R^n}, (F_0 F_0 + N) \phi)_{L^2(\Omega)} \\ &= 0 + \dots \end{aligned}$$

(iii) For the third term

$$-(A_1 u_0, \phi)_{L^2(\Omega)} = ((F' F_0 + F_0 F' + F_0 \beta F_0 + N \beta) u_0, \phi)_{L^2(\Omega)}$$

we have

$$\begin{aligned} (F' F_0 u_0, \phi)_{L^2(\Omega)} &= (F_1 F_0, \phi)_{L^2(\Omega)} + \dots, \\ (F_0 F' u_0, \phi)_{L^2(\Omega)} &= (F_0 F_1 u_0, \phi)_{L^2(\Omega)} + \dots, \\ (F_0 \beta F_0 u_0, \phi)_{L^2(\Omega)} &= (\beta F_0 y_0, F_0 \phi)_{L^2(\Omega)} \\ &= \int_{\Omega} \text{div } V(0, x) (F_0 u_0)(x) (F_0 \phi)(x) \, dx = 0 + \dots \end{aligned}$$

since $F_0 u_0, F_0 \phi \in H_0^1(\Omega)$.

Let us recall that by (4.10) it follows that $u_0 \in H_0^1(\Omega)$. Hence

$$\begin{aligned} (N \beta u_0, \phi)_{L^2(\Omega)} &= N \int_{\Omega} \text{div } V(0, x) u_0(x) \phi(x) \, dx \\ &= -N \int_{\Omega} \langle V(0, x), \text{grad } (u_0(x) \phi(x)) \rangle_{R^n} \, dx \\ &= 0 + \dots \quad \forall \phi \in H_0^1(\Omega); \end{aligned}$$

therefore,

$$-(A_1 u_0, \phi)_{L^2(\Omega)} = ((F_1 F_0 + F_0 F_1) u_0, \phi)_{L^2(\Omega)} + \dots$$

and in view of (4.15), we obtain the representation (4.42) of the distribution (4.44), provided the set $\{K(\Omega) - K(\Omega)\} \cap H_0^1(\Omega)$ is dense in the set $\{K(\Omega) - K(\Omega)\} \subset L^2(\Omega)$.

Remark 4. Let us present an equivalent form of the problem (P').

Problem (P'). Find an element $u'(\Omega) \in L^2(\Omega)$ which minimizes the cost functional

$$(4.45) \quad I(u) = \frac{1}{2} \int_{\Omega} (z(u; x))^2 \, dx + \int_{\partial \Omega} v_n(x) \frac{\partial z}{\partial n}(u; x) \frac{\partial p_0}{\partial n}(x) \, d + \frac{N}{2} \int_{\Omega} (u(x))^2 \, dx$$

over the cone $K(\Omega)$.

The element $z(u; \cdot) \in H^1(\Omega)$ is given by a unique weak solution of the following elliptic boundary value problem:

$$(4.46) \quad \begin{aligned} -\Delta z(u; x) &= u(x) \quad \text{in } \Omega, \\ z(u; x) &= v_n(x) \frac{\partial y_0}{\partial n}(u_0; x) \quad \text{on } \partial \Omega. \end{aligned}$$

The element $p_0 \in H_0^1(\Omega) \cap H^2(\Omega)$ is given by the solution of the adjoint-state equation in the system (4.8) for $\varepsilon = 0$.

Acknowledgment. The author is very indebted to Professor Kazimierz Malanowski for stimulating discussions concerning the sensitivity analysis of optimal control problems.

REFERENCES

- [1] A. V. Fiacco, *Introduction to Sensitivity and Stability Analysis in Nonlinear Programming*, Academic Press, New York, 1983.
- [2] A. HARAUX, *How to differentiate the projection on a convex set in Hilbert space. Some applications to variational inequalities*, J. Math. Soc. Japan, 29 (1977), pp. 615–631.
- [3] J. L. LIONS, *Contrôle optimal de systèmes gouvernés par des équations aux dérivées partielles*, Dunod, Paris, 1968.
- [4] J. L. LIONS AND E. MAGENES, *Problèmes aux limites non homogènes et applications*, Vol. 1, Dunod, Paris, 1968.
- [5] K. MALANOWSKI, *Differential sensitivity of solutions to convex, control constrained optimal control problems*, Appl. Math. Optim., 12 (1984), pp. 1–14.
- [6] K. MALANOWSKI AND J. SOKOŁOWSKI, *Sensitivity of solutions to convex, control constrained optimal control problems for distributed parameter systems*, J. Math. Anal. Appl., 120 (1986), pp. 240–263.
- [7] F. MIGNOT, *Contrôle dans les inéquations variationnelles elliptiques*, J. Funct. Anal., 22 (1976), pp. 130–185.
- [8] J. SOKOŁOWSKI, *Sensitivity analysis for a class of variational inequalities*, in Optimization of Distributed Parameter Structures, Vol. 2, E. J. Haug and J. Cea, eds., Sijthoff and Noordhoff, Alphen aan den Rijn, the Netherlands, 1981, pp. 1600–1609.
- [9] ———, *Conical differentiability of projection on convex sets—an application to sensitivity analysis of Signorini variational inequality*, Technical Report, Institute of Mathematics, University of Genoa, Italy, 1981.
- [10] ———, *Differential stability of solutions to constrained optimization problems*, Appl. Math. Optim., 13 (1985), pp. 97–115.
- [11] J. SOKOŁOWSKI AND J.-P. ZOLESIO: *Dérivation par rapport au domaine dans les problèmes unilatéraux*, INRIA, Rapport de Recherche No. 132, Rocquencourt, France, 1982.
- [12] ———, *Shape sensitivity analysis of unilateral problems*, Publication Mathématiques 67, Université de Nice, France, 1985; SIAM J. Math. Anal., 18 (1987), pp. 1416–1437.
- [13] ———, *Dérivée par rapport au domaine de la solution d'un problème unilatéral*, C.R. Acad. Sc. Paris 301 (1985), pp. 103–106.
- [14] J.-P. ZOLESIO, *The material derivative (or speed) method for shape optimization*, in Optimization of Distributed Parameter Structures, Vol. 2, E. J. Haug and J. Cea, eds., Sijthoff and Noordhoff, Alphen aan den Rijn, the Netherlands, 1981, pp. 1089–1151.
- [15] P. HOLNICKI, J. SOKOŁOWSKI AND A. ZOCHOWSKI, *Sensitivity analysis of an optimal control problem arising from air quality control in urban area*, in System Modelling and Optimization, Proc. 12th IFIP Conference, Budapest, Hungary, Lecture Notes in Control and Information Sciences 84, Springer-Verlag, 1986, pp. 331–339.
- [16] J. SOKOŁOWSKI, *Shape sensitivity analysis of boundary optimal control problems for parabolic systems*, to appear.

OPTIMAL PORTFOLIO AND CONSUMPTION DECISIONS FOR A "SMALL INVESTOR" ON A FINITE HORIZON*

IOANNIS KARATZAS†, JOHN P. LEHOCZKY‡ AND STEVEN E. SHREVE§

Abstract. A general consumption/investment problem is considered for an agent whose actions cannot affect the market prices, and who strives to maximize total expected discounted utility of both consumption and terminal wealth. Under very general conditions on the nature of the market model and on the utility functions of the agent, it is shown how to approach the above problem by considering separately the two more elementary ones of maximizing utility of consumption only and of maximizing utility of terminal wealth only, and then appropriately composing them. The optimal consumption and wealth processes are obtained quite explicitly. In the case of a market model with constant coefficients, the optimal portfolio and consumption rules are derived very explicitly in feedback form (on the current level of wealth).

Key words. portfolio and consumption processes, utility functions, stochastic control, martingale representation theorems, change of probability measure, Feynman-Kac Theorem

AMS(MOS) subject classifications. Primary 93E20; secondary 60G44, 90A16, 49B60

1. Introduction. This paper treats a very general consumption/investment decision problem for a single agent, endowed with some initial wealth, who can consume the wealth at some rate C_t and invest it in any of $d+1$ available assets. The agent is attempting to maximize a linear combination of two quantities, namely:

- (i) $E \int_0^T \exp(-\int_0^t \beta(s) ds) U_1(C_t) dt$, the total expected discounted *utility from consumption* over the time-interval $[0, T]$, and
- (ii) $E[\exp(-\int_0^T \beta(s) ds) U_2(X_T)]$, the expected *utility from terminal wealth*.

The $d+1$ *assets* or *securities* available to the agent in this paper are very general. One of them is a *bond*, a security whose instantaneous rate of return may fluctuate (possibly randomly), but which is otherwise riskless. The other assets are *stocks*, risky securities whose prices have randomly fluctuating mean rates of return $b_i(t)$ and dispersion coefficients $\sigma_{ij}(t)$. Section 2 provides a careful exposition of these matters. The stock prices are driven by d independent Wiener processes; these represent the sources of uncertainty in the market model, which we assume to be *complete* in the sense of Harrison and Pliska (1981) and Bensoussan (1984). In our context, completeness amounts to nondegeneracy of the "diffusion" matrix $a(t) = \sigma(t)\sigma^T(t)$, as imposed by condition (2.3).

This condition guarantees, roughly speaking, that "there are exactly as many stocks as there are sources of uncertainty in the market model." It also enables us to construct a new probability measure under which the stock prices, discounted at the rate $r(\cdot)$ of the bond, become local martingales; this fact is of great importance in the modern theory of financial economics, and we refer the reader to Harrison and Pliska (1981), (1983) for a fuller account of its ramifications.

* Received by the editors September 8, 1986; accepted for publication January 27, 1987. Part of this research was carried out while I. Karatzas was visiting the Center for Stochastic Processes, University of North Carolina in Chapel Hill during April-May of 1986, and also during a brief sojourn by I. Karatzas and S. Shreve at the Institute of Mathematics and its Applications, University of Minnesota in Minneapolis in June of 1986.

† Department of Statistics, Columbia University, New York, New York 10027. The work of this author was supported by National Science Foundation grant DMS-84-16734.

‡ Department of Statistics, Carnegie-Mellon University, Pittsburgh, Pennsylvania 15213. The work of this author was supported by National Science Foundation grant DMS-84-03166.

§ Department of Mathematics, Carnegie-Mellon University, Pittsburgh, Pennsylvania 15213. The work of this author was supported by National Science Foundation grant DMS-84-03166.

The processes $r(t)$, $b_i(t)$, $\sigma_{ij}(t)$; $1 \leq i, j \leq d$ and $\beta(t)$, the instantaneous discount rate in the economy, will be collectively referred to as the *coefficients of the market model*. We assume that our agent is a "small investor," in that his decisions do not influence the asset prices which are treated as exogenous.

Single-agent consumption/investment problems have been investigated by a number of authors. A significant plateau was reached by Merton (1969), (1971). In the special case of a market model with constant coefficients, he found closed-form solutions for the associated Bellman equations in the infinite-horizon and the zero-bequest, finite-horizon cases, when the utility of consumption belongs to the HARA class, i.e., $U_1(c) = ac^p$, $-1 < p < 1$, $p \neq 0$ or $U_1(c) = a \log c$. The infinite-horizon model was generalized by Karatzas, Lehoczky, Sethi and Shreve (1986) (abbreviated hereafter as KLSS (1986)) who presented closed-form expressions for the value function and the optimal consumption and investment policies, corresponding to general utility functions and general assumptions concerning the effect of *bankruptcy*, a possibility altogether ignored by previous authors. The work cited thus far allows short-selling of both the bond and stocks; indeed, such short-selling is mandated by the optimal investment process. A model in which such short-selling is prohibited, but in which the interest rate on the bond and the mean rate of return on the only stock are constant and equal, was studied by Lehoczky, Sethi and Shreve (1983).

The present paper generalizes previous work in two major ways. First, the time-horizon is finite and general utility functions for consumption and terminal wealth are allowed. Second, the coefficients of the market model are required only to be adapted and bounded processes. This means that stock prices can fluctuate in an almost arbitrary, not necessarily Markovian fashion. Such generality notwithstanding, explicit results for the solution of the problem are provided, and it is shown how to use the resulting formulas in order to derive more simply the results of Merton (1969) and KLSS (1986). The methodology that accounts for both the simplicity and generality is the *Girsanov change of probability measure*, which removes the differences in mean rates of return among the investments and thus endows certain processes with the martingale property. Such an idea appeared in the context of option pricing in Harrison and Kreps (1979) and was more fully developed by Harrison and Pliska (1981). To our knowledge, this is its first application to optimal consumption and investment.

In this paper, an eminent role is played by the process ζ of (4.6), which is closely related to the Radon-Nikodým derivative of the measure transformation mentioned earlier. This process starts at $\zeta_0 = 1$ and satisfies a *linear* stochastic differential equation (cf. (4.7)) with random coefficients, determined by the market model only and not by the utility functions of the investor. These functions determine in turn, in a surprisingly simple way, the number $y > 0$ that serves as the initial value of the process

$$Y_t = y\zeta_t,$$

in terms of which the optimal consumption and wealth processes C^* , X^* are constructed in (6.7) and (6.8). Thus, in a sense, the only real decision on the part of the investor takes place at $t = 0$ with the determination of the constant $y > 0$; cf. (1.3) below. Once this decision has been taken, the further evolution of the processes Y , C^* and X^* depends only on the market, and not on the investor's preferences (utility functions).

The article is organized as follows: § 2 contains a detailed discussion on consumption/portfolio process pairs which are *admissible* from the point of view of the investor, i.e., do not get him into debt (negative wealth) when his initial endowment is $x \geq 0$. Section 3 lists our assumptions on the utility functions. Section 4 discusses the optimization problem when *utility comes only from consumption*; we exhibit two strictly

decreasing functions G_1 and \mathcal{X}_1 on $(0, \infty)$, the latter with inverse \mathcal{Y}_1 , such that the optimal consumption and wealth processes are given by (4.9), (4.10) with $y = \mathcal{Y}_1(x)$ and the value function for this problem is given by the composition $V_1(x) = G_1(\mathcal{Y}_1(x))$.

Section 5 solves the “complementary” optimization problem, in which *utility is derived only from terminal wealth*; a version of this problem was discussed by Pliska (1986). Again, we produce two strictly decreasing functions G_2 , \mathcal{X}_2 and set $\mathcal{Y}_2 = \mathcal{X}_2^{-1}$; the optimal consumption process is then identically equal to zero, the optimal wealth is given by (5.8) with $y = \mathcal{Y}_2(x)$, and the value function for this problem is $V_2(x) = G_2(\mathcal{Y}_2(x))$.

In § 6 we take up the original problem, in which *utility comes from both consumption and terminal wealth*. Here the situation is most interesting: it is shown that the investor has to divide at time $t = 0$ his endowment $x \geq 0$ into two parts:

$$(1.1) \quad x_1 \geq 0, \quad x_2 \geq 0, \quad x_1 + x_2 = x.$$

From then on, he faces two separate problems: one with initial wealth x_1 and utility coming *only* from consumption, the other with initial wealth x_2 and utility coming *only* from terminal wealth. The “superposition” of his actions for these individual problems yields an optimal procedure for the original one, provided that the splitting (1.1) is done in such a way as to produce “equipartition in the y -scale”:

$$(1.2) \quad x_1 = \mathcal{X}_1(y), \quad x_2 = \mathcal{X}_2(y)$$

where y is now determined uniquely via

$$(1.3) \quad \mathcal{X}(y) \triangleq \mathcal{X}_1(y) + \mathcal{X}_2(y) = x;$$

the value function for this problem is then given as

$$(1.4) \quad V(x) = V_1(x_1) + V_2(x_2) = G(\mathcal{X}^{-1}(x))$$

where $G \triangleq G_1 + G_2$.

In § 7 we specialize this problem to the case of *constant coefficients*. We employ results from KLSS (1986) and ideas from the theory of option pricing in order to characterize the functions \mathcal{X} , G above in terms of suitable Cauchy problems for a degenerate parabolic equation, and we then obtain the solutions of these Cauchy problems in closed form (Proposition 7.3 and Remark 7.4). We also provide the optimal consumption and portfolio strategies in explicit “feedback” form (Theorem 7.7).

2. Portfolio and consumption processes. Let us consider a market in which $d + 1$ assets (or “securities”) are traded continuously on the fixed time-horizon $[0, T]$, $0 < T < \infty$. One of these assets, called the *bond*, has a price which evolves according to the differential equation

$$(2.1) \quad dP_0(t) = r(t)P_0(t)dt, \quad P_0(0) = p_0, \quad 0 \leq t \leq T.$$

The remaining d assets, called *stocks*, are risky; their prices are modelled by the linear stochastic differential equations:

$$(2.2) \quad dP_i(t) = P_i(t) \left[b_i(t) dt + \sum_{j=1}^d \sigma_{ij}(t) dW_t^{(j)} \right], \quad P_i(0) = p_i, \quad 0 \leq t \leq T$$

for $i = 1, 2, \dots, d$. Here $W = \{W_t = (W_t^{(1)}, \dots, W_t^{(d)})^T, \mathcal{F}_t; 0 \leq t \leq T\}$ is a d -dimensional Brownian motion on (Ω, \mathcal{F}, P) and the filtration $\{\mathcal{F}_t\}$ is the augmentation under P of $\mathcal{F}_t^W = \sigma(W_s; 0 \leq s \leq t)$, $0 \leq t \leq T$. The *interest rate* process $\{r(t), \mathcal{F}_t; 0 \leq t \leq T\}$, as well as the vector of *mean rates of return* $\{b(t) = (b_1(t), \dots, b_d(t))^T, \mathcal{F}_t; 0 \leq t \leq T\}$

and the *dispersion matrix* $\{\sigma(t) = (\sigma_{ij}(t))_{1 \leq i, j \leq d}, \mathcal{F}_t; 0 \leq t \leq T\}$ are assumed to be measurable, adapted and bounded, uniformly in $(t, \omega) \in [0, T] \times \Omega$. We introduce the *covariance matrix* $a(t) \triangleq \sigma(t)\sigma^T(t)$ and assume that for some $\varepsilon > 0$,

$$(2.3) \quad \xi^T a(t, \omega) \xi \geq \varepsilon \|\xi\|^2 \quad \forall \xi \in \mathbb{R}^d, \quad (t, \omega) \in [0, T] \times \Omega.$$

LEMMA 2.1. *Under the assumption (2.3), the matrices $\sigma^T(t, \omega)$ and $\sigma(t, \omega)$ are invertible and we have*

$$(2.4) \quad \|(\sigma^T(t, \omega))^{-1} \xi\| \leq \frac{1}{\sqrt{\varepsilon}} \|\xi\| \quad \forall \xi \in \mathbb{R}^d,$$

$$(2.5) \quad \|(\sigma(t, \omega))^{-1} \xi\| \leq \frac{1}{\sqrt{\varepsilon}} \|\xi\| \quad \forall \xi \in \mathbb{R}^d$$

for every $(t, \omega) \in [0, T] \times \Omega$.

Proof. For a $(d \times d)$ matrix Γ we recall the operator norm

$$\|\Gamma\| \triangleq \sup_{\xi \neq 0} \frac{\|\Gamma \xi\|}{\|\xi\|}$$

and can show easily that $\|\Gamma^T\| = \|\Gamma\|$. Now (2.3) implies that $\sigma^T(t, \omega)$ is nonsingular, for otherwise we could find a vector $\xi \in \mathbb{R}^d \setminus \{0\}$ which would make $\|\xi^T a(t, \omega) \xi\| = \|\sigma^T(t, \omega) \xi\|^2 = 0$. Letting $\xi = (\sigma^T(t, \omega))^{-1} \eta$, we may then rewrite (2.3) as

$$\|(\sigma^T(t, \omega))^{-1} \eta\| \leq \frac{1}{\sqrt{\varepsilon}} \|\eta\| \quad \forall \eta \in \mathbb{R}^d$$

for every $(t, \omega) \in [0, T] \times \Omega$, which is equivalent to $\|(\sigma^T(t, \omega))^{-1}\| \leq 1/\sqrt{\varepsilon}$. Because $\|(\sigma(t, \omega))^{-1}\| = \|(\sigma^T(t, \omega))^{-1}\|$, the condition (2.3) is equivalent to

$$\xi^T \sigma^T(t, \omega) \sigma(t, \omega) \xi \geq \varepsilon \|\xi\|^2 \quad \forall \xi \in \mathbb{R}^d$$

for every $(t, \omega) \in [0, T] \times \Omega$. From this relation we derive (2.5), in the same way that we established (2.4) from (2.3). \square

We envision now an investor who starts with some initial endowment $x \geq 0$ and invests it in the $d + 1$ assets described above. Let $N_i(t)$ denote the number of shares of asset i owned by the investor at time t . Then $X_0 = x = \sum_{i=0}^d N_i(0) p_i$, and the investor's *wealth* at time t is

$$(2.6) \quad X_t = \sum_{i=0}^d N_i(t) P_i(t).$$

If the trading of shares (and hence the adjustment of the portfolio) is allowed to take place only at discrete time points, say at $\dots t - h, t, t + h, \dots$ and there is no infusion or withdrawal of funds, then

$$(2.7) \quad X_{t+h} - X_t = \sum_{i=0}^d N_i(t) [P_i(t+h) - P_i(t)].$$

On the other hand, if the investor chooses at time $t + h$ to consume an amount hC_{t+h} and reduce the wealth accordingly, then (2.7) should be replaced by

$$(2.8) \quad X_{t+h} - X_t = \sum_{i=0}^d N_i(t) [P_i(t+h) - P_i(t)] - hC_{t+h}.$$

The continuous-time analogue of (2.8) is

$$dX_t = \sum_{i=0}^d N_i(t) dP_i(t) - C_t dt,$$

which becomes

$$(2.9) \quad dX_t = (r(t)X_t - C_t) dt + \sum_{i=1}^d (b_i(t) - r(t))\pi_i(t) dt + \sum_{i=1}^d \sum_{j=1}^d \pi_i(t)\sigma_{ij}(t) dW_t^{(j)}$$

if we take (2.1), (2.2), (2.6) into account and denote by $\pi_i(t) \triangleq N_i(t)P_i(t)$ the amount invested in the stock i , $1 \leq i \leq d$.

DEFINITION 2.2. A *portfolio process* $\pi = \{\pi(t) = (\pi_1(t), \dots, \pi_d(t))^T, \mathcal{F}_t; 0 \leq t \leq T\}$ is a measurable, adapted, \mathcal{R}^d -valued process for which

$$(2.10) \quad \sum_{i=1}^d \int_0^T \pi_i^2(t) dt < \infty \quad \text{a.s.}$$

A *consumption process* $C = \{C_t, \mathcal{F}_t; 0 \leq t < \infty\}$ is a measurable, adapted process with values in $[0, \infty)$ and

$$(2.11) \quad \int_0^T C_t dt < \infty \quad \text{a.s.}$$

Remark 2.3. Any component of the vector $\pi(t)$ may become negative, which is to be interpreted as short-selling that particular stock. The amount

$$\pi_0(t) \triangleq X_t - \sum_{i=1}^d \pi_i(t)$$

invested in the bond may also become negative, and this corresponds to borrowing at the interest rate $r(t)$.

The conditions (2.10), (2.11) guarantee that the stochastic differential equation (2.9) has the unique, strong solution

$$(2.12) \quad \begin{aligned} X_t = \exp \left(\int_0^t r(s) ds \right) & \left\{ x + \int_0^t \exp \left(- \int_0^s r(u) du \right) [\pi^T(s)(b(s) - r(s)\mathbf{1}) - C_s] ds \right. \\ & \left. + \int_0^t \exp \left(- \int_0^s r(u) du \right) \pi^T(s)\sigma(s) dW_s \right\}, \end{aligned}$$

$0 \leq t \leq T$

where $\mathbf{1}$ is the d -dimensional vector with every component equal to 1. All vectors are column vectors, and transposition is denoted by the superscript T .

DEFINITION 2.4. A pair (π, C) of portfolio and consumption processes is said to be *admissible for the initial endowment* $x \geq 0$ if the wealth process X of (2.12) satisfies

$$(2.13) \quad X_t \geq 0, \quad 0 \leq t \leq T \quad \text{a.s.}$$

We denote by $\mathcal{A}(x)$ the class of all such pairs. \square

If $b(t) = r(t)\mathbf{1}$; $0 \leq t \leq T$, then the discount factor $\exp \{-\int_0^t r(u) du\}$ exactly offsets the rate of growth of all assets, and (2.12) shows that

$$(2.14) \quad M_t \triangleq X_t \exp \left(- \int_0^t r(u) du \right) - x + \int_0^t C_s \exp \left(- \int_0^s r(u) du \right) ds$$

is a stochastic integral. In other words, the process consisting of current wealth plus cumulative consumption, both properly discounted, is a local martingale. When $b(t) \neq r(t)\mathbf{1}$, M of (2.14) is no longer a local martingale under P , but becomes one under a

new probability measure \tilde{P} that removes the drift term $\pi^T(t)(b(t) - r(t)\mathbf{1})$ from (2.9). More specifically, recall from Lemma 2.1 that the process

$$(2.15) \quad \theta(t) \triangleq (\sigma(t))^{-1}(b(t) - r(t)\mathbf{1}), \quad 0 \leq t \leq T$$

is bounded, and let

$$(2.16) \quad Z_t \triangleq \exp \left\{ - \sum_{i=1}^d \int_0^t \theta_i(s) dW_s^{(i)} - \frac{1}{2} \int_0^t \|\theta(s)\|^2 ds \right\}.$$

Then $\{Z_t, \mathcal{F}_t; 0 \leq t \leq T\}$ is a martingale; the new probability measure

$$(2.17) \quad \tilde{P}(A) \triangleq E[Z_T 1_A], \quad A \in \mathcal{F}_T$$

is such that P and \tilde{P} are mutually absolutely continuous on \mathcal{F}_T , and

$$(2.18) \quad \tilde{W}_t \triangleq W_t + \int_0^t \theta(s) ds, \quad 0 \leq t \leq T$$

is a standard, d -dimensional Brownian motion under \tilde{P} (Girsanov (1960) or Karatzas and Shreve (1987, § 3.5)). In terms of \tilde{W} , we may rewrite (2.12) as

$$(2.19) \quad \begin{aligned} X_t \exp \left(- \int_0^t r(u) du \right) + \int_0^t C_s \exp \left(- \int_0^s r(u) du \right) ds \\ = x + \int_0^t \exp \left(- \int_0^s r(u) du \right) \pi^T(s) \sigma(s) d\tilde{W}_s. \end{aligned}$$

For any $(\pi, C) \in \mathcal{A}(x)$, the left-hand side of (2.19) is nonnegative and the right-hand side is a local martingale under \tilde{P} . It follows that the left-hand side, and hence also $X_t \exp(-\int_0^t r(u) du)$, is a nonnegative supermartingale under \tilde{P} . Now with

$$(2.20) \quad \tau_0 = T \wedge \inf \{0 \leq t \leq T; X_t = 0\},$$

we have from Chung (1982, Thm. 1.4) or Karatzas and Shreve (1987, Problem 1.3.29) that

$$(2.21) \quad X_t = 0, \quad \tau_0 \leq t \leq T \quad \text{holds a.s. on } \{\tau_0 < T\}.$$

We say that *bankruptcy occurs at time* τ_0 on the event $\{\tau_0 < T\}$.

The supermartingale property yields in (2.19):

$$(2.22) \quad \tilde{E} \left[X_T \exp \left(- \int_0^T r(u) du \right) + \int_0^T C_t \exp \left(- \int_0^t r(u) du \right) dt \right] \leq x,$$

whence the following *necessary conditions for admissibility*:

$$(2.23) \quad \tilde{E} \int_0^T C_t \exp \left(- \int_0^t r(u) du \right) dt \leq x,$$

$$(2.24) \quad \tilde{E} \left[X_T \exp \left(- \int_0^T r(u) du \right) \right] \leq x.$$

These conditions will also turn out to be “sufficient” for admissibility, in the sense of Propositions 2.6 and 2.8.

DEFINITION 2.5. For a given $x \geq 0$, let

(i) $\mathcal{C}(x)$ (respectively, $\mathcal{D}(x)$) denote the class of consumption processes which satisfy (2.23) (resp., (2.23) as an equality),

(ii) $\mathcal{L}(x)$ (respectively, $\mathcal{M}(x)$) denote the class of nonnegative random variables B on $(\Omega, \mathcal{F}_T, \tilde{P})$ which satisfy

$$(2.25) \quad \tilde{E} \left[B \exp \left(- \int_0^T r(u) du \right) \right] \leq x$$

(resp., (2.25) as an equality),

(iii) $\mathcal{P}(x)$ denote the class of portfolio processes, such that $(\pi, 0) \in \mathcal{A}(x)$ and the corresponding terminal wealth X_T belongs to $\mathcal{M}(x)$. \square

We shall show that $\mathcal{C}(x)$ consists of exactly those “reasonable” consumption processes, for which an investor, starting out with wealth x at time $t=0$, is able to construct a portfolio that will avoid debt (i.e., negative wealth) on $[0, T]$, almost surely.

PROPOSITION 2.6. *For every given $C \in \mathcal{C}(x)$, there exists a portfolio process π such that $(\pi, C) \in \mathcal{A}(x)$.*

Proof. Let $D \triangleq \int_0^T C_t \exp \left(- \int_0^t r(u) du \right) dt$, and define the nonnegative process

$$(2.26) \quad \begin{aligned} \xi_t &\triangleq \tilde{E} \left[\int_t^T C_s \exp \left(- \int_t^s r(u) du \right) ds \middle| \mathcal{F}_t \right] + (x - \tilde{E}D) \cdot \exp \left(\int_0^t r(u) du \right) \\ &= \left\{ x + m_t - \int_0^t C_s \exp \left(- \int_0^s r(u) du \right) ds \right\} \exp \left(\int_0^t r(u) du \right), \end{aligned}$$

where

$$m_t \triangleq \tilde{E}(D | \mathcal{F}_t) - \tilde{E}D = \frac{E[DZ_T | \mathcal{F}_t]}{Z_t} - E(DZ_T),$$

thanks to the so-called “Bayes rule” (Karatzas and Shreve (1987, Lemma 3.5.3)). By choosing a proper modification, we may assume that the paths of the martingale

$$N_t \triangleq E[DZ_T | \mathcal{F}_t], \quad 0 \leq t \leq T$$

are P -a.s. right-continuous and admit finite left-hand limits (Karatzas and Shreve (1987, Thms. 1.3.8, 1.3.13)). Now the fundamental representation result for Brownian martingales (Karatzas and Shreve (1987, Problem 3.4.16)) guarantees the existence of a measurable, $\{\mathcal{F}_t\}$ -adapted and \mathcal{R}^d -valued process Y with

$$(2.27) \quad \sum_{j=1}^d \int_0^T Y_j^2(t) dt < \infty$$

and

$$(2.28) \quad N_t = E(DZ_T) + \sum_{j=1}^d \int_0^t Y_j(s) dW_s^{(j)}, \quad 0 \leq t \leq T$$

valid a.s. P . We conclude that $m_t = u(N_t, Z_t) - E(DZ_T)$, where $u(x, y) = x/y$, and from Itô’s rule we obtain with $\varphi(t) \triangleq (1/Z_t)(Y(t) + N_t \theta(t))$:

$$(2.29) \quad m_t = \sum_{j=1}^d \int_0^t \varphi_j(s) d\tilde{W}_s^{(j)}, \quad 0 \leq t \leq T.$$

We have used the relations

$$(2.30) \quad dZ_t = -Z_t \theta^T(t) dW_t$$

and (2.18). Now define

$$(2.31) \quad \pi(t) \triangleq \exp \left(\int_0^t r(u) du \right) (\sigma^T(t))^{-1} \varphi(t)$$

so that (2.26) becomes (2.19) when we make the identifications $\xi = X$, $m = M$. Condition (2.10) follows from (2.4), (2.27), the boundedness of $\|\theta\|$, and the path continuity of Z and N ; the latter is a consequence of (2.28). \square

Remark 2.7. The wealth process $X \equiv \xi$ in (2.26), corresponding to any $C \in \mathcal{D}(x)$, is given by

$$(2.32) \quad X_t = \tilde{E} \left[\int_t^T C_s \exp \left(- \int_t^s r(u) du \right) ds \middle| \mathcal{F}_t \right], \quad 0 \leq t \leq T.$$

In particular then, $X_T = 0$ almost surely.

PROPOSITION 2.8. *For every given $B \in \mathcal{L}(x)$, there exists a pair $(\pi, C) \in \mathcal{A}(x)$ with corresponding wealth process X , such that*

$$X_T = B \quad \text{a.s. } \tilde{P}.$$

Proof. Let $Q \triangleq B \exp \left(- \int_0^T r(u) du \right)$, and define the nonnegative process η by

$$(2.33) \quad \begin{aligned} \eta_t \exp \left(- \int_0^t r(u) du \right) &\triangleq \tilde{E}(Q | \mathcal{F}_t) + (x - \tilde{E}Q) \cdot \left(1 - \frac{t}{T} \right) \\ &= x + m_t - \rho t, \quad 0 \leq t \leq T \end{aligned}$$

where $m_t = \tilde{E}(Q | \mathcal{F}_t) - \tilde{E}Q$, $\rho = (\tilde{x} - EQ)/T$. Obviously, $\eta_0 = x$ and $\eta_T = B$ hold almost surely. We obtain a stochastic integral representation of the form (2.29) for the \tilde{P} -martingale m ; then (2.33) is cast in the form (2.19) once we take π as in (2.31),

$$C_t = \rho \cdot \exp \left(\int_0^t r(u) du \right), \quad 0 \leq t \leq T$$

and $X = \xi$.

COROLLARY 2.9. *For any given $B \in \mathcal{M}(x)$, there exists a portfolio process $\pi \in \mathcal{P}(x)$ with corresponding wealth process*

$$(2.34) \quad X_t = \tilde{E} \left[B \exp \left(- \int_t^T r(u) du \right) \middle| \mathcal{F}_t \right], \quad 0 \leq t \leq T. \quad \square$$

Proposition 2.8 and the relation (2.24) show that $\mathcal{L}(x)$ consists of precisely those “levels of terminal wealth” which are *attainable* from the initial endowment $x \geq 0$, via the usage of some portfolio/consumption pair that avoids debt. The terminology here is due to Pliska (1986). Corollary 2.9 shows that the “extreme” elements of $\mathcal{L}(x)$ are attainable by strategies that mandate zero consumption.

3. Utility functions. Consider a strictly increasing, strictly concave and C^1 function $U: (0, \infty) \rightarrow \mathcal{R}$ with $U(0) \triangleq \lim_{c \downarrow 0} U(c) \geq -\infty$, $U'(\infty) \triangleq \lim_{c \rightarrow \infty} U'(c) = 0$. We allow the possibility that $U'(0) \triangleq \lim_{c \downarrow 0} U'(c) = \infty$. A function with these properties will be called a *utility function* in the sequel.

Because $U': [0, \infty] \rightarrow_{\text{onto}} [0, U'(0)]$ is strictly decreasing, it has a strictly decreasing inverse $I: [0, U'(0)] \rightarrow_{\text{onto}} [0, \infty]$. We extend I to be a continuous function on the entirety of $[0, \infty]$ by setting $I(y) \equiv 0$ for $U'(0) \leq y \leq \infty$, and note that

$$(3.1) \quad U(I(y)) \geq U(c) + yI(y) - yc, \quad 0 \leq c < \infty, \quad 0 < y < \infty.$$

For part of our development in §§ 4–6, we shall need to impose the assumption

$$(3.2) \quad U \in C^2 \quad \text{and} \quad U'' \text{ is nondecreasing on } (0, \infty).$$

Under this condition, I is convex on $(0, \infty)$ and continuously differentiable on $(0, \infty) \setminus \{U'(0)\}$. If we define $I'(U'(0)) = 0$ in the case $U'(0) < \infty$, then because $I'(y) = 0$ for every $y > U'(0)$, the identity

$$(3.3) \quad (U(I(y)))' = yI'(y)$$

becomes valid on the entirety of $(0, \infty)$. This identity is also valid on $(0, \infty)$ if $U'(0) = \infty$.

4. Maximization of utility from consumption. In this section we formulate and study a particular stochastic control problem for the “small investor” model of § 2. Suppose that, in addition to the data given there, we have a measurable, $\{\mathcal{F}_t\}$ -adapted and uniformly bounded discount process $\{\beta(t), \mathcal{F}_t; 0 \leq t \leq T\}$, as well as a utility function U_1 . The objective will be to maximize the expected discounted utility from consumption

$$(4.1) \quad J_1(x; \pi, C) = E \int_0^T \exp\left(-\int_0^t \beta(s) ds\right) U_1(C_t) dt$$

with an initial endowment $x \geq 0$, over the class $\mathcal{A}_1(x)$ of portfolio/consumption pairs $(\pi, C) \in \mathcal{A}(x)$ for which

$$(4.2) \quad E \int_0^T \exp\left(-\int_0^t \beta(s) ds\right) U_1^-(C_t) dt < \infty.$$

The expectation in (4.1) is well defined for every pair $(\pi, C) \in \mathcal{A}_1(x)$; of course, $\mathcal{A}_1(x) \equiv \mathcal{A}(x)$ if $U_1(0) > -\infty$.

We denote by

$$(4.3) \quad V_1(x) \triangleq \sup_{(\pi, C) \in \mathcal{A}_1(x)} J_1(x; \pi, C)$$

the value function for this problem, which is trivial if $x = 0$. Indeed, the admissibility condition (2.23) implies then $V_1(0) = U_1(0) \cdot E \int_0^T \exp(-\int_0^t \beta(s) ds) dt$, and this obviously can be achieved by $\pi \equiv 0$, $C \equiv 0$. We concentrate, therefore, on $x > 0$.

Because utility comes only from consumption, it is plausible that one should strive to increase the net effect of the latter, up to the permissible limit dictated by (2.23), by considering only consumption processes C in $\mathcal{D}(x)$.

PROPOSITION 4.1. *For every $x > 0$ we have*

$$V_1(x) = \sup_{\substack{(\pi, C) \in \mathcal{A}_1(x) \\ C \in \mathcal{D}(x)}} J(x; \pi, C).$$

Proof. Take $(\pi, C) \in \mathcal{A}_1(x)$ and observe that the number $z \triangleq \tilde{E} \int_0^T \exp(-\int_0^t r(u) du) C_t dt$ lies in $[0, x]$. If $z > 0$, we may define $\hat{C}_t \triangleq (x/z) C_t$ so that $\hat{C} \in \mathcal{D}(x)$, and construct a portfolio process $\hat{\pi}$ such that $(\hat{\pi}, \hat{C}) \in \mathcal{A}(x)$; cf. Proposition 2.6 and Remark 2.7. Because $U_1(C_t) \leq U_1(\hat{C}_t)$; $0 \leq t \leq T$ and C satisfies (4.2), so does \hat{C} , i.e., $(\hat{\pi}, \hat{C}) \in \mathcal{A}_1(x)$ and we have

$$(4.4) \quad J_1(x; \pi, C) \leq J_1(x; \hat{\pi}, \hat{C}).$$

If $z = 0$ then $C_t = 0$, a.e. $t \in [0, T]$, almost surely, and the consumption process \hat{C} given by

$$(4.5) \quad \hat{C}_t \equiv \hat{c} \triangleq \frac{x}{\tilde{E} \int_0^T \exp(-\int_0^t r(u) du) dt}, \quad 0 \leq t \leq T$$

belongs to $\mathcal{D}(x)$. Again, (4.4) holds for some $\hat{\pi}$ chosen so that $(\hat{\pi}, \hat{C}) \in \mathcal{A}_1(x)$. \square

Recall the Radon–Nikodým derivative martingale Z of (2.16). The related process

$$(4.6) \quad \zeta_t \triangleq Z_t \exp \left(\int_0^t (\beta(u) - r(u)) du \right), \quad 0 \leq t \leq T$$

will be of fundamental significance in what follows. It determines the optimal consumption process, $C^{(1)}$ of (4.9) below, in a very direct and explicit manner. It is easily seen from (2.30) to satisfy the linear stochastic differential equation

$$(4.7) \quad d\zeta_t = (\beta(t) - r(t))\zeta_t dt - \zeta_t \theta^T(t) dW_t,$$

whose importance we already documented in KLSS (1986) in the context of constant coefficients and infinite-horizon.

We recall now from § 3 the notation I_1 for the inverse of the function U'_1 , and assume

$$(4.8) \quad \mathcal{X}_1(y) \triangleq \tilde{E} \int_0^T \exp \left(- \int_0^t r(u) du \right) I_1(y\zeta_t) dt < \infty, \quad \forall 0 < y < \infty.$$

LEMMA 4.2. *Under the condition (4.8), the function \mathcal{X}_1 defined there is continuous and strictly decreasing on $(0, \infty)$ with $\mathcal{X}_1(0) \triangleq \lim_{y \downarrow 0} \mathcal{X}_1(y) = \infty$, $\mathcal{X}_1(\infty) \triangleq \lim_{y \rightarrow \infty} \mathcal{X}_1(y) = 0$.*

Proof. Because I_1 is nonincreasing, and strictly decreasing on $(0, U'_1(0))$, we need only show that

$$P \left[Z_t \exp \left\{ \int_0^t (\beta(s) - r(s)) ds \right\} < \frac{U'_1(0)}{y}; \text{ some } 0 \leq t \leq T \right] > 0$$

holds for every fixed $y \in (0, \infty)$. But

$$\begin{aligned} & \log \left[Z_t \exp \left\{ \int_0^t (\beta(s) - r(s)) ds \right\} \right] \\ &= - \sum_{j=1}^d \int_0^t \theta_j(s) dW_s^{(j)} - \frac{1}{2} \int_0^t \|\theta(s)\|^2 ds + \int_0^t (\beta(s) - r(s)) ds \\ &= B_{A(t)} + \int_0^t \left\{ \beta(s) - r(s) - \frac{1}{2} \|\theta(s)\|^2 \right\} ds \end{aligned}$$

where B is a standard, one-dimensional Brownian motion and $A(t) = \int_0^t \|\theta(s)\|^2 ds$. The strict monotonicity of \mathcal{X} now follows easily from the properties of Brownian motion and the boundedness of the processes $\|\theta\|$, $\beta - r$. The other properties claimed for \mathcal{X} are also inherited from I_1 . \square

Let us denote by $\mathcal{Y}_1: [0, \infty] \xrightarrow{\text{onto}} [0, \infty]$ the inverse of the function \mathcal{X}_1 . For a fixed number $x_1 \in (0, \infty)$ we introduce the consumption process

$$(4.9) \quad C_t^{(1)} \triangleq I_1(\mathcal{Y}_1(x_1)\zeta_t), \quad 0 \leq t \leq T,$$

which belongs to $\mathcal{D}(x_1)$. Thanks to Proposition 2.6 and (2.32), (2.19), there exists a portfolio $\pi^{(1)}$ such that $(\pi^{(1)}, C^{(1)}) \in \mathcal{A}(x_1)$, and the corresponding wealth process $X^{(1)}$ is given by

$$\begin{aligned} (4.10) \quad X_t^{(1)} \exp \left(- \int_0^t r(u) du \right) &= \tilde{E} \left[\int_t^T C_s^{(1)} \exp \left(- \int_0^s r(u) du \right) ds \middle| \mathcal{F}_t \right] \\ &= x_1 - \int_0^t C_s^{(1)} \exp \left(- \int_0^s r(u) du \right) ds \\ &\quad + \int_0^t \exp \left(- \int_0^s r(u) du \right) (\pi^{(1)}(s))^T \sigma(s) d\tilde{W}_s. \end{aligned}$$

In particular, $X^{(1)}$ is nonnegative on $[0, T)$ and vanishes at $t = T$. If $U_1'(0) = \infty$, then $X^{(1)}$ is positive on $[0, T)$.

THEOREM 4.3. *Let us assume that (4.8) holds. Then for any $x_1 > 0$ and with $C^{(1)} \in \mathcal{D}(x_1)$ given by (4.9), the pair $(\pi^{(1)}, C^{(1)})$ constructed above belongs to $\mathcal{A}_1(x_1)$ and is optimal for the problem of (4.3):*

$$(4.11) \quad V_1(x_1) = E \int_0^T \exp\left(-\int_0^t \beta(s) ds\right) U_1(C_t^{(1)}) dt.$$

Proof. It suffices to show that $C^{(1)}$ satisfies (4.2) and to establish the comparison

$$(4.12) \quad E \int_0^T \exp\left(-\int_0^t \beta(s) ds\right) U_1(C_t) dt \leq E \int_0^T \exp\left(-\int_0^t \beta(s) ds\right) U_1(C_t^{(1)}) dt$$

for any other $C \in \mathcal{C}(x_1)$ with the property (4.2).

Recall now the process ζ of (4.6) and observe that, for any $C \in \mathcal{C}(x_1)$, the integral of ζC with respect to the measure

$$(4.13) \quad \nu(dt, d\omega) \triangleq \exp\left(-\int_0^t \beta(s, \omega) ds\right) dt P(d\omega)$$

on $\Theta = [0, T] \times \Omega$ is given by

$$(4.14) \quad \int_{\Theta} \zeta C d\nu = \tilde{E} \int_0^T \exp\left(-\int_0^t r(u) du\right) C_t dt \triangleq x_1,$$

with equality if $C \in \mathcal{D}(x_1)$. For any $C \in \mathcal{C}(x_1)$, the comparison (3.1) yields

$$(4.15) \quad U_1(C_t^{(1)}) \geq U_1(C_t) + \mathcal{Y}_1(x_1) \zeta_t C_t^{(1)} - \mathcal{Y}_1(x_1) \zeta_t C_t, \quad 0 \leq t \leq T$$

almost surely. Corresponding to the special choice $C = \hat{C} \in \mathcal{D}(x_1)$ of (4.5) with x replaced by x_1 , the right-hand side of (4.15) is integrable with respect to the measure ν of (4.13); recalling (4.14), we see that the value of this integral is actually $U_1(\hat{C}) \cdot E \int_0^T \exp\left(-\int_0^t \beta(s) ds\right) dt$. It follows then that $C^{(1)}$ satisfies (4.2).

Let us now take an arbitrary $C \in \mathcal{C}(x_1)$ with the property (4.2), integrate both sides of (4.15) against the measure ν of (4.13), and recall (4.14); we obtain the comparison (4.12). \square

In order to guarantee the finiteness of the value function V_1 in (4.3) and to obtain a useful representation for it, let us impose the condition

$$(4.16) \quad E \int_0^T \exp\left(-\int_0^t \beta(u) du\right) |U_1(I_1(y\zeta_t))| dt < \infty \quad \forall 0 < y < \infty.$$

We shall have more to say about the conditions (4.8), (4.16) in Lemma 4.6 below, as well as in § 7, where we specialize the model to the case of constant coefficients.

PROPOSITION 4.4. *Suppose that the utility function U_1 obeys the conditions (3.2), (4.8) and (4.16). Then the function $G_1: (0, \infty) \rightarrow \mathcal{R}$ given by*

$$(4.17) \quad G_1(y) \triangleq E \int_0^T \exp\left(-\int_0^t \beta(u) du\right) U_1(I_1(y\zeta_t)) dt, \quad 0 < y < \infty$$

is strictly decreasing, continuously differentiable, and satisfies

$$(4.18) \quad G_1'(y) = y\mathcal{X}'_1(y), \quad 0 < y < \infty$$

as well as

$$(4.19) \quad V_1(x) = G_1(\mathcal{Y}_1(x)), \quad 0 < x < \infty.$$

Proof. Under condition (3.2), the function \mathcal{X}_1 of (4.8) inherits the convexity on $(0, \infty)$ from I_1 . Fix $y > 0$ and observe that

$$\begin{aligned}\mathcal{X}_1(y) - \mathcal{X}_1(y - \delta) &= \tilde{E} \int_0^T \exp\left(-\int_0^t r(u) du\right) \{I_1(y\xi_t) - I_1((y - \delta)\xi_t)\} dt \\ &\leq \delta \cdot \tilde{E} \int_0^T \exp\left(-\int_0^t r(u) du\right) \xi_t I'_1(y\xi_t) dt\end{aligned}$$

holds for $0 < \delta < y$. It follows that the left-hand derivative $D^-\mathcal{X}_1(y)$ satisfies

$$-\infty < D^-\mathcal{X}_1(y) \leq \varphi(y) \triangleq E \int_0^T \exp\left(-\int_0^t \beta(u) du\right) \xi_t I'_1(y\xi_t) dt$$

where this last function $\varphi(\cdot)$ is easily seen to be nondecreasing and continuous. In a similar manner, one shows

$$\varphi(y) \leq D^+\mathcal{X}_1(y) < \infty.$$

A convex function of a real variable, whose left- and right-hand derivatives bound a continuous function $\varphi(\cdot)$ in such a way, must be continuously differentiable (Karatzas and Shreve (1987, Problem 3.6.20(iv))) with $\varphi(\cdot)$ as its derivative:

$$(4.20) \quad \mathcal{X}'_1(y) = E \int_0^T \exp\left(-\int_0^t \beta(u) du\right) \xi_t^2 I'_1(y\xi_t) dt.$$

On the other hand, the double inequality

$$\delta(y + \delta)\xi_t^2 I'_1(y\xi_t) \leq U_1(I_1((y + \delta)\xi_t)) - U_1(I_1(y\xi_t)) \leq \delta y \xi_t^2 I'_1((y + \delta)\xi_t)$$

follows easily from (3.3) and the convexity of I_1 , and yields, in conjunction with (4.16), (4.20):

$$\delta(y + \delta)\mathcal{X}'_1(y) \leq G_1(y + \delta) - G_1(y) \leq \delta y \mathcal{X}'_1(y + \delta).$$

The conclusion (4.18) now follows easily; (4.19) is a consequence of Theorem 4.3.

Remark 4.5. Differentiation in (4.19) shows that

$$V'_1(x) = \mathcal{Y}_1(x), \quad V''_1(x) = \mathcal{Y}'_1(x), \quad 0 \leq x < \infty.$$

Because \mathcal{X}_1 is strictly decreasing, \mathcal{Y}'_1 is negative and V_1 is strictly concave. Because $\mathcal{X}_1(\infty) = 0$, we also have $V'_1(0) = \infty$.

LEMMA 4.6. *If $U_1(0) > -\infty$, then (4.16) implies (4.8).*

Proof. From the concavity of U_1 we have in this case

$$(4.21) \quad 0 \leq cU'_1(c) \leq U_1(c) - U_1(0) \leq |U_1(c)| + |U_1(0)|$$

for every $0 \leq c < \infty$, whence

$$y\xi_t I_1(y\xi_t) \leq |U_1(I_1(y\xi_t))| + |U_1(0)|.$$

From this inequality it follows that

$$\begin{aligned}y\mathcal{X}_1(y) &= y \cdot E \int_0^T \exp\left(-\int_0^t \beta(u) du\right) \xi_t I_1(y\xi_t) dt \\ &\leq |U_1(0)| \cdot E \int_0^T \exp\left(-\int_0^t \beta(u) du\right) dt \\ &\quad + E \int_0^T \exp\left(-\int_0^t \beta(u) du\right) |U_1(I_1(y\xi_t))| dt < \infty\end{aligned}$$

holds for every $0 < y < \infty$. \square

Remark 4.7. In the important special case $U(c) = \log c$, the functions of (4.8), (4.17) are given by

$$\mathcal{X}_1(y) = \frac{\alpha_1}{y} \quad \text{and} \quad G_1(y) = -\alpha_1 \cdot \log y + \delta_1,$$

respectively, where

$$\alpha_1 \triangleq E \int_0^T \exp \left(- \int_0^t \beta(u) du \right) dt$$

and

$$\delta_1 \triangleq E \int_0^T \exp \left(- \int_0^t \beta(u) du \right) \cdot \left\{ \sum_{i=1}^d \int_0^t \theta_i(u) dW_u^{(i)} - \int_0^t \left(\beta(u) - r(u) - \frac{1}{2} \|\theta(u)\|^2 \right) du \right\} dt.$$

In particular, the conditions (4.8) and (4.16) are satisfied rather trivially, and the value function is given by (4.19) as

$$V_1(x) = \alpha_1 \cdot \log \left(\frac{x}{\alpha_1} \right) + \delta_1, \quad 0 < x < \infty.$$

5. Maximization of utility from terminal wealth. With a utility function U_2 as in § 3, the problem now is to maximize the expected discounted utility from terminal wealth

$$(5.1) \quad J_2(x; \pi, C) \triangleq E \left[\exp \left(- \int_0^T \beta(s) ds \right) U_2(X_T) \right]$$

over the class $\mathcal{A}_2(x)$ of pairs $(\pi, C) \in \mathcal{A}(x)$ for which

$$(5.2) \quad E \left[\exp \left(- \int_0^T \beta(s) ds \right) U_2^-(X_T) \right] < \infty.$$

Obviously $\mathcal{A}_2(x) \equiv \mathcal{A}(x)$ if $U_2(0) > -\infty$. We denote the value function of this problem by

$$(5.3) \quad V_2(x) \triangleq \sup_{(\pi, C) \in \mathcal{A}_2(x)} J_2(x; \pi, C).$$

Again, the case $x = 0$ is uninteresting: for every $(\pi, C) \in \mathcal{A}(x)$ we have $X_T = 0$ a.s. from (2.24), and thus $V_2(0) = U_2(0) \cdot E \exp \left(- \int_0^T \beta(u) du \right)$; this is achieved by $\pi \equiv 0, C \equiv 0$. We shall take $x > 0$ from now on.

This problem is analogous, and in a certain sense complementary, to that of the preceding section, with the same process ζ of (4.6) playing again a distinguished role.

Because utility comes now only from terminal wealth, it is quite reasonable that the latter should be increased, within the constraints mandated by the level of the initial endowment as quantified by (2.24), by considering portfolio processes π in the class $\mathcal{P}(x)$ of Definition 2.5.

The following is, then, an analogue of Proposition 4.1.

PROPOSITION 5.1. *For every $x > 0$ we have*

$$V_2(x) = \sup_{\substack{(\pi, 0) \in \mathcal{A}_2(x) \\ \pi \in \mathcal{P}(x)}} J_2(x; \pi, 0).$$

Proof. For any $(\pi, C) \in \mathcal{A}(x)$, the number $\lambda \triangleq \tilde{E}(X_T \exp(-\int_0^T r(u) du))$ is in $[0, x]$ by virtue of (2.24). If $\lambda > 0$, then $B \triangleq (x/\lambda)X_T$ belongs to $\mathcal{M}(x)$. From Proposition 2.8 and Corollary 2.9, there exists a portfolio $\hat{\pi} \in \mathcal{P}(x)$ with corresponding terminal wealth $\hat{X}_T = B \geq X_T$, a.s. Obviously then $(\hat{\pi}, 0) \in \mathcal{A}_2(x)$ and

$$(5.4) \quad J_2(x; \pi, C) \leq J_2(x; \hat{\pi}, 0).$$

On the other hand, if $\lambda = 0$, the role of the random variable B above can be played by the number

$$(5.5) \quad b \triangleq \frac{x}{\tilde{E}[\exp(-\int_0^T r(u) du)]} > 0,$$

and we also have $X_T = 0$, a.s. The same reasoning as before leads to (5.4) for some $\hat{\pi} \in \mathcal{P}(x)$. \square

Recalling from § 3 the notation I_2 for the inverse of U'_2 , we introduce now the condition

$$(5.6) \quad \mathcal{X}_2(y) \triangleq \tilde{E} \left[\exp \left(- \int_0^T r(s) ds \right) I_2(y \zeta_T) \right] < \infty \quad \forall 0 < y < \infty$$

and notice, just as in Lemma 4.2, that the function \mathcal{X}_2 is continuous and strictly decreasing on $(0, \infty)$, with $\mathcal{X}_2(0) \triangleq \lim_{y \downarrow 0} \mathcal{X}_2(y) = \infty$ and $\mathcal{X}_2(\infty) \triangleq \lim_{y \rightarrow \infty} \mathcal{X}_2(y) = 0$. Let us denote by $\mathcal{Y}_2: [0, \infty] \rightarrow_{\text{onto}} [0, \infty]$ the inverse of this function.

For a fixed number $x_2 \in (0, \infty)$, we introduce the random variable

$$(5.7) \quad X_T^{(2)} \triangleq I_2(\mathcal{Y}_2(x_2) \zeta_T).$$

Obviously $X_T^{(2)} \in \mathcal{M}(x_2)$, and from Corollary 2.9 we can find $\pi^{(2)} \in \mathcal{P}(x_2)$ with corresponding wealth process $X^{(2)}$ given, as in (2.34), by

$$(5.8) \quad \begin{aligned} X_t^{(2)} \exp \left(- \int_0^t r(u) du \right) &\triangleq \tilde{E} \left[X_T^{(2)} \exp \left(- \int_0^T r(u) du \right) \middle| \mathcal{F}_t \right] \\ &= x_2 + \int_0^t \exp \left(- \int_0^s r(u) du \right) (\pi^{(2)}(s))^T \sigma(s) d\tilde{W}_s. \end{aligned}$$

THEOREM 5.2. *Let us assume that (5.6) holds. Then for any $x_2 > 0$ the above pair $(\pi^{(2)}, 0)$ belongs to $\mathcal{A}_2(x_2)$ and is optimal for the problem (5.3):*

$$(5.9) \quad V_2(x_2) = E \left[\exp \left(- \int_0^T \beta(s) ds \right) U_2(X_T^{(2)}) \right].$$

Proof. It suffices to show that the random variable $X_T^{(2)} \in \mathcal{M}(x_2)$ of (5.7) satisfies (5.2), and that for any other random variable $X_T \in \mathcal{L}(x_2)$ satisfying this condition we have

$$(5.10) \quad E \left[\exp \left(- \int_0^T \beta(s) ds \right) U_2(X_T) \right] \leq E \left[\exp \left(- \int_0^T \beta(s) ds \right) U_2(X_T^{(2)}) \right].$$

The argument imitates the proof of Theorem 4.3; it hinges on the consequence of (3.1):

$$U_2(X_T^{(2)}) \geq U_2(X_T) + y \zeta_T X_T^{(2)} - y \zeta_T X_T \quad \text{a.s.}$$

which is valid for every $X_T \in \mathcal{L}(x_2)$ and is applied, first to the constant $b \in \mathcal{M}(x_2)$ of (5.5) with $x = x_2$, and then to an arbitrary $X_T \in \mathcal{L}(x_2)$ which satisfies (5.2). The details are omitted. \square

The condition

$$(5.11) \quad E \left[\exp \left(- \int_0^T \beta(u) du \right) |U_2(I_2(y\xi_T))| \right] < \infty \quad \forall y \in (0, \infty)$$

guarantees the finiteness of the value function V_2 .

PROPOSITION 5.3. *Let the utility function U_2 satisfy (3.2), (5.6) and (5.11). Then the function $G_2: (0, \infty) \rightarrow \mathcal{R}$ given by*

$$(5.12) \quad G_2(y) \triangleq E \left[\exp \left(- \int_0^T \beta(s) ds \right) U_2(I_2(y\xi_T)) \right], \quad 0 < y < \infty$$

is strictly decreasing, continuously differentiable, and satisfies

$$(5.13) \quad G_2'(y) = y\mathcal{X}_2'(y), \quad 0 < y < \infty,$$

$$(5.14) \quad V_2(x) = G_2(\mathcal{Y}_2(x)), \quad 0 < x < \infty.$$

The proof follows that of Proposition 4.4. Finally, just as in Lemma 4.6, if $U_2(0) > -\infty$ then (5.11) implies (5.6).

Remark 5.4. Just as in Remark 4.5, under the conditions of Proposition 5.3 the function V_2 is strictly concave with $V_2'(0) = \infty$. This follows from the formulas

$$V_2'(x) = \mathcal{Y}_2(x), \quad V_2''(x) = \mathcal{Y}_2'(x), \quad 0 \leq x < \infty.$$

Remark 5.5. In the special case $U_2(c) = \log c$, we have

$$\mathcal{X}_2(y) = \frac{\alpha_2}{y}, \quad G_2(y) = -\alpha_2 \cdot \log y + \delta_2,$$

$$V_2(x) = \alpha_2 \cdot \log \left(\frac{x}{\alpha_2} \right) + \delta_2$$

where

$$\alpha_2 \triangleq E \exp \left(- \int_0^T \beta(u) du \right),$$

$$\delta_2 \triangleq E \left[\exp \left(- \int_0^T \beta(u) du \right) \left\{ \sum_{i=1}^d \int_0^T \theta_i(u) dW_u^{(i)} - \int_0^T \left(\beta(u) - r(u) - \frac{\|\theta(u)\|^2}{2} \right) du \right\} \right].$$

6. Maximization of utility from both consumption and terminal wealth. A stochastic control problem, which is arguably more interesting than those studied in §§ 4 and 5, concerns the maximization of the total expected discounted utility

$$(6.1) \quad \begin{aligned} J(x; \pi, C) &\triangleq J_1(x; \pi, C) + J_2(x; \pi, C) \\ &= E \left[\int_0^T \exp \left(- \int_0^t \beta(s) ds \right) U_1(C_t) dt \right] \\ &\quad + E \left[\exp \left(- \int_0^T \beta(s) ds \right) U_2(X_T) \right] \end{aligned}$$

from both consumption and terminal wealth, over the class $\mathcal{A}_{1,2}(x) \triangleq \mathcal{A}_1(x) \cap \mathcal{A}_2(x)$:

$$(6.2) \quad V(x) \triangleq \sup_{(\pi, C) \in \mathcal{A}_{1,2}(x)} J(x; \pi, C).$$

Unlike the two problems studied already, this one calls for balancing *competing objectives*. Single-minded determination to “become rich” (i.e., to maximize $J_2(x; \pi, C)$)

mandates no consumption whatsoever—recall Theorem 5.2. On the other hand, single-minded “consumerism” (i.e., maximization of $J_1(x; \pi, C)$) will leave the investor broke at the end; cf. Remark 2.7 and Theorem 4.3. Both these alternatives are suboptimal in the present context.

We shall demonstrate that the proper compromise between these two competing objectives can be drawn in a very simple way. At time $t = 0$ the investor simply divides the endowment x into two parts $x_1 \geq 0$, $x_2 \geq 0$ with $x_1 + x_2 = x$; for the amount x_1 (respectively, x_2) the investor will face, from then on, an optimization problem with utility coming *only* from consumption (respectively, *only* from terminal wealth). It will be shown just how x_1 , x_2 should be determined, in order for the resulting procedure to be optimal.

Throughout this section it will be assumed that U_1 , U_2 are utility functions (§ 3) for which (3.2), (4.8), (4.16) and (5.6), (5.11) hold.

PROPOSITION 6.1. *For $x \geq 0$ and an arbitrary portfolio/consumption pair $(\pi, C) \in \mathcal{A}_{1,2}(x)$, let*

$$x_1 \triangleq \tilde{E} \int_0^T \exp \left(- \int_0^t \beta(u) du \right) C_t dt.$$

Then there exists a pair $(\tilde{\pi}, \tilde{C}) \in \mathcal{A}_{1,2}(x)$ such that

$$(6.3) \quad J(x; \pi, C) \leq J(x; \tilde{\pi}, \tilde{C}) = V_1(x_1) + V_2(x - x_1).$$

In particular,

$$(6.4) \quad V(x) \leq V_*(x) \triangleq \max_{\substack{x_1, x_2 \in [0, \infty) \\ x_1 + x_2 = x}} [V_1(x_1) + V_2(x_2)] = \max_{\substack{y_1, y_2 \in [0, \infty] \\ \mathcal{G}_1(y_1) + \mathcal{G}_2(y_2) = x}} [G_1(y_1) + G_2(y_2)].$$

Proof. From (2.22) we have that $x_1 \in [0, x]$ and

$$\tilde{E} \left[\exp \left(- \int_0^T r(u) du \right) X_T \right] \leq x_2 \triangleq x - x_1.$$

If $x_i = 0$, take $\pi^{(i)} \equiv 0$, $C^{(i)} \equiv 0$. If $x_1 > 0$, then the pair $(\pi^{(1)}, C^{(1)})$ of Theorem 4.3, with corresponding wealth process $X^{(1)}$ as in (4.10), is optimal for $V_1(x_1)$. Similarly, if $x_2 > 0$, the pair $(\pi^{(2)}, 0)$ of Theorem 5.2, with $y = \mathcal{Q}_2(x_2)$ and corresponding wealth process $X^{(2)}$ as in (5.8), is optimal for $V_2(x_2)$. Now let $\tilde{C} \triangleq C^{(1)}$, $\tilde{\pi} \triangleq \pi^{(1)} + \pi^{(2)}$ and observe from (4.10), (5.8) that $\tilde{X} \triangleq X^{(1)} + X^{(2)}$ can be written in the form (2.19):

$$\begin{aligned} \tilde{X}_t \exp \left(- \int_0^t r(u) du \right) &= x - \int_0^t \tilde{C}_s \exp \left(- \int_0^s r(u) du \right) ds \\ &\quad + \int_0^t \exp \left(- \int_0^s r(u) du \right) \tilde{\pi}^T(s) \sigma(s) d\tilde{W}_s \\ (6.5) \quad &= \tilde{E} \left[\int_t^T \tilde{C}_s \exp \left(- \int_0^s r(u) du \right) ds \right. \\ &\quad \left. + \tilde{X}_T \exp \left(- \int_0^T r(u) du \right) \middle| \mathcal{F}_t \right] \end{aligned}$$

because $X_T^{(1)} = 0$, a.s. The process \tilde{X} is nonnegative, and thus $(\tilde{\pi}, \tilde{C}) \in \mathcal{A}_{1,2}(x)$. Now C belongs to $\mathcal{D}(x_1)$ and satisfies (4.2), whereas X_T belongs to $\mathcal{L}(x_2)$ and satisfies (5.2);

from (4.11), (4.12) and (5.9), (5.10) we obtain the following comparisons:

$$E \int_0^T \exp \left(- \int_0^t \beta(u) du \right) U_1(C_t) dt \leq E \int_0^T \exp \left(- \int_0^t \beta(u) du \right) U_1(C_t^{(1)}) dt = V_1(x_1),$$

$$E \left[\exp \left(- \int_0^T \beta(u) du \right) U_2(X_T) \right] \leq E \left[\exp \left(- \int_0^T \beta(u) du \right) U_2(X_T^{(2)}) \right] = V_2(x_2).$$

Memberwise addition leads to (6.3). The identity in (6.4) is a consequence of Propositions 4.4 and 5.3. \square

Remarks 4.5 and 5.4 show that for any initial endowment $x > 0$, the maximum over $x_1, x_2 \in [0, \infty)$, $x_1 + x_2 = x$ indicated in (6.4) is obtained by $x_1 > 0$, $x_2 > 0$ which satisfy $V_1'(x_1) = V_2'(x_2)$. Recalling from these remarks that $\mathcal{V}_i(x_i) = V_i'(x_i)$; $i = 1, 2$, we conclude that the constrained maximization in the last expression of (6.4) is achieved by

$$y_1 = y_2 \equiv y$$

where this common value is determined uniquely by

$$(6.6) \quad \mathcal{X}_1(y) + \mathcal{X}_2(y) = x.$$

With y specified by (6.6), we recall the process ζ of (4.4) and set:

$$(6.7) \quad C_t^* \triangleq I_1(y\zeta_t), \quad X_T^* \triangleq I_2(y\zeta_T),$$

$$(6.8) \quad X_t^* \triangleq \tilde{E} \left[\int_t^T C_s^* \exp \left(- \int_t^s r(u) du \right) ds + X_T^* \exp \left(- \int_t^T r(u) du \right) \middle| \mathcal{F}_t \right]$$

for $0 \leq t \leq T$. The process X^* is nonnegative with

$$X_0^* = \tilde{E} \int_0^T I_1(y\zeta_s) \exp \left(- \int_0^s r(u) du \right) ds + \tilde{E} \left[\exp \left(- \int_0^T r(u) du \right) I_2(y\zeta_T) \right]$$

$$= \mathcal{X}_1(y) + \mathcal{X}_2(y) = x.$$

Remark 2.7, Corollary 2.9 and the idea of superposition of portfolio, consumption and wealth processes as in the proof of Proposition 6.1 can be employed to show that X^* is the wealth process corresponding to the pair $(\pi^*, C^*) \in \mathcal{A}_{1,2}(x)$, for a suitable portfolio π^* .

THEOREM 6.2. *The above pair $(\pi^*, C^*) \in \mathcal{A}_{1,2}(x)$ is optimal for the problem (6.2).*

Proof. From (4.17), (5.12) and (6.4) we have

$$(6.9) \quad V(x) \geq E \left[\int_0^T \exp \left(- \int_0^t \beta(u) du \right) U_1(C_t^*) dt \right. \\ \left. + \exp \left(- \int_0^T \beta(u) du \right) U_2(X_T^*) \right]$$

$$= E \left[\int_0^T \exp \left(- \int_0^t \beta(u) du \right) U_1(I_1(y\zeta_t)) dt \right]$$

$$+ E \left[\exp \left(- \int_0^T \beta(u) du \right) U_2(I_2(y\zeta_T)) \right]$$

$$= G_1(y) + G_2(y) = V_*(x) \geq V(x),$$

because the value of $y > 0$ determined by (6.6) achieves the maximum in (6.4). \square

Remark 6.3. The functions

$$(6.10) \quad \begin{aligned} \mathcal{X}(y) &\triangleq \mathcal{X}_1(y) + \mathcal{X}_2(y) \\ &= \tilde{E} \left[\int_0^T \exp \left(- \int_0^t r(u) du \right) I_1(y \zeta_t) dt + \exp \left(- \int_0^T r(u) du \right) I_2(y \zeta_T) \right], \end{aligned}$$

$$(6.11) \quad \begin{aligned} G(y) &\triangleq G_1(y) + G_2(y) \\ &= E \left[\int_0^T \exp \left(- \int_0^t \beta(u) du \right) U_1(I_1(y \zeta_t)) dt \right. \\ &\quad \left. + \exp \left(- \int_0^T \beta(u) du \right) U_2(I_2(y \zeta_T)) \right], \end{aligned}$$

are C^1 and strictly decreasing on $(0, \infty)$, and satisfy

$$(6.12) \quad G'(y) = y \mathcal{X}'(y)$$

there (cf. (4.18), (5.13)). The former actually maps $[0, \infty]$ onto itself, with $\mathcal{X}(0) = \infty$, $\mathcal{X}(\infty) = 0$; if we denote its inverse by $\mathcal{Y}: [0, \infty] \rightarrow^{\text{onto}} [0, \infty]$, then (6.9) and (6.6) show that the value function V of (6.2) is representable as

$$(6.13) \quad V(x) = G(\mathcal{Y}(x)), \quad 0 < x < \infty.$$

As in Remarks 4.5 and 5.4, V is strictly concave with $V'(0) = \infty$ and $V'(x) = \mathcal{Y}(x)$; $0 \leq x < \infty$. It is shown in the next section that the functions \mathcal{X} , G (and thus also V) above can be computed *in closed form* in the case of constant coefficients β , r , b and σ , for fairly general utility functions U_1 , U_2 .

Example 6.4. In the case $U_1(c) = U_2(c) = \log c$, the functions of (6.10), (6.11) and (6.13) are given by

$$\mathcal{X}(y) = \frac{\alpha}{y}, \quad G(y) = -\alpha \cdot \log y + \delta, \quad 0 < y < \infty$$

and

$$V(x) = \alpha \cdot \log \left(\frac{x}{\alpha} \right) + \delta, \quad 0 < x < \infty$$

where $\alpha \triangleq \alpha_1 + \alpha_2$, $\delta \triangleq \delta_1 + \delta_2$ in the notation of Remarks 4.7 and 5.5.

7. Model with constant coefficients. The developments in §§ 4–6 were based on martingale methods and provided useful and very explicit information about the optimal consumption and wealth processes, in the respective problems treated there. Concerning the optimal portfolio process, however, these methods were able to ascertain only its existence, without shedding much light on its properties or providing any useful characterizations. In order to amend this drawback, we shall specialize in this section the model to the case of *constant coefficients*

$$(7.1) \quad \beta(t) \equiv \beta, \quad r(t) \equiv r, \quad b(t) \equiv b, \quad \sigma(t) \equiv \sigma, \quad 0 \leq t \leq T$$

where β , r are given real number, b is a fixed vector in \mathcal{R}^d and σ is a constant, nonsingular $(d \times d)$ -matrix. The section culminates with Theorem 7.7, which provides the optimal portfolio/consumption pair in a closed and “feedback” form.

It will be assumed throughout that the utility functions U_j ; $j=1, 2$ satisfy the conditions of § 3, including (3.2); in addition, it will be supposed that they are of class C^3 and

$$(7.2) \quad \lim_{c \downarrow 0} \frac{(U'_j(c))^2}{U''_j(c)} \text{ exists,}$$

$$(7.3) \quad \lim_{c \rightarrow \infty} \frac{(U'_j(c))^\alpha}{U''_j(c)} = 0 \quad \text{for some } \alpha > 2,$$

$$(7.4) \quad U_j(0) > -\infty$$

hold for $j=1, 2$. These are certainly not the weakest conditions which permit the ensuing analysis, but they are convenient for our purposes. Furthermore, they include all of the so-called HARA functions (see KLSS (1986) for a definition) except for $U_j(c) = \log c$. We have dealt with this case separately, however, in Remarks 4.7 and 5.5.

The assumption of constant coefficients will permit us to employ "Markovian" methods, such as the Hamilton-Jacobi-Bellman (HJB) equation of Dynamic Programming (Proposition 7.6), in contrast to the "martingale" techniques of the previous sections. In this regard, it will be helpful to consider the problem of § 6 with initial times other than zero. Thus, for fixed $(t, x) \in [0, T] \times (0, \infty)$, we define the value function

$$(7.5) \quad V(t, x) = \sup_{(\pi, C) \in \mathcal{A}(t, x)} E \left[\int_t^T e^{-\beta s} U_1(C_s) ds + e^{-\beta T} U_2(X_T) \right]$$

where the class $\mathcal{A}(t, x)$ consists of those portfolio/consumption process pairs for which the corresponding wealth process

$$(7.6) \quad \begin{aligned} X_s = x + \int_t^s (rX_u - C_u) du + \sum_{i=1}^d \int_t^s (b_i - r) \pi_i(u) du \\ + \sum_{i=1}^d \sum_{j=1}^d \int_t^s \pi_i(u) \sigma_{ij} dW_u^{(j)}, \quad t \leq s \leq T \end{aligned}$$

remains nonnegative, almost surely. Corresponding to a given consumption process C , there exists a portfolio π with $(\pi, C) \in \mathcal{A}(t, x)$ if and only if

$$E \int_t^T Z_t^s e^{-r(s-t)} C_s ds \leq x$$

(Proposition 2.6), where we employ the notation

$$(7.7) \quad Z_s^t \triangleq \exp \{ -\theta^T (W_s - W_t) - \frac{1}{2} \|\theta\|^2 (s-t) \}, \quad \zeta_s^t \triangleq e^{(\beta-r)(s-t)} Z_s^t, \quad t \leq s \leq T.$$

Now for $(t, y) \in [0, T] \times (0, \infty)$ we consider the process

$$(7.8) \quad Y_s^{(t, y)} \triangleq y \zeta_s^t, \quad t \leq s \leq T$$

and the functions

$$(7.9) \quad G(t, y) \triangleq E \left[\int_t^T e^{-\beta(s-t)} U_1(I_1(Y_s^{(t, y)})) ds + e^{-\beta(T-t)} U_2(I_2(Y_T^{(t, y)})) \right],$$

$$(7.10) \quad S(t, y) \triangleq E \left[\int_t^T e^{-\beta(s-t)} Y_s^{(t, y)} I_1(Y_s^{(t, y)}) ds + e^{-\beta(T-t)} Y_T^{(t, y)} I_2(Y_T^{(t, y)}) \right],$$

$$(7.11) \quad \mathcal{X}(t, y) \triangleq \frac{S(t, y)}{y}.$$

We shall show (Lemmas 7.1, 7.2) that G and S are well defined and finite, and then just as in Lemma 4.2, we have that for every $0 \leq t < T$ the function $\mathcal{X}(t, \cdot)$ is continuous and strictly decreasing on $(0, \infty)$, with $\mathcal{X}(t, 0) \triangleq \lim_{y \downarrow 0} \mathcal{X}(t, y) = \infty$ and $\mathcal{X}(t, \infty) \triangleq \lim_{y \rightarrow \infty} \mathcal{X}(t, y) = 0$. We denote by $\mathcal{Y}(t, \cdot): [0, \infty] \rightarrow_{\text{onto}} [0, \infty]$ the inverse of $\mathcal{X}(t, \cdot)$, namely

$$(7.12) \quad \mathcal{Y}(t, \mathcal{X}(t, y)) = y, \quad 0 \leq t < T, \quad 0 \leq y \leq \infty.$$

If we define, for any given $x > 0$, the processes

$$(7.13) \quad \eta_s^{(t,x)} \triangleq \mathcal{Y}(t, x) \cdot \xi_s^t,$$

$$(7.14) \quad C_s^{(t,x)} \triangleq I_1(\eta_s^{(t,x)}), \quad t \leq s \leq T$$

and the random variable

$$(7.15) \quad X_T^{(t,x)} \triangleq I_2(\eta_T^{(t,x)}),$$

then we can show, as in Theorem 6.2, that

$$(7.16) \quad V(t, x) = E \left[\int_t^T e^{-\beta s} U_1(C_s^{(t,x)}) ds + e^{-\beta T} U_2(X_T^{(t,x)}) \right].$$

The new feature here is that, for $y = \mathcal{Y}(t, x)$, one has $\eta^{(t,x)} = Y^{(t,y)}$ from (7.8), (7.13), and consequently

$$(7.17) \quad V(t, x) = e^{-\beta t} G(t, \mathcal{Y}(t, x)), \quad 0 \leq t \leq T, \quad 0 < x < \infty.$$

On the other hand, by analogy with (6.8), the optimal wealth process is seen to be as follows:

$$(7.18) \quad \begin{aligned} X_s^{(t,x)} &\triangleq E \left[\int_s^T e^{-r(\theta-s)} Z_\theta^s I_1(\eta_\theta^{(t,x)}) d\theta + e^{-r(T-s)} Z_T^s I_2(\eta_T^{(t,x)}) \middle| \mathcal{F}_s \right] \\ &= \frac{1}{Y_s^{(t,y)}} E \left[\int_s^T e^{-\beta(\theta-s)} Y_\theta^{(t,y)} I_1(Y_\theta^{(t,y)}) d\theta + e^{-\beta(T-t)} Y_T^{(t,y)} I_2(Y_T^{(t,y)}) \middle| \mathcal{F}_s \right] \\ &\equiv \frac{S(s, Y_s^{(t,y)})}{Y_s^{(t,y)}} = \mathcal{X}(s, \eta_s^{(t,x)}), \quad t \leq s \leq T, \end{aligned}$$

almost surely. In (7.18) we used the “Bayes rule” for conditional expectations under the measure $Z_\theta^t dP$ (cf. proof of Proposition 2.6) and the Markov property for $\{Y_s^{(t,y)}, \mathcal{F}_s; t \leq s \leq T\}$.

Our program now is to characterize G, S of (7.9), (7.10) in terms of two Cauchy problems involving the *linear* differential operator

$$L\varphi(t, y) \triangleq \gamma y^2 \frac{\partial^2 \varphi(t, y)}{\partial y^2} + (\beta - r)y \frac{\partial \varphi(t, y)}{\partial y} - \beta \varphi(t, y),$$

where

$$\gamma \triangleq \frac{1}{2} \|\theta\|^2, \quad \theta \triangleq \sigma^{-1}(b - r\mathbf{1})$$

in accordance with (2.15). We shall produce *closed-form* solutions for these Cauchy problems (7.23), (7.24) and (7.25), (7.26), and thus also for the function $\mathcal{X}(t, y)$ of (7.11). In this manner, an expression for the value function $V(t, x)$ will be made available via (7.17).

We shall need the following Feynman–Kac result, whose proof is deferred to the Appendix.

LEMMA 7.1. Let the real-valued function $q(t, y)$ be defined and continuous on $[0, T] \times (0, \infty)$, as well as Hölder continuous in y uniformly with respect to (t, y) on compact subsets of the domain. Also, let $f: (0, \infty) \rightarrow \mathbb{R}$ be continuous, and assume that both q, f satisfy a growth condition of the form

$$(7.19) \quad \max_{0 \leq t \leq T} |u(t, y)| \leq K(1 + y^\alpha + y^{-\alpha}), \quad 0 < y < \infty$$

for some $K > 0$, $\alpha > 0$. Then the function

$$(7.20) \quad H(t, y) = E \left[\int_t^T e^{-\beta(s-t)} q(s, Y_s^{(t,y)}) ds + e^{-\beta(T-t)} f(Y_T^{(t,y)}) \right]$$

solves the Cauchy problem

$$(7.21) \quad \left(\frac{\partial}{\partial t} + L \right) H(t, y) + q(t, y) = 0, \quad 0 \leq t < T, \quad 0 < y < \infty$$

$$(7.22) \quad H(T, y) = f(y), \quad 0 < y < \infty,$$

and is actually the unique solution of (7.21), (7.22) in the class of $C([0, T] \times (0, \infty)) \cap C^{1,2}([0, T) \times (0, \infty))$ functions which satisfy a growth condition of the type (7.19). \square

We shall use Lemma 7.1 to characterize G and S as the unique solutions to the Cauchy problems

$$(7.23) \quad \left(\frac{\partial}{\partial t} + L \right) G(t, y) + U_1(I_1(y)) = 0, \quad 0 \leq t < T, \quad 0 < y < \infty,$$

$$(7.24) \quad G(T, y) = U_2(I_2(y)), \quad 0 < y < \infty,$$

and

$$(7.25) \quad \left(\frac{\partial}{\partial t} + L \right) S(t, y) + yI_1(y) = 0, \quad 0 \leq t < T, \quad 0 < y < \infty,$$

$$(7.26) \quad S(T, y) = yI_2(y), \quad 0 < y < \infty,$$

respectively. We will then solve these Cauchy problems explicitly. For this second task, we introduce the function

$$(7.27) \quad v(t, y, \xi) \triangleq \begin{cases} \xi e^{-\beta(T-t)} \Phi(-\mu_-(T-t, y, \xi)) - y e^{-r(T-t)} \Phi(-\mu_+(T-t, y, \xi)), & 0 \leq t < T, \\ (\xi - y)^+, & t = T, \end{cases}$$

for $0 < y, \xi < \infty$, where

$$\Phi(z) \triangleq \frac{1}{\sqrt{2\pi}} \int_{-\infty}^z e^{-x^2/2} dx, \quad \mu_{\pm}(t, y, \xi) = \frac{1}{\sqrt{2\gamma t}} \left[\log \left(\frac{y}{\xi} \right) + (\beta - r \pm \gamma)t \right].$$

Straightforward computation shows that, for fixed ξ , v is of class $C([0, T] \times (0, \infty)) \cap C^{1,2}([0, T) \times (0, \infty))$ and solves the Cauchy problem

$$(7.28) \quad \left(\frac{\partial}{\partial t} + L \right) v(t, y, \xi) = 0, \quad 0 \leq t < T, \quad 0 < y < \infty,$$

$$(7.29) \quad v(T, y, \xi) = (\xi - y)^+, \quad 0 < y < \infty.$$

It is apparent that for $0 \leq t \leq T$,

$$(7.30) \quad \overline{\lim}_{y \rightarrow \infty} \frac{|v(t, y, \xi)|}{y} < \infty, \quad \overline{\lim}_{\xi \rightarrow \infty} \frac{|v(t, y, \xi)|}{\xi} < \infty,$$

and from the Mills' ratio (e.g. McKean (1969, p. 4))

$$\sqrt{2\pi}\Phi(-z) \leq \frac{1}{z} e^{-z^2/2}$$

one can also establish that for every $\alpha > 0$,

$$(7.31) \quad \overline{\lim}_{y \downarrow 0} y^{-\alpha} |v(t, y, \xi)| < \infty, \quad \overline{\lim}_{\xi \downarrow 0} \xi^{-\alpha} |v(t, y, \xi)| < \infty.$$

In particular, for each fixed $0 < \xi < \infty$, Lemma 7.1 can be invoked to provide the stochastic representation

$$(7.32) \quad v(t, y, \xi) = E[e^{-\beta(T-t)}(\xi - Y_T^{(t,y)})^+].$$

We also introduce the functions

$$(7.33) \quad g(y) \triangleq \frac{U_1(I_1(y))}{\beta} - \frac{1}{\gamma(\lambda_+ - \lambda_-)} \left\{ \frac{y^{1+\lambda_+}}{1+\lambda_+} J_+(y) - \frac{y^{1+\lambda_-}}{1+\lambda_-} J_-(y) \right\},$$

$$(7.34) \quad s(y) \triangleq \frac{yI_1(y)}{r} - \frac{1}{\gamma(\lambda_+ - \lambda_-)} \left\{ \frac{y^{1+\lambda_+}}{\lambda_+} J_+(y) - \frac{y^{1+\lambda_-}}{\lambda_-} J_-(y) \right\},$$

where

$$(7.35) \quad J_+(y) = \int_0^{I_1(y)} \frac{d\eta}{(U'_1(\eta))^{\lambda_+}}, \quad J_-(y) = \int_1^{I_1(y)} \frac{d\eta}{(U'_1(\eta))^{\lambda_-}},$$

and λ_+ , λ_- are the positive and negative roots, respectively, of the quadratic equation

$$(7.36) \quad \gamma\lambda^2 - (r - \beta - \gamma)\lambda - r = 0.$$

Note that $\lambda_+\lambda_- = -r/\gamma$, $(1+\lambda_+)(1+\lambda_-) = -\beta/\gamma$, and $\lambda_- < -1$. Direct computation shows that

$$(7.37) \quad g'(y) = -\frac{1}{\gamma(\lambda_+ - \lambda_-)} \{y^{\lambda_+} J_+(y) - y^{\lambda_-} J_-(y)\},$$

$$(7.38) \quad g''(y) = -\frac{1}{\gamma(\lambda_+ - \lambda_-)} \{\lambda_+ y^{\lambda_+-1} J_+(y) - \lambda_- y^{\lambda_--1} J_-(y)\},$$

$$(7.39) \quad s'(y) = \frac{I_1(y)}{r} - \frac{1}{\gamma(\lambda_+ - \lambda_-)} \left\{ \frac{1+\lambda_+}{\lambda_+} y^{\lambda_+} J_+(y) - \frac{1+\lambda_-}{\lambda_-} y^{\lambda_-} J_-(y) \right\},$$

$$(7.40) \quad s''(y) = -\frac{1}{\gamma(\lambda_+ - \lambda_-)} \{(1+\lambda_+)y^{\lambda_+-1} J_+(y) - (1+\lambda_-)y^{\lambda_--1} J_-(y)\},$$

and

$$(7.41) \quad Lg(y) + U_1(I_1(y)) = 0,$$

$$(7.42) \quad Ls(y) + yI_1(y) = 0.$$

Our arguments will utilize the following growth, integrability, and limit properties, whose proofs are deferred to the Appendix.

LEMMA 7.2. *Each of the functions $g(y)$, $s(y)$, $U_j(I_j(y))$, $yI_j(y)$; $j = 1, 2$, satisfies the growth condition (7.19) for some $K > 0$, $\alpha > 0$. Furthermore, each of these functions is of class $C^2(0, \infty)$ and satisfies the integrability conditions*

$$(7.43) \quad \int_1^\infty \xi |f''(\xi)| d\xi < \infty, \quad \int_0^1 \xi^\alpha |f''(\xi)| d\xi < \infty$$

for some $\alpha > 0$. Finally, the limit property

$$(7.44) \quad \lim_{\xi \rightarrow \infty} \xi f'(\xi) = 0$$

holds for each of the above-named functions. For $s(y)$, $yI_j(y)$; $j = 1, 2$ we have

$$(7.45) \quad \lim_{\xi \rightarrow \infty} f(\xi) = 0,$$

whereas

$$(7.46) \quad \lim_{\xi \rightarrow \infty} g(\xi) = \frac{U_1(0)}{\beta}, \quad \lim_{\xi \rightarrow \infty} U_j(I_j(\xi)) = U_j(0), \quad j = 1, 2. \quad \square$$

Lemmas 7.1 and 7.2 imply that G and S , defined by (7.9) and (7.10), are indeed characterized by the Cauchy problems (7.23), (7.24) and (7.25), (7.26).

PROPOSITION 7.3. *The functions G , S of (7.9), (7.10) admit the following stochastic representations:*

$$(7.47) \quad G(t, y) = g(y) + E[e^{-\beta(T-t)}\{U_2(I_2(Y_T^{(t,y)})) - g(Y_T^{(t,y)})\}],$$

$$(7.48) \quad S(t, y) = s(y) + E[e^{-\beta(T-t)}\{Y_T^{(t,y)}I_2(Y_T^{(t,y)}) - s(Y_T^{(t,y)})\}],$$

and are expressible in closed form as

$$(7.49) \quad G(t, y) = g(y) + \left(U_2(0) - \frac{U_1(0)}{\beta} \right) e^{-\beta(T-t)} + \int_0^\infty (U_2(I_2(\xi)) - g(\xi))'' v(t, y, \xi) d\xi,$$

$$(7.50) \quad S(t, y) = s(y) + \int_0^\infty (\xi I_2(\xi) - s(\xi))'' v(t, y, \xi) d\xi.$$

Proof. We deal with the function G only; the treatment for S is similar. Because of the growth assertions in Lemma 7.2 and the special form (7.7), (7.8) of $Y^{(t,y)}$, the functions $G_1(t, y) \triangleq g(y)$, $G_2(t, y) \triangleq E[e^{-\beta(T-t)}U_2(I_2(Y_T^{(t,y)}))]$ and $G_3(t, y) \triangleq E[-e^{-\beta(T-t)}g(Y_T^{(t,y)})]$, as well as their sum $\hat{G} \triangleq G_1 + G_2 + G_3$, satisfy growth conditions of the type (7.19). By virtue of (7.41) and Lemma 7.1, these functions satisfy in $[0, T) \times (0, \infty)$ the equations

$$\left(\frac{\partial}{\partial t} + L \right) G_1(t, y) + U_1(I_1(y)) = 0, \quad \left(\frac{\partial}{\partial t} + L \right) G_j(t, y) = 0, \quad j = 2, 3,$$

and the terminal conditions

$$G_2(T, y) = U_2(I_2(y)), \quad G_1(T, y) = -G_3(T, y) = g(y).$$

Thus \hat{G} solves the Cauchy problem (7.23), (7.24) which characterizes G ; (7.47) follows.

Turning to (7.49), we note first that the integral on the right-hand side is defined because of (7.30), (7.31), (7.43). In conjunction with Fubini's Theorem, (7.32) yields

$$(7.51) \quad \begin{aligned} & \int_0^\infty (U_2(I_2(\xi)) - g(\xi))'' v(t, y, \xi) d\xi \\ &= E \int_0^\infty e^{-\beta(T-t)} (U_2(I_2(\xi)) - g(\xi))'' (\xi - Y_T^{(t,y)})^+ d\xi. \end{aligned}$$

Integration by parts shows that

$$\int_0^\infty f''(\xi)(\xi - y)^+ d\xi = f(y) + \lim_{\xi \rightarrow \infty} \xi f'(\xi) - \lim_{\xi \rightarrow \infty} f(\xi)$$

for any function f for which the limits are defined. Relations (7.44)–(7.46) thus allow us to evaluate (7.51) and discover that (7.49) is equivalent to (7.47). \square

Remark 7.4. The nonnegative function $\mathcal{X}(t, y)$ of (7.11) satisfies the Cauchy problem

$$(7.52) \quad \mathcal{X}_t(t, y) + \gamma y^2 \mathcal{X}_{yy}(t, y) + (\beta - r + 2\gamma)y \mathcal{X}_y - r \mathcal{X}(t, y) + I_1(y) = 0, \\ 0 \leq t < T, \quad 0 < y < \infty,$$

$$(7.53) \quad \mathcal{X}(T, y) = I_2(y), \quad 0 < y < \infty.$$

For its inverse $\mathcal{Y}(t, x)$, as in (7.21), we have

$$(7.54) \quad \mathcal{Y}_x(t, \mathcal{X}(t, y)) = \frac{1}{\mathcal{X}_y(t, y)} < 0, \quad \mathcal{Y}_t(t, \mathcal{X}(t, y)) = -\frac{\mathcal{X}_t(t, y)}{\mathcal{X}_y(t, y)}.$$

Furthermore, it can be shown by analogy with (6.12) that

$$(7.55) \quad G_y(t, y) = y \cdot \mathcal{X}_y(t, y), \quad 0 \leq t < T, \quad 0 < y < \infty.$$

Example 7.5. In the special case $U_1(c) = U_2(c) = c^\delta$ for some $\delta \in (0, 1)$, we set

$$k = \frac{1}{1-\delta} \left(\beta - r\delta - \frac{\gamma\delta}{1-\delta} \right)$$

and

$$p(t) = \begin{cases} \frac{1 - e^{-k(T-t)}}{k} + e^{-k(T-t)}, & k \neq 0, \\ 1 + T - t, & k = 0. \end{cases}$$

Then

$$G(t, y) = p(t) \left(\frac{y}{\delta} \right)^{\delta/(\delta-1)}, \quad S(t, y) = \delta G(t, y), \quad \mathcal{X}(t, y) = p(t) \left(\frac{y}{\delta} \right)^{1/(\delta-1)},$$

and

$$V(t, x) = e^{-\beta t} (p(t))^{1-\delta} x^\delta.$$

PROPOSITION 7.6. *The function $V(t, x) : [0, T] \times \mathcal{R} \rightarrow \mathcal{R}$ of (7.16) satisfies the following initial-boundary value problem for the Hamilton–Jacobi–Bellman (HJB) equation of dynamic programming, associated with the stochastic control problem (7.5) and the dynamics (7.6):*

(7.56)

$$V_t(t, x) + \max_{\substack{c \geq 0 \\ \pi \in \mathcal{R}^d}} \left[\frac{1}{2} \|\sigma^T \pi\|^2 \cdot V_{xx}(t, x) + \{(rx - c) + \pi^T(b - r\mathbf{1})\} \cdot V_x(t, x) + e^{-\beta t} U_1(c) \right] = 0 \\ \text{in } [0, T) \times (0, \infty),$$

(7.57)

$$V(T, x) = e^{-\beta T} U_2(x), \quad 0 \leq x < \infty,$$

(7.58)

$$V(t, 0) = \left(U_2(0) - \frac{U_1(0)}{\beta} \right) e^{-\beta T} + \frac{U_1(0)}{\beta} e^{-\beta t}, \quad 0 \leq t \leq T.$$

Proof. The last two conditions are quite obvious; on the other hand, if we recall the expression (7.17) for V , we can cast the HJB equation (7.56), with the help of (7.54), (7.55), in the form

$$(7.59) \quad G_t(t, \mathcal{Y}(t, x)) - \beta G(t, \mathcal{Y}(t, x)) + G_y(t, \mathcal{Y}(t, x)) \mathcal{Y}_t(t, x) + rx \mathcal{Y}(t, x) \\ + \max_{c \geq 0} [U_1(c) - c \mathcal{Y}(t, x)] + \max_{\pi \in \mathcal{R}^d} \left[\frac{1}{2} \|\sigma^T \pi\|^2 \mathcal{Y}_x(t, x) + \pi^T(b - r\mathbf{1}) \mathcal{Y}(t, x) \right] = 0.$$

The maximizations over $c \geq 0$, $\pi \in \mathcal{R}^d$ are achieved by

$$(7.60) \quad \hat{c} = I_1(\mathcal{Y}(t, x)),$$

and

$$(7.61) \quad \hat{\pi} = -(\sigma\sigma^T)^{-1}(b - r\mathbf{1}) \frac{\mathcal{Y}(t, x)}{\mathcal{Y}_x(t, x)},$$

respectively. Substitution of these values into the right-hand side of (7.59) makes the latter read as follows:

$$\begin{aligned} G_t(t, \mathcal{Y}(t, x)) - \beta G(t, \mathcal{Y}(t, x)) + G_y(t, \mathcal{Y}(t, x)) \cdot \mathcal{Y}_t(t, x) + \{rx - I_1(\mathcal{Y}(t, x))\} \mathcal{Y}(t, x) \\ - \gamma \frac{\mathcal{Y}^2(t, x)}{\mathcal{Y}_x(t, x)} + U_1(I_1(\mathcal{Y}(t, x))), \end{aligned}$$

or equivalently, with the change of variable $y = \mathcal{Y}(t, x)$ and the help of (7.23):

$$\begin{aligned} G_t(t, y) - \beta G(t, y) - y\mathcal{X}_t(t, y) + ry\mathcal{X}(t, y) - yI_1(y) - \gamma y^2\mathcal{X}_y(t, y) + U_1(I_1(y)) \\ = y[-\mathcal{X}_t(t, y) + r\mathcal{X}(t, y) - (\beta - r + 2\gamma)y\mathcal{X}_y(t, y) - \gamma y^2\mathcal{X}_{yy}(t, y) - I_1(y)]. \end{aligned}$$

But this last expression vanishes, thanks to (7.52). \square

Notice that we have achieved a closed-form solution of the *strongly nonlinear* HJB equation (7.56), by solving instead the two *linear* equations (7.23), (7.25) subject to the appropriate initial and growth conditions, and then performing the composition (7.17).

The functional form of the maximizers in (7.60), (7.61) now suggests very strongly the nature of an optimal portfolio/consumption process pair in *feedback form*, i.e., in terms of the current level of wealth. Let us recall from (7.18) the optimal wealth process $X^{(t,x)}$ for the problem (7.5).

THEOREM 7.7. *The pair $(\pi^{(t,x)}, C^{(t,x)})$ given by*

$$(7.62) \quad C_s^{(t,x)} = I_1(\mathcal{Y}(s, X_s^{(t,x)})),$$

$$(7.63) \quad \pi_s^{(t,x)} \triangleq -(\sigma\sigma^T)^{-1}(b - r\mathbf{1}) \frac{\mathcal{Y}(s, X_s^{(t,x)})}{\mathcal{Y}_x(s, X_s^{(t,x)})}$$

for $t \leq s \leq T$, belongs to $\mathcal{A}(t, x)$ and is optimal for the stochastic control problem of (7.5).

Proof. It suffices to show that $X^{(t,x)}$ is the wealth process corresponding to the pair $(\pi^{(t,x)}, C^{(t,x)})$ above. Now (7.62) is just a restatement of (7.14), because of (7.18); applying Itô's rule to the latter, in conjunction with

$$d\eta_s^{(t,x)} = (\beta - r)\eta_s^{(t,x)} ds - \eta_s^{(t,x)} \theta^T dW_s, \quad t \leq s \leq T, \quad \eta_t^{(t,x)} = x$$

and (7.52), we obtain:

$$\begin{aligned} dX_s^{(t,x)} &= \mathcal{X}_t(s, \eta_s^{(t,x)}) ds + \mathcal{X}_y(s, \eta_s^{(t,x)}) d\eta_s^{(t,x)} + \gamma(\eta_s^{(t,x)})^2 \mathcal{X}_{yy}(s, \eta_s^{(t,x)}) ds \\ &= (rX_s^{(t,x)} - C_s^{(t,x)}) ds + (b - r\mathbf{1})^T \pi_s^{(t,x)} ds + (\pi_s^{(t,x)})^T \sigma dW_s. \end{aligned}$$

But this is equation (7.6) in differential form. \square

Example 7.5 (continued). In this case (7.62), (7.63) read as follows:

$$C_s^{(t,x)} = \frac{X_s^{(t,x)}}{p(s)}, \quad \pi_s^{(t,x)} = (\sigma\sigma^T)^{-1}(b - r\mathbf{1}) \frac{X_s^{(t,x)}}{1 - \delta}, \quad t \leq s \leq T.$$

8. Appendix.

Proof of Lemma 7.1. For any solution $H(t, y)$ of (7.21), (7.22) in the indicated class, the function $\mathcal{H}(t, z) \triangleq H(t, e^z)$:

- (i) belongs to $C([0, T] \times \mathcal{R}) \cap C^{1,2}([0, T] \times \mathcal{R})$,
- (ii) satisfies a growth condition of the type

$$(8.1) \quad \max_{0 \leq t \leq T} |\mathcal{H}(t, z)| \leq M e^{\alpha|z|}, \quad z \in \mathcal{R}$$

for some $M > 0$, $\alpha > 0$, and

(iii) solves the Cauchy problem with constant coefficients

$$(8.2) \quad \mathcal{H}_t + \gamma \mathcal{H}_{zz} + (\beta - r - \gamma) \mathcal{H}_z - \beta \mathcal{H} + q(t, e^z) = 0, \quad [0, T] \times \mathcal{R},$$

$$(8.3) \quad \mathcal{H}(T, z) = f(e^z), \quad z \in \mathcal{R}.$$

Conversely, if $\mathcal{H}(t, z)$ is a function which satisfies (i)–(iii) above, then $H(t, y) \triangleq \mathcal{H}(t, \log y)$ belongs to $C([0, T] \times (0, \infty)) \cap C^{1,2}([0, T] \times (0, \infty))$, satisfies the Cauchy problem (7.21)–(7.22), as well as a growth condition of the type (7.19).

For every given $(t, z) \in [0, T] \times \mathcal{R}$, let us now consider the process

$$L_s^{(t,z)} \triangleq z + (\beta - r - \gamma)(s - t) - \theta^T(W_s - W_t), \quad t \leq s \leq T.$$

Friedman (1975, Thm. 6.4.6, p. 142) proves the existence of a function $\mathcal{H}(t, z)$ with properties (i) and (iii) and satisfying the growth condition

$$\max_{0 \leq t \leq T} |\mathcal{H}(t, z)| \leq M e^{\varepsilon z^2}, \quad z \in \mathcal{R}$$

for every $\varepsilon > 0$ and some $M = M(\varepsilon)$. According to Karatzas and Shreve (1987, Problem 5.7.7), this function is unique and has the stochastic representation

$$\mathcal{H}(t, z) = E \left[\int_t^T e^{-\beta(s-t)} q(s, \exp(L_s^{(t,z)})) ds + e^{-\beta(T-t)} f(\exp(L_T^{(t,z)})) \right],$$

from which one can obtain the growth condition (8.1). If we take $\log y = z$, then $\log Y_s^{(t,y)} = L_s^{(t,z)}$, and $H(t, y) \triangleq \mathcal{H}(t, \log y)$ is given by the right-hand side of (7.20) and satisfies the assertions of the lemma. \square

Proof of Lemma 7.2. We simplify the notation by writing $U = U_1$, $I = I_1$. We assume that

$$(8.4) \quad U'(0) = \infty,$$

for when $U'(0) < \infty$ we have $I(y) = 0$ for y sufficiently large, and then the assertions of the lemma are more easily proved. We observe from concavity that $cU'(c) \leq U(c) - U(0)$, so according to (7.4),

$$(8.5) \quad \lim_{c \downarrow 0} cU'(c) = 0.$$

From (8.4), (8.5) and the Fundamental Theorem of Calculus,

$$\lim_{c \downarrow 0} \frac{U''(c)}{(U'(c))^2} = \lim_{c \downarrow 0} \frac{1}{c} \int_0^c \frac{U''(\eta)}{(U'(\eta))^2} d\eta = -\lim_{c \downarrow 0} \frac{1}{cU'(c)} = \infty,$$

so in the presence of (8.4), (8.5) the existence of the limit in (7.2) implies the stronger statement:

$$(8.6) \quad \lim_{c \downarrow 0} \frac{(U'(c))^2}{U''(c)} = 0.$$

Finally, from L'Hôpital's rule and (7.3) we obtain

$$(8.7) \quad \lim_{c \rightarrow \infty} (U'(c))^\alpha U(c) = 0 \quad \text{for some } \alpha > 0.$$

To study the properties of $U(I(y))$, $yI(y)$, $g(y)$ and $s(y)$, we make the change of variable $y = U'(c)$. In particular,

$$\lim_{y \downarrow 0} y^\alpha U(I(y)) = \lim_{c \rightarrow \infty} (U'(c))^\alpha U(c) = 0$$

for some $\alpha > 0$, by (8.7). This establishes (7.19) for $U(I(y))$. Similarly,

$$\begin{aligned} \lim_{y \downarrow 0} y^{\alpha+1} I(y) &= \lim_{c \rightarrow \infty} c (U'(c))^{\alpha+1} \leq \lim_{c \rightarrow \infty} (U(c) - U(0)) (U'(c))^\alpha \\ &= 0, \end{aligned}$$

so $yI(y)$ also satisfies (7.19). L'Hôpital's rule yields

$$\begin{aligned} \lim_{y \downarrow 0} y^{\alpha-1+\lambda_+} J_+(y) &= \lim_{c \rightarrow \infty} (U'(c))^{\alpha-1+\lambda_+} \int_0^c \frac{d\eta}{(U'(\eta))^{\lambda_+}} \\ &= \frac{1}{1-\lambda_+-\alpha} \lim_{c \rightarrow \infty} \frac{(U'(c))^\alpha}{U''(c)} = 0 \quad \text{for some } \alpha > 2. \end{aligned}$$

Likewise,

$$(8.8) \quad \lim_{y \downarrow 0} y^{\alpha-1+\lambda_+} J_-(y) = 0 \quad \text{for some } \alpha > 2.$$

Finally,

$$\begin{aligned} (8.9) \quad \lim_{y \rightarrow \infty} y^{1+\lambda_-} J_-(y) &= \lim_{c \downarrow 0} (U'(c))^{1+\lambda_-} \int_1^c \frac{d\eta}{(U'(\eta))^{\lambda_-}} \\ &= -\frac{1}{1+\lambda_-} \lim_{c \downarrow 0} \frac{(U'(c))^2}{U''(c)} = 0 \end{aligned}$$

by (8.6). This concludes the proof that (7.19) is satisfied by all the functions under consideration. On the other hand, just as in (8.9) we have

$$(8.10) \quad \lim_{y \rightarrow \infty} y^{1+\lambda_+} J_+(y) = 0,$$

and taken together, these relations give us (7.44) when $f = g$ or $f = s$. The proof of (7.45) hinges on the observation from (8.5) that

$$(8.11) \quad \lim_{\xi \rightarrow \infty} \xi I(\xi) = \lim_{c \downarrow 0} c U'(c) = 0.$$

For the function $U(I(\xi))$, (7.44) becomes

$$(8.12) \quad \lim_{\xi \rightarrow \infty} \xi \frac{d}{d\xi} U(I(\xi)) = \lim_{\xi \rightarrow \infty} \xi^2 I'(\xi) = \lim_{c \downarrow 0} \frac{(U'(c))^2}{U''(c)} = 0.$$

For the function $\xi I(\xi)$, we have

$$\xi \frac{d}{d\xi} (\xi I(\xi)) = \xi I(\xi) + \xi^2 I'(\xi),$$

and (7.44) follows from (8.11), (8.12). It is now easy to verify (7.46).

It remains only to establish (7.43) for the functions under consideration, and for this it suffices to prove that

$$(8.13) \quad \int_1^\infty \xi |I'(\xi)| d\xi < \infty,$$

$$(8.14) \quad \int_1^\infty \xi^2 |I''(\xi)| d\xi < \infty,$$

$$(8.15) \quad \int_1^\infty \xi^{\lambda_+} |J_+(\xi)| d\xi < \infty,$$

$$(8.16) \quad \int_1^\infty \xi^{\lambda_-} |J_-(\xi)| d\xi < \infty,$$

and for some $\alpha > 0$:

$$(8.17) \quad \int_0^1 \xi^\alpha |I'(\xi)| d\xi < \infty,$$

$$(8.18) \quad \int_0^1 \xi^{\alpha+1} |I''(\xi)| d\xi < \infty,$$

$$(8.19) \quad \int_0^1 \xi^{\alpha-1+\lambda_+} |J_+(\xi)| d\xi < \infty,$$

and

$$(8.20) \quad \int_0^1 \xi^{\alpha-1+\lambda_-} |J_-(\xi)| d\xi < \infty.$$

In each of these integrals, we will make the change of variables $c = I(\xi)$, so $d\xi = U''(c) dc$, $I'(\xi) = 1/U''(c) \leq 0$, and $I''(\xi) = -U'''(c)/(U''(c))^3 \geq 0$. Beginning with (8.13), we write

$$\int_1^\infty \xi |I'(\xi)| d\xi = \int_0^{I(1)} U'(c) dc = U(I(1)) - U(0) < \infty.$$

Because of (8.13), finiteness of the integral in (8.14) is equivalent to finiteness of

$$\begin{aligned} \int_1^\infty [\xi^2 I''(\xi) + 2\xi I'(\xi)] d\xi &= \int_0^{I(1)} \frac{d}{dc} \left[\frac{(U'(c))^2}{U''(c)} \right] dc \\ &= \frac{1}{U''(I(1))} - \lim_{c \downarrow 0} \frac{(U'(c))^2}{U''(c)} = \frac{1}{U''(I(1))}, \end{aligned}$$

thanks to (8.6). Because U' is decreasing, we may bound (8.15) by

$$\begin{aligned} \int_1^\infty \xi^{\lambda_+} |J_+(\xi)| d\xi &= \int_0^{I(1)} (U'(c))^{\lambda_+} |U''(c)| \int_0^c \frac{d\eta}{(U'(\eta))^{\lambda_+}} dc \\ &\leq - \int_0^{I(1)} c U''(c) dc \\ &= -c U'(c) \Big|_{c=0}^{c=I(1)} + \int_0^{I(1)} U'(c) dc. \end{aligned}$$

These last expressions are finite because of (8.5) and (7.4). In (8.16) we integrate by parts to obtain

$$\begin{aligned} \int_{U'(1)}^\infty \xi^{\lambda_-} |J_-(\xi)| d\xi &= - \int_0^1 (U'(c))^{\lambda_-} U''(c) \int_c^1 \frac{d\eta}{(U'(\eta))^{\lambda_-}} dc \\ &= - \frac{1}{1+\lambda_-} (U'(c))^{1+\lambda_-} \int_c^1 \frac{d\eta}{(U'(\eta))^{\lambda_-}} \Big|_{c=0}^{c=1} - \frac{1}{1+\lambda_-} \int_0^1 U'(c) dc. \end{aligned}$$

We conclude from (8.9) that the resulting expression is finite.

Now let us take up (8.17)–(8.20). For (8.17) we write

$$(8.21) \quad \int_0^1 \xi^\alpha |I'(\xi)| d\xi = \int_{I(1)}^\infty (U'(c))^\alpha dc \leq \text{const.} \int_{I(1)}^\infty U''(c) dc < \infty$$

where we have used (7.3). Because of (8.17), finiteness of the integral in (8.18) is equivalent to finiteness of

$$\begin{aligned} \int_0^1 [\xi^{\alpha+1} I''(\xi) + (\alpha+1) \xi^\alpha I'(\xi)] d\xi &= \int_{I(1)}^\infty \frac{d}{dc} \left[\frac{(U'(c))^{\alpha+1}}{U''(c)} \right] dc \\ &= \frac{1}{U''(c)} - \lim_{c \rightarrow \infty} \frac{(U'(c))^{\alpha+1}}{U''(c)}, \end{aligned}$$

and again we use (7.3). From the monotonicity of U' , we may bound the integral in (8.19) by

$$\begin{aligned} \int_0^1 \xi^{\alpha-1+\lambda_+} |J_+(\xi)| d\xi &= \int_{I(1)}^\infty (U'(c))^{\alpha-1+\lambda_+} |U''(c)| \int_0^c \frac{d\eta}{(U'(\eta))^{\lambda_+}} dc \\ &\leq - \int_{I(1)}^\infty c (U'(c))^{\alpha-1} U''(c) dc \\ &= -\frac{c}{\alpha} (U'(c))^\alpha \Big|_{c=I(1)}^{c=\infty} + \frac{1}{\alpha} \int_{I(1)}^\infty (U'(c))^\alpha dc \\ &\leq \frac{I(1)}{\alpha} + \frac{1}{\alpha} \int_{I(1)}^\infty (U'(c))^\alpha dc; \end{aligned}$$

we conclude as in (8.21). For (8.20) we write

$$\int_0^{U(1)} \xi^{\alpha-1+\lambda_-} |J_-(\xi)| d\xi = \int_1^\infty (U'(c))^{\alpha-1+\lambda_-} |U''(c)| \int_1^c \frac{d\eta}{(U'(\eta))^{\lambda_-}} dc.$$

According to (8.8), $(U'(c))^{\alpha-1+\lambda_-} \int_1^c (d\eta/(U'(\eta))^{\lambda_-})$ is bounded for $c \geq 1$, and (8.20) follows from the integrability on $[1, \infty)$ of U'' . \square

Acknowledgments. We wish to thank the Center for Stochastic Processes, University of North Carolina in Chapel Hill and the Institute of Mathematics and its Applications, University of Minnesota for their hospitality.

Note added in proof. Lemma 4.2 notwithstanding, the function \mathcal{X}_1 defined by (4.8) might not be strictly decreasing on $(0, \infty)$, because we may have $A_\infty \triangleq \int_0^\infty \|\theta(s)\|^2 ds < \infty$, a.s. However, \mathcal{X}_1 is strictly decreasing on $(0, \bar{y}_1)$, as is G_1 defined by (4.17), where

$$\bar{y}_1 \triangleq \sup \{ y > 0 : \mathcal{X}_1(y) > 0 \}$$

is in $(0, \infty]$. Consequently, we must restrict \mathcal{X}_1 to $(0, \bar{y}_1]$ in order to obtain an “inverse” $\mathcal{Y}_1 : [0, \infty] \rightarrow_{\text{onto}} [0, \bar{y}_1]$. The same comments apply to \mathcal{X}_2 defined by (5.6), to its “inverse” $\mathcal{Y}_2 : [0, \infty] \rightarrow_{\text{onto}} [0, \bar{y}_2]$, where

$$\bar{y}_2 \triangleq \sup \{ y > 0 : \mathcal{X}_2(y) > 0 \}$$

and to G_2 defined by (5.12).

The functions \mathcal{X} of (6.10) and G of (6.11) are strictly decreasing on $(0, \bar{y}_1 \vee \bar{y}_2)$. We should define $\mathcal{Y} : [0, \infty] \rightarrow_{\text{onto}} [0, \bar{y}_1 \vee \bar{y}_2]$ to be the inverse of \mathcal{X} restricted to

$[0, \bar{y}_1 \vee \bar{y}_2]$.

In § 7, the assumption should be made that there is at least one stock whose mean rate of return is different from the interest rate, i.e., $\theta \triangleq \sigma^{-1}(b - r\mathbf{1})$ is nonzero. Under this assumption Lemma 4.2 and its proof are correct, as are the conclusions drawn from it in §§ 5–7.

REFERENCES

- A. BENSOUSSAN (1984), *On the theory of option pricing*, Acta Applicandae Mathematicae 2, pp. 139–158.
- K. L. CHUNG (1982), *Lectures from Markov Processes to Brownian Motion*, Springer-Verlag, Berlin.
- A. FRIEDMAN (1975), *Stochastic Differential Equations and Applications*, Volume I, Academic Press, New York.
- I. V. GIRSANOV (1960), *On transforming a certain class of stochastic processes by absolutely continuous substitution of measures*, Theory Probab. Appl., 5, pp. 285–301.
- J. M. HARRISON AND D. M. KREPS (1979), *Martingales and arbitrage in multiperiod security markets*, J. Econom. Theory, 20, pp. 381–408.
- J. M. HARRISON AND S. R. PLISKA (1981), *Martingales and stochastic integrals in the theory of continuous trading*, Stochastic Process. Appl., 11, pp. 215–260.
- (1983), *A stochastic calculus model of continuous trading: complete markets*, Stochastic Process. Appl., 15, pp. 313–316.
- I. KARATZAS, J. P. LEHOCZKY, S. P. SETHI AND S. E. SHREVE (1986), *Explicit solution of a general consumption/investment problem*, Math. Oper. Res., 11, pp. 261–294.
- I. KARATZAS AND S. E. SHREVE (1987), *Brownian Motion and Stochastic Calculus*, Springer-Verlag, to appear.
- J. P. LEHOCZKY, S. P. SETHI AND S. E. SHREVE (1983), *Optimal consumption and investment policies allowing consumption constraints and bankruptcy*, Math. Oper. Res., 8, pp. 613–636.
- H. P. MCKEAN, JR. (1969), *Stochastic Integrals*, Academic Press, New York.
- R. C. MERTON (1969), *Lifetime portfolio selection under uncertainty: the continuous-time case*, Rev. Econom. Statist., 51, pp. 247–257.
- (1971), *Optimum consumption and portfolio rules in a continuous-time model*, J. Econom. Theory, 3, pp. 373–413. Erratum: ibid. 6 (1973), pp. 213–214.
- S. R. PLISKA (1986), *A stochastic calculus model of continuous trading: optimal portfolio*, Math. Oper. Res., 11, pp. 371–382.

OPTIMAL STABILIZATION OF FAMILIES OF LINEAR STOCHASTIC DIFFERENTIAL EQUATIONS WITH JUMP COEFFICIENTS AND MULTIPLICATIVE NOISE*

WILLIAM E. HOPKINS, JR.†

Abstract. Stabilization of families of finite-dimensional linear systems by constant linear feedback is studied by introducing a sequence of finite horizon parameter optimization problems. If the set of stabilizing constant linear feedback controllers is not empty and a regularity condition holds, a subsequence of solutions to the optimization problems converges to a constant stabilizing controller.

Key words. robust stabilization, stochastic stabilization, stochastic control, parameter optimization

AMS(MOS) subject classifications. 49B60, 93E15, 93E20

1. Introduction. The design of controllers for deterministic linear systems with parameter uncertainties has been studied using guaranteed cost control [4], zero sensitivity [17], geometric [20] and minimax [9], [16] methods, bounding methods for Lyapunov equations [14], [19], nonlinear controllers [11], and dynamic compensators [5], [6]. Quadratic mean stabilizability for families of stochastic systems with multiplicative noise has also been studied [7], [20], [21]. For families of deterministic linear systems, a parameter optimization problem was formulated in [8] that has a solution if and only if there exists a stabilizing state feedback controller, which may be computed as a solution of the first order necessary conditions. The contributions of this paper are: (i) the generalization of that result to output feedback control of systems with jump Markov coefficients and additive and multiplicative white Gaussian noise, and (ii) the definition of a sequence of finite time horizon optimization problems whose solutions converge to solutions of the infinite horizon problem whenever the latter exists and a regularity condition holds. If the set of stabilizing controllers is not empty, the latter result permits the computation of stabilizing controllers without any prior knowledge of that set.

To illustrate these results, consider the problem of stabilizing the family of linear systems $(d/dt)\xi = A_i\xi + Bu$, where i runs over a finite index set, by finding a constant control $u = -K\xi$ that minimizes the functional $\sum_i \int_0^T (\xi^T \xi + u^T u) dt$. If the initial condition is a zero-mean Gaussian random variable with unit covariance, the expected value of this functional is

$$(1) \quad \begin{aligned} & \sum_i \text{trace}(P_i(T)) \quad \text{if } T < \infty, \\ & \sum_i \text{trace}(Y_i) \quad \text{if } T = \infty, \end{aligned}$$

subject to the constraint that $P_i(t)$ be the solution of the matrix differential equation

$$\frac{d}{dt} P_i = (A_i - BK)^T P_i + P_i (A_i - BK) + I + K^T K, \quad P_i(0) = 0$$

* Received by the editors February 18, 1986; accepted for publication (in revised form) January 27, 1987. This work was supported by the National Science Foundation under grant ECS-8351621.

† Department of Electrical Engineering, Princeton University, Princeton, New Jersey 08544.

and Y_i be the solution of the corresponding Lyapunov equation. This problem may be solved by introducing multiplier matrices $W_i(t)$, Z_i , adjoining the constraint to the cost functional with the inner product

$$\sum_i \int_0^T \text{trace} (P_i(t) W_i(t)) dt \quad \text{if } T < \infty,$$

$$\sum_i \text{trace} (Y_i Z_i) \quad \text{if } T = \infty,$$

and computing Fréchet derivatives. The resulting necessary condition for a minimum is that K be a solution of the nonlinear equation

$$(2) \quad K = \begin{cases} \left(\sum_i \int_0^T B^T P_i(t) W_i(t) dt \right) \left(\sum_i \int_0^T W_i(t) dt \right)^{-1} & \text{if } T < \infty, \\ \left(\sum_i B^T Y_i Z_i \right) \left(\sum_i Z_i \right)^{-1} & \text{if } T = \infty \end{cases}$$

where the multiplier matrices are defined by

$$-\frac{d}{dt} W_i = (A_i - BK) W_i + W_i (A_i - BK)^T, \quad W_i(T) = I,$$

$$(A_i - BK) Z_i + Z_i (A_i - BK)^T + I = 0.$$

For this example, the results of this paper are as follows:

- (i) For each $T < \infty$, (1) is minimized by a solution K_T^* of (2).
- (ii) If the pairs (A_i, B) can be simultaneously stabilized by constant state feedback and $T = \infty$, then (1) is minimized by a (stabilizing) solution of (2) which may be computed from K_T^* in the limit $T \rightarrow \infty$.

The proofs involve constrained optimization and rely on the divergence of (1) as the norm of K becomes large. The novel idea here is that restricting the optimization to *constant* controls for $T < \infty$ allows convergence to be proven by showing that

$$\lim_{T \rightarrow \infty} \int_0^T W_i(t) dt = Z_i, \quad \lim_{T \rightarrow \infty} \int_0^T P_i(t) W_i(t) dt = Y_i Z_i.$$

These limits may be computed by interpreting the integrals as responses of a stable linear system having the impulse response $W_i(T-t)$ and driven by convergent forcing functions.

An interesting feature of this example is that it provides an algorithm for solving matrix Riccati equations. If there is only one system and it is controllable, then (2) may be rewritten as follows:

$$(3a) \quad K = B^T \int_0^T P(t) W(t) dt \left(\int_0^T W(t) dt \right)^{-1} \quad \text{if } T < \infty,$$

$$(3b) \quad A^T Y + Y A + I - Y B B^T Y = 0 \quad \text{if } T = \infty.$$

Therefore, the unique, positive definite solution of the Riccati equation (3b) may be computed as the limit of solutions of (3a) as $T \rightarrow \infty$.

2. A parameter optimization problem. The model of parameter uncertainty used in this paper is parametric dependence on a random variable a taking values in a closed and bounded subset S of a Euclidean space according to a known probability distribution $\mu(da)$. The linear system consists of an autonomous, finite-state jump Markov process $\beta(t) \in \{1, 2, \dots, N\}$ with prior distribution $\nu \in R^N$, $\nu_i > 0$, and generator $\pi \in R^{N \times N}$ which depends continuously on the random variable, and a controlled, stochastic differential equation with additive and multiplicative white Gaussian noise whose coefficients and prior distribution depend on both the jump process and the random variable:

$$(4) \quad \begin{aligned} d\xi &= [A_\beta \xi + B_\beta u] dt + \sum_{j=1}^q D_\beta^j \xi dv_j + \delta F_\beta dv_0, \quad t > 0, \\ \xi(0) &= x \in R^n. \end{aligned}$$

Here $\delta \in \{0, 1\}$, $v_0 \in R^n$, $v_i \in R^1$, $1 \leq i \leq q$ are independent standard Wiener processes, and the coefficient matrices A_i , D_i^j , $F_i \in R^{n \times n}$, $B_i \in R^{n \times m}$ are continuous functions on S . It is assumed that $F_i F_i^T > 0$, the zero eigenvalue of π has multiplicity one, and the random variable x has a positive-definite second moment σ .

An optimization problem is defined by considering the expected value of the finite-horizon, quadratic cost functional

$$V_T(a, i, x) = \frac{1}{1 + \delta T} E \left\{ \int_0^T e^{(1-\delta)\lambda_\beta t} (\xi^T Q_\beta \xi + u^T u) dt \mid \xi(0) = x, \beta(0) = i \right\}$$

and seeking constant output feedback controls to minimize the averages

$$J_T(K) = E\{V_T(a, i, x)\} \quad \text{for } T < \infty, \quad J_\infty(K) = \lim_{T \rightarrow \infty} J_T(K).$$

Here $Q_i \in R^{n \times n}$, $Q_i = Q_i^T \geq 0$, and $\lambda_i \in R$ are continuous functions on S . The parameter δ allows the simultaneous study of both infinite horizon ($\delta = 0$) and average cost ($\delta = 1$) criteria for the limit problem.

The admissible controls are restricted to the gain schedule $u = -K_\gamma \eta$, where γ , η are noiseless observations of β , ξ . Here $\eta(t) = C\xi(t)$ where $C \in R^{r \times n}$ is of full rank, and $\gamma(t) = j$ whenever $\beta(t) \in I_j$ where $\bigcup_{j=1}^M I_j$ is a disjoint partition of $\{1, 2, \dots, N\}$. (For example, β , γ might correspond to a multi-dimensional jump process, only some of whose components are measured.)

Computation of the optimal feedback gains involves solving systems of transcendental equations. For $T < \infty$, the minimization of J_T is carried out by transforming the problem into a parameter optimization problem with terminal cost. For notational convenience, denote $K = (K_1, \dots, K_M)$, $U = \prod_{i=1}^M R^{m \times r}$, define an indicator function by $\theta_i = j$ whenever $i \in I_j$, and introduce $G_i = A_i - B_i K_{\theta_i} C$. For $P \in R^{n \times n}$, define

$$(5) \quad L_i(P) = G_i^T P + P G_i + \sum_{j=1}^q (D_i^j)^T P D_i^j, \quad L_i^*(P) = P G_i^T + G_i P + \sum_{j=1}^q D_i^j P (D_i^j)^T.$$

PROPOSITION 1.

$$J_T(K) = \frac{1}{1 + \delta T} \sum_{i=1}^N \int_S \nu_i (\text{trace}(P_i(T, T)\sigma) + p_i(T, T)) \mu(da)$$

where $P_i(t, T)$, $p_i(t, T)$, $1 \leq i \leq N$, are defined for $K \in U$ by

$$\begin{aligned} \frac{d}{dt} P_i &= L_i(P_i) + \sum_{j=1}^N \pi_{ij} P_j + (Q_i + C^T K_{\theta_i}^T K_{\theta_i} C) e^{(1-\delta)\lambda_i(T-t)}, \quad 0 < t \leq T, \\ P_i(0, T) &= 0, \\ \frac{d}{dt} p_i &= \sum_{j=1}^N \pi_{ij} p_j + \delta \operatorname{trace}(F_i^T P_i F_i), \quad 0 < t \leq T, \\ p_i(0, T) &= 0. \end{aligned} \quad (6)$$

To state the first order necessary conditions for a minimum, define adjoint variables $W_i(t, T)$, $w_i(t, T)$, $1 \leq i \leq N$, by

$$\begin{aligned} -\frac{d}{dt} W_i &= L_i^*(W_i) + \sum_{k=1}^N \pi_{ki} W_k + \delta w_i F_i F_i^T, \quad 0 \leq t < T, \\ W_i(T, T) &= \nu_i \sigma, \\ -\frac{d}{dt} w_i &= \sum_{k=1}^N \pi_{ki} w_k, \quad 0 \leq t < T, \\ w_i(T, T) &= \nu_i \end{aligned} \quad (7)$$

and set $f^T = (f_1^T, \dots, f_M^T)$ where

$$\begin{aligned} f_i^T(K) &= \left(\sum_{i \in I_i} \int_0^T \int_S B_i^T P_i W_i C^T \mu(da) dt \right) \\ &\cdot \left(\sum_{i \in I_i} \int_0^T \int_S e^{(1-\delta)\lambda_i(T-t)} C W_i C^T \mu(da) dt \right)^{-1}. \end{aligned} \quad (8)$$

THEOREM 1. *The equation $K = f^T(K)$ has a solution K_T^* that minimizes J_T over U .*

The limit problem of minimizing J_∞ requires further assumptions and a more complicated notation. Define $\Gamma_1 \in R^{n^2 N \times n^2 N}$ by

$$\Gamma_1 = \pi \otimes I_{n^2} + \operatorname{diag} \left[I_n \otimes (G_i)^T + (G_i)^T \otimes I_n + \sum_{j=1}^q (D_i^j)^T \otimes (D_i^j)^T \right]$$

and define the open sets

$$U_0 = \bigcap_{a \in S} U_0(a), \quad U_0(a) = \{K \in U: \max \{\operatorname{real}(\sigma(\Gamma_1))\} + (1-\delta) \max_i \{\lambda_i\} < 0\}.$$

PROPOSITION 2. *If $K \in U_0$ then*

$$J_\infty(K) = (1-\delta) \sum_{i,j=1}^N \int_S \nu_i \operatorname{trace}(Y_{ij} \sigma) \mu(da) + \delta N \sum_{i,j=1}^N \int_S \alpha_i \operatorname{trace}(F_i^T Y_{ij} F_i) \mu(da)$$

where α is the normalized left zero eigenvector of π and Y_{ij} , $1 \leq i, j \leq N$, are defined for $K \in U_0$ by

$$L_i(Y_{ij}) + \sum_{k=1}^N \pi_{ik} Y_{kj} + (1-\delta)\lambda_j Y_{ij} + \chi_{\{i=j\}}(Q_i + C^T K_{\theta_i}^T K_{\theta_i} C) = 0.^1 \quad (9)$$

(As will be seen in Lemma 1, § 4, $U_0(a)$ is a set of feedback gains for which (9) uniquely defines Y_{ij} as a positive-semidefinite, symmetric matrix.)

¹ $\chi_{\{i=j\}} = \begin{cases} 0 & \text{if } i \neq j, \\ 1 & \text{if } i = j. \end{cases}$

To state the first order necessary conditions for a minimum, define adjoint matrices Z_{ij} , $1 \leq i, j \leq N$, for the limit problem by

$$(10) \quad L_i^*(Z_{ij}) + \sum_{k=1}^N \pi_{ki} Z_{kj} + (1-\delta) \lambda_j Z_{ij} + (1-\delta) \nu_i \sigma + \delta N \alpha_i F_i F_i^T = 0$$

and for $K \in U_0$ set $f^\infty = (f_1^\infty, \dots, f_M^\infty)$ where

$$(11) \quad f_i^\infty(K) = \left(\sum_{i \in I_i} \int_S \sum_{j=1}^N B_i^T Y_{ij} Z_{ij} C^T \mu(da) \right) \left(\sum_{i \in I_i} \int_S C Z_{ii} C^T \mu(da) \right)^{-1}.$$

Finally, the following regularity hypothesis implies J_∞ is differentiable at its minimum: There exists a symmetric matrix $\bar{Q} \in R^{n \times n}$ satisfying $0 \leq \bar{Q} \leq Q_i$ such that $X_{ij} \in R^{n \times n}$, $1 \leq i, j \leq N$, defined for $a \in S$, $K \in U_0$ by

$$(12) \quad L_i(X_{ij}) + \sum_{k=1}^N \pi_{ik} X_{kj} + (1-\delta) \lambda_j X_{ij} + \chi_{\{i=j\}} \bar{Q} = 0$$

satisfies

$$(13) \quad \lim_{K \rightarrow K_0} \sum_{i,j=1}^N \int_S \text{trace}(X_{ij}) \mu(da) = +\infty \quad \text{for all } K_0 \in \partial U_0.$$

THEOREM 2. *If U_0 is not empty and (13) holds then the equation $K = f^\infty(K)$ has a solution K_∞^* that minimizes J_∞ over U_0 . Conversely, if $K = f^\infty(K)$ has a solution then U_0 is not empty.*

THEOREM 3. *Suppose U_0 is not empty and (13) holds. Let T_i be an unbounded sequence of positive numbers, and let $K_{T_i}^*$ be a solution of $K = f^{T_i}(K)$ that minimizes J_{T_i} . Then $K_{T_i}^*$ has a subsequence which converges to a solution of $K = f^\infty(K)$ that minimizes J_∞ over U_0 .*

The proofs of Theorems 1–3 will be given in § 4. Using integral representations of $P_i(T, T)$, $p_i(T, T)$, and Y_{ij} , the optimization problems will be recast as ordinary minimization problems for scalar functions of several variables. The equations $K = f^T(K)$, $K = f^\infty(K)$, correspond to the requirement that the first derivatives of J_T , J_∞ be equal to zero. The convergence properties will follow from limits to be proven in Proposition 5. (An equivalent derivation of equations (7), (8) (respectively (10), (11)) consists of using $\sum_i \int_0^T \int_S \text{trace}(M_i N_i) \mu(da) dt$ (respectively, $\sum_i \int_S \text{trace}(M_i N_i) \mu(da)$) to adjoin equations (6) (respectively, (9)) to the cost J_T (respectively, J_∞) and computing Fréchet derivatives.)

3. Discussion of results. Theorem 2 is important because of the following connection between stability and the set U_0 . (The proof is given in Appendix A.)

PROPOSITION 3. *Assume $K \in U_0$. If $\delta = 0$ and $\max \{\lambda_i\} \geq 0$ then ξ is stable in quadratic mean. If $\delta = 1$ then ξ has an invariant probability distribution.*

Calling the system (4) stable if either of the conclusions of the proposition holds, we find that Theorems 2 and 3 characterize stabilizing controls when $(1-\delta) \max \{\lambda_i\} \geq 0$. If the model is deterministic apart from the initial data and the parameter a , then quadratic mean stability is the same as Lyapunov stability for every $a \in S$. Therefore, the theorems apply to problems of simultaneous stabilization of families of linear, deterministic systems. If the model includes Gaussian noise but no jump process, the theorems apply to problems of quadratic mean stabilization in the case of multiplicative noise, and to weak stochastic stabilization in the case of additive noise.

Theorem 2 may also be interpreted as a generalization of the result that stabilizability and detectability imply the existence of a unique, positive semi-definite solution of the matrix, algebraic Riccati equation [24]: The equation $K = f^\infty(K)$ generalizes the

Riccati equation, the hypothesis $U_0 \neq \emptyset$ plays the role of stabilizability, and the regularity hypothesis plays the role of detectability. Subject to the regularity hypothesis, f^∞ has a fixed point if and only if U_0 is not empty. There is no claim of uniqueness. The novelty of Theorem 1 is the restriction of admissible controls to *constant* rather than time-varying feedback gains. Although J_T can be redefined for, and minimized with respect to, time-varying controls, study of the passage to the limit appears to be technically intractable unless S is a singleton set. In that case, results are known for some nonlinear systems [18]. Theorem 3 is important for computing stabilizing controllers. Since gradient methods for computing fixed points of f^∞ require an initial guess $K_0 \in U_0$, they are not practical if the set U_0 is unknown. Theorem 3 ensures that at least one sequence of fixed points of f^T converges to U_0 as $T \rightarrow \infty$, and since $\lim_{T \downarrow 0} K_T^* = 0$, that sequence may be computed without any "initial guess." However, rates of convergence are not known. Numerical algorithms could be based on solving $K = f^T(K)$ for $T \rightarrow \infty$, or Theorem 3 could be used to compute starting points for gradient search algorithms.

When S is a singleton set, some special cases of Theorem 2 are known. The case when (4) is deterministic and $\delta = 0$ recovers the output feedback regulator problem studied by Levine and Athans [12]. State feedback regulation ($\delta = 0$) of a linear differential equation modulated by a jump process, and some state feedback, long run average cost problems ($\delta = 1$) with no jump process were studied by Wonham [23], [22]. Recently, Mariton and Bertrand have studied the former problem when the control is independent of the jump process [13], and Hyland and Bernstein [10] have studied extensions of the latter problem with dynamic output feedback control. In [2], the state feedback problem with no jump process and only multiplicative noise was studied as an approximation for a control problem with wide bandwidth disturbances.

Whereas most previous work on stabilization of uncertain systems has focused on developing verifiable sufficient conditions, Theorem 2 gives an analytical characterization as solvability of a set of transcendental equations. The difficult issue of verifiable criteria is hidden in the regularity hypothesis and the assumption that U_0 is nonempty. Many sufficient conditions for U_0 to be nonempty are given in the references cited in the Introduction. The regularity hypothesis (13) is the weaker assumption that the solution of (12) has singularities which are not integrable.

PROPOSITION 4. *Suppose $Q_i > 0$ and either $N = 1$ or $\pi_{ij} > 0$, $i \neq j$. The regularity hypothesis (13) holds if either (i) S is a finite set and μ is positive on S , or (ii) S is a finite interval, $\mu(da) \geq \phi(a) da$ where $\phi > 0$, and L_i, λ_i, π are polynomial functions of a .*

The proof of Proposition 4 is given in Appendix A. It is not known if (13) holds for polynomial parametrizations when S is a hypercube. Also, the regularity hypothesis may fail when some discount rates are negative and the jump process is not irreducible.

Example 1. For the scalar problem $(d/dt)\xi = A_\beta \xi + u$, $u = -K\xi$, consider choosing K independent of the jump process β to minimize $E \int_0^\infty (\xi^2 + u^2) dt$. (Here $\delta = 0$ and $\theta_i = 1$.) Then $U_0 = (K_0, \infty)$ where K_0 is the maximum real part of any eigenvalue of $\frac{1}{2}\pi + \text{diag}[A_i]$, and the regularity hypothesis holds if

$$(14) \quad \lim_{K \downarrow K_0} \int_S \text{trace} \left(KI_N - \frac{1}{2} \pi - \text{diag}[A_i] \right)^{-1} \mu(da) = +\infty.$$

The equation $K = f^\infty(K)$ may be rewritten as follows:

$$(15) \quad \frac{2K}{1 + K^2} = \frac{\int_S \nu^T(KI_N - \frac{1}{2}\pi - \text{diag}(A_i))^{-2} J\mu(da)}{\int_S \nu^T(KI_N - \frac{1}{2}\pi - \text{diag}(A_i))^{-1} J\mu(da)}$$

where $J = (1, \dots, 1)^T$. By calculus, (14) implies (15) has a solution in U_0 . Depending on the data, (15) may also have solutions in $(-\infty, K_0)$. (This example was studied in [8] in the case $N = 1$.)

Example 2. Suppressing subscripts and superscripts let $d\xi = (A\xi + Bu) dt + D\xi dv$, where $u = -KC\xi$,

$$A = \begin{pmatrix} 0 & 1 \\ 0 & -1 \end{pmatrix}, \quad B = \begin{pmatrix} 0 \\ 1 \end{pmatrix}, \quad D = \sqrt{a} \begin{pmatrix} 2 \\ -1 \end{pmatrix} (\varepsilon 2).$$

For the case $C = I_2$ it was shown in [20] that (i) for any $\varepsilon \geq 0$, $a_0 > 0$, there exists a feedback K which stabilizes the system for all $a \in [0, a_0)$, and (ii) if $\varepsilon = 1$, then one may take $a_0 = \infty$. For the case $C = (1, 0)$, let $\varepsilon \in R$, $S = [0, a_2]$, $a_2 > 0$, $\mu(da) = da/a_2$, $\delta = \lambda = 0$, $Q = \bar{Q} = \sigma = I_2$. Then $\det(\Gamma_1) = b_2 K^2 + b_1 K + b_0$, where $b_2 = -32a$, $b_1 = 4 - 8a(\varepsilon^2 - 4\varepsilon + 1)$, $b_0 = -2a\varepsilon^2$. Let $\bar{a}(1) = +\infty$ and for $\varepsilon \neq 1$ let $\bar{a}(\varepsilon)$ denote the unique positive solution of $(b_1)^2 - 4b_2b_0 = 0$. Applying the Routh-Hurwitz stability criterion to the characteristic polynomial of Γ_1 , it may be shown that U_0 is nonempty if and only if $a_2 < \bar{a}(\varepsilon)$. If $a_2 < \bar{a}(\varepsilon)$ then $U_0 = (K_-, K_+) \subseteq (0, \infty)$ where K_{\pm} are the ordered roots of $\det(\Gamma_1)|_{a=a_2} = 0$, $U_0 \rightarrow (0, \infty)$ as $a_2 \downarrow 0$, and $U_0 \rightarrow \emptyset$ as $a_2 \uparrow \bar{a}(\varepsilon)$. Equation (11) has the form

$$K = \int_0^{a_2} \frac{\sum_{i=0}^4 c_i K^i}{(b_2 K^2 + b_1 K + b_0)^2} da \bigg/ \int_0^{a_2} \frac{d_1 K + d_0}{b_2 K^2 + b_1 K + b_0} da$$

where c_i, d_i are quadratic polynomials in a for every ε . By Proposition 4 the regularity hypothesis holds and by Theorem 2 equation (11) has a stabilizing solution $K_{\infty}^* \in U_0$ if $a_2 < \bar{a}(\varepsilon)$.

4. Proofs. The compactness of S and the continuity of $P_i, W_i, Y_{ij}, Z_{ij}, X_{ij}$ with respect to a justify the frequent use of Fubini's theorem and the dominated convergence theorem. Several of the proofs use the fact that $\int_0^{\infty} e^{At} C e^{Bt} dt$ is the unique solution of $AX + XB + C = 0$ if A, B are square matrices and the sum of the maximum real parts of their eigenvalues is negative. Also, since the zero eigenvalue of π has multiplicity one, there exist $\alpha \in R^N$, $\alpha_i > 0$, $\Lambda(t) \in R^{N \times N}$, continuous on S , such that $e^{\pi t} = J\alpha^T + \Lambda$, where $J = (1, \dots, 1)^T$ and $\|\Lambda\| \leq \kappa e^{-t/\kappa}$, $\kappa > 0$. The proofs of Propositions 2 and 5 use the fact that if A is a stable matrix and $g(t)$ is continuously differentiable, then by a version of l'Hospital's rule [15, p. 375]

$$\lim_{T \rightarrow \infty} \frac{1}{1 + \delta T} \int_0^T e^{A(T-t)} g(t) dt = -A^{-1} \lim_{t \rightarrow \infty} \left((1 - \delta)g(t) + \delta \frac{d}{dt} g(t) \right)$$

whenever the limit on the right-hand side exists.

LEMMA 1. (i) If $K \in U$ then $P_i(t, T)$ are symmetric, positive-semidefinite matrices and $W_i(t, T)$ are symmetric positive-definite matrices. (ii) If $K \in U_0$, then Y_{ij}, X_{ij} are symmetric, positive-semidefinite matrices uniquely defined by (9), (12) respectively, and Z_{ij} are symmetric, positive-definite matrices uniquely defined by (10).

Proof of part (ii). (The proof of part (i) is similar.) Define $\bar{X}, \bar{Y}, \bar{Z} \in R^{n^2 N \times N}$ by $\bar{X} = (\text{vec}(X_{ij}))_{i,j=1}^N$, $\bar{Y} = (\text{vec}(Y_{ij}))_{i,j=1}^N$, $\bar{Z} = (\text{vec}(Z_{ij}))_{i,j=1}^N$, and define $\Gamma_2 \in R^{N \times N}$, $\Gamma_4 \in R^{n^2 N \times N}$, $\Gamma_5, \Gamma_6 \in R^{n^2 N}$ for $a \in S$, $K \in U_0$, by²

$$\begin{aligned} \Gamma_2 &= \text{diag}[\lambda_i], & \Gamma_4 &= \text{diag}[\text{vec}(Q_i + C^T K_{\theta_i}^T K_{\theta_i} C)], \\ \Gamma_5 &= (\alpha_i \text{vec}(F_i F_i^T))_{i=1}^N, & \Gamma_6 &= (\nu_i \text{vec}(\sigma))_{i=1}^N. \end{aligned}$$

² For any matrix E , $\text{vec}(E)$ denotes the vector of stacked columns of E [3].

By properties of Kronecker products [3], (9), (10), (12) are equivalent to

$$(9') \quad \Gamma_1 \bar{Y} + (1 - \delta) \bar{Y} \Gamma_2 + \Gamma_4 = 0,$$

$$(10') \quad \Gamma_1^T \bar{Z} + (1 - \delta) \bar{Z} \Gamma_2 + \delta N \Gamma_5 J^T + (1 - \delta) \Gamma_6 J^T = 0,$$

$$(12') \quad \Gamma_1 \bar{X} + (1 - \delta) \bar{X} \Gamma_2 + \text{diag} [\text{vec} (\bar{Q})] = 0.$$

Therefore, the existence of unique, symmetric solutions follows from the definition of U_0 and the symmetry of (5).

To prove $Z_{ij} > 0$, define $E_{ij} = (1 - \delta) \nu_i \sigma + \delta N \alpha_i F_i F_i^T > 0$,

$$\bar{\pi} = \pi^T - \text{diag} \left[\sum_{j=i}^N \pi_{ji} \right], \quad \bar{\lambda}_{ij} = \lambda_j + \sum_{j=i}^N \pi_{ji},$$

let $\bar{\beta}$ be a Markov jump process on $\{1, 2, \dots, N\}$ with generator $\bar{\pi}$, and let ξ^j be the solution starting at x of

$$d\xi^j = \left(G_{\bar{\beta}} + \frac{1}{2} (1 - \delta) \bar{\lambda}_{\bar{\beta}} I_n \right)^T \xi^j dt + \sum_{k=1}^q (D_{\bar{\beta}}^k)^T \xi^j dv_k.$$

Then $Z_{ij} > 0$ since for fixed j the solution of (10) has the representation

$$E \left\{ \int_0^\infty (\xi^j)^T E_{\bar{\beta}j} \xi^j dt \mid \xi^j(0) = x, \bar{\beta}(0) = i \right\} = x^T Z_{ij} x.$$

The semi-definiteness of Y_{ij} , X_{ij} is proved similarly.

Proof of Proposition 1. V_T is a quadratic function of x and $V_T = v(0, x)/(1 + \delta T)$, where v satisfies the backward system of linear, partial differential equations ($x \in R^n$)

$$\frac{\partial}{\partial t} v(t, x) + Lv(t, x) + \text{diag} [x^T (Q_i + C_i^T K_{\theta_i}^T K_{\theta_i} C_i) x] e^{(1-\delta)\lambda_i t} = 0, \quad 0 \leq t < T,$$

$$v(T, x) = 0.$$

Here L is the generator of the joint process β, ξ :

$$L = \pi + \text{diag} \left[x^T G_i^T \frac{\partial}{\partial x} + \frac{1}{2} \text{trace} \left(\left(\delta F_i F_i^T + \sum_{j=1}^q D_i^j x x^T (D_i^j)^T \right) \frac{\partial^2}{\partial x^2} \right) \right].$$

Therefore, by a standard calculation $v_i(0, x) = x^T P_i(T, T)x + p_i(T, T)$ and the result follows.

Proof of Theorem 1. Since $P_i \geq 0$, Proposition 1 and a comparison theorem imply there exists a constant κ , independent of T , such that

$$(16) \quad J_T(K) \geq \frac{\kappa}{1 + \delta T} \sum_{i=1}^N \int_S \left(\text{trace} (\hat{P}_i(T)) + \delta \int_0^T \text{trace} (\hat{P}_i(t)) \right) \mu(da)$$

where $\hat{P}_i(t)$ are defined by

$$\frac{d}{dt} \hat{P}_i = \left(G_i + \frac{1}{2} \pi_{ii} \right)^T \hat{P}_i + \hat{P}_i \left(G_i + \frac{1}{2} \pi_{ii} \right) + C^T K_{\theta_i}^T K_{\theta_i} C e^{(1-\delta)\lambda_i(T-t)}, \quad 0 < t \leq T,$$

$$\hat{P}_i(0) = 0.$$

Therefore, $\lim_{\|K\| \rightarrow \infty} J_T = \infty$, implying J_T is minimized by some $K_T^* \in U$. The fact that K_T^* is a fixed point of f^T follows from Proposition B1, Appendix B, by a long calculation using Kronecker products to express (6) and (7) in vector form. The details are omitted.

Proof of Proposition 2. Define $P_{ij}(t, T)$, $p_{ij}(t, T)$, by

$$(17) \quad \begin{aligned} \frac{d}{dt} P_{ij} &= L_i(P_{ij}) + \sum_{k=1}^N \pi_{ik} P_{kj} + \chi_{\{i=j\}} (Q_i + C^T K_{\theta_i}^T K_{\theta_i} C) e^{(1-\delta)\lambda_i(T-t)}, \quad 0 < t \leq T, \\ P_{ij}(0, T) &= 0, \\ \frac{d}{dt} p_{ij} &= \sum_{k=1}^N \pi_{ik} p_{kj} + \delta \operatorname{trace}(F_i^T P_{ij} F_i), \\ p_{ij}(0, T) &= 0. \end{aligned}$$

Then $P_i = \sum_{j=1}^N P_{ij}$, $p_i = \sum_{j=1}^N p_{ij}$, and

$$J_T(K) = \frac{1}{1+\delta T} \sum_{i,j=1}^N \int_S \nu_i(\operatorname{trace}(P_{ij}(T, T)\sigma) + p_{ij}(T, T)) \mu(da).$$

The formula then holds by the following calculations. Let $\bar{P}(t, T) = (\operatorname{vec}(P_{ij}))_{i,j=1}^N$, $\bar{p}(t, T) = (p_{ij})_{i,j=1}^N$, and define $\Gamma_3, \bar{\Gamma}_3(t) \in \mathbb{R}^{N \times N}$ by

$$(\Gamma_3)_{ij} = \operatorname{trace}(F_i^T Y_{ij} F_i), \quad (\bar{\Gamma}_3)_{ij} = \operatorname{trace}(F_i P_{ij}(t, T) F_i).$$

By properties of Kronecker products [3], (17) is equivalent to

$$(18) \quad \begin{aligned} \frac{d}{dt} \bar{P} &= \Gamma_1 \bar{P} + \Gamma_4 e^{(1-\delta)\Gamma_2(T-t)}, \quad 0 < t \leq T, \\ \bar{P}(0, T) &= 0. \end{aligned}$$

Therefore,

$$\bar{P}(t, T) = \bar{P}(t, t) e^{(1-\delta)\Gamma_2(T-t)}, \quad \bar{P}(T, T) = \int_0^T e^{\Gamma_1 t} \Gamma_4 e^{(1-\delta)\Gamma_2 t} dt,$$

$$\bar{p}(T, T) = \delta \int_0^T e^{\pi(T-t)} \bar{\Gamma}_3(t) dt$$

and $\bar{p}(t, t)$ satisfies $(d/dt)\bar{p} = \pi\bar{p} + (1-\delta)\bar{p}\Gamma_2 + \delta\bar{\Gamma}_3$. If $K \in U_0$, then

$$\lim_{T \rightarrow \infty} \frac{1}{1+\delta T} \bar{P}(T, T) = (1-\delta) \int_0^\infty e^{\Gamma_1 t} \Gamma_4 e^{(1-\delta)\Gamma_2 t} dt = (1-\delta) \bar{Y}.$$

If $\delta = 0$, then $\bar{p}(T, T) = 0$ and if $\delta = 1$, then

$$\lim_{T \rightarrow \infty} \frac{1}{T} \bar{p}(T, T) = \lim_{T \rightarrow \infty} \frac{1}{T} \int_0^T (J\alpha^T + \Lambda(T-t)) \bar{\Gamma}_3(t) dt = J\alpha^T \Gamma_3.$$

Proof of Theorem 2. $\bar{G}_i = G_i + \frac{1}{2}(\pi_{ii} + (1-\delta)\lambda_i)$ is stable if $K \in U_0$. Therefore, by Lemma 1, Proposition 2, and a comparison theorem,

$$J_\infty(K) \geq \kappa_1 \sum_{i=1}^N \int_S \operatorname{trace}(\hat{X}_i) \mu(da), \quad \kappa_1 > 0$$

where \hat{X}_i are the unique solutions of $\bar{G}_i^T \hat{X}_i + \hat{X}_i \bar{G}_i + C^T K_{\theta_i}^T K_{\theta_i} C = 0$. Therefore, $\lim_{\|K\| \rightarrow \infty} J_\infty = \infty$, implying J_∞ is minimized by some $K_\infty^* \in \text{closure}(U_0)$. Also,

$$J_\infty(K) \geq \kappa_2 \sum_{i,j=1}^N \int_S \operatorname{trace}(X_{ij}) \mu(da), \quad \kappa_2 > 0.$$

Therefore, the regularity hypothesis (13) implies $K_\infty^* \notin \partial U_0$. The fact that K_∞^* is a fixed point of f^∞ follows from Proposition B1, Appendix B, by a long calculation using Kronecker products to express (9), (10) in vector form. The details are omitted.

PROPOSITION 5. Let $E \in R^{n \times n}$ and $K_T \in U$ for $T > 0$. If $\lim_{T \rightarrow \infty} K_T = K_\infty \in U_0$, then

$$(i) \quad \lim_{T \rightarrow \infty} \frac{1}{1 + \delta T} \int_0^T e^{(1-\delta)\lambda_i(T-t)} W_i(t, T) dt|_{K=K_T} = \left(1 - \delta + \frac{\delta}{N}\right) Z_{ii}|_{K=K_\infty};$$

$$(ii) \quad \lim_{T \rightarrow \infty} \frac{1}{1 + \delta T} \int_0^T P_{ij}(t, T) E W_i(t, T) dt|_{K=K_T} = \left(1 - \delta + \frac{\delta}{N}\right) Y_{ij} E Z_{ij}|_{K=K_\infty}.$$

Proof. By the continuity of W_i , Z_{ij} , P_{ij} , Y_{ij} with respect to t , a , K , there is no loss of generality in assuming $K_T = K_\infty$. Define $\bar{\Gamma}_5(t) \in R^{n^2 N}$ by $\bar{\Gamma}_5 = ((e^{\pi^T t} \nu)_i \text{vec}(F_i F_i^T))_{i=1}^N$ and note $\lim_{t \rightarrow \infty} \bar{\Gamma}_5(t) = \bar{\Gamma}_5$. Define $\bar{W}_i = W_i$ and let $\bar{W}(t, T) = (\text{vec}(W_{ij}))_{i,j=1}^N$. Then (7) is equivalent to

$$-\frac{d}{dt} \bar{W} = \Gamma_1^T \bar{W} + \delta \bar{\Gamma}_5(T-t) J^T, \quad 0 \leq t < T,$$

$$\bar{W}(T, T) = \Gamma_6 J^T$$

and (10) is equivalent to

$$(19) \quad \bar{Z} = \int_0^\infty e^{\Gamma_1^T t} ((1-\delta)\Gamma_6 J^T + \delta N \Gamma_5 J^T) e^{(1-\delta)\Gamma_2 t} dt.$$

Therefore

$$\frac{1}{1 + \delta T} \int_0^T \bar{W}(t, T) e^{(1-\delta)\Gamma_2(T-t)} dt = I_1 + \delta I_2,$$

$$I_1 = \frac{1}{1 + \delta T} \int_0^T e^{\Gamma_1^T t} \Gamma_6 J^T e^{(1-\delta)\Gamma_2 t} dt,$$

$$I_2 = \frac{1}{1 + T} \int_0^T \int_0^t e^{\Gamma_1^T(t-s)} \bar{\Gamma}_5(s) J^T ds dt.$$

To prove part (i) it suffices to show that $I_1 + \delta I_2$ converges to (19). If $\delta = 0$, then I_1 converges to (19), and if $\delta = 1$, then I_1 converges to zero and I_2 converges to the solution of $\Gamma_1^T z + \Gamma_5 J^T = 0$.

The proof of part (ii) depends on the equations for Y_{ij} , Z_{ij} decoupling with respect to j . Define $\bar{P}_i = \bar{P} e_i$, $\bar{W}_i = \bar{W} e_i$, $\bar{Y}_i = \bar{Y} e_i$, $\bar{Z}_i = \bar{Z} e_i$.³ Then by (18) and (19)

$$\bar{P}_i(t, T) = \bar{P}_i(t, t) e^{(1-\delta)\lambda_i(T-t)}, \quad \bar{P}_i(t, t) = \int_0^t e^{\Gamma_1^s \Gamma_4} e_i e^{(1-\delta)\lambda_i s} ds,$$

$$\bar{Y}_i = \int_0^\infty e^{\Gamma_1^s \Gamma_4} e_i e^{(1-\delta)\lambda_i s} ds,$$

$$\bar{W}_i(t, T) = e^{\Gamma_1^T(T-t)} \Gamma_6 + \delta \int_t^T e^{\Gamma_1^T(s-t)} \bar{\Gamma}_5(T-s) ds,$$

$$\bar{Z}_i = \int_0^\infty e^{\Gamma_1^T t} ((1-\delta)\Gamma_6 + \delta N \Gamma_5) e^{(1-\delta)\lambda_i t} dt.$$

³ e_i denotes the vector consisting of zeros except for a one in the i th position.

By a change of coordinates and order of integration,

$$\begin{aligned} \frac{1}{1+\delta T} \int_0^T \bar{W}_i(t, T) \bar{P}_i^T(t, T) dt &= I_3 + \delta I_4, \\ I_3 &= \frac{1}{1+\delta T} \int_0^T e^{\Gamma_1^T(T-t)} \Gamma_6 e^{(1-\delta)\lambda_i(T-t)} \bar{P}_i^T(t, t) dt, \\ I_4 &= \frac{1}{1+T} \int_0^T e^{\Gamma_1^T(T-t)} \left(\int_0^t \bar{\Gamma}_5(t-s) \bar{P}_i^T(s, s) ds \right) dt. \end{aligned}$$

To prove part (ii) it suffices to show that $I_3 + \delta I_4$ converges to $\bar{Z}_i \bar{Y}_i^T$. If $\delta = 0$, then I_3 converges to the solution of $\Gamma_1^T x + \lambda_i x + \Gamma_6 \bar{Y}_i^T = 0$. If $\delta = 1$, then I_3 converges to zero and, by repeated application of l'Hospital's rule, I_4 converges to the solution of $\Gamma_1^T x + \Gamma_5 \bar{Y}_i^T = 0$.

LEMMA 2. Let $K_T \in U$ for $T > 0$. If U_0 is not empty, the regularity hypothesis (13) holds, and $\lim_{T \rightarrow \infty} K_T = K_\infty \notin U_0$, then

$$\lim_{T \rightarrow \infty} \sum_{i,j=1}^N \int_S \text{trace}(P_{ij}(T, T)) \mu(da)|_{K=K_T} = +\infty.$$

Proof. Let $U_0^1 \subseteq U_0$, $U_0^2 \subseteq U - U_0$ be bounded. Since S is compact and Γ_1 is a continuous function of a, K , by (18) there exist positive constants κ, τ such that for any $T \geq \tau$, $K^1 \in U_0^1$, $K_2 \in U_0^2$,

$$\text{trace}(P_{ij}(T, T))|_{K=K^1} \leq \kappa \text{trace}(P_{ij}(T, T))|_{K=K^2}.$$

The result follows by (13) since $\lim_{T \rightarrow \infty} \text{trace}(P_{ij}(T, T))|_{K=K^1} \geq \text{trace}(X_{ij})|_{K=K^1}$.

Proof of Theorem 3. Let $\varepsilon > 0$. By Propositions 1 and 2 and (17) there are constants κ, τ_0 such that for $T \geq \tau_0$,

$$\begin{aligned} J_\infty(K_\infty^*) + \varepsilon &\geq J_T(K_\infty^*) \geq J_T(K_T^*) \\ (20) \quad &\geq \frac{\kappa}{(1+\delta T)} \sum_{i,j=1}^N \int_S \left(\text{trace}(P_{ij}(T, T)) \right. \\ &\quad \left. + \delta \int_0^T \text{trace}(P_{ij}(t, t)) dt \right) \mu(da)|_{K=K_T^*}. \end{aligned}$$

Therefore, there exist constants ρ, τ_1 such that $K_{T_i}^* \in U_0$ and $\|K_{T_i}^*\| < \rho$ for $T_i \geq \tau_1$, since otherwise equation (16) or Lemma 2 would contradict (20). Hence (20) and Proposition 5 imply the limit of each convergent subsequence minimizes J_∞ and is a fixed point of f^∞ .

5. Conclusion. For a large class of families of linear systems with uncertainty, this paper has shown that, subject to a mild regularity hypothesis, stabilizability is equivalent to solvability of a system of transcendental equations. These are the stationary equations for a related parameter optimization problem. A procedure for their solution has been given.

Several minor extensions of Theorems 1-3 are possible. For instance, if (4) is generalized to include control dependent noise in the form

$$d\xi = [A_\beta \xi + B_\beta u] dt + \sum_{j=1}^q [D_\beta^j \xi + E_\beta^j u] dv_j + \delta F_\beta dv_0,$$

then Theorems 1-3 hold with D_i^j replaced by $D_i^j - E_i^j K_\theta C$ in (5) and the right-hand sides of (8), (11) replaced by the unique solutions X of the linear equations

$$(8') \quad \sum_{i \in I_1} \int_0^T \int_S \left(I_m e^{(1-\delta)\lambda_i(T-t)} + \sum_{k=1}^q (E_i^k)^T P_i E_i^k \right) X C W_i C^T \mu(da) dt$$

$$= \sum_{i \in I_1} \int_0^T \int_S \left(B_i^T P_i + \sum_{k=1}^q (E_i^k)^T P_i D_i^k \right) W_i C^T \mu(da) dt,$$

$$(11') \quad \sum_{i \in I_1} \int_S \left(X C Z_{ii} C^T + \sum_{j=1}^N \sum_{k=1}^q (E_i^k)^T Y_{ij} E_i^k X C Z_{ij} C^T \right) \mu(da)$$

$$= \sum_{i \in I_1} \int_S \sum_{j=1}^N \left(B_i^T Y_{ij} + \sum_{k=1}^q (E_i^k)^T Y_{ij} D_i^k \right) Z_{ij} C^T \mu(da).$$

The results may also be extended to certain cases when $F_i F_i^T$ is not positive definite, the observation structures C, I_j depend on a , and the control is affine in the null space of B_i . Open problems include study of the solution space of (11), study of numerical methods for solving (8), (11), including determination of rates of convergence of K_T^* , and extensions to jump processes with continuous state space.

Appendix A.

Proof of Proposition 3. Since $(1-\delta) \max \{\lambda_i\} \geq 0$ implies Γ_1 is stable, let \tilde{X}_i , $1 \leq i \leq N$, denote the unique solution of $L_i(\tilde{X}_i) + \sum_{j=1}^N \pi_{ij} \tilde{X}_j + I_n = 0$. If $a \in S$ and $\delta = 0$, then as in the proof of Proposition 2

$$E \left\{ \int_0^\infty \xi^T \xi dt \mid \xi(0) = x, \beta(0) = i \right\} = x^T \tilde{X}_i x,$$

implying $E\{\|\xi(t)\|^2 \mid \xi(0) = x, \beta(0) = i\} \rightarrow 0$ as $t \rightarrow \infty$. If $\delta = 1$, then

$$\lim_{T \rightarrow \infty} E \left\{ \frac{1}{T} \int_0^T \xi^T \xi dt \right\} = N \sum_{i=1}^N \int_S \alpha_i \text{trace}(F_i^T \tilde{X}_i F_i) \mu(da).$$

Therefore, $\xi(t)$ has bounded second moments, implying the existence of an invariant probability distribution [1].

Proof of Proposition 4. Choose $\bar{Q} = \varepsilon I_n$, $\varepsilon > 0$ and let $K_0 \in U_0$. Since S is closed and bounded, $K_0 \in \partial U_0(a_0)$ for some $a_0 \in S$. Furthermore, it may be shown that $\partial U_0(a_0)$ is a subset of the curve $\{K \in U: \det(\Gamma_1) + (1-\delta) \max_i \{\lambda_i\} \mid_{a=a_0} = 0\}$ and $\lim_{\kappa \rightarrow \kappa_0} \sum_{i,j=1}^N \text{trace}(X_{ij}) = +\infty$ for $a = a_0$. The result follows trivially in (i) while in (ii) it follows from the inequality:

$$\sum_{i,j=1}^N \int_S \text{trace}(X_{ij}) \mu(da) \geq \kappa \int_S (1+g)/h da$$

where $\kappa > 0$ and g, h are polynomials in (a, K) vanishing at (a_0, K_0) .

Appendix B. The notation does not correspond to that used in the body of the paper. Let S be a closed and bounded subset of a Euclidean space. Let $A \in R^{n \times n}$, $b(t)$, \bar{b} , $c \in R^n$ be continuous functions of $t \in [0, T]$, $a \in S$, $k \in R^m$, which are continuously differentiable in k . Let $\Omega = \{k: \text{real}(\sigma(A)) < 0 \text{ for all } a \in S\}$. Define

$$J_T(k) = \int_S c^T g(T) \mu(da), \quad g(t) = \int_0^t e^{A(t-s)} b(s) ds, \quad h(t) = e^{A^T(T-t)} c$$

and for $k \in \Omega$ define

$$J_{\infty}(k) = \int_S c^T \bar{g} \mu(da), \quad \bar{g} = -A^{-1} \bar{b}, \quad \bar{h} = -(A^T)^{-1} c.$$

PROPOSITION B1.

(i) If J_T has a local minimum at $k^* \in R^m$ then

$$\int_0^T \int_S h^T(t) \left(\frac{\partial A}{\partial k_i} g(t) + \frac{\partial b}{\partial k_i}(t) \right) \mu(da) dt \Big|_{k=k^*} = 0, \quad 1 \leq i \leq m.$$

(ii) If J_{∞} has a local minimum at $k^* \in \Omega$ then

$$\int_S \bar{h}^T \left(\frac{\partial A}{\partial k_i} \bar{g} + \frac{\partial \bar{b}}{\partial k_i} \right) \mu(da) \Big|_{k=k^*} = 0, \quad 1 \leq i \leq m.$$

Proof. Because Ω is open, the formulas follow from a computation of the gradient of $J(k)$, utilizing the formulas $(\partial/\partial k_i)A^{-1} = -A^{-1}(\partial A/\partial k_i)A^{-1}$ and (27) of [3],

$$\frac{\partial}{\partial k_i} e^{At} = \int_0^t e^{A(t-s)} \frac{\partial A}{\partial k_i} e^{As} ds.$$

REFERENCES

- [1] V. E. BENES, *Finite regular invariant measures for Feller processes*, J. Appl. Probab., 5 (1968), pp. 203–209.
- [2] G. L. BLANKENSHIP AND G. C. PAPANICOLAOU, *Stability and control of stochastic systems with wide-band noise disturbances*, II, preprint.
- [3] J. W. BREWER, *Kronecker products and matrix calculus in system theory*, IEEE Trans. Circuits and Systems, 25 (1978), pp. 772–781.
- [4] S. S. L. CHANG AND T. K. C. PENG, *Adaptive guaranteed cost control of systems with uncertain parameters*, IEEE Trans. Automat. Control, AC-17 (1972), pp. 474–483.
- [5] E. EMRE, *Simultaneous stabilization with fixed closed loop characteristic polynomial*, IEEE Trans. Automat. Control, AC-28 (1983), pp. 103–104.
- [6] B. K. GHOSH AND C. I. BYRNES, *Simultaneous stabilization and simultaneous pole-placement by nonswitching dynamic compensation*, IEEE Trans. Automat. Control, AC-28 (1983), pp. 735–741.
- [7] U. G. HAUSSMANN, *Stability of linear systems with control dependent noise*, SIAM J. Control., 11 (1973), pp. 382–394.
- [8] W. E. HOPKINS, JR., *Optimal control of linear systems with parameter uncertainty*, IEEE Trans. Automat. Control, 31 (1986), pp. 72–74.
- [9] H. P. HORISBERGER AND P. R. BELANGER, *Regulators for linear, time invariant plants with uncertain parameters*, IEEE Trans. Automat. Control, AC-21 (1976), pp. 705–708.
- [10] D. C. HYLAND AND D. S. BERNSTEIN, *The optimal projection equations for fixed-order dynamic compensation*, IEEE Trans. Automat. Control, 29 (1984), pp. 1034–1037.
- [11] G. LEITMANN, *Guaranteed ultimate boundedness for a class of uncertain linear dynamical systems*, IEEE Trans. Automat. Control, AC-23 (1978), pp. 1109–1110.
- [12] W. S. LEVINE AND M. ATHANS, *On the determination of the optimal output feedback gains for linear multivariable systems*, IEEE Trans. Automat. Control, AC-15 (1970), pp. 44–48.
- [13] M. MARITON AND P. BERTRAND, *Non-switching control strategies for continuous-time jump linear quadratic systems*, Proc. 24th IEEE Conf. Decision Control, Ft. Lauderdale, FL, December 1985, pp. 916–921.
- [14] I. R. PETERSEN AND C. V. HOLLOT, *A Riccati equation approach to the stabilization of uncertain linear systems*, preprint.
- [15] J. F. RANDOLPH, *Basic Real and Abstract Analysis*, Academic Press, New York, 1968.
- [16] D. M. SALMON, *Minimax controller design*, IEEE Trans. Automat. Control, AC-13 (1968), pp. 369–376.
- [17] U. SHAKED, *The design of multivariable systems having zero sensitive poles*, IEEE Trans. Automat. Control, AC-24 (1979), pp. 117–119.
- [18] R. TARRES, *Asymptotic evolution of a stochastic control problem*, this Journal, 23 (1985), pp. 614–631.
- [19] J. S. THORP AND B. R. BARMISH, *On guaranteed stability of uncertain linear systems via linear control*, J. Optim. Theory Appl., 35 (1981), pp. 559–579.

- [20] J. L. WILLEMS AND J. C. WILLEMS, *Robust stabilization of uncertain systems*, this Journal, 21 (1983), pp. 352-374.
- [21] ———, *Feedback stabilizability for stochastic systems with state and control dependent noise*, Automatica, 12 (1976), pp. 277-283.
- [22] W. M. WONHAM, *Optimal stationary control of a linear system with state-dependent noise*, SIAM J. Control., 5 (1967), pp. 486-500.
- [23] ———, *Random differential equations in control theory*, in Probabilistic Methods in Applied Mathematics, Vol. 2, A. T. Bharucha-Reid, ed., Academic Press, New York, 1970, pp. 131-212.
- [24] ———, *Linear Multivariable Control: A Geometric Approach*, Springer-Verlag, New York, 1985.

INFINITE HORIZON OPTIMIZATION FOR FINITE STATE MARKOV CHAIN*

ARIE LEIZAROWITZ†

Abstract. We consider the infinite horizon optimal control of a finite state Markov chain from the point of view of overtaking optimality and the long-run average cost. The stochastic model is cast into a deterministic framework by considering the distribution of the original state as a new state. Using known results about deterministic control systems we obtain short proofs of existence and characterization of stationary overtaking optimal strategies for the stochastic problem.

We characterize and prove existence of stationary strategies which have a minimal cost growth rate in the class of all nonanticipative strategies. Restricting our attention only to stationary strategies we show that for every given initial state there exists an overtaking optimal strategy. Finally, under more restrictive conditions, we establish the existence of a stationary overtaking optimal strategy for all the initial values.

Key words. infinite horizon control problem, long-run minimal cost, overtaking optimal, finite state Markov chain

AMS(MOS) subject classifications. 93E20, 60J10

1. Introduction. This paper studies infinite horizon control of a finite state Markov chain from the point of view of the overtaking optimality criterion and the long-run average cost. The stochastic model is cast into a deterministic framework by considering the distribution of the original states as the new state. Using known results about deterministic control systems will enable us to give short proofs of existence and characterization of stationary overtaking optimal strategies for the stochastic problem.

We shall consider *nonanticipative strategies*, namely strategies for which the value of the control applied at time k depend on the values of the controlled process up to time k . Here $k = 0, 1, 2, \dots$ is the discrete time variable. A subclass of this set is composed of the *Markov strategies* for which the value of the control at the time k is a function of k and the current state of the controlled process at this time. Of special interest are the Markov strategies which are *stationary strategies*, namely, the value of the control u which is applied at time k depends only on the current state of the controlled process, and not on the time variable k .

There are two main results in this paper. One is concerned with the existence and characterization of stationary strategies which have a minimal cost growth rate in the class of all nonanticipative strategies, while the second is about overtaking optimal strategies in the class of all stationary strategies. In the first main result we prove the existence of a stationary strategy with a minimal long-run average cost and we show how it can be computed. Existence results about strategies with a minimal long-run average cost have been displayed in the literature under various assumptions. In Whittle's book [10] it is assumed that the set of action controls is finite. In Kushner's study [6] it is assumed that the controls take values in a countable set. In our framework the set of admissible controls is considerably larger, and the result about strategies with a minimal long-run average cost will be established while assuming that the set of admissible control values is convex and compact. Moreover, any two states communicate after a finite number of steps under the action of a stationary admissible strategy. These assumptions are denoted in the sequel as Assumption B. The convexity assumption is natural since we allow *mixing of controls*. Namely, we assume that if v_1, \dots, v_k

* Received by the editors March 31, 1986; accepted for publication (in revised form) February 5, 1987. This research was supported in part by the Institute for Mathematics and Its Applications with funds provided by the National Science Foundation.

† Department of Mathematics, Carnegie-Mellon University, Pittsburgh, Pennsylvania 15213.

are some possible action controls and if $\lambda_i > 0$ are such that $\sum_{i=1}^k \lambda_i = 1$, then the action of choosing v_i with probability λ_i is a possible control also.

In [1] Borkar proved the existence of stationary strategies with a minimal cost growth rate for Markov chains. The framework of his treatment is different from ours. The state space in [1] is denumerable rather than finite, and the optimization is a.s. rather than of the expected average cost. However, the cost expression considered in [1] is a function of the current state only, while we consider a cost expression which depends on both the current state and the applied control.

The other main result is concerned with overtaking optimality of stationary strategies. We will establish the existence of a stationary overtaking optimal strategy for every initial state. Moreover, the existence of a stationary overtaking optimal strategy for all the initial states will be established under the following assumption. Let $\{1, 2, \dots, n\}$ be the state space. Then for every initial state i at time $k=0$ and a distribution $p = (p_1, \dots, p_n)$ with $p_i > 0$, $1 \leq i \leq n$ there is a control u for which the distribution of the process at time $k=1$ is p . We will demonstrate in an example that this assumption, which looks quite restrictive, holds naturally in certain systems. We want to emphasize that this assumption, which is denoted in the sequel as Assumption A, is needed only in the part of the paper which deals with overtaking optimality in § 6. The existence of optimal strategies with a minimal long-run average cost is proved, in § 5, under the relaxed Assumption B described above. The Assumption A is employed merely as an auxiliary tool in § 4.

We will show that, under Assumption A, the stationary strategy with a minimal cost growth rate possesses an additional optimality property. Denote this strategy by σ_0 and let σ be any nonanticipative strategy. Then there is a constant K such that the cost of σ_0 on the time interval $[0, k]$ is not larger than that of σ on this interval by more than K , and this for every $k \geq 1$. This property clearly implies that σ_0 has a minimal cost growth rate. One can, however, construct two nonanticipative strategies, both having a minimal long-run average cost. Nevertheless, the difference between their respective costs on the time intervals $[0, k]$ diverge to infinity as k grows to infinity.

The fact that the minimal cost growth is attained by a stationary strategy gives some justification to restricting the discussion to the class of stationary strategies. Moreover, we assume that the cost incurred when applying a certain control while the system occupies a given state does not depend on the time when this control is employed. Therefore one expects that strategies which are optimal in a reasonable sense would be stationary. Indeed, we will prove that under Assumption B there exists an overtaking optimal stationary strategy, for every given initial state. However, under Assumption A, we will establish the following much stronger property: Every stationary strategy with a minimal long-run average cost is in fact overtaking optimal, for all the initial states, in the class of stationary strategies. This main result will also provide us with a computational tool for overtaking optimal stationary strategies. In the course of proving our main result we will show that the deterministic problem associated with our stochastic problem has the remarkable property of having overtaking optimal solutions with respect to a *vector cost criterion*.

We will consider now some possible extensions and discuss some inherent limitations of our approach. Concerning overtaking optimal strategies, there are several possible extensions. One is to establish the existence of overtaking stationary strategies in the class of all the *Markov strategies*, rather than in the smaller class of stationary strategies. Another possible extension is to relax the controllability assumption that every distribution can be obtained in a unit time from any initial state. Our approach has the inherent limitation of requiring this in order to exploit results about the

deterministic control systems. It would be desirable to prove existence and characterize overtaking optimal strategies under the natural assumption that the constraint set for admissible controls is convex and compact. Such a result is motivated by the two main results of this paper. Another possible extension is to consider controlled Markov chains with infinite denumerable state space.

Our approach is to consider the distribution of the states as a new state, and accordingly we study the expected cost values. Thus there is no hope to get a.s. optimality results. This, however, is not much of a limitation as far as overtaking optimality is concerned, since, in general, one can hardly expect a.s. optimization in this context.

The paper is organized as follows. In § 2 we describe the stochastic control problem and display examples which satisfy the assumptions of our model. In § 3 we associate a deterministic control problem with the original stochastic one. The existence and computation of a stationary optimal strategy with a minimal long run average cost is proved in § 4 under Assumption A and in § 5 under the relaxed Assumption B.

In § 6 we restrict our attention to stationary strategies and show that, under Assumption B, there exists an overtaking optimal stationary strategy for every initial state. Furthermore, assuming Assumption A, we will show that every stationary strategy with a minimal long-run average cost is in fact overtaking optimal for every initial state.

2. The stochastic control problem. We consider the controlling of a discrete time stochastic process which is defined on a probability space (Ω, F, P) . The state space of the controlled process is the finite set $\{1, 2, \dots, n\}$. We denote the discrete time variable by $k = 0, 1, 2, \dots$. If we apply a control u at time k while the system occupies the state i , $1 \leq i \leq n$, then as a result the system will occupy the state j at time $k+1$ with probability $p_{ij}(u)$, which does not depend on time k . We assume that the class of controls is *closed under mixing*, namely: If u_1, \dots, u_k are admissible controls, and $\lambda_i \geq 0$, $\sum_{i=1}^k \lambda_i = 1$, then the control u which is obtained by choosing u_i with probability λ_i is an admissible control also.

We denote by S the following subset of R^n :

$$S = \left\{ q \in R^n : q_i > 0, \sum_{i=1}^n q_i = 1 \right\}$$

where q_i denotes the i th component of q . S is a convex subset of R^n , and let \tilde{S} be the smallest affine subset of R^n which contains S . Then we denote by $\text{rel int } S$ and by ∂S , respectively, the interior and boundary of S as a subset of \tilde{S} (see Rockafellar [8]).

Suppose that when the system occupies the state i then a cost $\phi(i, u)$ incurred if the control u is applied. If the control u is obtained by mixing the controls u_1, \dots, u_k with the respective probabilities $\lambda_1, \dots, \lambda_k$, then the cost of using it when the system occupies the state i is $\sum_{j=1}^k \lambda_j \phi(i, u_j)$. For a given state i and $q \in S$, let $U_i(q)$ be the set of all the controls u such that $p_{ij}(u) = q_j$ for all $1 \leq j \leq n$ (q_j is the j th component of q). Then we define

$$\phi_i(q) = \inf \{ \phi(i, u) : u \in U_i(q) \}$$

agreeing that the infimum over the empty set is $+\infty$. $\phi_i(q)$ is the infimal cost for obtaining the distribution q at the next sampling time if the present state is i . We thus consider the set of all stochastic matrices $M = (M_{ij})$, $1 \leq i, j \leq n$, as the set of our admissible controls. The controller chooses at every instant of time such a matrix M , and if the system occupies the state i at that time then the incurred cost is $\phi_i(M_i)$, where M_i is the i th row of M .

Remark. Because we allow mixing of controls in the original framework of the model, the set of admissible stochastic matrices is convex.

We say that $q \in R^n$ is positive and denote $q > 0$ if $q_j > 0$ for every $1 \leq j \leq n$. We shall assume the following.

ASSUMPTION A.

- (i) Let $q \in S$ be positive. Then for every i there is a control u such that $p_{ij}(u) = q_j$ for all $1 \leq j \leq n$.
 - (ii) The cost expression $q \rightarrow \phi_i(q)$ satisfies
- $$(2.1) \quad \lim \phi_i(q) = \infty \quad \text{as } q \rightarrow \partial S$$
- for every $1 \leq i \leq n$.

Remark. The meaning of (i) is that every distribution $q > 0$ can be achieved in a unit time from every initial state by a choice of a suitable control. Example 2.2 will display systems for which this holds under natural assumptions.

To explain the meaning of (ii) we observe that if q belongs to ∂S then at least one state is a.s. unoccupied. Thus assuming (ii) means that all the states communicate in the unit time interval, under the action of every admissible control. Moreover, an action which guarantees a high probability for the nonoccupancy of a certain state would also be very expensive and the relation between these two quantities is expressed in (2.1). These remarks are further demonstrated in the following example.

Example 2.1. We consider a continuous time Markov chain with the state space $\{1, \dots, n\}$. If $t \rightarrow A(t)$ is its generator and $t \rightarrow q(t)$ is its distribution at time t , then

$$(2.2) \quad \frac{dq}{dt} = A^T(t)q(t), \quad q(0) = q_0$$

where q_0 is the initial distribution at time $t = 0$, and A^T is the transpose of A (see e.g. Elliott [4]). The entries $A_{ij}(t)$ of the matrix $A(t)$ determine the rates of transition from i to j at the time t . In the case of a controlled Markov chain the controller can influence these values. The function $A(\cdot)$ is thus the control function and we assume in this example that the class of admissible controls is composed of all the continuous functions $t \rightarrow A(t)$ such that $A_{ij}(t) \geq 0$ for $i \neq j$ and $A_{ii}(t) = -\sum_{j \neq i} A_{ij}(t)$. This assumption will be much relaxed in Example 2.2. Let \mathcal{A} denote the set of $n \times n$ matrices A such that $A_{ij} \geq 0$ whenever $i \neq j$ and $A_{ii} = -\sum_{j \neq i} A_{ij}$. Let $X \subset R^n$ be the subspace $\{x \in R^n: \sum_{i=1}^n x_i = 0\}$ where x_i is the i th component of x . Then for every $q \in \text{rel int } S$ we have

$$\{A^T q: A \in \mathcal{A}\} = X,$$

as can easily be seen from the definition of \mathcal{A} and the fact that $A^T q$ is a convex combination of the columns of A^T with the weights $\{q_1, \dots, q_n\}$. From this it follows that for every $q \in \text{rel int } S$ and every initial distribution q_0 there is an admissible control $A(\cdot)$ such that the solution $q(\cdot)$ of (2.2) satisfies $q(1) = q$. Thus Assumption A(i) holds.

The following discussion serves as motivation and explanation of A(ii). If for two different states i and j the value of $A_{ij}(t)$ is very small then this means that transitions from i to j in a short time interval after t are very unlikely to occur. On the other hand very large values of $A_{ij}(t)$ mean that there is a high probability that the state will change from i to j during such a time interval. Thus we see that extreme values of A_{ij} correspond to a more predictable behavior of the system. It is, however, reasonable to expect that controls which achieve a less random behavior of the system will also be the more expensive ones. Thus a reasonable cost expression in the present model is

$$(2.3) \quad \sum_{i \neq j} \phi(A_{ij}(t)) dt$$

where $\phi: R^+ \rightarrow R^+$ is a convex function satisfying $\lim \phi(\xi) = \infty$ as $\xi \rightarrow \infty$ or $\varepsilon \rightarrow 0+$. (The convexity is natural in light of the possibility of mixing controls.)

For a given i let $C_i(\alpha, \beta)$ be the set of all values $q(1)$, where $q(\cdot)$ is a solution of (2.2) and where $A(\cdot)$ runs over all the possible measurable functions $t \rightarrow A(t)$ such that $0 < \alpha \leq A_{ij} \leq \beta$ for all $0 \leq t \leq 1$ and all $i \neq j$, $1 \leq i, j \leq n$. It is not hard to see that $C_i(\alpha, \beta)$ is a closed set and $C_i(\alpha, \beta) \subset \text{rel int } S$, namely $C_i(\alpha, \beta)$ is bounded away from ∂S . (Consider $t \rightarrow A(t)$ as matrices whose entries are bounded elements of $L^2[0, 1]$. Then the solutions of (2.2) are uniformly continuous. For a sequence $\{A_k(\cdot)\}_{k=0}^\infty$ of controls with a corresponding solution $\{q_k(\cdot)\}_{k=0}^\infty$ we can assume that $A_k(\cdot) \rightarrow A_0(\cdot)$ weakly and $q_k(\cdot) \rightarrow q_0(\cdot)$ strongly, both in $L_2[0, 1]$. Therefore $q_0(\cdot)$ is the solution corresponding to $A_0(\cdot)$.) This means that if q is sufficiently close to ∂S , then there must be values of $A_{ij}(t)$ which are either very small or very large, thus causing the cost expression in (2.3) to become large.

The following example relaxes the assumption on \mathcal{A} in Example 2.1.

Example 2.2. The controlled process is as described in Example 2.1. Let $X \subset R^n$ be the subspace $\{x \in R^n: \sum_{i=1}^n x_i = 0\}$ and let a_1, \dots, a_n be in X such that $(a_i)_j > 0$ whenever $j \neq i$. Denote by K the convex hull of a_1, \dots, a_n (namely the smallest convex set in X which contains $\{a_1, \dots, a_n\}$). We assume that K has a nonempty interior as a subset of X , and moreover, zero belongs to this interior. Then every $x \in X$ can be written as

$$(2.4) \quad x = \sum_{i=1}^n \alpha_i a_i$$

for some $\alpha_i \geq 0$, $1 \leq i \leq n$. Let the set of all admissible values for the controls \mathcal{A}_0 be defined by

$$\mathcal{A}_0 = \left\{ A: A = \begin{pmatrix} - & u_1 a_1^T & - \\ & \vdots & \\ - & u_n a_n^T & - \end{pmatrix} u_i > 0, 1 \leq i \leq n \right\}$$

where the n vectors a_i , $1 \leq i \leq n$, are as above. An admissible control is thus a continuous function $t \rightarrow A(t)$ such that $A(t) \in \mathcal{A}_0$ for every $0 \leq t \leq 1$.

Thus the situation is that, for every given state, the rates of transition to the other states during short time intervals have fixed ratios on which the controller does not have any influence at all. The controller can choose the "intensities" of the transitions from each state and at all the times, and thus can either increase or decrease all the transition rates from a certain state by the same factor. In this framework a control function can be identified with a continuous function $u: [0, 1] \rightarrow R^n$ such that each component $u_i(t)$ is nonnegative for all $0 \leq t \leq 1$.

Let $q \in \text{rel int } S$ and consider the set

$$C_q = \{A^T q: A \in \mathcal{A}_0\}.$$

Then we claim that $C_q = X$. Let $x \in X$; then it can be written as in (2.4) for some nonnegative α_i , $1 \leq i \leq n$, and then

$$x = \begin{pmatrix} | & & | \\ u_1 a_1 & \cdots & u_n a_n \\ | & & | \end{pmatrix} q$$

where $u_i = \alpha_i/q_i$, which proves that $x \in C_q$. Therefore for every initial distribution q_0 and every $q \in \text{rel int } S$ there is an admissible control $t \rightarrow A(t)$ such that the solution $q(1)$ of (2.2) satisfies $q(1) = q$, namely Assumption A(i) holds.

PROPOSITION 2.3. *Let Assumption A hold. Then $q \rightarrow \phi_i(q)$ is a convex function on $\text{rel int } S$ for every $1 \leq i \leq n$. For every $q \in \text{rel int } S$ and $1 \leq i \leq n$, we have $\phi_i(q) > -\infty$.*

Proof. It follows from the possibility of mixing controls that whenever q_1 and q_2 are in $\text{rel int } S$ and $q = \frac{1}{2}(q_1 + q_2)$, then $\phi_i(q) \leq \frac{1}{2}[\phi_i(q_1) + \phi_i(q_2)]$, which proves the convexity of $\phi_i(\cdot)$. If a convex function is larger than $-\infty$ somewhere in the relative interior of its effective domain then it is greater than $-\infty$ everywhere (see Rockafellar [8]). Thus the second assertion of the proposition follows from (2.1). \square

We consider the control problem on an infinite time interval. The controller chooses a strategy σ in order to control the system's dynamics. In our framework a strategy is a sequence $\{M^{(k)}\}_{k=1}^{\infty}$ of stochastic matrices. Thus if

$$M^{(k)} = \begin{pmatrix} - & M_1^{(k)} & - \\ & \vdots & \\ - & M_n^{(k)} & - \end{pmatrix}$$

where $M_j^{(k)}$ is the j th row of $M^{(k)}$, then $M_i^{(k)}$ is the distribution of the states of the system at time $k+1$ if it occupies the state i at time k .

DEFINITION 2.4. Let the random variable $\zeta^{(k)}$ be the state of the system at time k . A strategy $\sigma = \{M^{(k)}\}_{k=0}^{\infty}$ is nonanticipative if $M^{(k)}$ is a function of $\zeta^{(0)}, \dots, \zeta^{(k)}$ only. A Markov strategy is a sequence $\{M^{(k)}\}_{k=0}^{\infty}$ of stochastic matrices (with no functional dependence of $M^{(k)}$ on $\zeta^{(0)}, \dots, \zeta^{(k-1)}$). A Markov strategy σ is a stationary strategy if $\sigma = \{M\}_{k=0}^{\infty}$ for some fixed stochastic matrix M .

With every initial state i and every Markov strategy $\sigma = \{M^{(k)}\}_{k=0}^{\infty}$ there is associated a Markov chain $\{\zeta^{(k)}\}_{k=0}^{\infty}$ where

$$P(\zeta^{(0)} = i) = 1 \quad \text{and} \quad \zeta^{(k)} \in \{1, 2, \dots, n\}.$$

The conditional distribution of $\zeta^{(k+1)}$ given that $\zeta^{(k)} = j$ is given by $M_j^{(k)}$, the j th row of $M^{(k)}$

$$P(\zeta^{(k+1)} = l | \zeta^{(k)} = j) = M_{jl}^{(k)}.$$

We denote by $x^{(k)}$ the distribution of $\zeta^{(k)}$ and thus have the relation

$$(2.5) \quad x^{(k+1)} = (M^{(k)})^T x^{(k)}$$

(where M^T is the transpose of M).

If the system occupies the state i then the cost $\phi_i(M_i)$ is incurred if the control M is applied. Thus the expected cost on the $[0, N]$ time interval which corresponds to the strategy σ is

$$(2.6) \quad C_N(i, \sigma) = E_i \sum_{k=0}^{N-1} \phi_{\zeta^{(k)}}(M_{\zeta^{(k)}}^{(k)}).$$

We call the sequence $\{C_N(i, \sigma)\}_{N=1}^{\infty}$ the cost flow which corresponds to σ , and are interested in the behavior of this sequence as N tends to infinity. We will be interested in finding a sequence which is minimal according to one of the following criterions: The minimal long-run average cost, or the overtaking optimality criterion. A cost flow $\{C_N\}$ has a minimal long-run average if $\lim_{N \rightarrow \infty} (1/N)C_N$ exists and is not larger than $\lim_{N \rightarrow \infty} \inf (1/N)C'_N$ for every other cost flow $\{C'_N\}_{N=1}^{\infty}$. (The overtaking optimality criterion will be defined in the next section.) We write (2.6) as

$$(2.7) \quad C_N(i, \sigma) = \sum_{k=0}^{N-1} \left[\sum_{j=1}^n x_j^{(k)} \phi_j(M_j^{(k)}) \right]$$

where

$$x_j^{(k)} = P(\zeta^{(k)} = j | \zeta^{(0)} = i, \sigma),$$

namely $x^{(k)}$, is the distribution of $\zeta^{(k)}$ given the initial state i and the strategy σ . Motivated by (2.5) and (2.7) we define the following function: Let x and y be in $\text{rel int } S$, then

$$(2.8) \quad u(x, y) = \inf \left\{ \sum_{j=1}^n x_j \phi_j(M_j) : M^T x = y \right\}$$

where the infimum is carried over all the stochastic matrices M which satisfy the equation $M^T x = y$. It is easy to see that $u(x, y)$ is finite whenever y is in $\text{rel int } S$.

3. The associated deterministic problem. We now relate a deterministic control problem to the stochastic control problem described in the previous section. To every sequence $\bar{x} = \{x_k\}_{k=0}^\infty$, where $x_k \in S$, there corresponds the cost flow

$$(3.1) \quad C_N(\bar{x}) = \sum_{k=0}^{N-1} u(x_k, x_{k+1})$$

where u is the function defined in (2.8). Let the initial value x_0 be fixed. We want to find a sequence \bar{x} for which the cost flow $\{C_N(\bar{x})\}$ is minimal, either according to the long-run average cost criterion, or according to the overtaking optimality criterion. More precisely, if \bar{x} is such that

$$\lim_{N \rightarrow \infty} \frac{1}{N} C_N(\bar{x}) \leq \liminf_{N \rightarrow \infty} \frac{1}{N} C_N(\bar{y})$$

for every other \bar{y} with $y_0 = x_0$, then we say that \bar{x} has a *minimal cost growth rate*. As mentioned, it can also be minimal in a more refined sense, namely in the *overtaking optimality sense*. This notion was introduced in the economic literature by Gale [5], von Weizsacker [9] and used in the control literature by, e.g., Brock and Haurie [2] and Carlson [3]. A version of this notion is given by the following.

DEFINITION 3.1. Let the initial value x_0 be fixed. We say that \bar{x} is an overtaking optimal sequence if for every \bar{y} with $y_0 = x_0$ and every $\varepsilon > 0$ there exists an N_0 such that

$$C_N(\bar{x}) < C_N(\bar{y}) + \varepsilon$$

for all $N \geq N_0$.

Namely, up to an arbitrarily small $\varepsilon > 0$ the sequence \bar{x} is better than any other sequence from a certain time on. In particular, it follows that \bar{x} has a minimal cost growth rate.

PROPOSITION 3.2. For every $x, y \in \text{rel int } S$ the infimum in (2.8) is attained for some stochastic matrix with positive rows. The function $(x, y) \rightarrow u(x, y)$ is continuous on $(\text{rel int } S) \times (\text{rel int } S)$.

Proof. For given $x, y \in \text{rel int } S$, it follows from (2.1) and (2.8) that there is a minimizing sequence of stochastic matrices $\{M_j\}_{j=1}^\infty$ such that $(M_j)_l \in K \subset \text{rel int } S$ for every $j \geq 1$ and $1 \leq l \leq n$, where K is a compact set, and

$$u(x, y) = \lim_{j \rightarrow \infty} \sum_{i=1}^n x_i \phi_i((M_j)_i).$$

(Here $(M_j)_l$ is the l th row of M_j .) We can assume that $M_j \rightarrow M_0$ as $j \rightarrow \infty$ and then $u(x, y) = \sum_{i=1}^n x_i \phi_i((M_0)_i)$, proving the first assertion. A similar argument proves that $(x, y) \rightarrow u(x, y)$ is a lower semicontinuous function in $(\text{rel int } S) \times (\text{rel int } S)$.

Let $(x, y) \in (\text{rel int } S) \times (\text{rel int } S)$ and let $u(x, y) = \sum_{i=1}^n x_i \phi_i(M_i)$ where M is a stochastic matrix with positive rows M_i , $i \leq i \leq n$. Given an $\varepsilon > 0$ we can find a $\delta > 0$ such that whenever $x', y' \in \text{rel int } S$ satisfy $|x - x'| + |y - y'| < \delta$ then there is a stochastic matrix M' with $\|M' - M\| < \varepsilon$ satisfying $y' = (M')^T x'$ (here $\|\cdot\|$ is a norm on the space of $n \times n$ matrices). This follows from the observation that $y = M^T x$ is equivalent to saying that y is a convex combination of the rows of M with the weights x_1, \dots, x_n . Therefore the function $(x, y) \rightarrow u(x, y)$ is also upper semicontinuous in $(\text{rel int } S) \times (\text{rel int } S)$. This concludes the proof of the proposition. \square

Thus there is a correspondence between the original stochastic control problem with state space consisting of n points and the deterministic control problem of minimizing the cost flow $C_N(\bar{x})$ in (3.1), for sequences $\bar{x} = \{x_k\}_{k=0}^\infty$ in S . As described above every initial state i and a strategy σ induces a sequence \bar{x} in S . On the other hand, for the sequence $\{x_k\}_{k=0}^\infty$ in $\text{rel int } S$ there exist, by Proposition 3.2, matrices $\{M^{(k)}\}_{k=0}^\infty$ such that $x_{k+1} = (M^{(k)})^T x_k$ for $k \geq 0$, and x_k is the distribution of $\zeta^{(k)}$ where $\{\zeta^{(k)}\}_{k=0}^\infty$ is the Markov chain associated with the initial distribution x_0 and the strategy $\sigma = \{M^{(k)}\}_{k=0}^\infty$. The cost flows corresponding to \bar{x} and to σ and x_0 are the same.

4. Strategies with a minimal cost growth rate. The deterministic problem which was described in § 3 was studied by Leizarowitz [7] in light of the overtaking criterion. The framework is to consider sequences $\bar{x} = \{x_k\}_{k=0}^\infty$ in a compact set D in R^n and study the associated cost flows

$$C_N(\bar{x}) = \sum_{k=0}^{N-1} u(x_k, x_{k+1}).$$

The function $u(\cdot, \cdot)$ is assumed to be continuous on $D \times D$. The following result was established there [7, Thm. 3.1].

THEOREM 4.1. *There exist constants $K > 0$ and μ such that:*

(i) *For every \bar{x} in D*

$$(4.1) \quad \sum_{k=0}^{N-1} [u(x_k, x_{k+1}) - \mu] > -K \quad \text{for all } N \geq 1.$$

(ii) *There is a sequence \bar{x}^* in D such that*

$$(4.2) \quad \left| \sum_{k=0}^{N-1} [u(x_k^*, x_{k+1}^*) - \mu] \right| < K \quad \text{for all } N \geq 1.$$

Remark. It is clear that the constant μ is unique, and it describes the minimal cost growth rate possible for sequences in D . The sequence \bar{x}^* which is mentioned in the theorem indeed has a cost flow with this minimal cost growth rate. It is not true, however, that any sequence with a minimal cost growth rate satisfies (4.2). It is true for stationary strategies: If σ is stationary and of a minimal cost growth, then $|C_N(x_0, \sigma) - \mu N| < K$ for some K and all $N \geq 1$. This is easily obtained for $\sigma = \{M\}_{k=0}^\infty$ by considering $C_N(z_0, \sigma)$ where z_0 satisfies $M^T z_0 = z_0$ and $u(z_0, z_0) = \mu$, realizing that $C_n(z_0, \sigma) = \mu N$ for every $N \geq 1$ and $C_n(i, \sigma) - \mu N > -K$ for some K and all $N \geq 1$ and i .

It is also proved in § 8 of [7] that if $u(\cdot, \cdot)$ is continuous on $R^n \times R^n$ and $u(x, y) \rightarrow \infty$ as $|x| + |y| \rightarrow \infty$ the conclusions of Theorem 4.1 remain true for sequences in R^n . In fact, the same proof as that in [7] implies that these conclusions still hold under the assumption that $u(x, y) \rightarrow \infty$ as $|y| \rightarrow \infty$ uniformly for x in R^n . Thus, since $\text{rel int } S$ is homeomorphic to R^{n-1} and $u(\cdot, \cdot)$ in (2.8) is such that

$$\lim u(x, y) = \infty \quad \text{as } y \rightarrow \partial S$$

uniformly for $x \in S$ we conclude that the relations (4.1) and (4.2) hold when we consider sequences in $\text{rel int } S$.

Let σ be a nonanticipative strategy and let, as in (2.6), $C_N(i, \sigma)$ be the expected cost while using σ on the $[0, N]$ time interval beginning at the initial state i . If, however, the initial state is i with probability x_i , $1 \leq i \leq n$, then the expected cost is given by

$$(4.3) \quad \sum_{i=1}^n x_i C_N(i, \sigma) = x^T \cdot C_N(\sigma)$$

where we denote $x = (x_i)_{i=1}^n$ in R^n and $C_N(\sigma)$ is the point in R^n whose i th component is $C_N(i, \sigma)$. We have the following result.

PROPOSITION 4.2. *There is a Markov strategy σ_0 such that the following holds:*

$$(4.4) \quad C_N(i, \sigma_0) < C_N(i, \sigma) + K \quad \text{for all } N \geq 1$$

for some constant K and for every initial state i and a nonanticipative strategy σ . In particular the strategy σ_0 has a minimal cost growth rate

$$\lim_{N \rightarrow \infty} \frac{1}{N} C_N(i, \sigma_0) \leq \liminf_{N \rightarrow \infty} \frac{1}{N} C_N(i, \sigma)$$

for every nonanticipative σ and every $1 \leq i \leq n$.

Proof. Let $u(\cdot, \cdot)$ be the function defined in (2.8). It follows from Theorem 4.1 and the discussion following it that there is a sequence $\bar{x}^* = \{x_k^*\}_{k=0}^\infty$ in $\text{rel int } S$ for which (4.1) and (4.2) hold. It follows from Proposition 3.2 that there is a sequence of matrices $\{M^{(k)}\}_{k=0}^\infty$ such that

$$(4.5) \quad x_{k+1}^* = (M^{(k)})^T x_k^* \quad \text{for all } k \geq 0$$

and

$$(4.6) \quad u(x_k^*, x_{k+1}^*) = \sum_{j=1}^n (x_k^*)_j \phi_j(M_j^{(k)}).$$

Let σ_0 be the Markov strategy $\{M^{(k)}\}_{k=0}^\infty$. Using the notation employed in (4.3), we get from (4.5) and (4.6) that

$$(4.7) \quad -K' \leq (x_0^*)^T \cdot C_N(\sigma_0) - \mu N \leq K'$$

for some constant K' and all $N \geq 1$. Since the following holds for every $1 \leq i \leq n$ and every $N \geq 1$

$$C_N(i, \sigma_0) - \mu N \geq -K',$$

it follows from (4.7) and the fact that $x_0^* \in \text{rel int } S$ that there is a constant K'' such that

$$(4.8) \quad |C_N(i, \sigma_0) - \mu N| \leq K''$$

for all $N \geq 1$ and all $1 \leq i \leq n$.

We now claim that

$$(4.9) \quad C_N(i, \sigma) - \mu N \geq -K'$$

for every nonanticipative strategy σ , all $1 \leq i \leq n$ and all $N \geq 1$. This is true for every Markov strategy, as a consequence of Theorem 4.1. Now let σ be any nonanticipative strategy, say $\sigma = \{N^{(k)}\}_{k=0}^\infty$, where each $N^{(k)}$ is now a random matrix which is measurable with respect to the σ -algebra that is generated by the random variables $\zeta^{(0)}, \dots, \zeta^{(k)}$. Let the distribution $x^{(0)}$ of $\zeta^{(0)}$ be given and let $x^{(k)} \in R^n$ be defined by

$$(x^{(k)})_j = P(\zeta^{(k)} = j).$$

Thus $\{x^{(k)}\}_{k=0}^\infty$ is the sequence of distributions of $\{\zeta^{(k)}\}_{k=0}^\infty$ as determined by the strategy $\sigma = \{N^{(k)}\}_{k=0}^\infty$. Let us define also

$$(4.10) \quad L^{(k)} = EN^{(k)}.$$

Then $\bar{\sigma} = \{L^{(k)}\}_{k=0}^\infty$ is a Markov strategy and we will show that

$$(4.11) \quad C_N(i, \sigma) \geq C_N(i, \bar{\sigma})$$

for every $N \geq 1$ and every $1 \leq i \leq n$, which will imply the validity of (4.9) for σ , since (4.9) holds for Markov strategies.

To prove (4.11) we will show that $\{L^{(k)}\}_{k=1}^\infty$ in (4.10) and $\{x^{(k)}\}_{k=0}^\infty$ are related by

$$(4.12) \quad x^{(k+1)} = (L^{(k)})^T x^{(k)},$$

which means that $\{x^{(k)}\}_{k=0}^\infty$ is the sequence of distributions of the Markov chain $\{\bar{\zeta}^{(k)}\}_{k=0}^\infty$ which corresponds to the Markov strategy $\bar{\sigma}$. We have

$$\begin{aligned} P(\zeta^{(k+1)} = j_{k+1} | \zeta^{(k)} = j_k) \\ &= E[P(\zeta^{(k+1)} = j_{k+1} | \zeta^{(0)}, \zeta^{(1)}, \dots, \zeta^{(k)}) | \zeta^{(k)} = j_k] \\ &= E[N_{\zeta^{(k)}, j_{k+1}}^{(k)}(\zeta^{(0)}, \dots, \zeta^{(k)}) | \zeta^{(k)} = j_k] = L_{j_k, j_{k+1}}^{(k)}. \end{aligned}$$

Therefore

$$\begin{aligned} P(\zeta^{(k+1)} = j_{k+1}) &= \sum_{j_k=1}^n P(\zeta^{(k+1)} = j_{k+1} | \zeta^{(k)} = j_k) \cdot P(\zeta^{(k)} = j_k) \\ &= \sum_{i=1}^n x_i^{(k)} L_{i, j_{k+1}}^{(k)} \end{aligned}$$

which is the equality (4.12). To obtain the estimate (4.11) we examine a typical term in the cost expression (2.6). Thus we have that

$$\begin{aligned} E\phi_{\zeta^{(k)}}(N_{\zeta^{(k)}}^{(k)}) &= E(E(\phi_{\zeta^{(k)}}(N_{\zeta^{(k)}}^{(k)}) | \zeta^{(k)})) \\ &= \sum_{i=1}^n x_i^{(k)} E(\phi_i(N_i^{(k)}) | \zeta^{(k)} = i) \\ &\geq \sum_{i=1}^n x_i^{(k)} \phi_i(E(N_i^{(k)} | \zeta^{(k)} = i)) \end{aligned}$$

where we used Jensen's inequality in the last step. But since $E(N_i^{(k)} | \zeta^{(k)} = i) = (L^{(k)})_i$, we get the estimate

$$(4.13) \quad E\phi_{\zeta^{(k)}}(N_{\zeta^{(k)}}^{(k)}) \geq \sum_{i=1}^n x_i^{(k)} \phi_i((L^{(k)})_i).$$

Summing (4.13) for $k=0, 1, \dots, N-1$ and recalling (4.12), we get (4.11), since the whole discussion holds for every initial state $1 \leq i < n$. As explained above this proves the validity of (4.9) for every nonanticipative strategy σ . Now the inequality (4.4) is a consequence of (4.8) and (4.9). Dividing (4.4) by N and letting $N \rightarrow \infty$ we get that σ_0 has a minimal cost growth rate. By (4.7) $\lim_{N \rightarrow \infty} (1/N)C_N(i, \sigma_0)$ exists and is equal to μ . \square

The Markov strategy σ_0 which has a minimal cost growth rate is not guaranteed to be a stationary strategy, namely such that $M^{(k)} = M$ for some fixed stochastic matrix and all $k \geq 0$. However, since the problem is time homogeneous and defined on an infinite horizon, one expects the existence of a stationary strategy with a minimal long-run average cost. This we will now prove.

To this end we will carefully examine the constant μ which is guaranteed in Theorem 4.1. It is proved in [7] that

$$(4.14) \quad \mu = \inf \left\{ \frac{1}{N} \sum_{k=0}^{N-1} u(x_k, x_{k+1}) \right\}$$

where the infimization is carried over all the integers $N \geq 1$ and all the possible choices of sequences $\{x_k\}_{k=0}^N$ such that $x_0 = x_N$. For a general discrete time deterministic control system, the infimum in (4.14) is not realizable for a finite N and a finite sequence $\{x_k\}_{k=0}^N$. However, we will see that for the function $u(\cdot, \cdot)$ in (2.8) which is associated with our stochastic control problem, the infimum in (4.14) is realized for $N = 1$, namely there is a point $\theta \in \text{rel int } S$ such that

$$(4.15) \quad \mu = u(\theta, \theta).$$

This will be demonstrated while proving Theorem 4.4. Assume now that (4.15) holds and let M be a stochastic matrix with positive rows such that

$$(4.16) \quad \mu = \sum_{j=1}^n \theta_j \phi_j(M_j) \quad \text{and} \quad \theta = M^T \theta.$$

The existence of such an M follows from Proposition 3.2.

PROPOSITION 4.3. *Assume that $\theta \in \text{rel int } S$ satisfies (4.15) and M is a stochastic matrix for which (4.16) holds. Let σ^* be the stationary strategy*

$$\sigma^* = \{M\}_{k=0}^{\infty}.$$

Then σ^ has a minimal cost growth rate for every initial state i and*

$$\lim_{N \rightarrow \infty} \frac{1}{N} C_N(i, \sigma^*) = \mu.$$

Moreover, there is a constant K such that

$$(4.17) \quad C_N(i, \sigma^*) \leq C_N(i, \sigma) + K$$

for all the nonanticipative strategies σ , and all $N \geq 1$, $1 \leq i \leq n$.

Proof. Let θ and M satisfy (4.15) and (4.16). The constant sequence

$$\bar{x}^* = \{\theta\}_{k=0}^{\infty}$$

is such that

$$\sum_{k=0}^N [u(x_k^*, x_{k+1}^*) - \mu] = 0 \quad \text{for all } N \geq 0.$$

Therefore we can choose, in the proof of Proposition 4.2, the strategy σ_0 to be equal to σ^* . Hence σ^* possesses all the properties of σ_0 in Proposition 4.2, which completes the proof of the proposition. \square

THEOREM 4.4. *There is a stationary strategy σ^* with a minimal cost growth and which satisfies relation (4.17). Such is the strategy $\sigma^* = \{M\}_{k=0}^{\infty}$ where M satisfies (4.16).*

Proof. As discussed above the only thing that remains to prove is the existence of a point $\theta \in \text{rel int } S$ satisfying equation (4.15). Let $N \geq 1$ and let $x^{(0)}, \dots, x^{(N)}$ be a sequence in $\text{rel int } S$ such that $x^{(0)} = x^{(N)}$. To prove the existence of a θ as in (4.15) it is enough to show that there is a $z \in \text{rel int } S$ satisfying

$$(4.18) \quad u(z, z) \leq \frac{1}{N} \sum_{k=0}^{N-1} u(x^{(k)}, x^{(k+1)}).$$

This will show that the infimum in (4.14) can be computed using $N = 1$, which is our assertion. The right-hand side of (4.18) is

$$(4.19) \quad \frac{1}{N} \left\{ \sum_{j=1}^n x_j^{(0)} \phi_j(M_j^{(0)}) + \sum_{j=1}^n x_j^{(1)} \phi_j(M_j^{(1)}) + \cdots + \sum_{j=1}^n x_j^{(N-1)} \phi_j(M_j^{(N-1)}) \right\}$$

where the stochastic matrices $\{M^{(k)}\}_{k=0}^{N-1}$ are such that

$$(4.20) \quad x^{(k+1)} = (M^{(k)})^T x^{(k)}, \quad k = 0, 1, \dots, N-1.$$

Changing the order of summation in (4.19) we get

$$(4.21) \quad \frac{1}{N} \left\{ \sum_{j=1}^n \sum_{k=0}^{N-1} (x^{(k)})_j \phi_j(M_j^{(k)}) \right\} = \sum_{j=1}^n z_j \sum_{k=0}^{N-1} \frac{x_j^{(k)}}{N z_j} \phi_j(M_j^{(k)})$$

where we denoted $z_j = (1/N) \sum_{k=0}^{N-1} x_j^{(k)}$. Then $z = \{z_j\}_{j=1}^n$ is in $\text{rel int } S$. For every fixed j the numbers $\{x_j^{(k)} / N z_j\}_{k=0}^{N-1}$ are positive and sum up to 1. Therefore, by the convexity of $\phi_j(\cdot)$, the expression in the right-hand side of (4.21) is not less than

$$(4.22) \quad \sum_{j=1}^n z_j \phi_j \left(\frac{1}{N z_j} \sum_{k=0}^{N-1} x_j^{(k)} M_j^{(k)} \right) = \sum_{j=1}^n z_j \phi_j(q_j)$$

where $q_j = (1/N z_j) \sum_{k=0}^{N-1} x_j^{(k)} M_j^{(k)}$ is a convex combination of points in $\text{rel int } S$ and hence is in $\text{rel int } S$ itself. Let Q be the stochastic matrix whose j th row is q_j . Then we claim that

$$(4.23) \quad Q^T z = z.$$

The last equation is equivalent to saying that z is a convex combination of the rows of Q with the weights $\{z_j\}_{j=1}^\infty$. But

$$\begin{aligned} \sum_{j=1}^n z_j q_j &= \frac{1}{N} \sum_{j=1}^n \sum_{k=0}^{N-1} x_j^{(k)} M_j^{(k)} \\ &= \frac{1}{N} \sum_{k=0}^{N-1} \sum_{j=1}^n x_j^{(k)} M_j^{(k)} = \frac{1}{N} \sum_{k=0}^{N-1} x^{(k+1)} \end{aligned}$$

where the last equality follows from (4.20). However,

$$\frac{1}{N} \sum_{k=0}^{N-1} x^{(k+1)} = \frac{1}{N} \sum_{k=0}^{N-1} x^{(k)} = z$$

since $x^{(0)} = x^{(N)}$. This proves the equality (4.23). It now follows from (4.21), (4.22) and (4.23) that

$$u(z, z) \leq \frac{1}{N} \sum_{k=0}^{N-1} u(x^{(k)}, x^{(k+1)})$$

and, as explained above, this completes the proof of the theorem. \square

5. Stationary strategies with a minimal cost-growth rate—a general constraint set for the controls. In the discussion up to now we assumed Assumption A(i) namely that every distribution can be realized from every initial state in a unit time by some control. This assumption will now be relaxed and we will assume that the controls take values in some convex and compact set. The convexity assumption is natural since we allow, in the original framework, mixing of controls.

Let \mathcal{M} be the set of stochastic matrices which are admissible controls for our system. If the system occupies the state i and the control $M \in \mathcal{M}$ is employed, then the cost $\phi_i(M_i)$ is incurred. We assume the following:

ASSUMPTION B.

- (i) The set \mathcal{M} is convex and compact.
- (ii) The functions $\phi_i: \mathcal{M} \rightarrow R^1 \cup \{\infty\}$ are lower semicontinuous and bounded below for every $1 \leq i \leq n$.
- (iii) If $M \in \mathcal{M}$, then any two states communicate on some time interval $[0, n_0]$ under the action of the stationary strategy $\sigma = \{M\}_{k=0}^\infty$, namely $((M^T)^{n_0})_{ij} > 0$ for some n_0 .

We shall prove in this section that for a constraint set \mathcal{M} as in Assumption B there is a stationary strategy $M \in \mathcal{M}$ which is of a minimal cost growth rate among all the nonanticipative strategies. For every $M \in \mathcal{M}$ let $\psi(M)$ be the point whose i th component is $\phi_i(M_i)$. The functions $\phi_i(\cdot)$ can take the value $+\infty$, and we agree to interpret scalar products $x^T \cdot \psi(M)$ as equal to

$$\sum_{x_i > 0} x_i \phi_i(M_i)$$

and possibly equal to $+\infty$. We define the following:

$$(5.1) \quad u(x, y) = \min \{x^T \cdot \psi(M) : M^T x = y, M \in \mathcal{M}\}.$$

PROPOSITION 5.1. *The function $u(\cdot, \cdot)$ is lower semicontinuous.*

Proof. Let $(x^i, y^i) \rightarrow (x^0, y^0)$ as $i \rightarrow \infty$, and let $M^i \in \mathcal{M}$ be such that

$$u(x^i, y^i) = (x^i)^T \psi(M^i) \quad \text{and} \quad (M^i)^T x^i = y^i.$$

Then there is an $M^0 \in \mathcal{M}$ such that $(M^0)^T x^0 = y^0$. Let j be such that $(x^0)_j > 0$. Then we claim that

$$(5.2) \quad (x^0)_j \phi_j((M^0)_j) \leq \liminf_{i \rightarrow \infty} (x^i)_j \phi_j((M^i)_j).$$

This is clear if $\phi_j((M^i)_j) \rightarrow \infty$ as $i \rightarrow \infty$, and it follows from the lower semicontinuity of $\phi_j(\cdot)$ if $\{\phi_j((M^i)_j)\}_{i=1}^\infty$ is bounded. If $(x^0)_j = 0$, then (5.2) holds again since $\phi_j(\cdot)$ is bounded below. Thus we have

$$u(x^0, y^0) \leq (x^0)^T \psi(M^0) \leq \liminf_{i \rightarrow \infty} u(x^i, y^i)$$

proving the assertion of the proposition. \square

Let σ be any nonanticipative strategy with control values in \mathcal{M} . Then, by the same proof as that of (4.11), there is a Markov strategy $\bar{\sigma}$ such that (4.11) holds for every $N \geq 1$. The proof of (4.11) did not make use of Assumption A, while the convexity assumption on \mathcal{M} guarantees that the expectation of the random controls with values in \mathcal{M} are also in \mathcal{M} . Thus we can consider from now on only Markov strategies and establish the existence of a stationary strategy with a minimal long-run average cost is this smaller class.

The constant μ which is the minimal cost growth rate is given by

$$(5.3) \quad \mu = \inf \left\{ \liminf_{N \rightarrow \infty} \frac{1}{N} \sum_{k=0}^{N-1} u(x_k, x_{k+1}) \right\}$$

where N varies over all the integers and $\{x_k\}_{k=1}^\infty$ varies over all the sequences in S such that $x_{k+1} = M^T x_k$, $M \in \mathcal{M}$ (which depends on x_k, x_{k+1}).

PROPOSITION 5.2. *For every $\varepsilon > 0$ there are points $z_1, z_2 \in S$ such that*

$$|z_1 - z_2| < \varepsilon \quad \text{and} \quad u(z_1, z_2) < \mu + \varepsilon.$$

Proof. Given ε , there is by (5.3) a sequence $\{x_k\}_{k=0}^N \subset S$ such that

$$\frac{1}{N} \sum_{k=0}^{N-1} u(x_k, x_{k+1}) < \mu + \varepsilon$$

where N can be taken arbitrarily large. We define

$$(5.4) \quad z_1 = \frac{1}{N} \sum_{k=0}^{N-1} x_k, \quad z_2 = \frac{1}{N} \sum_{k=1}^N x_k.$$

Repeating the calculation between (4.18) and (4.23) in the proof of Theorem 4.4, while using the fact that \mathcal{M} is convex, we get that there is a matrix $M \in \mathcal{M}$ such that

$$(5.5) \quad z_2 = M^T z_1 \quad \text{and} \quad \sum_{i=1}^n (z_1)_i \phi_i(M_i) \leq \mu + \varepsilon$$

and therefore $u(z_1, z_2) \leq \mu + \varepsilon$. But if N is large enough, then it follows from (5.4) that $|z_1 - z_2| < \varepsilon$, which concludes the proof. \square

THEOREM 5.3. *Let \mathcal{M} satisfy Assumption B. Then there is a stationary strategy $\sigma_0 = \{M_0\}_{k=0}^\infty$ which is of a minimal cost growth rate in the class of all nonanticipative strategies with control values in \mathcal{M} . M_0 is any matrix such that*

$$(5.6) \quad M_0^T z_0 = z_0, \quad z_0^T \cdot \psi(M_0) = u(z_0, z_0), \quad M_0 \in \mathcal{M}$$

and

$$(5.7) \quad u(z_0, z_0) = \min_{z \in S} u(z, z).$$

Proof. As remarked above, it is enough to establish the existence of an optimal stationary strategy as in the theorem in the smaller class of Markov strategies $\{M^{(k)}\}_{k=0}^\infty$, $M^{(k)} \in \mathcal{M}$ for all $k \geq 0$. It follows from Proposition 5.2 that there is a z_0 such that $u(z_0, z_0) = \mu$, thus z_0 also satisfies (5.7).

Let M_0 be such that (5.6) is satisfied and consider the stationary strategy $\sigma_0 = \{M_0\}_{k=0}^\infty$. It follows from Assumption B(iii) that the corresponding Markov chain has a unique equilibrium distribution z_0 . Thus, if $x^{(k)}$ is the distribution of the Markov chain at the time k , then

$$\frac{1}{N} \sum_{k=0}^{N-1} u(x^{(k)}, x^{(k+1)}) \rightarrow u(z_0, z_0) = \mu$$

proving that σ_0 is of a minimal cost growth rate. \square

6. Overtaking optimal strategies. In this section we confine our attention to stationary strategies, i.e., $\sigma = \{M\}_{k=0}^\infty$ for some stochastic matrix M . In this situation it is easy to derive simple expressions for the cost flows. We will prove that under Assumption B for every initial state there exists an overtaking optimal strategy (recall Definition 3.1). The difficulty is to obtain a stationary strategy which is overtaking optimal for all the initial states. We will show that under Assumption A every stationary strategy with a minimal long-run average cost is in fact overtaking optimal for all the initial states.

We assume now that Assumption B holds. Let $M \in \mathcal{M}$, $\sigma = \{M\}_{k=0}^\infty$, $\{\zeta^{(k)}\}_{k=0}^\infty$ be the Markov chain associated with this strategy and the initial distribution $x^{(0)}$, and $\{x^{(k)}\}_{k=0}^\infty$ is the corresponding distributions. Then $x^{(k+1)} = M^T x^{(k)}$ and it is well known that

$$(6.1) \quad x^{(k)} \rightarrow z \quad \text{exponentially fast as } k \rightarrow \infty$$

for some $z \in \text{rel int } S$ which satisfies $z = M^T z$ (see Kushner [6, Lemma 3, p. 54]). Thus we get a mapping

$$(6.2) \quad M \rightarrow \phi(M) = z^M$$

which associates with every $M \in \mathcal{M}$ the equilibrium distribution z^M , and we claim the following.

PROPOSITION 6.1. *The mapping in (6.2) is continuous.*

Proof. Let $\{M_l\}$ and $\{z_l\}$, $l \geq 1$, be related by $z_l = (M_l)^T z_l$, and suppose that $M_l \in \mathcal{M}$ and $M_l \rightarrow M_0$ as $l \rightarrow \infty$. Then the uniqueness of z_0 implies that every convergent subsequence of $\{z_l\}_{l=1}^\infty$ must coverage to z_0 , implying that $z_l \rightarrow z_0$ as $l \rightarrow \infty$ and proving the proposition. \square

For a stationary strategy $\sigma = \{M\}_{k=0}^\infty$ and an initial state i we have

$$(6.3) \quad \frac{1}{N} C_N(i, \sigma) \rightarrow (z^M)^T \cdot \phi(M) \quad \text{as } N \rightarrow \infty$$

where we denote $(\phi(M))^T = (\phi_1(M_1), \dots, \phi_n(M_n))$. By Assumption B every $\phi_j(\cdot)$ is a lower semicontinuous function. Thus it follows that the right-hand side of (6.3) is a lower semicontinuous function of M , and it obtains its minimal value on \mathcal{M} . Let the minimal value be μ and the compact set of matrices where it is obtained be $\mathcal{M}_0 \subset \mathcal{M}$. Thus we have for every $M \in \mathcal{M}_0$

$$\mu = (z^M)^T \cdot M^k \phi(M) \quad \text{for every } k \geq 0.$$

For a stationary $\sigma = \{M\}_{k=0}^\infty$ and an initial state i we consider the *modified cost flow*

$$\tilde{C}_N(i, \sigma) = C_N(i, \sigma) - \mu N$$

which is equal to $\sum_{k=0}^{N-1} (x^{(k)} - z^M)^T \cdot \phi(M)$, namely

$$(6.4) \quad \tilde{C}_N(i, \sigma) = (e_i - z^M)^T \cdot \sum_{k=0}^{N-1} M^k \phi(M)$$

(e_i is the i th vertex of S). By (6.1) the limit in (6.4) exists as $N \rightarrow \infty$ and we denote

$$\psi(i, M) = \lim_{N \rightarrow \infty} \tilde{C}_N(i, \sigma)$$

and

$$(6.5) \quad \psi(i, M) = \lim_{N \rightarrow \infty} (e_i - z^M)^T \cdot \sum_{k=0}^{N-1} M^k \phi(M)$$

and $(\psi(M))^T = (\psi(1, M), \dots, \psi(n, M))$. For an initial distribution $x^{(0)}$ the modified expected cost is then $(x^{(0)})^T \cdot \psi(M)$. In order for a strategy $\sigma = \{M^i\}_{k=0}^\infty$ to be overtaking optimal for the initial state i it must satisfy $\psi(i, M^i) \leq \psi(i, M)$ for every $M \in \mathcal{M}$. If $\sigma = \{M^*\}_{k=0}^\infty$ is to be overtaking optimal for all the initial states, then it must satisfy $\psi(M^*) \leq \psi(M)$ for every $M \in \mathcal{M}$, the inequality being understood componentwise.

The discussion which leads to (6.4) shows that for every $y \in R^n$ the limit

$$(6.6) \quad \lim_{N \rightarrow \infty} (x - z^M)^T \cdot \sum_{k=0}^{N-1} M^k y$$

exists for every $x \in S$. If we denote

$$X = \left\{ w \in R^n : \sum_{i=1}^n w_i = 0 \right\}$$

then the limit in (6.6) defines a linear function on X which can be represented uniquely as

$$w \rightarrow w^T \cdot \eta(y)$$

for some $\eta(y) \in X^n$. Then $\eta(y)$ depends linearly on y , therefore, it depends on it continuously. It follows that we can write (6.5) as

$$\psi(i, M) = (e_i - z^M)^T \cdot \eta(\phi(M)),$$

hence $\psi(i, M)$ depends continuously on M and it attains its minimal value on \mathcal{M}_0 , say at M^i . Then $\sigma = \{M^i\}_{k=0}^\infty$ is an overtaking optimal strategy for the initial state i . We thus have proved the following.

THEOREM 6.2. *Let Assumption B hold and let the initial state be i . Then there is a stationary strategy $\sigma = \{M^i\}_{k=0}^\infty$ which is overtaking optimal in the class of all stationary strategies.*

We will assume now that Assumption A holds and will establish the existence of a stationary strategy which is overtaking optimal for all the initial states. In fact we will prove the stronger result, that is, that every stationary strategy with a minimal cost growth rate is overtaking optimal for all the initial states.

Let $\{M_i\}_{k=0}^\infty$ be the overtaking optimal strategy guaranteed by Theorem 6.2 for the initial state i . We assume that at least one of the matrices M_i , $1 \leq i \leq n$, is not overtaking optimal for all the initial states. We shall now show that this assumption leads into a contradiction. We define the matrix M^* as the stochastic matrix whose i th row is equal to the i th row of M_i . We define the following strategies: $\sigma^* = \{M^*\}_{k=0}^\infty$ is a stationary strategy. Let σ_N , $N \geq 1$, be the nonanticipative strategy which coincides with σ^* as long as there occurred no more than N changes of states. If after N changes of states the state is j , then from that time on σ_N coincides with the stationary strategy $\{M_j\}_{k=0}^\infty$. Note that σ_N is not a Markov strategy. Finally, let σ_0 be a strategy which coincides with the stationary strategy $\{M_i\}_{k=0}^\infty$ if the initial state is i , and where the initial state is a random variable $\zeta^{(0)}$.

We also define the sequence of random times $k_1 < k_2 < \dots < k_l < \dots$ where changes of states occur, namely, $\zeta_r = i$ for $0 \leq r < k_1$, and $\zeta_{k_1} \neq i$, $\zeta_r = \zeta_{k_1}$ for $k_1 \leq r < k_2$ and $\zeta_{k_2} \neq \zeta_{k_1}$, and so on.

The expected modified cost of σ_1 if the initial state is i is given by

$$(6.7) \quad \sum_{j \neq i} \{[\phi_i((M_i)_i) - \mu]E(k_1 | \zeta_{k_1} = j) + \psi(M_j, j)\}P(\zeta_{k_1} = j)$$

(here $(M_i)_i$ is the i th row of M_i). This expression is not greater than

$$(6.8) \quad \sum_{j \neq i} [\phi_i((M_i)_i) - \mu]E(k_1 | \zeta_{k_1} = j) + \psi(M_i, j)P(\zeta_{k_1} = j)$$

for every $1 \leq i \leq n$, since $\psi(M_j, j) \leq \psi(M_i, j)$. However, since at least one of the strategies $\{M_i\}_{k=0}^\infty$ is not overtaking optimal for every initial state, there is some i for which the expression in (6.7) is strictly smaller than that in (6.8). This discussion implies the following.

PROPOSITION 6.3. *Assume that at least one of the strategies $\{M_i\}_{k=0}^\infty$ is not overtaking optimal for every initial state. Let the initial state $\zeta^{(0)}$ have a distribution $\theta \in \text{rel int } S$. Then the cost of applying the strategy σ_1 is strictly smaller than that of applying the strategy σ_0 .*

Proof. The assertion follows from the above discussion and from the fact that with a positive probability the initial state is one for which a strict inequality between the expression (6.7) and (6.8) will occur. \square

We are looking now for a special probability distribution for $\zeta^{(0)}$, say θ^* , which will have the following property: If $\zeta^{(k)}$ is the Markov chain which is obtained while employing the strategy σ^* then the distribution of $\zeta^{(k_l)}$ is θ^* for every $l \geq 1$, where $\{k_l\}_{l=1}^\infty$ are the random times of changes of states.

LEMMA 6.4. *There exists a $\theta^* \in \text{rel int } S$ as asserted above.*

Proof. The probability that $\zeta^{(k_l)} = j$ if the initial state is i and we employ the strategy σ^* is given by

$$N_{ij} = \begin{cases} 0 & \text{if } j = i, \\ \frac{M_{ij}^*}{1 - M_{ii}^*} & \text{if } j \neq i. \end{cases}$$

If $n = 2$ (the number of states) then

$$N = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \quad \text{and} \quad \theta^* = \begin{pmatrix} \frac{1}{2} \\ \frac{1}{2} \end{pmatrix}$$

satisfies the assertion of the theorem. If $n > 2$ then N^2 has only positive entries and there is a unique $\theta^* \in \text{rel int } S$ which satisfies $N^T \theta^* = \theta^*$. This θ^* is the one we are looking for. \square

THEOREM 6.5. *If the stationary strategy $\{M_i\}_{k=0}^\infty$ is overtaking optimal for the initial state i then it is overtaking optimal for every initial state. In particular there is an overtaking optimal strategy, for all the initial states, in the class of stationary strategies.*

Proof. Assume that the assertion of the theorem is false. Let θ^* be as in Lemma 6.4 and assume that it is the initial distribution. If we employ the strategy σ_1 then the distribution of $\zeta^{(k_1)}$ will be θ^* also. Therefore the expected modified cost of applying σ_1 , after time k , is equal to the expected modified cost of applying σ_0 from the initial time. It thus follows from Proposition 6.3 that the expected modified cost of applying σ_1 up to time k_1 is negative, say $-\varepsilon$. Then, since the distribution of $\zeta^{(k_1)}$ is given by θ^* , it follows that the expected modified cost of applying σ_2 on the time interval $[k_1, k_2]$ is $-\varepsilon$ also; and, in fact, the expected modified cost of applying σ_2 on $[0, k_2]$ is -2ε . It then follows that the expected modified cost of applying the nonanticipative strategy σ_N on the time interval $[0, k_N]$ is $-N\varepsilon$, which tends to $-\infty$ as N tends to $+\infty$. This, however, is a contradiction to Proposition 4.2 and the inequality (4.9), which implies that there is some lower bound on the modified costs of all the nonanticipative strategies and for all the times. This concludes the proof of the theorem. \square

We shall now establish our main result.

THEOREM 6.6. *Let $\sigma = \{M\}_{k=0}^\infty$ be a stationary strategy which has a minimal cost growth rate. Then σ is overtaking optimal for every initial state, in the class of all stationary strategies.*

Proof. By Theorem 6.5 there exists an overtaking optimal strategy $\sigma^* = \{M^*\}_{k=0}^\infty$ and if σ is not overtaking optimal then

$$(6.9) \quad \psi(M^*, i) < \psi(M, i)$$

for some initial state i .

It follows from (6.9) that there is an integer L such that the following holds:

$$(6.10) \quad C_N(i, \sigma^*) < C_N(i, \sigma) - \varepsilon$$

for some $\varepsilon > 0$, for this i , and for all $N > L$.

Let i be the initial state, and let $\{\zeta^{(k)}\}_{k=0}^\infty$ and $\{\eta^{(k)}\}_{k=0}^\infty$ be the Markov chains corresponding to σ^* and σ , respectively. Let k_1 be the first time k which is greater than L where the equality $\zeta^{(k)} = \eta^{(k)} = i$ occurs. Inductively, let k_{r+1} be the first time

k which is larger than $k_r + L$ where the equality $\zeta^{(k)} = \eta^{(k)} = i$ occurs. Clearly $P(k_r < \infty) = 1$ for every $r \geq 1$. It follows from (6.10) that

$$E_i C_{k_i}(i, \sigma^*) \leq E_i C_{k_i}(i, \sigma) - \varepsilon.$$

It also follows from (6.10) that for every $r \geq 1$

$$(6.11) \quad E_{\zeta^{(k_r)}} C_{k_{r+1}-k_r}(\zeta^{(k_r)}, \sigma^*) < E_{\eta^{(k_r)}} C_{k_{r+1}-k_r}(\eta^{(k_r)}, \sigma) - \varepsilon.$$

Adding the inequalities (6.11) for $1 \leq r \leq l-1$, we get

$$(6.12) \quad E_i C_{k_l}(i, \sigma^*) \leq E_i C_{k_l}(i, \sigma) - l\varepsilon.$$

The strategy σ is such that

$$(6.13) \quad |C_N(i, \sigma) - \mu N| < K$$

for some $K > 0$ and all $N \geq 1$, since it is stationary and of a minimal cost growth rate. Also, for every strategy σ'

$$C_N(i, \sigma') - \mu N \geq -K$$

for all $N \geq 1$. Therefore, also for every random time $k(\omega)$

$$(6.14) \quad E[C_k(i, \sigma') - k\mu] \geq -K, \quad E[C_k(i, \sigma) - k\mu] \leq K.$$

Taking $\sigma' = \sigma^*$ and letting $l \rightarrow \infty$, however, we find that (6.12) contradicts (6.14). This concludes the proof of the theorem. \square

Theorem 6.6 provides us with a tool of computing overtaking optimal strategies. We have to compute the function $(x, y) \rightarrow u(x, y)$ and consider its restriction to the diagonal $z \rightarrow u(z, z)$. Let z_0 be in $\text{rel int } S$ such that

$$u(z_0, z_0) \leq u(z, z)$$

for every $z \in S$. We have then to compute the stochastic matrix M_0 which satisfies the relations

$$M_0^T z_0 = z_0, \quad \sum_{i=1}^n (z_0)_i \phi_i((M_0)_i) = u(z_0, z_0).$$

This matrix M_0 is such that $\sigma_0 = \{M_0\}_{k=0}^\infty$ is overtaking optimal for every initial state.

Acknowledgment. I would like to thank Professor Steven Orey for many fruitful discussions concerning this work.

REFERENCES

- [1] V. S. BORKAR, *On minimum cost per unit time control of Markov chains*, this Journal, 22 (1984), pp. 965-978.
- [2] W. A. BROCK AND A. HAURIE, *On existence of overtaking optimal trajectories over an infinite time horizon*, Math. Oper. Res., 1 (1976), pp. 337-346.
- [3] D. A. CARLSON, *On the existence of catching-up optimal solutions for Lagrange problems defined on unbounded intervals*, J. Optimization Theory Appl., 49 (1986), pp. 207-225.
- [4] R. J. ELLIOTT, *Smoothing for a finite state Markov process*, in Lecture Notes in Control and Information Sciences, 69, M. Metevier and E. Pardoux, eds., Springer-Verlag, New York, 1984.
- [5] D. GALE, *On optimal development in a multi-sector economy*, Rev. Econom. Stud., 34 (1967), pp. 1-19.
- [6] H. KUSHNER, *Introduction to Stochastic Control*, Holt Rinehart and Winston, New York, 1971.
- [7] A. LEIZAROWITZ, *Infinite horizon autonomous systems with unbounded cost*, Appl. Math. Optim., 13 (1985), pp. 19-43.
- [8] R. T. ROCKAFELLAR, *Convex Analysis*, Princeton University Press, Princeton, NJ, 1970.
- [9] C. C. VON WEIZSACKER, *Existence of optimal programs of accumulation for an infinite horizon*, Rev. Econom. Stud., 32 (1965), pp. 85-104.
- [10] P. WHITTLE, *Optimization Over Time*, Vol. II, John Wiley, New York, 1983.